

**Structural Dynamics of Thrombin-Binding DNA Aptamer
d(GGTTGGTGTGGTTGG) Quadruplex DNA Studied by
Large-Scale Explicit Solvent Simulations**

Roman Reshetnikov*

*Department of Bioengineering and Bioinformatics, Lomonosov Moscow State
University, GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation*

Andrey Golovin

*Department of Bioengineering and Bioinformatics, Lomonosov Moscow State
University, GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation*

Vera Spiridonova

*A.N.Belozersky Institute of Physical Chemical Biology, Lomonosov Moscow State
University, GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation*

Alexei Kopylov

*Chemistry Department, Lomonosov Moscow State University, Gsp-1, Leninskie Gory,
Moscow, 119991, Russian Federation*

Jiří Šponer

*Institute of Biophysics, Academy of Sciences of the Czech Republic, Královopolská
135, 61265 Brno, Czech Republic*

Received May 13, 2010

Abstract: The thrombin-binding aptamer (15-TBA) is a 15-mer DNA oligonucleotide with sequence d(GGTTGGTGTGGTTGG). 15-TBA folds into a quadruplex DNA (G-DNA) structure with two planar G-quartets connected by three single-stranded loops. The arrangement of the 15-TBA-thrombin complex is unclear, particularly with respect to the precise 15-TBA residues that interact with the thrombin structure. Our present understanding suggests either the 15-TBA single stranded loops containing sequential thymidines (TT) or alternatively a single-stranded loop, containing a guanine flanked by 2 thymidines (TGT), physically associates with thrombin protein. In the present study, the explicit solvent molecular dynamics (MD) simulation method was utilized to further analyze the 15-TBA-thrombin three-dimensional structure. Functional annotation of the loop residues was made with long simulations in the parmbsc0 force field. In total, the elapsed time of simulations carried out in this study exceeds 12 microseconds, substantially surpassing previous G-DNA simulation reports. Our simulations suggest that the TGT-loop function is to stabilize the structure of the aptamer, while the TT-loops participate in direct binding to thrombin. The findings of the present report advance our understanding of the molecular structure of the 15-TBA-thrombin structure further enabling the construction of biosensors for aptamer bases and the development of anticoagulant agents.

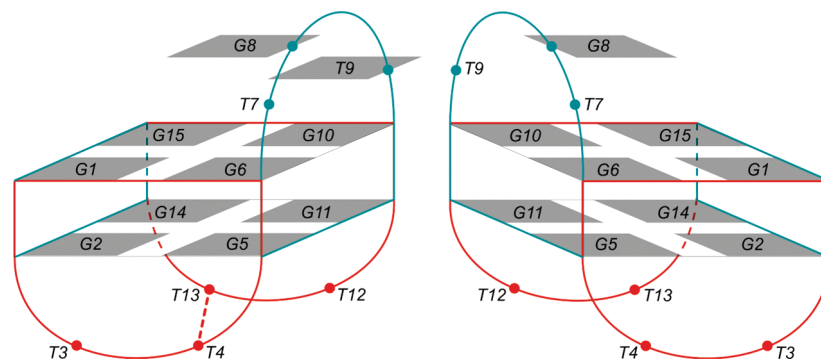


Figure 1. Schematic representation of 15-TBA. Left: NMR-based model. Right: X-ray-based model. Two G-quartets, upper (G1, G6, G10, G15) and lower (G2, G5, G11, G14), form G-quadruplex. The remaining nucleotides form three lateral loops, one TGT and two TT. An approximate 2-fold axis of symmetry relates the two halves of the G-quadruplex, resulting in two symmetric wide grooves (blue) and two symmetric narrow grooves (red). The 15-TBA models differ in chain direction and in loop topology. In the NMR-based model, two nucleic bases from the TGT-loop, G8 and T9, are stacked with the upper G-quartet. Stacking nucleotides from the TGT-loop are shown by gray tetragons. There are T4-T13 pair interactions between the TT-loops. In the X-ray-based model, only the G8 base is stacked with the upper G-quartet with no interactions between the TT-loops.

Introduction

Aptamers are synthetic oligonucleotides that specifically bind with high affinity a wide range targets, from small molecules to whole cells.^{1,2} Aptamers have been developed through the use of Systematic Evolution of Ligands by Exponential Enrichment (SELEX).³ Among the first successful SELEX targets was the serine protease thrombin, which plays a key role in blood coagulation cascade.⁴ Thrombin is a globular protein with two positively charged substrate (ligand) binding domains positioned on opposite sides of the protein surface.⁵ These substrate binding domains are termed fibrinogen-binding site (exosite I) and heparin-binding site (exosite II) (Supporting Information, Figure S1). The most widely studied thrombin-binding DNA aptamer is the 15-mer oligonucleotide with sequence d(GGTTGGTGTGGTTGG) (15-TBA).^{6–10} 15-TBA forms secondary structure consisting of two planar G-quartets, one over another (G-quadruplex or G-stem), connected by three intervening lateral loops. Two of these loops consist of a pair of thymidine bases (TT), while the third loop consists of two thymidines flanking a central guanine base (TGT) (Figure 1).

Despite numerous reports on the structure of 15-TBA, the precise structure and points of interaction with thrombin remain poorly resolved. Both NMR^{6–9} and X-ray¹⁰ structures have been reported for 15-TBA. The respective models are mutually inconsistent however, differing both in chain direction and loop geometry. The NMR resolved structure is widely favored over the X-ray structure. Indeed, the NMR structure is in better agreement with the raw X-ray data because of the R-factor and real space correlation coefficient.¹¹ The NMR-based and X-ray-based 15-TBA models differ with respect to which bases associate directly with thrombin. In the NMR-based model, 15-TBA binds the exosite-I site of thrombin through the TT-loops. Alternatively, the X-ray resolved model suggests that it is the TGT-loop associating directly with exosite-I.⁹ Mutational analysis of 15-TBA demonstrates that modification of the TT-loops

diminishes thrombin binding.^{12,13} However, modification of the TGT-loop sequence of 15-TBA adversely affects thrombin inhibition activity.¹⁴ The stoichiometry of the thrombin-aptamer complex also remains unclear. Bock⁴ and Tasset¹⁵ with colleagues have assumed 1:1 stoichiometry of the aptamer-thrombin complex. In contrast, crystallographic data from Padmanabhan et al.¹⁰ suggest that 15-TBA can also interact with exosite-II of a symmetry-related molecule of neighboring thrombin. The crystallographic data is consistent with the isothermal titration calorimetry (ITC) results reported by Pagano et al.,¹⁶ which accordingly demonstrate 2:1 stoichiometry for the thrombin-aptamer complex.

Molecular dynamics simulation (MD) is a valuable tool for investigating G-quadruplex-containing structures.^{17–22} Current force fields, such as the parm99^{23,24} version of the Cornell et al. force field,²⁵ can be readily used to provide descriptions of G-stem structures.^{17,18} In contrast, diagonal and propeller loops of G-quadruplex structures have previously been difficult targets in molecular modeling approaches.^{17–19} In 2007, the parmbsc0 version of the nucleic acids force field was released.²⁶ The parmbsc0 has since been used to describe correctly a wide range of canonical and noncanonical nucleic acid structures.^{26,27} The parmbsc0 version enabled a dramatically improved description of B-DNA structure (which was unstable in longer simulations with earlier versions of the Cornell et al. force field). Parmbsc0 also improves the description of single stranded DNA loop structures (such as those of G-DNA). Albeit, there remain limitations with respect to the capacity of the parmbsc0 force field to yield complete loop descriptions.¹⁹

Two previously published studies have applied short MD simulations to 15-TBA. Pagano et al.¹⁶ reported that MD simulations of 15-TBA and its derivatives produce stable 5 ns-long trajectories in the parm98 force field. Importantly, the resulting average structure agrees well with the published, NMR-based structure that is the starting model for MD. The lateral loops of 15-TBA are represented well in the parm98 force field though the short simulation time scale precludes definitive conclusions. In a simulation published by Jayapal et al.,²⁸ predistorted 15-TBA recovers a structure similar to

* To whom correspondence should be addressed. E-mail: r.reshetnikov@gmail.com.

the initial 15-TBA structure during 2 ns of MD in an OPLS-AA force field using entries for nucleic acids added by Golovin and Polyakov.²⁹ The 2 ns simulations, however, provide only limited insights.

The determination of function for each nucleotide of 15-TBA will resolve many of the aforementioned disparities among published reports. In the present study, we provide functional annotation of 15-TBA residues resulting from the use of long MD simulations. We evaluated the viability of a variety of G-quadruplex-containing structures including G-DNA stems as well as complexes of 15-TBA with thrombin. Herein, we report the use of two force fields to resolve these structures, parm99 and parmbsc0, with simulation times from 600 to 900 ns in individual runs. The combined data from these MD simulations exceeds 12 μ s surpassing the duration of any currently published simulations of G-DNA structure.

Data from this study suggest that the NMR-based conformation is the only viable 15-TBA structure, either in its free state or in complex with thrombin. These data further suggest that the X-ray resolved conformation is unstable. MD simulations of loop-free two-quartet G-stem, 15-TBA in a free state, and 15-TBA complexed with thrombin show that the TT-loops substantially influence the twist of the G-stem (as compared to simulation with loop-free stem). Interestingly, subsequent binding of thrombin reduces the structural strain on the stem, clearly suggesting a mutual adaptation of the TT-loops, the stem, and the protein. These data further suggest that the principle function of the TGT-loop is to stabilize the G-stem.

Materials and Methods

Computer Modeling. The X-ray-based structure of 15-TBA was taken from the structure of the complex between thrombin and the aptamer, PDB ID 1hut.¹⁰ The NMR-based structure of 15-TBA was taken from PDB entry 148d, eighth frame.⁶ The structure of the four-stranded stem consisting of two G-quartets was obtained from the NMR-based structure by removing loop residues. We have also studied NMR and X-ray models of 15-TBA with modified conformations of the TGT-loop named TG(-T), T(-GT), and TG(+T). Here, signs “-” or “+” denote residues whose position was changed to either disrupt (-) or establish (+) base stacking with the upper G-quartet. These models were obtained from the initial conformations by rotating the G8 and T9 locations around dihedral angles γ , ϵ , and χ using the Pymol, version 1.1, software program.³⁰ The energy minimization procedure, with the quasi-Newtonian limited memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm,³¹ was applied for modified conformations to remove strain. For simulations of thrombin-aptamer complexes (PDB ID of NMR-based model 1hao,⁹ X-ray-based model 1hut¹⁰), Asp, Glu, and His residues of thrombin were protonated according to the determination by Ahmed et al.³² Models of 1:2 complexes of 15-TBA with thrombin were obtained from models of the 1:1 complexes (PDB IDs 1hao, 1hut) by generating the packing interactions within 4 Å from the

models in the Pymol, version 1.1, software program. The thrombin molecule interacting with the aptamer of the initial model through its exosite-II was then used as a second protein in the 1:2 complex (Supporting Information, Figure S2).

Molecular Dynamics Simulation. The GROMACS 4.0 software package^{33,34} was used for simulation and analysis of MD trajectories using the AMBER-99 φ and parmbsc0²⁶ force fields. The AMBER-99 φ force field is an improved version of the parm99²³ force field with reconsidered φ torsion potential, developed and adapted to GROMACS by Sorin and Pande.³⁵ The parmbsc0²⁶ force field was ported onto GROMACS by us through a modification of the AMBER-99SB³⁶ force field entries for nucleic acids. Explicit solvent simulations were performed at $T = 300$ K with a time constant for coupling of 0.1 ps under the control of a velocity rescaling thermostat,³⁷ isotropic constant-pressure boundary conditions under the control of the Berendsen algorithm of pressure coupling³⁸ with a time constant of 5 ps and application of the particle mesh Ewald method for electrostatic interactions (PME)³⁹ with grid spacing of 0.178 nm and interpolation order 6. A triclinic box of TIP4P⁴⁰ water molecules was added around the DNA to a depth of 15 Å on each side of the solute. Negative charges were neutralized with the addition of sodium cations and positive charges by chloride ions. Additional NaCl was added to a final concentration of 0.1 M to protein-containing systems. In each of the simulations, there were two temperature coupling groups, the first consisting of DNA with K⁺ ion and the second consisting of water with Na⁺ and Cl⁻ ions. Protein atoms, when present, were added to the first group. The time step for integration in all simulations was 3 fs. Coordinates were written to output a trajectory file every 6 ps. Data extraction from the trajectory file for analysis was made with a time step of 150 ps. Stabilization of DNA models, except of the specially stipulated simulations, was made by placing of potassium cation in the geometrical center of the G-quadruplex stem, coordinates of the center were calculated as an arithmetic mean value from O6 atoms positions of the quadruplex stem guanines. We used standard AMBER potassium (radius 0.2658 Å and well depth 0.00137 kJ/mol), sodium (radius 0.1868 Å, well depth 0.01589 kJ/mol) and chloride (radius 0.2470 Å, well depth 0.41840 kJ/mol) parameters. Sodium and chloride ions were added to the systems by replacing water molecules at random positions with minimal distance between ions equal to 6 Å. All simulations were done on a “Chebyshev” supercomputer provided by SRCC of Moscow State University. Information about run times and parallelization of the simulations on the supercomputer is provided in Supporting Information (Figure S3). The simulations parameters are provided in Table 1. Analysis of the trajectories was also performed using the GROMACS 4.0 software package. Hydrogen bonds were treated as existing if their lifetime was greater than 50% of the trajectory length. H-bonds were counted for

Table 1. Simulation Parameters

solute ^a	force field	trajectory length (ns)	quantity of atoms					
			DNA	protein	water	Na ⁺	K ⁺	Cl ⁻
G-stem	parmbsc0	700	260	0	12516	3	1	0
G-stem without ion in the center	parmbsc0	10	260	0	12516	4	0	0
X-ray 15-TBA	parm99	900	488	0	22516	13	1	0
X-ray 15-TBA	parmbsc0	900	488	0	19240	13	1	0
NMR 15-TBA	parm99	900	488	0	22524	13	1	0
NMR 15-TBA	parmbsc0	900	488	0	22532	13	1	0
NMR 15-TBA with Na ⁺ in the center	parmbsc0	900	488	0	22532	14	0	0
NMR 15-TBA without ion in the center	parmbsc0	900	488	0	22528	14	0	0
modified TGT conformations								
TG(+T) X-ray 15-TBA ^b	parmbsc0	900	488	0	18688	13	1	0
TG(-T) NMR 15-TBA ^b	parmbsc0	900	488	0	18312	13	1	0
TG(-T) NMR ^{eq} 15-TBA ^{b,c}	parm99	900	488	0	18968	13	1	0
T(-GT) NMR ^{eq} 15-TBA ^{b,c}	parmbsc0	900	488	0	20532	13	1	0
complexes with thrombin								
X-ray thrombin-aptamer complex	parmbsc0+parm99SB	600	488	4657 ^d	104964	49	1	45
NMR thrombin-aptamer complex	parmbsc0+parm99SB	600	488	4658 ^d	105104	48	1	45
1:2 X-ray aptamer-thrombin complex	parmbsc0+parm99SB	600	488	9314	155404	71	1	78
1:2 NMR aptamer-thrombin complex	parmbsc0+parm99SB	600	488	9316	155732	71	1	78

^a One K⁺ ion in the central cavity of the G-stem if not specified otherwise. ^b See the text for explanation of the abbreviation. ^c NMR^{eq} indicates that the starting 15-TBA structure is based on preceding MD simulation of the NMR-based structure. The last 100 ns of the 900 ns simulation were used to create the averaged structure which then was manipulated to arrive at the TG(-T) or T(-GT) arrangement. ^d Thrombin molecules in 1hao and 1hut structures differ from each other in aminoacid sequence.

donor–acceptor distances shorter than 3.5 Å with an acceptor–donor-hydrogen angle cutoff of 30°.

Results

Comparison of the NMR-Based and X-ray-Based Models of Free 15-TBA. The relative stability of the alternative models was determined with 900 ns MD simulation runs in both parm99 and parmbsc0 force fields. In this simulation, the X-ray model was treated as a potentially correct conformation of free 15-TBA. The X-ray resolved model completely lost its G-quartets in both force fields however. In either force field, destruction of the G-quadruplex began with the formation of stacking interactions between T4 and T13 under the G-stem with consequent disruption of the lower G-quartet planarity. In the parm99 force field, the structure maintained characteristics of a G-quadruplex structure until reaching 170 ns of MD trajectory. In the case of the parmbsc0 force field, the G-quadruplex collapsed during the first 10 ns (Supporting Information, Figure S4). In sharp contrast the NMR-based structure remained topologically similar to that of the initial structure in both force fields with the exception of some rearrangement of the T-T interactions below the G-stem. In the parmbsc0 force field, T4-T13 hydrogen bonding interactions switched to T4-T12 and T3-T13 interactions, both under the lower G-quartet of the stem. In the parm99 force field, T4-T13 interactions switched to T4-T12, while T13 remained under the lower G-quartet, T3 remained exposed to solution (Figure 2).

Why is the NMR-Based Structure Stable while the X-ray-Based Structure Is Not? The X-ray resolved model of 15-TBA is presumed incorrect. Therefore, the instability of this model in our simulations is not surprising. However, the collapse of the entire X-ray structure that we observed is of interest. The observed instability in either force field

suggests that this 15-TBA arrangement is absolutely unstable at the level of a single molecule, a conclusion that can not be directly derived from the experimental data. Because this collapse did not occur in the case of the NMR-based model, we have concluded not only that the NMR-based model is a better candidate structure, but also that the X-ray model is intrinsically not viable at all. The stability of the NMR structure in our exceptionally long simulations indicates reasonable performance in these force fields. Further, the long simulations used in these studies enable us to better understand the forces and factors that shape the 15-TBA molecule relative to available experimental data.

We have specifically attempted to understand the roles of the stem, the loops and the stem-loop mutual influence. The main structural element of 15-TBA is the G-stem, consisting of two G-quartets. In the NMR resolved model, this stem is shielded from water by G8 and T9 above and by the T4–T13 pair below. In the X-ray resolved structure, the stem is shielded only by G8 above and two thymidines that do not interact with each other below. Taken together, there are four structural features that likely influence the stability of the G-stem: the T4–T13 pair under the G-stem, the G8 and T9 bases above the G-stem and a cation residing between the G-quartets. The NMR-based structure incorporates each of these features. The X-ray-based structure, however, has only the T9 base above the G-stem and a cation situated between the G-quartets. The “minimal” two-quartet stem is itself of interest. While stability of the four-quartet stem has been intensely studied by simulations,⁴¹ the two quartet stem likely possesses a substantially different balance of stabilizing forces. Table 2 summarizes simulations carried out to estimate the relative influence of each of the a-fore-mentioned stabilizing factors.

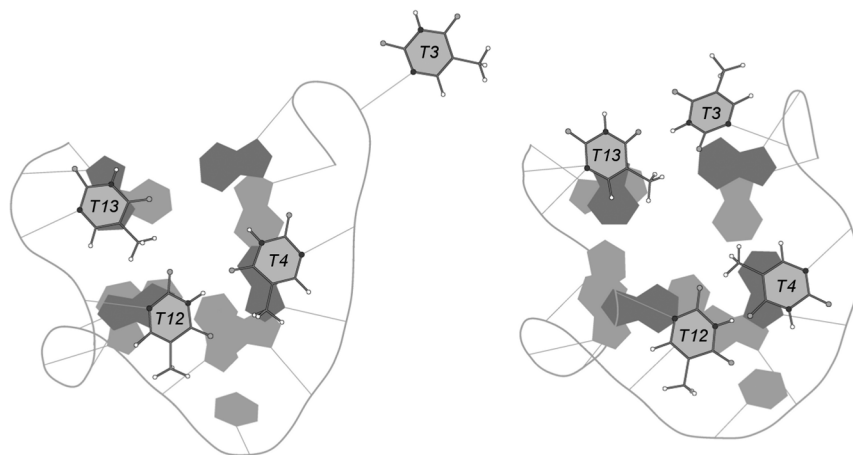


Figure 2. Bottom view of the final MD structure of the NMR-based model of 15-TBA. The structural organization of the TT-loops of the NMR-based model differs depending on the force field used for simulation. In parm99 (left), the T3 thymidine is exposed to solution. After rearrangement, the TT-loops in the parm99 force field acquired, at ~200 ns of MD trajectory, the same geometry as in parmbc0 force field simulation (right). However, the structure was further changed to the final state at 820 ns of the simulation. In parmbc0, the four thymidines are stacked with the lower G-quartet (shown in dark gray) forming T3-T13 and T4-T12 pair interactions. This geometry was adopted by loops in the beginning of the simulation and remained unchanged during the entire 900 ns of MD trajectory. Note that the loops are less accurately described by the force fields compared to stems and the results may be force field dependent.¹⁸ Further, the loops may sample multiple conformations so that long simulations may not be statistically converged.

Table 2. Simulations Carried Out to Test the Importance of Specific Factors That May Contribute to the Stability of 15-TBA

system	what was estimated	result
four-stranded G-stem consisting of two G-quartets with K^+ ion between the G-quartets	the influence of the loops on the G-stem stability	the G-stem was stable during all 700 ns of simulation
four-stranded G-stem without stabilizing cation between the G-quartets ^a	importance of stabilizing cation for the G-stem viability	the G-stem was disrupted during the first ns of simulation
NMR 15-TBA model with substitution of stabilizing K^+ ion to Na^+	the influence of different cation parameters on the 15-TBA behavior and geometry	no significant difference between behavior of Na^+ - or K^+ -stabilized NMR structures
NMR-based 15-TBA model without stabilizing cation between the G-quartets ^a	importance of stabilizing ion for the 15-TBA viability	the model successfully survived, despite fluctuation, until 72 ns of the simulation when bulk Na^+ cation penetrated the center of the G-stem from the bottom through the pore between TT loops (see Figures 2 and 3), fully stabilizing the molecule
NMR-based 15-TBA model with T9 base reoriented away from stacking with the upper G-quartet (TG(-) NMR)	would the NMR model having only G8 in stacking with the upper G-quartet be viable	the G-quadruplex of the model collapsed with loss of G-quartets, despite of T4-T13 pair in initial structure
X-ray TBA model with T9 base reoriented to establish stacking with the upper G-quartet (TG(+) X-ray)	would the X-ray model having additional T9 in stacking with the upper G-quartet be viable	the model survived until 789 ns of simulation. Than T4 and T13 formed stacking interactions with each other, which disturbed planarity of the lower G-quartet and resulted in the collapse of the model

^a That is, there was no ion initially in the channel, while there were obviously ions present in the bulk solvent enabling the stem to capture ions.⁴²

Complexes between Thrombin and 15-TBA.

1:1 Complexes. The X-ray-based conformation of 15-TBA simulated above was derived from the structure of the thrombin-aptamer complex (PDB ID 1hut). It is possible that the protein may have influenced the starting structure and simulation behavior of the oligonucleotide in this case. Thus, the dynamic behavior of the NMR-based and X-ray-based models of the thrombin-aptamer complex was tested with 600 ns of MD in the parmbc0 force field. In the initial structure of the X-ray model of the thrombin-aptamer complex, 15-TBA is anchored on thrombin through the TGT-

loop with the TT-loops exposed to solution. The H-bond donor residues of thrombin formed H-bonds not only with the anchored TGT-loop, but also with the T3 nucleotide from the TT-loops. Simulations using this conformation resulted in collapse of the 15-TBA structure with subsequent loss of the G-quartets. The map of hydrogen bonds between the protein and aptamer changed dramatically. There were only 17% of the initially formed H-bonds remaining during simulation (Table 3).

In the initial structure of the NMR-based complex, 15-TBA is anchored to thrombin through the TT-loops with its

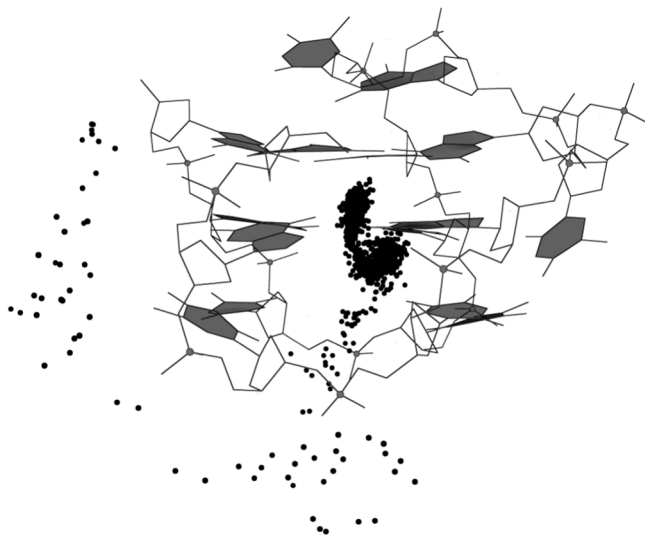


Figure 3. Travel of the Na^+ ion inside the aptamer in the simulation of 15-TBA without stabilizing cation between the G-quartets. This figure represents a period of MD simulation between 60 and 80 ns. The cation (black dots, starting from the left), moving along the phosphodiester backbone of the aptamer, penetrated the interior of 15-TBA between the TT-loops. The Na^+ cation then passed into the G-quadruplex through the lower G-quartet and subsequently remained between the G-quartets for the duration of the simulation. The phosphate atoms of DNA backbone are shown by gray spheres.

top exposed to solution. The resulting MD structure agrees well with the initial structure, with the exception of the orientation of T7 (Figure 4).

The mapping of hydrogen bonds between thrombin and the aptamer reveals much smaller changes of H-bonding pattern than in the simulation using the X-ray-based complex (Table 4). Thirty % of the initially formed H-bonds from the NMR-based structure remained during MD simulation. T3, which is not listed in Table 4, also interacts with thrombin through stacking interactions with Tyr76. Note that the initial structure of the NMR complex is a model.⁹ Consequently, loss of some H-bonds is not surprising as they are not based directly on experimentally derived measurement. Importantly, the simulation easily finds alternative H-bonds with no large shift in the overall structure.

Table 3. H-Bonds Map of the X-ray Complex^a

initial model			dynamical model ^b		
donor	hydrogen	acceptor	donor	hydrogen	acceptor
N3 (T9)	H3 (T9)	OH (Tyr117)	<i>N3 (T7)</i>	<i>H3 (T7)</i>	<i>OH (Tyr117)</i>
N (Ile79)	H (Ile79)	O1P (T9)	<i>N3 (T3)</i>	<i>H3 (T3)</i>	<i>OG (Ser72)</i>
N (Asn78)	H (Asn78)	O1P (T9)	<i>ND2 (Asn78)</i>	<i>HD22 (Asn78)</i>	<i>O3' (T7)</i>
NH2 (Arg77A)	HH22 (Arg77A)	O4' (G10)	<i>ND2 (Asn78)</i>	<i>HD22 (Asn78)</i>	<i>O2P (G8)</i>
NH1 (Arg77A)	HH12 (Arg77A)	O4' (G10)	<i>N (Asn78)</i>	<i>H (Asn78)</i>	<i>O2P (G8)</i>
N (Arg77A)	H (Arg77A)	O3' (G8)	<i>N (Tyr76)</i>	<i>H (Tyr76)</i>	<i>O6 (G8)</i>
NH2 (Arg75)	HH22 (Arg75)	N7 (G1)	NH2 (Arg75)	HH22 (Arg75)	O1P (G8)
NH2 (Arg75)	HH22 (Arg75)	O1P (G8)	NH1 (Arg75)	HH12 (Arg75)	O1P (G8)
NH1 (Arg75)	HH12 (Arg75)	O1P (G8)	<i>N (Arg75)</i>	<i>H (Arg75)</i>	<i>O4 (T3)</i>
NE (Arg75)	HE (Arg75)	N9 (G1)	<i>OG1 (Thr74)</i>	<i>HG1 (Thr74)</i>	<i>O3' (G1)</i>
NE (Arg75)	HE (Arg75)	N7 (G1)	<i>N (Thr74)</i>	<i>H (Thr74)</i>	<i>O4 (T3)</i>
ND1 (Hys71)	HD1 (Hys71)	O2P (G8)			

^a H-bonds that were kept during simulation are marked by bold; new H-bonds are marked by italic. ^b Criterion of H-bond existence in MD are described in Materials and Methods section.

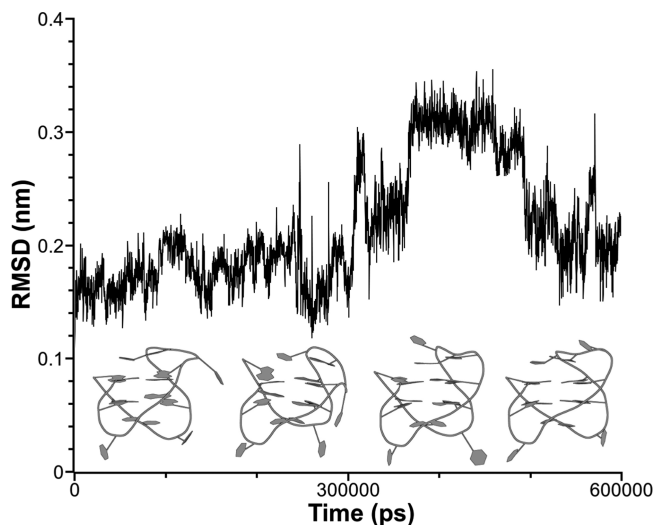


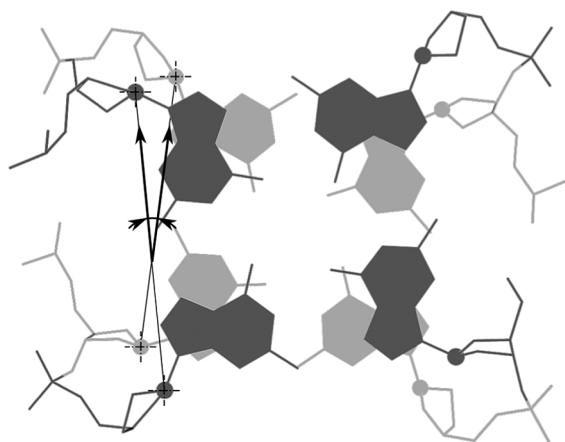
Figure 4. Dynamic behavior of NMR-based 15-TBA in 1:1 complex with thrombin. Snapshots of the aptamer structure are placed at corresponding moments of the trajectory on the rmsd graph. The initial structure of 15-TBA as part of the complex was taken as the reference structure for rmsd calculation, which was made for all atoms of the aptamer. The shift of the rmsd value at 400 ns is related to rearrangement of the TGT-loop; T7 found a new position in stacking with the upper G-quartet.

1:2 Complexes. MD simulation of aptamer-thrombin complexes with 1:2 stoichiometry resulted in only minor changes from the initial models, especially in the case of the NMR-based 15-TBA model. Denoting the thrombin protein from the structure of the aptamer-thrombin complex with 1:1 stoichiometry as thrombin A, and a symmetry-related protein molecule from the neighboring crystal lattice cell as thrombin B, the NMR model of 15-TBA interacts with exosite-I of thrombin A through its TT-loops and with exosite-II of thrombin B through its TGT-loop. These interactions remain unchanged through 600 ns of MD trajectory. T7, which was unbound in the 1:1 complex simulations, interacted with thrombin B. G8 and T9 perfectly shielded the upper G-quartet of the G-stem from charged amino acids and H-bond donors from exosite-II of thrombin B.

Table 4. H-Bond Map of the NMR Complex^a

initial model			dynamical model ^b		
donor	hydrogen	acceptor	donor	hydrogen	acceptor
N3 (T12)	H3 (T12)	OE2 (Glu77)	<i>ND2 (Asn 78)</i>	<i>HD22</i>	<i>O3' (T13)</i>
N3 (T12)	H3 (T12)	O (Glu77)	<i>NH2 (Arg 77A)</i>	<i>HH22</i>	<i>O1P (G14)</i>
OG (Ser153)	HG (Ser153)	O4 (T7)	<i>NH2 (Arg 77A)</i>	<i>HH22</i>	<i>O4' (G14)</i>
OH (Tyr117)	HH (Tyr117)	O1P (T13)	<i>NH1 (Arg 77A)</i>	<i>HH12</i>	<i>O (T13)</i>
N (Asn78)	H (Asn78)	O (T13)	<i>NH1 (Arg 77A)</i>	<i>HH12</i>	<i>O5' (G14)</i>
NE (Arg77A)	HE (Arg77A)	O1P (G14)	<i>NH1 (Arg 77A)</i>	<i>HH12</i>	<i>O4' (G14)</i>
N (Tyr76)	H (Tyr76)	O4' (T4)	N (Tyr 76)	H	O4' (T4)
NH2 (Arg75)	HH22 (Arg75)	O (T4)	NH2 (Arg 75)	HH22	O4 (T13)
NH1 (Arg75)	HH12 (Arg75)	O4 (T13)	NH1 (Arg 75)	HH12	O (T4)
NE (Arg75)	HE (Arg75)	O (T4)	<i>NH1 (Arg 75)</i>	<i>HH12</i>	<i>O4 (T13)</i>

^a H-bonds that were kept during simulation are marked by bold; new H-bonds are marked by italic. ^b Criterion of H-bond existence in MD are described in Materials and Methods section.

**Figure 5.** Definition of the twist angle between the two quartets.

In a similar simulation, the X-ray-based model of 15-TBA interacts with exosite-I of thrombin A through its TGT-loop and with exosite-II of thrombin B through its TT-loops. The G-quartet planarity of the aptamer was disrupted early in the MD trajectory. This disturbance did not however result in overall unfolding of the 15-TBA structure. Multiple contacts of the TT-loops with residues of exosite-II of thrombin B anchored this pole of the aptamer structure preventing structural collapse on the observed simulation time scale (Supporting Information, Figure S5).

Structural Dynamics of the G-Stem. Twist Values Indicate Structural Strain. The twist between two adjacent G-tetrads was chosen as an important structural element of the stem. Twist is represented by the angle between two vectors using C1' atoms of adjacent guanines as the initial and terminal points (Figure 5). We compared the twist angle values and their fluctuations in simulations of free 15-TBA, 15-TBA-protein complexes, and two-quartet stem simulated without the loops (loop-free system). We assumed that the loop-free structure reflects the ideal twist between the two quartets when the stem is not perturbed by other forces.

There are large differences in twist values between the individual structures and in the range of values sampled reflecting the flexibility of the stem (Figure 6). Additionally, there were substantial differences in twist values measured across different grooves. When taking the loop-free stem as a reference, the differences seen in other simulations highlight the influence of the loops on the stem as well as that of

protein binding. It is interesting to note that the range of twist values in the deposited 15-TBA NMR-based structure (12 structures, horizontal blue lines in Figure 6) is quite different from values sampled in the simulation of 15-TBA (black time course). This potentially reflects the effect of the simulation force field, which could shift the optimal twist value relative to experimental structures, as is known to be the case for B-DNA simulations. It is important to note that even if the force field is systematically shifting the absolute twist values the simulations would still properly reflect the relative twist values of different structures. However, the twist range in the NMR-based structure could also be affected by the NMR structure refinement protocol, which consists of a simulated annealing run from 1000 to 75 K in 1000 cycles, followed by energy minimization in the X-PLOR 3.1 system. Note also that the NMR-based structure has substantial deformation (nonplanarity) of the quartets which may indicate some inherent conflicts in the NMR-based data (Figure 7 left), while the MD simulation yields regular quartets.

The 15-TBA loops influence the twist value by substantially restricting the structure and flexibility of the stem. As a result, there is a dramatic reduction in the twist value as well as a reduction in its variability (cf. black and gray lines in Figure 6). Indeed, there is nearly no overlap in twist values sampled in simulations of the loop-free stem and complete 15-TBA. Considering that a 100 ns simulation can, assuming Arrhenius kinetics, sample events differing by as much as 7–8 kcal/mol from the free energy minimum, the influence of the loops is significant. Importantly, complex formation of 15-TBA with thrombin mitigates the influence of the loops on the twist value.

Thrombin's influence on the twist angles corresponding to the narrow grooves of 15-TBA (Figure 6A and C, black and brown lines) is slight. Deviations of the twist angle, both in free 15-TBA and 15-TBA complexed with thrombin, is similar when compared to the loop-free G-stem. Thrombin, however, significantly influences the twist angles corresponding to the wide grooves of 15-TBA (Figure 6, B and D). 15-TBA, complexed with thrombin, has a twist angle that is much closer to that of the loop-free G-stem than to that of the free 15-TBA. In the 15-TBA-thrombin complex, thrombin compensates for the influence of the loops on the stem structure at the wide groove, while there is no similar

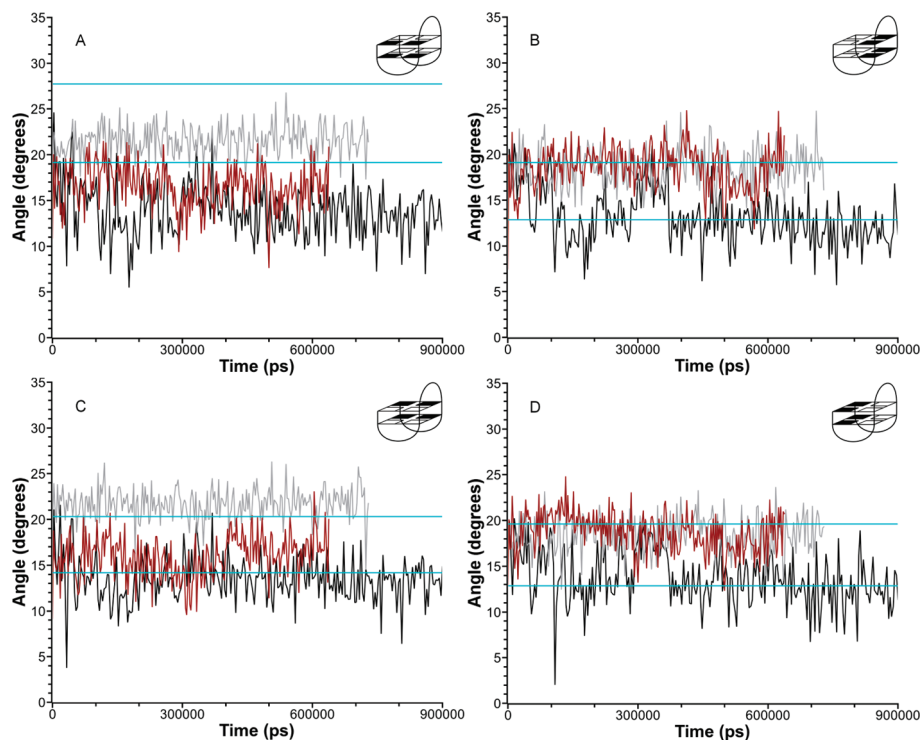


Figure 6. Timed development of the twist angle between the quartets. Twist angle was monitored in different simulations for different grooves, as indicated by the inset structures. Data were smoothed with spline interpolation. Gray: The twist between the two quartets of the four-stranded G-stem without the loops. Black: NMR-based model of 15-TBA in free state. Brown: NMR model in 1:1 complex with thrombin. The two horizontal blue lines demarcate the range of twist values seen in the 12 structures representing the NMR-based model of 15-TBA in PDB entry 148d.

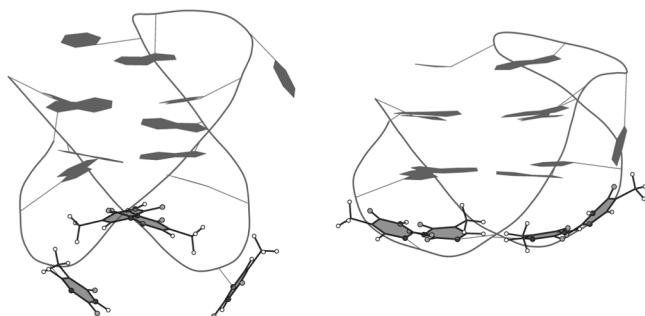


Figure 7. NMR model of 15-TBA before and after MD in parmesc0 force field. Left: Structure of NMR-based model of 15-TBA from PDB entry 148d. Right: Final MD structure of NMR-based model of 15-TBA. Thymidines from the TT-loops are outlined.

compensation in the narrow groove. The only structural elements of 15-TBA that could influence the geometry of the stem at the narrow groove are the TT-loops which display the greatest rearrangement among the structural elements of 15-TBA during simulations (Figure 7). It seems that the origin of the strains that led to collapse of the X-ray-based 15-TBA structure, as well as the modified TG(-T) NMR-based structure, is the initial geometry of the TT-loops. Two modified 15-TBA models were simulated to test this possibility: 1) the resulting MD NMR-based structure with T9 subsequently reoriented away from stacking with the upper G-quartet (TG(-T) NMR^{eq}) and 2) the resulting MD NMR-based structure with T9 and G8 bases reoriented away from stacking with the upper G-quartet (T(-GT) NMR^{eq}). In the case of both starting structures, the TT-loops are in a

conformation that is equilibrated (by the preceding simulations) for minimal negative influence on the G-stem. Both structures successfully survived simulation until the reoriented bases returned to form stacking interactions with the upper G-quartet.

Discussion

The thrombin-binding aptamer (15-TBA) is an intriguing example of a G-DNA containing structure. In addition to its intrinsic affinity for thrombin and potential medicinal value, 15-TBA also represents an important system to study the basic physical chemistry of G-DNA folding and the stabilizing balance of forces. 15-TBA contains the minimum number of G-quartets (just two), raising an interesting question with respect to how is the 15-TBA structure is stabilized? The stability of G-DNA originates primarily from cation-stabilized G-quartet stems. However, the ions within the stem exchange with the bulk solvent. Since the molecule must be regularly exposed to periods when no ion is left in its single ion-binding cavity in the stem, formation of a stable stem with only two quartets is somewhat surprising.

15-TBA contains three short single-stranded loops that maintain proximity between the G-stretches and are essential for thrombin binding. Besides their obvious entropic role (i.e., the difference between forming the stem from either a single strand or from four separate strands), these loops also may exert a direct influence on the stem that may be either stabilizing or destabilizing. Destabilizing effects can arise when the length of the loop is in conflict with the optimal

stem structure while stabilizing effects could be caused by molecular interactions such as base stacking.

Complicating our understanding of 15-TBA structure, conflicting reports based upon X-ray data for the thrombin-aptamer complex and NMR data for free aptamer have been reported. In these disparate reports, there resides a discrepancy in chain orientation. There is also uncertainty with respect to the exact coordination of potassium ions. Marathias and Bolton⁸ suggested that two potassium ions bind 15-TBA, while several studies indicate that a single potassium ion binds 15-TBA.^{43–45} The putative 2:1 binding stoichiometry may represent an additional stabilizing influence for the two-quartet stem during ion exchange with bulk. The MD simulation technique is a well established tool for the study of monovalent ion binding to nucleic acids, including G-DNA. However, there are inherent limitations in the use of MD for this purpose.^{18,19} For example, simple pair-additive force field cannot model polarization of electron clouds, which adversely affects our ability to describe the coordination of ions. Because of this limitation, we elected not to study the detailed difference between the influences of Na⁺ and K⁺ ions on the 15-TBA structure or analyze exact ion binding patterns, that is, the 2:1 versus 1:1 binding. Nevertheless, we did not observe any population of 2:1 binding in our simulations.

In principle, a desirable, precise evaluation of the described interactions could be achieved using combined quantum/molecular mechanics methods.^{46,47} These methods however, are limited in the duration of simulation (dozens of ps) due to substantial computational demand. Though accurate, standard gas phase QM computations on small models are not likely to provide a correct description of the balance of interactions in G-DNA stems. Despite its limitations, MD simulation is capable of substantially contributing to our understanding of the basic role of charge in the 15-TBA quadruplex channel. In the present study, we focused on the comparative study of 15-TBA with published chain orientations, 15-TBA loop-free analogue (two-quartet stem with no loops), and 15-TBA in complex with thrombin using very long MD simulations reaching 12 microsecond of simulation time in total. Our study has two basic parts. We initially investigated the basic properties of free 15-TBA models. We then studied 15-TBA-thrombin complexes using essentially all structural data available in the literature.

The free NMR-based model of 15-TBA is viable in MD simulations, while the X-ray-based model is not. The reason for this difference relates to intramolecular interactions that can either stabilize or disrupt the G-quadruplex structure. Stabilizing interactions refer to the functions of the TGT-loop and a cation that resides within the G-stem. However, the relative importance of these contributions remains uncertain. The NMR-based model of free 15-TBA was viable even when simulated initially without a stabilizing cation. Moreover, the NMR-based model was capable of spontaneously capturing a cation from the bulk to achieve full stabilization. In contrast, the two quartet loop-free stem simulated initially without a bound ion collapsed immediately. However, the two-quartet, loop-free stem was stable if the ion was initially placed into its cavity. Taken

together, these simulations reveal the following stability order of the structures used in simulations: 15-TBA NMR-based structure > two quartet stem > 15-TBA X-ray-based structure. These data clearly show that the overall effect of the three loops is direct energy stabilization of the 15-TBA NMR-based structure.

When we tried to specifically weaken the structure by initially shifting T9 from the TGT loop to disrupt its stacking with the stem, the molecule was destabilized even in the presence of an internal cation. This observation indicates (albeit does not prove) a potentially important stabilizing role for T9. Obviously, the artificial intervention into the starting structure could introduce undesired destabilization (high-energy deformation) because of strained topology of the TGT loop. The simulation might be subsequently unable to repair such a destabilizing interaction. The dynamics of destabilization indicate that the cause of structural collapse is primarily associated with the conformation of the TT-loops. Once the T9 is unstacked, the TGT loop is unable to counterbalance the strain associated with the TT-loops. However, the electronic part of coordination interactions between the ion and solute could not be included via the force field model. Consequently, the description of the ion's stabilizing role may be imprecise.

Our simulations indicate that there should be at least two nucleotides (G8 and T9) stacked with the upper G-quartet for the molecule's viability. The only remaining residue from the TGT-loop that, according to MD, does not take part in any intramolecular interactions is T7. It seems that this residue merely functions to extend the length of the loop. Indeed, Smirnov and Shafer previously reported that three nucleotides is the optimal length for the central loop.⁴⁸ This finding also correlates well with example 11 in the U.S. patent of Griffin et al.,⁴⁹ in which modified forms of 15-TBA containing an abasic nucleotide at each position were synthesized. The only mutant with increased thrombin clotting time relative to the unmodified form was the nonbasic T7 substitution (161 s versus 136 s). Each of the remaining mutants demonstrated decreased thrombin clotting time about 30–50 s.

We have several lines of evidence to support the destabilizing influence of the TT-loops on the 15-TBA structure highlighted by the significant effect of the loops on the twist value of the stem (see below and Figure 6). The loops may be too short and cause strain within 15-TBA. Interestingly, in simulations of the 15-TBA NMR-based model, geometries of the TT-loops are substantially remodeled (Figure 7), indicating that their starting structures are suboptimal and relax during simulation. There are two potential explanations to account for the observed remodeling that we cannot readily discern. First, the starting NMR-based geometry is not perfect and the simulation is remodeling to arrive at the correct structure. Second, the geometry of G-DNA loops can also be influenced by force field approximations, as demonstrated in the literature.¹⁹ This would mean that the simulation again increases the stability of the molecule relative to the starting structure. This improvement, however, would be obtained within the approximation of the force field. Interestingly, when we repeated simulation with T9 unstacking using the

15-TBA starting structure equilibrated by our 900 ns simulation (i.e., having relaxed TT loops), the molecule did not collapse as a result of being weakened by shifting of either T9, or both T9 and G8 from their stacking positions. This 15-TBA starting structure survived until the displaced 15-TBA bases returned to their initial stacking positions.

The TT-loops also serve a critical function in the 15-TBA complex with thrombin. The X-ray-based conformation, if correct, would interact with the protein through its TGT-loop. Despite multiple contacts of the top region of the aptamer with exosite-1 that could stabilize the structure of the oligonucleotide (Table 3), the destabilizing influence of the TT-loops was of greater magnitude. In contrast, in the NMR-based model, the TT-loops are in contact with thrombin, rendering the aptamer stem structure less strained (as indicated by the twist values) than the free structure (Figure 6A–D). This observation not only favors the NMR-based model of the complex but also corresponds with data showing that thrombin can serve as a molecular chaperone for 15-TBA in the absence of stabilizing ions.⁵⁰

MD simulations of aptamer-thrombin complexes with 1:2 stoichiometry demonstrate that these structures are stable and may exist. The 1:2 complex based on the X-ray model of 15-TBA confirmed the importance of the TT-loops for thrombin binding. Multiple contacts of the TT-loops with thrombin exosite II anchored the structure of the aptamer, preventing disintegration during the 600 ns of MD trajectory. Tsiang et al. showed that substitution of the thrombin exosite II residues does not affect inhibition of thrombin activity by the aptamer⁵¹ suggesting that 15-TBA either does not bind exosite II of thrombin or binding of 15-TBA to exosite II is not inhibitory. We are examining the later suggestion in ongoing studies.

The stem in the NMR-based 15-TBA structure is evidently strained (deformed) by the presence of TT-loops as is evidenced by the substantial change of the stem twist angle relative to simulation involving a fully relaxed stem without loops. However, the effect of the loops can be complex, as noted above. Data from these simulations are consistent with the suggestion that the TGT-loop stabilizes the stem while the TT-loops induce strain in the stem structure. In agreement with the study of Baldrich and O'Sullivan,⁵⁰ binding of thrombin apparently reduces the structural strain exerted by the TT-loops on the stem. Notably, each of these considerations is based indirectly on the analysis of structural dynamics. We did not attempt any free energy calculations as we are unaware of a straightforward procedure to perform the necessary calculations using contemporary simulation methods in a reliable manner.

The present study is based on simulations that are 1–2 orders of magnitude longer than those in the preceding G-DNA simulation studies. Such long simulations give us considerably more confidence in the validity of the results. We have seen several changes after extending the individual simulations beyond 50 ns, so this extension is useful and in any case brings a substantial improvement in the reliability of the simulations, at least as far as the sampling is concerned. Still, we think it would not be appropriate to make any definitive conclusions about convergence of the results,

since even 0.1–1.0 μ s simulations are short compared to real conformational changes. Further, for the specific system studied here we do not have the highest-resolution X-ray structures available that would be necessary to rigorously benchmark the simulation data. For the present system, actually, there has been a literature controversy about correctness of some of the experimental structures (our long simulations speak clearly in favor of the NMR structure).

In summary, our simulations suggest the following conclusions. The loops have a stabilizing influence on the 15-TBA molecule. However, the TT-loops (although they help to keep the GG-stretches together) have a destabilizing influence on the stem structure. The TGT-loop, in contrast, appears to be in all aspects stabilizing. However, the TT-loops mediate thrombin binding, an interaction that in addition appears to reduce the conflict between the optimal structure of the stem and the short TT-loops. The simulations described herein strongly support the NMR-based model of 15-TBA. The results provided by this study can aid in the construction of biosensors. A potential design of such a biosensor based upon the results of this study would involve the immobilization of the aptamer through its TGT loop. The exposed TT-loops would then project into solution to bind thrombin.

Acknowledgment. Computer resources were provided by the Research Computing Center of Moscow State University. The supercomputer, “Chebyshev”, was used for all modeling studies. J.S. was supported by the Grant Agency of the Academy of Sciences of the Czech Republic grant IAA400040802, Grant Agency of the Czech Republic grant 203/09/1476, Ministry of Education of the Czech Republic grant LC06030 and Academy of Sciences of the Czech Republic, Grants AV0Z50040507 and AV0Z50040702. R.R. is grateful to Arthur Zalevsky for his help in the preparation of simulations. This research was also supported by Russian Foundation for Basic Research Grants 08-04-01244-a and 08-04-01540-a and Ministry of Education and Science of the Russian Federation Grant 02.512.11.2242.

Supporting Information Available: Representation of thrombin functional sites, illustration of the creation of the 1:2 complex model, “Chebyshev” supercomputer performance characteristics, collapse of the X-ray model in parmbsc0 force field with rmsd plot, the X-ray-based model of 1:2 complex after MD simulation in parmbsc0 force field and a plot of the radius of gyration of G-quadruplex structures in several simulations. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

References

- (1) Nimjee, S. M.; Rusconi, C. P.; Sullenger, B. A. Aptamers: An emerging class of therapeutics. *Annu. Rev. Med.* **2005**, *56*, 555–583.
- (2) Shamah, S. M.; Healy, J. M.; Cload, S. T. Complex target SELEX. *Acc. Chem. Res.* **2008**, *41*, 130–138.
- (3) Tuerk, C.; Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **1990**, *249*, 505–510.
- (4) Bock, L. C.; Griffin, L. C.; Latham, J. A.; Vermaas, E. H.; Toole, J. J. Selection of single-stranded DNA molecules that

- bind and inhibit human thrombin. *Nature*. **1992**, 355, 564–566.
- (5) Di Cera, E. Thrombin. *Mol. Aspects Med.* **2008**, 29, 203–254.
- (6) Schultze, P.; Macaya, R. F.; Feigon, J. Three-dimensional solution structure of the thrombin-binding DNA aptamer d(GGTTGGTGTGGTTGG). *J. Mol. Biol.* **1994**, 235, 1532–1547.
- (7) Mao, X.; Marky, L. A.; Gmeiner, W. H. NMR structure of the thrombin-binding DNA aptamer stabilized by Sr^{2+} . *J. Biomol. Struct. Dyn.* **2004**, 22, 25–33.
- (8) Marathias, V. M.; Bolton, P. H. Structures of the potassium-saturated, 2:1, and intermediate, 1:1, forms of a quadruplex DNA. *Nucleic Acids Res.* **2000**, 28, 1969–1977.
- (9) Padmanabhan, K.; Tulinsky, A. An ambiguous structure of a DNA 15-mer thrombin complex. *Acta Crystallogr. D* **1996**, 52, 272–282.
- (10) Padmanabhan, K.; Padmanabhan, K. P.; Ferrara, J. D.; Sadler, J. E.; Tulinsky, A. The structure of alpha-thrombin inhibited by a 15-mer single-stranded DNA aptamer. *J. Biol. Chem.* **1993**, 268, 17651–17654.
- (11) Kelly, J. A.; Feigon, J.; Yeates, T. O. Reconciliation of the X-ray and NMR structures of the thrombin-binding aptamer d(GGTTGGTGTGGTTGG). *J. Mol. Biol.* **1996**, 256, 417–422.
- (12) Heckel, A.; Mayer, G. Light regulation of aptamer activity: An anti-thrombin aptamer with caged thymidine nucleobases. *J. Am. Chem. Soc.* **2005**, 127, 822–823.
- (13) Mendelboum Raviv, S.; Horváth, A.; Aradi, J.; Bagoly, Z.; Fazakas, F.; Batta, Z.; Muszbek, L.; Hársfalvi, J. 4-Thio-deoxyuridylate-modified thrombin aptamer and its inhibitory effect on fibrin clot formation, platelet aggregation and thrombus growth on subendothelial matrix. *J. Thromb. Haemost.* **2008**, 6, 1764–1771.
- (14) Ikebukuro, K.; Okumura, Y.; Sumikura, K.; Karube, I. A novel method of screening thrombin-inhibiting DNA aptamers using an evolution-mimicking algorithm. *Nucleic Acids Res.* **2005**, 33, e108–e108.
- (15) Tasset, D. M.; Kubik, M. F.; Steiner, W. Oligonucleotide inhibitors of human thrombin that bind distinct epitopes. *J. Mol. Biol.* **1997**, 272, 688–698.
- (16) Pagano, B.; Martino, L.; Randazzo, A.; Giancola, C. Stability and binding properties of a modified thrombin binding aptamer. *Biophys. J.* **2008**, 94, 562–569.
- (17) Fadrná, E.; Špačková, N.; Štefl, R.; Koča, J.; Cheatham, T. E.; Šponer, J. Molecular dynamics simulations of guanine quadruplex loops: advances and force field limitations. *Biophys. J.* **2004**, 87, 227–242.
- (18) Šponer, J.; Špačková, N. Molecular dynamics simulations and their application to four-stranded DNA. *Methods* **2007**, 43, 278–290.
- (19) Fadrná, E.; Špačková, N.; Sarzyńska, J.; Koča, J.; Orozco, M.; Cheatham, T. E.; Kulinski, T.; Šponer, J. Single stranded loops of quadruplex DNA as key benchmark for testing nucleic acids force fields. *J. Chem. Theory Comput.* **2009**, 5, 2514–2530.
- (20) Haider, S.; Parkinson, G. N.; Neidle, S. Molecular dynamics and principal components analysis of human telomeric quadruplex multimers. *Biophys. J.* **2008**, 95, 296–311.
- (21) Hazel, P.; Parkinson, G. N.; Neidle, S. Predictive modelling of topology and loop variations in dimeric DNA quadruplex structures. *Nucleic Acids Res.* **2006**, 34, 2117–2127.
- (22) Cavallari, M.; Calzolari, A.; Garbesi, A.; Di Felice, R. Stability and migration of metal ions in G4-wires by molecular dynamics simulations. *J. Phys. Chem. B.* **2006**, 110, 26337–26348.
- (23) Cheatham, T. E.; 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, 845–862.
- (24) Wang, J.; Cieplak, P.; Kollman, P. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules. *J. Comput. Chem.* **2000**, 21, 1049–1074.
- (25) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1996**, 118, 2309.
- (26) Pérez, A.; Marchán, I.; Svozil, D.; Šponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophys. J.* **2007**, 92, 3817–3829.
- (27) Pérez, A.; Luque, F. J.; Orozco, M. Dynamics of B-DNA on the microsecond time scale. *J. Am. Chem. Soc.* **2007**, 129, 14739–14745.
- (28) Jayapal, P.; Mayer, G.; Heckel, A.; Wennmohs, F. Structure-activity relationships of a caged thrombin binding DNA aptamer: Insight gained from molecular dynamics simulation studies. *J. Struct. Biol.* **2009**, 166, 241–250.
- (29) Golovin, A.; Polyakov, N. OPLS-AA/L force field entries for nucleic acids. <http://mp-group.genebee.msu.su/3d/ff.htm> (accessed Feb 22, 2005).
- (30) *The PyMOL Molecular Graphics System*, version 1.1, Schrödinger LLC. <http://www.schrodinger.com/> (accessed Sep 20, 2010).
- (31) Byrd, R. H.; Lu, P.; Nocedal, J.; Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* **1995**, 16, 1190.
- (32) Ahmed, H. U.; Blakeley, M. P.; Cianci, M.; Cruickshank, D. W. J.; Hubbard, J. A.; Helliwell, J. R. The determination of protonation states in proteins. *Acta Crystallogr. D* **2007**, 63, 906–922.
- (33) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005**, 26, 1701–1718.
- (34) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Comput. Chem.* **2008**, 4, 435–447.
- (35) Sorin, E. J.; Pande, V. S. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys. J.* **2005**, 88, 2472–2493.
- (36) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, 65, 712–725.
- (37) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, 126, 014101–014107.

- (38) 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.
- (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) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (41) Štefl, R.; Cheatham, T. E.; Špačková, N.; Fadrná, E.; Berger, I.; Koča, J.; Šponer, J. Formation pathways of a guanine-quadruplex DNA revealed by molecular dynamics and thermodynamic analysis of the substates. *Biophys. J.* **2003**, *85*, 1787–1804.
- (42) Špačková, N.; Berger, I.; Šponer, J. Structural dynamics and cation interactions of DNA quadruplex molecules containing mixed guanine/cytosine quartets revealed by large-scale MD simulations. *J. Am. Chem. Soc.* **2001**, *123*, 3295–3307.
- (43) Hud, N. V.; Smith, F. W.; Anet, F. A.; Feigon, J. The selectivity for K^+ versus Na^+ in DNA quadruplexes is dominated by relative free energies of hydration: a thermodynamic analysis by 1H NMR. *Biochemistry*. **1996**, *35*, 15383–15390.
- (44) Vairamani, M.; Gross, M. L. G-quadruplex formation of thrombin-binding aptamer detected by electrospray ionization mass spectrometry. *J. Am. Chem. Soc.* **2003**, *125*, 42–43.
- (45) Majhi, P. R.; Qi, J.; Tang, C.; Shafer, R. H. Heat capacity changes associated with guanine quadruplex formation: an isothermal titration calorimetry study. *Biopolymers*. **2008**, *89*, 302–309.
- (46) Car, R.; Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.
- (47) Eichinger, M.; Tavan, P.; Hutter, J.; Parrinello, M. A hybrid method for solutes in complex solvents: Density functional theory combined with empirical force fields. *J. Chem. Phys.* **1999**, *110*, 10452–10467.
- (48) Smirnov, I.; Shafer, R. H. Effect of loop sequence and size on DNA aptamer stability. *Biochemistry* **2000**, *39*, 1462–1468.
- (49) Griffin, L. C.; Albrecht, G.; Latham, J. A.; Leung, L.; Vermaas, E.; Toole, J. J. Aptamers specific for biomolecules and methods of making. U.S. Patent 5756291, May 26, 1998.
- (50) Baldrich, E.; O'Sullivan, C. K. Ability of thrombin to act as molecular chaperone, inducing formation of quadruplex structure of thrombin-binding aptamer. *Anal. Biochem.* **2005**, *341*, 194–197.
- (51) Tsiang, M.; Jain, A. K.; Dunn, K. E.; Rojas, M. E.; Leung, L. L. K.; Gibbs, C. S. Functional mapping of the surface residues of human thrombin. *J. Biol. Chem.* **1995**, *270*, 16854–16863.

CT100253M

JCTC

Journal of Chemical Theory and Computation

Direct Dynamics Implementation of the Least-Action Tunneling Transmission Coefficient. Application to the $\text{CH}_4/\text{CD}_3\text{H}/\text{CD}_4 + \text{CF}_3$ Abstraction Reactions

Rubén Meana-Pañeda,[†] Donald G. Truhlar,[‡] and Antonio Fernández-Ramos^{*†}

Department of Physical Chemistry and Center for Research in Biological Chemistry and Molecular Materials, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain and Department of Chemistry and Supercomputing Institute, University of Minnesota, 207 Pleasant Street SE, Minneapolis, Minnesota 55455-0431

Received May 31, 2010

Abstract: We present two new direct dynamics algorithms for calculating transmission coefficients of polyatomic chemical reactions by the multidimensional least-action tunneling approximation. The new algorithms are called the interpolated least-action tunneling method based on one-dimensional interpolation (ILAT1D) and the double interpolated least-action tunneling (DILAT) method. The DILAT algorithm, which uses a one-dimensional spline under tension to interpolate both of the effective potentials along the nonadiabatic portions of tunneling paths and the imaginary action integrals as functions of tunneling energies, was designed for the calculation of multidimensional LAT transmission coefficients for very large polyatomic systems. The performance of this algorithm has been tested for the $\text{CH}_4/\text{CD}_3\text{H}/\text{CD}_4 + \text{CF}_3$ hydrogen abstraction reactions with encouraging results, i.e., when the fitting is performed using 13 points, the algorithm is about 30 times faster than the full calculation with deviations that are smaller than 5%. This makes direct dynamics least-action tunneling calculations practical for larger systems, higher levels of electron correlation, and/or larger basis sets.

1. Introduction

Hydrogen and proton transfer reactions are among the most prominent reactive processes in chemistry and biology.^{1,2} These reactions are often dominated by quantum mechanical tunneling because the hydrogen atom, due to its small mass, can readily pass through classically forbidden regions of a potential energy surface (PES). Tunneling effects can be taken into account by rigorous quantum mechanical methods,^{3–15} which are only applicable to systems with a small number of atoms, or by Wentzel–Kramers–Brillouin (WKB)-like semiclassical methods,^{16–28} which can handle a large number of atoms. Among the semiclassical methods, variational transition-state theory with multidimensional tunneling corrections (VTST/MT)^{29–42} is the best validated practical

choice for the study of chemical reactions with several atoms because, on the one hand, it has proved to be very accurate when compared with quantum mechanical dynamics calculations^{42,43} and, on the other hand, it needs only semiglobal information about the PES and in many cases is sensitive to the PES only near to the minimum-energy path.

The simplest case for VTST is when the transition-state dividing surface (which is the dynamical bottleneck for reaction) is located at a saddle point and the quantum effects on the reaction coordinate are negligible; in such a case, all the information required for the evaluation of thermal rate constants can be obtained from the reactants and the conventional transition state. In this case VTST/MT can be safely replaced by conventional transition-state theory.⁴⁴ Unfortunately, this is hardly ever the case for hydrogen transfer reactions, which, unless they have no barrier, are usually dominated by tunneling even up to temperatures well above room temperature.^{40,42,43,45–49}

* Corresponding author. E-mail: qf.ramos@usc.es.

[†] University of Santiago de Compostela.

[‡] University of Minnesota.

Even when variational effects (i.e., effects due to the variational transition state not being located at a saddle point) are negligible, the incorporation of quantum effects in the VTST/MT treatment of generalized transition states requires more information about the PES than just reactants and transition-state properties. Quantum effects are incorporated differently for the reaction coordinate, which—for overbarrier processes—is the mode with an imaginary frequency at the saddle point and for the $F - 1$ normal modes of bound motion perpendicular to the reaction coordinate (where $F - 1$ equals $3N - 7$ for nonlinear transition states and $3N - 6$ for linear transition states, where N is the number of atoms, and the reaction coordinate is labeled as mode F). The thermal rate constant calculated by taking into account only the quantum effects on the coordinates in which motion is bound is called quasiclassical, and it is obtained by replacing the classical vibrational partition functions by quantum mechanical ones.^{33,44,50} The quantum effects on the reaction coordinate are taken into account through a transmission coefficient^{21,33,34,43,50–53} that multiplies the quasiclassical thermal rate constant. The evaluation of the transmission coefficient requires the selection of a tunneling path or paths.

As a zeroth approximation, one may assume that the tunneling path coincides with the minimum-energy path (MEP) (the union of the steepest-descent paths in isoinertial coordinates down from the saddle point to reactants and that down to products).^{52,54,55} When zero-point effects are taken into account for bound motions transverse to the MEP, this assumption yields the zero-curvature tunneling (ZCT) approximation.⁵² The signed distance from the saddle point along the MEP will be called the reaction coordinate, even though the dominant dynamical path may be offset from the MEP. The MEP is tangent to the imaginary frequency normal mode at the saddle point, so this definition coincides with defining the reaction coordinate in the vicinity of the saddle point as the distance along that mode.

It has been shown that the ZCT path, i.e., the MEP, is a poor choice as a tunneling path^{42,56} since it does not take account of the MEP's curvature, which couples the reaction coordinate to the other vibrational modes. The curvature has the effect that the dominant tunneling path is on the concave side of the reaction path, and depending on the magnitude of the curvature, tunneling is better treated by the small-curvature tunneling (SCT) approximation^{57–60} or by the large-curvature tunneling (LCT) approximation^{23,37,41,59,61–65} for the cases of small and large couplings, respectively (for collinear atom–diatom reactions with very small curvature one could also use the Marcus–Coltrin approximation).²⁰ The path implied by the SCT approximation is not uniquely defined because the calculation is carried out in terms of an effective mass for tunneling along the MEP rather than using the true reduced mass along a tunneling path; the curvature-dependent effective mass is smaller than the true reduced mass to account for shortening of the tunneling path by corner cutting. The LCT approximation, in contrast, involves for every energy an explicit sequence of paths chosen as the straight lines that join equipotential points on the reactant and product sides of the vibrationally adiabatic potential curves along the MEP. Neither the SCT nor the LCT is

variational; rather they represent limiting cases. However, the tunneling fluxes predicted by the SCT and LCT approximations roughly overlap for intermediate curvature, so they more or less cover the whole range of curvatures. It is reasonable to define a new tunneling probability that, at every tunneling energy, gives the larger of the SCT and LCT tunneling probabilities. This result is called the (microcanonically) optimized multidimensional tunneling probability (μ OMT or, for short, OMT).⁶⁴ Note that the ZCT, SCT, LCT, and OMT tunneling approximations are all multidimensional in that they all include the important effect that the vibrational zero point energy (or, in the LCT and OMT approximations, also excited-state quantized vibrational energies) depends upon the distance along the reaction path or tunneling path; thus, the reaction coordinate is not separable in these approximations, and this mimics VTST in removing one of the major approximations of conventional transition-state theory. For this reason, it is most appropriate to apply these approximations in the context of VTST rather than conventional transition-state theory. The SCT, LCT, and OMT approximations include multidimensional effects not only in the vibrational energy requirements along the tunneling path but also in the choice of the tunneling path.

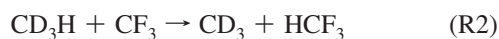
Very recently²⁸ we have generalized to polyatomic reactions the least-action tunneling (LAT) path, which was initially developed by Garrett and Truhlar for atom–diatom reactions.²² In this approximation, one considers, for each tunneling energy and final vibrational state, a sequence of paths parametrized by a unique parameter. These paths are all located at or between the MEP and LCT paths. At every tunneling energy, the path is variationally optimized within this sequence by choosing it as the path with the largest tunneling probability. For this reason, in principle, the LAT transmission coefficients should be more accurate than those obtained by the μ OMT approximation.

The current version of the LAT method for ground-state transmission coefficients (which are used to calculate thermally averaged rate constants)⁴¹ may be called the least-action ground-state tunneling method, version 4, (LAG4) because we always base the transmission coefficients for thermal reactions on a ground-state transmission coefficient (computed in the exoergic direction) and because the LCT-like portions of the calculation are based on version 4 of the LCT method.⁶⁵ (Note that, although the reactant is in the ground state for the prototype tunneling process on which the calculation of the thermally averaged rate is based, a range of vibrational states is populated in the product of the tunneling event, if the energy is high enough to populate dynamically coupled states in the product valley, and tunneling for excited-state reactants is approximated in terms of the ground-state tunneling probabilities and the quantized threshold energies at the variational transition state.) The present article is concerned with the calculation of LAG4 transmission coefficients, and we will simply abbreviate them as LAT. Similarly we use LCT as shorthand for LCG4.

The LCT, μ OMT, and LAT transmission coefficients are more computationally intensive than the SCT one because, whereas the SCT transmission coefficient can be obtained from a very limited knowledge of the PES, i.e., from

information calculated along the MEP (including its curvature and local force constants for motion transverse to the MEP), the calculation of the other transmission coefficient approximations requires information not only along the MEP but also in the wide region on the concave side of the MEP. This region is called the reaction swath,^{66,67} and it is the region through which LCT and LAT tunneling paths pass. The LCT, μ OMT, and LAT transmission coefficients involve the calculation of a potentially large number of points of the PES in the reaction swath. The development of faster computers and more accurate density functionals has made it possible in many cases to evaluate the energy reasonably accurately at those geometries by direct dynamics, which allows “the calculation of rates or other observables directly from electronic structure information without intermediacy of fitting the electronic energies in the form of a potential energy surface”.⁶⁸ Direct dynamics together with VTST/MT is a powerful combination that, for instance, is being widely used in the evaluation of thermal rate constants and kinetic isotope effects (KIEs) of many enzymatic reactions.⁶⁹

However, the calculation of LCT and LAT transmission coefficients by direct dynamics is still computationally very demanding if one uses the original algorithms. For that reason we developed an algorithm (called ILCT2D) based on a two-dimensional spline-under-tension,⁷⁰ to evaluate LCT tunneling probabilities with a reduction in the computer time by a factor of about 30.⁷¹ The error with respect to the full calculation is less than 1%. It is the objective of the present work to present an analogous efficient algorithm based on spline-under-tension interpolations for calculation of LAT transmission coefficients using direct dynamics, and we will present two such algorithms. To show the performance of the new algorithms, we have chosen the following set of hydrogen abstraction reactions:



which we have previously studied using the SCT, LCT (with the ILCT2D algorithm), and μ OMT approximations for tunneling.⁷¹ The calculated thermal rate constants were in good agreement with experimental data. However, the calculated KIEs were too low, particularly those for the ratio R2/R3. In this paper, in addition to developing a more efficient algorithm for LAT calculations, we use it to apply the LAT approximation to these reactions to see if this method improves the previous results.

Section 2 presents a general description of the evaluation of tunneling transmission coefficients and presents the new interpolation schemes used for efficient direct dynamics calculations of LAT transmission coefficients. Section 3 describes the performance of those interpolation schemes for reactions R1–R4. Section 4 has concluding remarks.

2. Methodology

The VTST/MT thermal rate constant^{37,41} can be written as the quasiclassical canonical variational theory (CVT) rate constant k^{CVT} multiplied by the tunneling transmission coefficient, $\kappa^{\text{CVT/X}}$, i.e.,

$$k^{\text{CVT/X}}(T) = \kappa^{\text{CVT/X}}(T)k^{\text{CVT}}(T) \quad (1)$$

where X stands for the ZCT,⁵² SCT,⁶⁰ LCT,^{22,64,65} μ OMT,⁶⁴ or LAT^{22,28} approximations for tunneling. In general $\kappa^{\text{CVT/X}}$ is equal to a so-called CAG factor (called $\kappa^{\text{CVT/CAG}}$ and almost always very close, within about 15%, to unity) times a more universal transmission coefficient called κ^{X} . Since the CAG factor is explained in detail elsewhere,^{21,37,41} we concentrate here on κ^{X} .

The ZCT approximation assumes that the reaction coordinate is adiabatically separated from the $F - 1$ other degrees of freedom and that all the excited-state vibrationally adiabatic potentials that significantly contribute to tunneling have the same shape as the ground-state vibrationally adiabatic potential $V_a^{\text{G}}(s)$, which is given by

$$V_a^{\text{G}}(s) = V_{\text{MEP}}(s) + \varepsilon_{\text{vib}}^{\text{G}}(s) \quad (2)$$

where s is the reaction coordinate mentioned in Section 1 (it measures progress along the isoenergetic MEP, being negative on the reactants side, zero at the saddle point, and positive on the products side, where the isoenergetic coordinates are scaled to a reduced mass of μ); $V_{\text{MEP}}(s)$ is the potential along the MEP; and $\varepsilon_{\text{vib}}^{\text{G}}(s)$ is the local zero-point vibrational energy. The other tunneling approximations also involve the $V_a^{\text{G}}(s)$ but in more complicated ways.

The lowest energy possible to have tunneling is the energy of the reactant zero-point level when the reaction is written in the exoergic direction; this is called E_0 . The transmission coefficient is given by

$$\kappa^{\text{X}}(T) = \beta \exp(\beta V_a^{\text{AG}}) \int_{E_0}^{\infty} dE P^{\text{X}}(E) \exp(-\beta E) \quad (3)$$

where $\beta = (k_{\text{B}}T)^{-1}$, k_{B} is the Boltzmann constant, and T is the temperature; V_a^{AG} is the maximum of the ground-state vibrationally adiabatic potential; and $P^{\text{X}}(E)$ is the ground-state semiclassical probability at energy E , which is approximated in the ZCT and SCT approximations as

$$P^{\text{X}}(E) = \begin{cases} 0, & E < E_0 \\ \{1 + \exp[2\theta(E)]\}^{-1}, & E_0 \leq E \leq V_a^{\text{AG}} \\ 1 - P^{\text{X}}(2V_a^{\text{AG}} - E), & V_a^{\text{AG}} \leq E \leq 2V_a^{\text{AG}} - E_0 \\ 1, & 2V_a^{\text{AG}} - E_0 < E \end{cases} \quad (4)$$

where $\theta(E)$ is the imaginary part of the action integral. When $X = \mu$ OMT, the tunneling probabilities are obtained as⁶⁴

$$P^{\mu\text{OMT}} = \max_E \left\{ P^{\text{SCT}}(E), P^{\text{LCT}}(E) \right\} \quad (5)$$

where P^{LCT} is obtained from a more complicated expression than P^{SCT} .

In the LCT and LAT approximations, one must sum over tunneling probabilities from the ground state of the reactants to all accessible diabatic vibrational states of the product. In many cases, only the ground-state-to-ground-state process needs to be considered. Even when the excited states of the product must be considered, it is sufficient to consider the ground-state-to-ground-state case to explain the new algorithms being introduced here, and so we limit our consideration to the ground-state-to-ground-state case. (We previously found⁷¹ that tunneling into excited vibrational states does not make a large contribution for the reactions under consideration here.)

For a given tunneling path, the imaginary part of the action integral is given by

$$\theta(E) = \hbar^{-1} \int_{\xi_0}^{\xi_1} \text{Im } p(\xi) d\xi \quad (6)$$

where ξ is a progress variable along the tunneling path; ξ_0 and ξ_1 mark the beginning and end of the tunneling path, respectively; and $\text{Im } p(\xi)$ is the imaginary part of the momentum in the tunneling direction, which is written as

$$p(\xi) = \{2\mu_{\text{eff}}(\xi)[E - V_{\text{eff}}(\xi)]\}^{1/2} \quad (7)$$

where $\mu_{\text{eff}}(\xi)$ and $V_{\text{eff}}(\xi)$ are respectively the effective reduced mass and the effective potential along the tunneling path. The calculation of the transmission coefficient of eq 3 requires the evaluation of tunneling probabilities at several energies, and these depend on the tunneling paths. For X = ZCT the tunneling path coincides with the MEP, and therefore the progress variable along the path is s , and the effective potential is given by the ground-state vibrationally adiabatic potential given by eq 2. The effective mass $\mu_{\text{eff}}(s) = \mu$. Therefore, in the ZCT approximation the action integral, at every tunneling energy, is given by

$$\theta(E) = \hbar^{-1} \int_{\tilde{s}_0}^{\tilde{s}_1} ds \{2\mu(V_a^G(s) - E)\}^{1/2} \quad (8)$$

where \tilde{s}_0 and \tilde{s}_1 are the classical turning points in the reactant and product valleys, respectively. Both turning points obey the resonance condition:

$$V_a^G(\tilde{s}_0) = V_a^G(\tilde{s}_1) = E \quad (9)$$

and therefore it is equivalent to write $\theta(E)$ or $\theta(\tilde{s}_0)$ in eq 8.

The coupling between the reaction coordinate and the $F - 1$ other modes produces an internal centrifugal effect that shortens the dominant tunneling path at a given energy by displacing it toward the concave side of the MEP. The SCT approximation incorporates this effect in the effective mass for tunneling without an explicit evaluation of the tunneling path. The SCT action integral is given by

$$\theta(E) = \hbar^{-1} \int_{\tilde{s}_0}^{\tilde{s}_1} ds \{2\mu_{\text{eff}}(s)(V_a^G(s) - E)\}^{1/2} \quad (10)$$

It should be noticed that now the effective mass depends on the progress along the MEP and that $\mu_{\text{eff}} \leq \mu$. For this reason the SCT transmission coefficient is always larger or equal to the ZCT transmission coefficient.

To evaluate the LAT tunneling probability, one must calculate the action integrals of a family of tunneling paths that depend on a parameter α . These paths correspond to the MEP when $\alpha = 0$ and to the LCT path, which is a straight path, for $\alpha = 1$. Let $\xi_p(0)$ be the length of the tunneling path along the MEP from \tilde{s}_0 to \tilde{s}_1 (this is equal to $\tilde{s}_1 - \tilde{s}_0$), and let $\xi_p(1)$ be the length of the straight-line path, which is shorter. Then, the geometry of a point on the path with parameter α is given by

$$\mathbf{x}[\alpha, \xi(\alpha), \tilde{s}_0] = (1 - \alpha)\mathbf{x}[0, \xi(0), \tilde{s}_0] + \alpha\mathbf{x}[1, \xi(1), \tilde{s}_0] \quad (11)$$

where $\mathbf{x}[0, \xi(0), \tilde{s}_0]$ and $\mathbf{x}[1, \xi(1), \tilde{s}_0]$ are respectively geometries on the MEP and on the straight path; thus $\xi(1)$ is equal to $\xi(0)$ times the ratio of $\xi_p(1)$ to $\xi_p(0)$. Consequently, the progress variable ξ depends on the value of the α parameter, and $\xi(1)$ is less than or equal to $\xi(\alpha)$, which is less than or equal to $\xi(0)$. The probabilities along the series of paths of eq 11 may involve regions of the PES that are vibrationally nonadiabatic (see refs. 28 and 41 for details), so in general the action integral is split into three terms:

$$\theta(\alpha, E) = \theta_I(\alpha, E) + \theta_{\text{II}}(\alpha, E) + \theta_{\text{III}}(\alpha, E) \quad (12)$$

The action integrals $\theta_i(\alpha, E)$, with $i = \text{I}$ and III , correspond to the adiabatic regions on the reactants ($i = \text{I}$) and products ($i = \text{III}$) side, respectively, and they are given by the following expressions:

$$\theta_I(\alpha, E) = \hbar^{-1} \int_0^{\xi_i(\alpha)} d\xi(\alpha) \{V_a^G[s_I(0, \xi(0)); \tilde{s}_0]\} - V_a^G(\tilde{s}_0)^{1/2} \cos \chi_0 \quad (13)$$

$$\theta_{\text{III}}(\alpha, E) = \hbar^{-1} \int_{\xi_{\text{III}}(\alpha)}^{\xi_p(\alpha)} d\xi(\alpha) \{V_a^G[s_{\text{III}}(0, \xi(0)); \tilde{s}_0]\} - V_a^G(\tilde{s}_0)^{1/2} \cos \chi_1 \quad (14)$$

The total length of the path is $\xi_p(\alpha)$; and the values $\xi_i(\alpha)$ $i = \text{I}$ and III indicate boundaries of the adiabatic region. Each of the $s_i(0, \xi(0))$, $i = \text{I}$ and III values needed for the evaluation of the vibrationally adiabatic potentials $V_a^G[s_i(0, \xi(0)); \tilde{s}_0]$, is obtained in such a way that the vector defined by the geometry $\mathbf{x}[\alpha, \xi(\alpha), \tilde{s}_0]$ and the reaction path geometry $\mathbf{x}[0, \xi(0), \tilde{s}_0]$ is perpendicular to the derivative of $\mathbf{x}[0, \xi(0), \tilde{s}_0]$ with respect to s at that s value, i.e.,

$$\{\mathbf{x}[\alpha, \xi(\alpha), \tilde{s}_0] - \mathbf{x}[0, \xi(0), \tilde{s}_0]\} \cdot \frac{d\mathbf{x}[0, \xi(0), \tilde{s}_0]}{ds} = 0 \quad (15)$$

The angles between the gradient and tangent vector to the path at \tilde{s}_0 and \tilde{s}_1 are χ_0 and χ_1 , respectively. If the entire path is adiabatic (i.e., if there is no region II), then there will be overlap between regions I and III in the interval $\xi_{\text{III}}(\alpha) \leq \xi(\alpha) \leq \xi_I(\alpha)$, and the vibrationally adiabatic potential in that region is taken to be

$$\min\{V_a^G[s_I(0, \xi(0)); \tilde{s}_0], V_a^G[s_{\text{III}}(0, \xi(0)); \tilde{s}_0]\} \quad (16)$$

The action integral through the nonadiabatic region is given by

$$\theta_{\text{II}}(\alpha, E) = \hbar^{-1} \int_{\xi_1(\alpha)}^{\xi_{\text{III}}(\alpha)} d\xi(\alpha) \{ V_{\text{eff}}^{\text{II}}(\alpha, \xi(\alpha), \tilde{s}_0) - V_a^G(\tilde{s}_0) \}^{1/2} \quad (17)$$

The effective potential $V_{\text{eff}}^{\text{II}}(\alpha, \xi(\alpha), \tilde{s}_0)$ is obtained from

$$V_{\text{eff}}^{\text{II}}(\alpha, \xi(\alpha), \tilde{s}_0) = V\{\mathbf{x}[\alpha, \xi(\alpha), \tilde{s}_0]\} + V_{\text{corr}}^{\text{I}}(\alpha, \xi_1(\alpha), \tilde{s}_0) + V_{\text{anh}}^{\text{I}}(\alpha, \tilde{s}_0) + \frac{\xi(\alpha) - \xi_1(\alpha)}{\xi_{\text{III}}(\alpha) - \xi_1(\alpha)} [V_{\text{corr}}^{\text{III}}(\alpha, \xi_{\text{III}}(\alpha), \tilde{s}_0) - V_{\text{corr}}^{\text{I}}(\alpha, \xi_1(\alpha), \tilde{s}_0) + V_{\text{anh}}^{\text{III}}(\alpha, \tilde{s}_0) - V_{\text{anh}}^{\text{I}}(\alpha, \tilde{s}_0)] \quad (18)$$

In this expression the potentials $V_{\text{corr}}^i(\alpha, \xi_i(\alpha), \tilde{s}_0)$, $i = \text{I}$ and III correct for the zero-point energy in the modes that still behave adiabatically. The potentials $V_{\text{anh}}^i(\alpha, \tilde{s}_0)$ incorporate anharmonic nonquadratic corrections to the effective potential in the same way as in eq 5 of the LCT method.⁶⁵ The geometries $\mathbf{x}[\alpha, \xi(\alpha), \tilde{s}_0]$, needed for the evaluation of the classical potential $V\{\mathbf{x}[\alpha, \xi(\alpha), \tilde{s}_0]\}$, are obtained from the straight path joining the geometries $\mathbf{x}[\alpha, \xi_1(\alpha), \tilde{s}_0]$ and $\mathbf{x}[\alpha, \xi_{\text{III}}(\alpha), \tilde{s}_0]$, i.e.,

$$\mathbf{x}[\alpha, \xi(\alpha), \tilde{s}_0] = \mathbf{x}[\alpha, \xi_1(\alpha), \tilde{s}_0] + \frac{\xi(\alpha) - \xi_1(\alpha)}{\xi_{\text{III}}(\alpha) - \xi_1(\alpha)} (\mathbf{x}[\alpha, \xi_{\text{III}}(\alpha), \tilde{s}_0] - \mathbf{x}[\alpha, \xi_1(\alpha), \tilde{s}_0]) \quad (19)$$

We note that the LCT expressions are obtained for the effective potential of eq 18 and the action integrals of eqs 12–14 and 17 when $\alpha = 1$.

Converged probabilities at every tunneling energy are obtained by the numerical integration of eqs 13, 14, and 17 at N points along the path; for the present work we set $N = 180$. If there is a nonadiabatic region, then N_i of those N points belong to region I, N_{II} are in region II, and N_{III} are in region III. The potential in regions I or III is obtained from the vibrationally adiabatic potential along the MEP. However, the effective potential at the geometries obtained from eq 19 requires single-point calculations of the potential energy at points $\xi_i(\alpha)$, $i = 1, \dots, N_{\text{II}}$, where $\xi_1(\alpha) = \xi_1(\alpha)$ and $\xi_{N_{\text{II}}}(\alpha) = \xi_{\text{III}}(\alpha)$. We found that the evaluation of the LCT transmission coefficients by the interpolated large-curvature tunneling algorithm based on one-dimensional interpolation (ILCT1D)⁷² of these potential energies almost perfectly reproduces the specifically calculated potentials $V\{\mathbf{x}[1, \xi(1), \tilde{s}_0]\}$ when the N_{II} points are replaced by $N_s = 9$ equally spaced points, which are interpolated by a one-dimensional spline-under-tension.^{73,74} If, at a given tunneling energy, $N_{\text{II}} < 9$, then no interpolation is carried out along the nonadiabatic region of the tunneling path. Similarly, the above procedure can be used to obtain the LAT transmission coefficients; the only difference being that now N_s points are used to evaluate the α -dependent effective potential of eq 18. We call this algorithm the interpolated least-action tunneling method based on one-dimensional interpolation (ILAT1D).

The calculation of transmission coefficients by the full-LAT method (without any interpolation) using direct dynamics requires a large amount of computer time, so we tested the performance of the ILAT1D algorithm by using analytical PESs. We used the same analytical PESs as for the testing of the ILCT1D method and found that the mean unsigned

percentage error (MUPE) of the ILAT1D algorithm with respect to a full LAT calculations (as a reference) is smaller than 0.20% in the interval from $T = 200$ –400 K (see Supporting Information for further details of these tests). Therefore, we believe that the transmission coefficients obtained by the ILAT1D algorithm can safely replace full LAT calculations without loss of accuracy. Hereafter, we use the ILAT1D algorithm as a reference in the development of more approximate algorithms, as discussed below.

The ILAT1D algorithm is still very expensive in computer time, since the value, $\tilde{\alpha}$, of α that minimizes the action of eq 12 is obtained by a golden section search⁷⁵ at every tunneling energy. Therefore, we developed another, even more efficient algorithm that further reduces the number of tunneling energies at which the least-action integral, $\theta(\tilde{\alpha}, E)$, has to be explicitly computed. The method is described next.

In the ILAT1D algorithm the least-action integral is evaluated at tunneling energies E_i , $i = 1, \dots, M$, with E_1 being the lowest energy at which it is possible to locate the classical turning points that determine the straight path and M being the number of tunneling energies (all below the maximum of the vibrationally adiabatic potential) at which the tunneling probabilities are calculated. In general, one has $40 \leq M \leq 80$. The tunneling energies that are computationally most demanding are those for which there is a nonadiabatic region for some of the possible tunneling paths. It should be noticed that the absence of nonadiabatic region at a given tunneling energy along the LCT path means also the absence of nonadiabatic regions at any of the α -dependent paths at that tunneling energy and makes the LAT and LCT probabilities coincide. The potential in the adiabatic region is readily available, because it can be obtained from information along the MEP. Therefore the effort in developing the more efficient algorithm is focused on the E_i , $i = 1, \dots, M_{\text{II}}$ tunneling energies with nonadiabatic regions along the straight path, where $E_{i=M_{\text{II}}}$ is the highest tunneling energy for which there is a nonadiabatic region along the LCT path. It is possible to reduce computer time by explicitly evaluating the least-action integral at M_s tunneling energies instead of at M_{II} tunneling energies. The remaining least-action integrals are obtained implicitly by interpolation with a one-dimensional spline-under-tension. The M_s tunneling energies are chosen in such a way that E_1 and $E_{M_{\text{II}}}$ are the first and last energies of the fit, respectively, and the remaining $M_s - 2$ energies are taken as equally spaced between those two values. (We also considered interpolating $\tilde{\alpha}$ instead of $\theta(\tilde{\alpha}, E)$, but we found, as shown in Section 3, that the latter is a much better choice because $\theta(\tilde{\alpha}, E)$ changes smoothly with the tunneling energy.) In summary, the new algorithm uses one-dimensional interpolation for both the tunneling paths and the optimized action integrals, and therefore we call the method the double interpolated least-action tunneling (DILAT) method.

The remaining steps are explained fully in previous discussions of the LCT and LAT methods^{28,41,59,76} and so are only briefly summarized here. The least-action integrals obtained at every tunneling energy are used to compute tunneling amplitudes defined by

$$T_{\text{tun}}^{\text{LAT}}(\tilde{\alpha}, \tilde{s}_0) = T_{\text{tun}}^{\text{LAT}}(\tilde{\alpha}, \tilde{s}_1) = \exp[-\theta(\tilde{\alpha}, \tilde{s}_0)] \quad (20)$$

The LAT primitive probability at every tunneling energy, using either the ILAT1D or DILAT algorithm, is obtained from the tunneling amplitude of eq 20 plus the contribution due to the vibrational motion perpendicular to the reaction coordinate along the incoming and outgoing trajectories. The LAT primitive probability is then uniformized such that it goes to half of the maximum of the ground-state vibrationally adiabatic potential. The resulting LAT tunneling probabilities are also used for the calculation of the nonclassical over-the-barrier tunneling probabilities, as in eq 4.

Note that the μOMT transmission coefficient is always greater than or equal to both the SCT and LCT ones, and the LAT transmission coefficient is always greater than or equal to the LCT one. However, the LAT transmission coefficient can be either greater or smaller than the SCT one, because the LAT paths lie between the MEP and LCT paths, but the LAT method does not incorporate the small-curvature limit explicitly.

A full calculation of the LAT transmission coefficients scales as $M \times N \times L$, since for each of the M tunneling energies, we need to perform L iterations to obtain a converged least-action integral on tunneling paths obtained with N single-point calculations. Typical values for these parameters are $M = 60$, $N = 180$, and $L = 25$, which involves approximately 3×10^5 single-point energy calculations. Many of those points fall in the adiabatic regions, so they can be readily calculated from the information available along the MEP, and only the evaluation of the effective potential of nonadiabatic region II requires additional direct dynamics electronic structure calculations. The number of points in the nonadiabatic region in the full calculation would be $M_{\text{II}} \times \bar{N}_{\text{II}} \times L$, where \bar{N}_{II} is the average of nonadiabatic points at every tunneling energy. The size of the nonadiabatic region depends on the PES, but reasonable numbers for M_{II} and \bar{N}_{II} are 40 and 50, respectively, and therefore the number of single-point calculations in the nonadiabatic region is approximately 5×10^4 . The ILAT1D algorithm reduces the number of single-point calculations to $M_{\text{II}} \times N_s \times L$ such that it requires approximately 9000 single-point calculations in the nonadiabatic region. The DILAT algorithm reduces further the number of single-point calculations by the ratio M_{II}/M_s .

The ILAT1D and the DILAT algorithms for the calculation of LAT transmission coefficients using direct dynamics have been implemented in a development version of POLYRATE,⁷⁷ and we plan that they will be made available in an upcoming new release version of the program.

3. Applications

The electronic structure calculations needed for the evaluation of LAT transmission coefficients for reactions R1–R4 using the ILAT1D and DILAT algorithms were performed with the MPWB1K⁷⁸ density functional using the 6-31+G(d,p) basis set.⁷⁹ The details of the electronic structure calculations can be found in ref 71. The previous calculations showed that the maximum of the vibrationally adiabatic potential

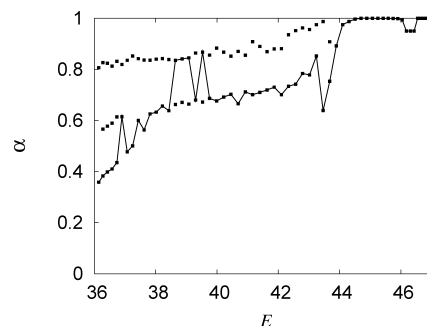


Figure 1. Plot of the α parameter versus the tunneling energy (E , in $\text{kcal}\cdot\text{mol}^{-1}$) relative to the reactants at their equilibrium separation without zero-point energy. The dots correspond to the values of α at which there are local minima of the imaginary action integral at all of the calculated tunneling energies. The solid line joins the global minimum of the α parameter, $\tilde{\alpha}$, at every tunneling energy.

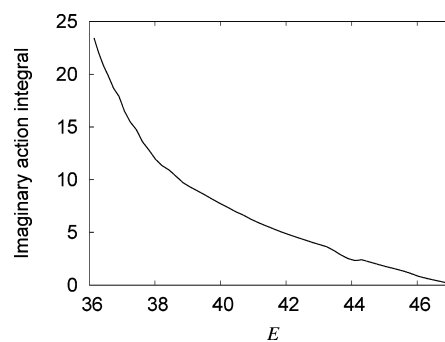


Figure 2. The solid line shows the variation of the imaginary action integral, $\theta(\tilde{\alpha}, E)$, of eq 12 with the tunneling energy (in $\text{kcal}\cdot\text{mol}^{-1}$).

occurs at the saddle point, that is, $V_a^{\text{AG}} = V_a^{\ddagger\text{G}}$, and that there are no variational effects in the interval of temperatures between 200 and 700 K, so the variational dividing surface is located at the conventional transition state. Besides, $\kappa^{\text{CVT}/\text{X}}$ of eq 1 equals κ^{X} because $\kappa^{\text{CVT}/\text{CAG}} = 1$.

Figure 1 shows that the values of $\tilde{\alpha}$ at different tunneling energies may change abruptly, which makes a fit of $\tilde{\alpha}$ as a function of tunneling energy quite difficult and inaccurate, so instead we chose to interpolate the action integrals corresponding to $\tilde{\alpha}$. It is noteworthy that there are several tunneling energies with two or even three local minima for the imaginary action integral. However, even in this difficult case, an interpolation of the least-action integral as a function of tunneling energy is easier to perform than an interpolation of $\tilde{\alpha}$ due to the smooth behavior of $\theta(\tilde{\alpha}, E)$, as shown in Figure 2.

Table 1 lists the number of single-point calculations needed to evaluate the effective potential of eq 18 for R1 for the calculation of the LAT transmission coefficients with the ILAT1D and DILAT algorithms (the number of points for R2–R4 is similar to R1 and is not shown in the table). We use as reference calculations those obtained by the ILAT1D algorithm. The ILAT1D algorithm allows one to obtain LAT transmission coefficients 7.5 times faster than the full LAT calculation. In this case the DILAT algorithm is 50 and 20

Table 1. Number of Single-Point Calculations (NSP) in the Nonadiabatic Region Needed for the Calculation of the LAT Transmission Coefficients with Selected Values of M_S for Reaction R1

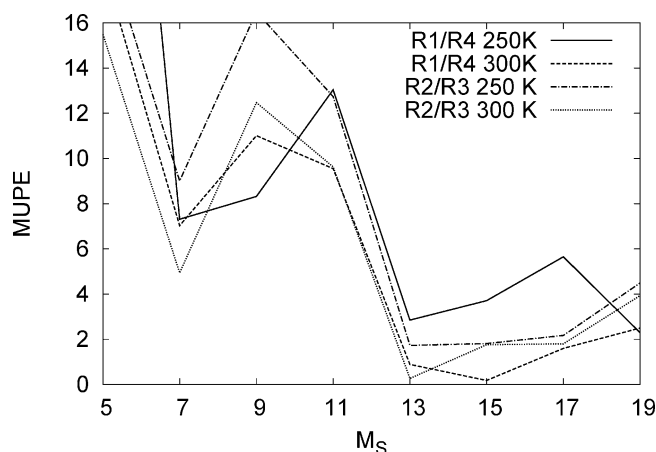
level	M_S	NSP
full	43	137 895
ILAT1D	43	18 564
DILAT	7	2809
	9	3699
	13	5165
	17	7264

Table 2. MUPes for Reactions R1–R4 of the LAT Transmission Coefficients Obtained by DILAT for Different Numbers of Fitting Points, M_S , When Compared with ILAT1D Reference Values

M_S	MUPE			
	R1	R2	R3	R4
$T = 250$ K				
5	16.56	17.76	1.70	18.11
7	7.88	7.12	2.08	0.61
9	9.54	11.32	6.17	1.33
11	9.00	12.12	0.54	3.58
13	0.51	1.28	3.06	2.27
15	1.74	2.59	0.77	5.25
17	0.51	2.15	0.02	4.86
19	3.68	5.10	0.63	5.83
$T = 300$ K				
5	13.06	14.12	1.19	4.62
7	6.06	5.20	0.27	1.04
9	9.85	9.90	2.93	1.31
11	8.79	9.31	0.29	0.69
13	1.30	1.67	1.39	0.40
15	1.39	2.13	0.36	1.21
17	0.61	1.79	0.01	0.97
19	4.02	4.28	0.36	1.56

times faster than the full calculation for $M_S = 7$ and 17 points, respectively. The next step would be to test the accuracy of the LAT transmission coefficients by the DILAT algorithm by finding the optimum number of M_S points that give the best compromise between accuracy and computational cost. The procedure to obtain the transmission coefficients was the one described in the previous section, i.e., a set of $\theta(\tilde{\alpha}_r, E_r)$ values at energies E_r , $r = 1, \dots, M_S$ is chosen, with $E_{r=1}$ being the lowest tunneling energy at which it is possible to locate the classical turning points on the MEP for defining the straight path and $E_{r=M_S}$ being the last tunneling energy at which there is a nonadiabatic region along the straight path.

The deviation from the ILAT1D values of the DILAT transmission coefficients for different numbers of fitting points is given in Table 2 and plotted in Figure 3 for reactions R1–R4. For the present study, we have considered temperatures from 250 to 400 K, which for many practical applications is the temperature range for which one needs to evaluate the tunneling. At $T = 250$ K the smallest value of M_S that yields an accuracy better than 5% is $M_S = 13$ (and an interpolation with this value is hereafter called DILAT(13)). However, the interpolation using $M_S = 7$ (hereafter DILAT(7)), although it gives MUPES about three times larger than DILAT(13), yields small errors when

**Figure 3.** MUPes for $\eta_{1,4}$ and $\eta_{2,3}$ KIEs obtained by the DILAT algorithm using different fitting points to a spline-under-tension with respect to those LAT values obtained with the ILAT1D algorithm. The solid line plots the MUPE obtained at $T = 250$ and 300 K.**Table 3.** Transmission Coefficients for Reactions R1–R4

reaction	T (K)	LCT ^a			LAT ^b		
		SCT	ILCT2D	μ OMT ^a	DILAT(7)	DILAT(13)	ILAT1D
R1	250	21.9	28.1	31.4	39.4	36.1	36.3
	300	9.18	9.26	10.7	11.6	10.7	10.9
	350	5.32	4.97	5.78	5.75	5.43	5.52
	400	3.69	3.37	3.88	3.73	3.57	3.62
	500	2.37	2.16	2.42	2.28	2.22	2.25
R2	250	16.1	15.8	18.2	19.6	18.0	18.2
	300	7.40	6.83	7.91	7.79	7.27	7.39
	350	4.54	4.13	4.74	4.50	4.27	4.33
	400	3.27	2.98	3.36	3.16	3.03	3.07
	500	2.18	2.02	2.22	2.09	2.03	2.06
R3	250	15.2	10.7	16.0	11.8	11.7	12.1
	300	6.52	4.60	6.68	4.84	4.76	4.83
	350	3.94	2.92	3.98	3.00	2.98	2.99
	400	2.85	2.22	2.86	2.26	2.24	2.25
	500	1.95	1.64	1.95	1.66	1.65	1.65
R4	250	14.3	10.3	15.3	11.7	11.9	11.6
	300	6.29	4.49	6.47	4.65	4.68	4.70
	350	3.85	2.87	3.90	2.92	2.92	2.83
	400	2.80	2.20	2.82	2.22	2.22	2.22
	500	1.93	1.63	1.93	1.64	1.64	1.64

^a From ref 71. There are errors in Table 7 of ref 71; the correct values of the LCT and μ OMT transmission coefficients are listed here. ^b LAT transmission coefficients.

compared with $M_S = 9$ or 11 points and is about two times faster than DILAT(13), so comparisons involving both of them are interesting. At $T = 300$ K, all the MUPES are smaller, with the largest errors for DILAT(7) and DILAT(13) being 6% and 1.7%, respectively. For reactions R1–R4, Table 3 shows the DILAT(7) and DILAT(13) transmission coefficients together with the reference ILAT1D transmission coefficients.

In general Table 2 shows convergence to about 5% and 1% at 250 and 300 K, respectively.

The transmission coefficients are compared in Table 3. To compute the KIEs $\eta_{1,4} \equiv k_{R1}/k_{R4}$ and $\eta_{2,3} \equiv k_{R2}/k_{R3}$ using several tunneling approximations, we have factored them into two contributions

$$\eta^{\text{TST}/X} = \eta_{\text{tun}}^X \eta^{\text{TST}} \quad (21)$$

Table 4. Calculated KIEs Using Various Approximations for Tunneling^a

T (K)	η^{TST}	$\eta^{\text{TST/SCT}}$	$\eta^{\text{TST/LCT}}$	$\eta^{\text{TST}/\mu\text{OMT}}$	$\eta^{\text{TST/LAT}}$ ^a			η_{exp} ^b
R1/R4								
250	7.8	11.9	21.3	16.0	26.3	23.7	24.4	—
300	5.8	8.4	11.9	9.6	14.5	13.3	13.5	—
350	4.6	6.4	8.0	6.8	9.1	8.6	9.0	—
400	3.9	5.1	6.0	5.4	6.5	6.3	6.4	6.5 ^c
500	3.0	3.7	4.0	3.8	4.2	4.1	4.1	4.8 ^c
R2/R3								
250	5.7	6.0	8.4	6.5	9.4	8.8	8.6	—
300	4.4	5.0	6.5	5.2	7.1	6.7	6.7	—
350	3.7	4.3	5.3	4.4	5.7	5.3	5.4	13.0
400	3.2	3.6	4.3	3.2	4.5	4.3	4.4	8.5
500	2.6	2.9	3.2	3.0	3.3	3.2	3.2	5.0

^a The last column lists the experimental KIEs. LAT transmission factors obtained with DILAT(7), DILAT(13), and ILAT1D algorithms are listed in columns 6–8, respectively. ^b From refs 80 and 81. ^c Erratum in Table 8 of ref 71; the correct values of the experimental KIEs are listed here.

In eq 21, $\eta_{\text{un}}^{\text{X}}$ includes quantum effects (tunneling plus nonclassical reflection) on the reaction coordinate using the approximation X, where X = SCT, LCT, μOMT , or LAT, and η^{TST} includes the symmetry numbers, the classical translational and rotational contributions, and the quasiclassical quantized vibrational contribution. (There is no potential energy contribution in the cases considered here because the variational transition state is the conventional transition state for these reactions.)

Table 4 lists the KIEs obtained by the different tunneling approximations together with the experimental^{80,81} data. In general, all the methods underestimate the observed KIEs, although the ones obtained with the LAT approximation for tunneling are in better agreement with experimental values. The LCT transmission coefficients may underestimate the tunneling contribution in some cases, as was pointed out by Sansón et al.,⁸² however the KIEs obtained by this approximation are quite similar to the LAT ones. The μOMT approximation gives similar results to those obtained with the LAT approximation for the hydrogen abstraction processes, but it gives larger values for the deuterium transfer. Therefore, this discrepancy is due to the magnitude of the SCT transmission coefficients for deuterium transfer, which has the effect of decreasing the calculated KIEs. In any case the $\eta_{2,3}$ KIEs calculated using the LAT approximation for tunneling are still too small when compared to the available experimental data. From these values we arrive to the same conclusions as in ref 71, i.e., at the moment we cannot explain this discrepancy, and we encourage further experiments on these systems.

Finally, it is interesting to analyze the errors (with respect to a ILAT1D calculation) of the DILAT method, not just in the case of the transmission coefficients but also in the context of the KIEs. In the worst scenario, the largest error in the evaluation of the KIEs would be the sum of the MUPes, i.e., assuming no error cancellation. Using this worst-case possibility, we establish a maximum error of the DILAT(7) algorithm at $T = 250$ K of 9% for $\eta_{1,4}$ and $\eta_{2,3}$. For DILAT(13), these errors go down to 3% and 4% for $\eta_{1,4}$ and $\eta_{2,3}$, respectively. In round

numbers, the errors of the DILAT(7) and DILAT(13) algorithms, at $T = 250$ K, are smaller than 10% and 5%, respectively. In fact the errors, as shown in Figure 3, due to error cancellation are 7% and 4% for $\eta_{1,4}$ and 9% and 2% for $\eta_{2,3}$ using DILAT(7) and DILAT(13), respectively.

At $T = 300$ K, if we assume no error cancellation, the MUPes for $\eta_{1,4}$ and $\eta_{2,3}$ would be about 7% and 3% for DILAT(7) and DILAT(13), respectively. Similar calculated errors are obtained for DILAT(7), but when using DILAT(13), the calculated MUPes are less than 1% for both of the two evaluated KIEs. These results are very encouraging, especially when we take into account that the DILAT(7) and DILAT(13) methods are, respectively, 6.6 and 3.6 times faster than ILAT1D and about 50 and 30 times faster than the full (uninterpolated) calculation. It should also be noticed that this is a difficult case with two or three minima in the action integral at every tunneling energy, so for reactions with a less abrupt PES, the errors are expected to be smaller. The present results show that the DILAT(13) algorithm is reliable above $T = 250$ K to within 5% for the cases studied, although more testing would be needed to make broadly applicable statements of this nature.

4. Concluding Remarks

We have presented two algorithms for efficient direct dynamics evaluation of the least-action tunneling (LAT) transmission coefficients for polyatomic reactions. The interpolated least-action tunneling method based on one-dimensional interpolation (ILAT1D) uses the same philosophy as the previous ILCT1D algorithm; in particular, both make use of spline-under-tension interpolations for the effective potentials in the nonadiabatic regions of the tunneling paths. This algorithm, depending on the system, is about 5–10 times faster than the full calculation without loss of accuracy. However, the ILAT1D procedure is still quite expensive for polyatomic systems, so we have developed a much less expensive algorithm called double interpolated least-action tunneling, DILAT, which employs one-dimensional interpolations of not only the effective potential along nonadiabatic portions of the tunneling paths but also of the values of the action integrals as functions of energy. This even simpler method still provides quite accurate results. The performance of the DILAT algorithm was tested for four hydrogen/deuterium abstraction reactions, and we found that the optimum number of effective potential energies to be calculated in the nonadiabatic region is $M_S = 13$. The DILAT method based on 13 tunneling energies can be from 3 to 5 times faster than the ILAT1D algorithm, depending on the characteristics of the nonadiabatic region, but with an error of less than 5%. The method is being incorporated into the POLYRATE computer program.

The LAT calculations do not account for the discrepancy from experimental $\eta_{2,3}$ KIEs of the previously computed KIEs that were based on the less accurate large-curvature

tunneling (LCT) approximation. This discrepancy remains unexplained.

Glossary

This glossary contains an explanation of all acronyms used in this article. The acronyms are explained at first use, but this is an extra guide for convenience.

CAG:	classical adiabatic ground-state, a factor (usually close to unity) that makes the a transmission coefficient based on the vibrationally adiabatic ground-state potential curve consistent with a quasiclassical transition state calculation that implies a different threshold energy.
CVT:	canonical variational theory, VTST applied to a canonical ensemble.
DILAT:	doubly interpolated LAT.
ILAT1D:	interpolated LAT method based on one-dimensional interpolation.
ILCT1D:	interpolated LCT method based on one-dimensional interpolation.
ILCT2D:	interpolated LCT method based on two-dimensional interpolation.
KIE:	kinetic isotope effect, the ratio of rate constants for two reactions differing by isotopic substitution or isotopic placement.
LAG4:	least action ground state, 4, version 4 of the LAT method when applied with the ground-state tunneling approximation.
LAT:	least-action tunneling, a dynamical approximation for computing tunneling probabilities based on minimizing the magnitude of the imaginary part of the action integral along the tunneling path.
LCT:	large-curvature tunneling, a dynamical approximation for computing tunneling probabilities that is appropriate when the MEP has large curvature in the tunneling region.
MEP:	minimum-energy path in isoinertial coordinates.
μ OMT:	microcanonical OMT, a dynamical approximation for computing tunneling probabilities in which the choice between SCT and LCT tunneling is optimal at each tunneling energy (may be considered to be a poor person's version of LAT).
OMT:	shorthand for μ OMT.
PES:	potential energy surface, same as potential energy function.
SCT:	small-curvature tunneling, a dynamical approximation for computing tunneling probabilities that is appropriate when the MEP has only small curvature in the tunneling region.
VTST:	variational transition-state theory, a theory for calculating absolute reaction rates from PES information.

VTST/
MT: VTST with multidimensional tunneling, for example, with tunneling computed by the ZCT, SCT, LCT, μ OMT, or LAT tunneling approximation.

ZCT: zero-curvature tunneling, a dynamical approximation for computing tunneling probabilities that assumes that the tunneling path is a straightened MEP and that the vibrational motions orthogonal to the tunneling path are adiabatic.

Acknowledgment. A. F.-R and R. M.-P. thank Xunta de Galicia for financial support through Axuda para a Consolidación e Estructuración de unidades de investigación competitivas do Sistema Universitario de Galicia, 2007/50, cofinanciada polo FEDER 2007–2013. This work was supported in part by the U.S. Department of Energy, Office of Basic Energy Sciences, under grant no. DE-FG02-86ER13579.

Supporting Information Available: The reactions used for the tests of the ILAT1D algorithm against full-LAT calculations. Number of single-point calculations (NSP) in the nonadiabatic region needed for the evaluation of the transmission coefficients by the full-LAT (taken as reference) and ILAT1D methods for reactions R5 to R9. Transmission coefficients, κ , evaluated at T = 200, 300, and 400 K by the full-LAT (taken as reference) and ILAT1D methods for reactions R5 to R9. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) *Isotope Effects in Chemistry and Biology*; Kohen, A., Limbach, H. H., Eds.; CRC: Boca Raton, FL, 2006.
- (2) *Hydrogen-Transfer Reactions*; Hynes, J. T., Schowen, R. L., Klinman, J. P., Limbach, H. H., Eds.; Wiley-VCH: Weinheim, Germany, 2007.
- (3) Yamamoto, T. *J. Chem. Phys.* **1960**, *33*, 281.
- (4) Kuppermann, A.; Greene, E. *J. Chem. Educ.* **1968**, *45*, 361.
- (5) Manolopoulos, D. E.; Clary, D. C. *Annu. Rep. Prog. Chem., Sect. C: Phys. Chem.* **1989**, *86*, 95.
- (6) Park, T. J.; Light, J. C. *J. Chem. Phys.* **1989**, *91*, 974.
- (7) Truhlar, D. G.; Schwenke, D. W.; Kouri, D. J. *J. Phys. Chem.* **1990**, *94*, 7346.
- (8) Gray, S. K.; Goldfield, E. M.; Schatz, G. C.; Balint-Kurti, G. G. *Phys. Chem. Chem. Phys.* **1999**, *1*, 1141.
- (9) Mielke, S. L.; Lynch, G. C.; Truhlar, D. G.; Schwenke, D. W. *J. Phys. Chem.* **1994**, *98*, 8000.
- (10) Bowman, J. M.; Wang, D.; Huang, X.; Huarte-Larrañaga, F.; Manthe, U. *J. Chem. Phys.* **2001**, *114*, 9683.
- (11) Mielke, S. J.; Schwenke, D. W.; Garrett, B. C.; Truhlar, D. G.; Michael, J. V.; Su, M.-C.; W., S. J. *Phys. Rev. Lett.* **2003**, *91*, 63201.
- (12) Balucani, N.; Skouteris, D.; Capozza, G.; Segoloni, E.; Casavecchia, P.; Alexander, M.; Capecchi, G.; Werner, H.-J. *Phys. Chem. Chem. Phys.* **2004**, *6*, 5007.
- (13) Ceotto, C. S.; Yang, S.; Miller, W. H. *J. Chem. Phys.* **2005**, *122*, 044109.

- (14) Chakraborty, A.; Truhlar, D. G. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 10774.
- (15) Nyman, G.; van Harrevelt, R.; Manthe, U. *J. Phys. Chem. A* **2007**, *111*, 10331.
- (16) Wentzel, G. Z. *Phys.* **1926**, *38*, 518.
- (17) Kramers, H. A. Z. *Phys.* **1960**, *39*, 828.
- (18) Brillouin, L. *Comptes Rendus* **1926**, *24*, 183.
- (19) Kemble, E. C. *The Fundamental Principles of Quantum Mechanics with Elementary Applications*; Dover Publications: New York, 1937.
- (20) Marcus, R. A.; Coltrin, M. E. *J. Chem. Phys.* **1977**, *67*, 2609.
- (21) Garrett, B.; Truhlar, D. G.; Grev, R.; Magnuson, A. *J. Phys. Chem.* **1980**, *84*, 1730.
- (22) Garrett, B. C.; Truhlar, D. G. *J. Chem. Phys.* **1983**, *79*, 4931.
- (23) Garrett, B. C.; Abusalbi, N.; Kouri, D. J.; Truhlar, D. G. *J. Chem. Phys.* **1985**, *83*, 2252.
- (24) Lynch, G. C.; Truhlar, D. G.; Garrett, B. C. *J. Chem. Phys.* **1989**, *90*, 3102.
- (25) Taketsugu, T.; Hirao, K. *J. Chem. Phys.* **1997**, *107*, 10506.
- (26) Tautermann, C. S.; Voegelé, A. F.; Loerting, T.; Liedl, K. R. *J. Chem. Phys.* **2002**, *117*, 1962.
- (27) Yamamoto, T.; Miller, W. H. *J. Chem. Phys.* **2003**, *118*, 2135.
- (28) Meana-Pañeda, R.; Truhlar, D. G.; Fernández-Ramos, A. *J. Chem. Theory Comput.* **2010**, *6*, 6.
- (29) Wigner, E. *J. Chem. Phys.* **1937**, *5*, 720.
- (30) Horiuti, J. *Bull. Chem. Soc. Jpn.* **1938**, *13*, 210.
- (31) Keck, J. C. *Adv. Chem. Phys.* **1967**, *13*, 85.
- (32) Garrett, B. C.; Truhlar, D. G. *J. Chem. Phys.* **1979**, *70*, 1593.
- (33) Garrett, B. C.; Truhlar, D. G. *Acc. Chem. Res.* **1980**, *13*, 440.
- (34) Pechukas, P. *Annu. Rev. Phys. Chem.* **1981**, *32*, 159.
- (35) Truhlar, D. G.; Hase, W. L.; Hynes, J. T. *J. Phys. Chem.* **1983**, *87*, 2664.
- (36) Truhlar, D. G.; Garrett, B. C. *Annu. Rev. Phys. Chem.* **1984**, *35*, 159.
- (37) Truhlar, D. G.; Isaacson, A. D.; Garrett, B. C. In *Theory of Chemical Reaction Dynamics*; Baer, M., Ed.; CRC: Boca Raton, FL, 1985; Vol. 4, p 65.
- (38) Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. *J. Phys. Chem.* **1996**, *100*, 12771.
- (39) Garrett, B. C.; Truhlar, D. G. In *Theory and Applications of Computational Chemistry: The First Forty Years*; Dykstra, C. E.; Frenking, G.; Kim, K. S., Scuseria, G. E., Eds.; CRC: Boca Raton, FL, 2005; p 67.
- (40) Truhlar, D. G.; Garrett, B. C. In *Hydrogen-Transfer Reactions*; J. T. Hynes, H.-H. L.; Klinman, J. P., Schowen, R. L., Eds.; Wiley-VCH: Weinheim, Germany, 2007; Vol. 2, p 833.
- (41) Fernández-Ramos, A.; Ellingson, A.; Garrett, B. C.; Truhlar, D. G. *Rev. Comput. Chem.* **2007**, *23*, 125.
- (42) Allison, T. C.; Truhlar, D. G. In *Modern Methods for Multidimensional Dynamics Computations in Chemistry*; Thompson, D. L., Ed.; World Scientific: Singapore, 1998; p 618.
- (43) Pu, J.; Gao, J.; Truhlar, D. G. *Chem. Rev.* **2006**, *106*, 3140.
- (44) Eyring, H. *J. Chem. Phys.* **1935**, *3*, 107.
- (45) Garrett, B. C.; Truhlar, D. G. *J. Chem. Phys.* **1980**, *72*, 3460.
- (46) Garrett, B. C.; Truhlar, D. G.; Bowman, J. M.; Wagner, A. F.; Robie, D.; Arepalli, S.; Presser, N.; Gordon, R. J. *J. Am. Chem. Soc.* **1986**, *108*, 3515.
- (47) Fernández-Ramos, A.; Smedarchina, Z.; Rodríguez-Otero, J. *J. Chem. Phys.* **2001**, *114*, 1567.
- (48) Alhambra, C.; Sánchez, M. L.; Corchado, J. C.; Gao, J.; Truhlar, D. G. *Chem. Phys. Lett.* **2002**, *355*, 388.
- (49) Nagel, Z.; Klinman, J. *Chem. Rev.* **2006**, *106*, 3095.
- (50) Johnston, H. S. *Gas Phase Reaction Rate Theory*; Ronald Press: New York, 1966.
- (51) Hirschfelder, J. O.; Wigner, E. *J. Chem. Phys.* **1939**, *7*, 616.
- (52) Truhlar, D. G.; Kupperman, A. *J. Am. Chem. Soc.* **1971**, *93*, 1840.
- (53) Kuppermann, A. *J. Phys. Chem.* **1979**, *83*, 171.
- (54) Marcus, R. A. *J. Chem. Phys.* **1966**, *45*, 4493.
- (55) Fukui, K.; Kato, S.; Fujimoto, H. *J. Am. Chem. Soc.* **1975**, *97*, 1.
- (56) Truhlar, D. G.; Kuppermann, A. *J. Chem. Phys.* **1972**, *56*, 2232.
- (57) Skodje, R. T.; Truhlar, D. G.; Garrett, B. C. *J. Phys. Chem.* **1981**, *85*, 3019.
- (58) Skodje, R. T.; Truhlar, D. G.; Garrett, B. C. *J. Chem. Phys.* **1982**, *77*, 5955.
- (59) Lu, D.-h.; Truong, T. N.; Melissas, V. S.; Lynch, G. C.; Liu, Y.-P.; Garrett, B. C.; Steckler, R.; Isaacson, A. D.; Rai, S. N.; Hancock, G. C.; Lauderdale, J. G.; Joseph, T.; Truhlar, D. G. *Comput. Phys. Commun.* **1992**, *71*, 235.
- (60) Liu, Y.-P.; Lynch, G. C.; Truong, T. N.; Lu, D.-h.; Truhlar, D. G. *J. Am. Chem. Soc.* **1993**, *115*, 2408.
- (61) Garrett, B. C.; Truhlar, D. G.; Wagner, A. F.; Dunning Jr, T. H. *J. Chem. Phys.* **1983**, *78*, 4400.
- (62) Garrett, B. C.; Joseph, T.; Truong, T. N.; Truhlar, D. G. *Chem. Phys.* **1989**, *136*, 271.
- (63) Truong, T. N.; Lu, D.-h.; Lynch, G. C.; Liu, Y.-P.; Melissas, V. S.; Stewart, J. J. P.; Steckler, R.; Garrett, B. C.; Isaacson, A. D.; González-Lafont, A.; Rai, S. N.; Hancock, G. C.; Joseph, T.; Truhlar, D. G. *Comput. Phys. Commun.* **1993**, *75*, 143.
- (64) Liu, Y.-P.; Lu, D.-h.; González-Lafont, A.; Truhlar, D. G.; Garrett, B. C. *J. Am. Chem. Soc.* **1993**, *115*, 7806.
- (65) Fernández-Ramos, A.; Truhlar, D. G. *J. Chem. Phys.* **2001**, *114*, 1491.
- (66) Truhlar, D. G.; Brown, F. B.; Steckler, R.; Isaacson, A. D. In *The Theory of Chemical Reaction Dynamics*; Clary, D. C., Ed.; D. Reidel: Dordrecht, The Netherlands, 1986; p 285.
- (67) Truhlar, D. G.; Gordon, M. S. *Science* **1990**, *249*, 491.
- (68) González-Lafont, A.; Truong, T. N.; Truhlar, D. G. *J. Phys. Chem.* **1991**, *95*, 4618.
- (69) *Quantum Tunneling in Enzyme-Catalysed Reactions*; Allemann, K.; Scrutton, N. S., Eds.; RSC Publishing: Cambridge, UK, 2009.
- (70) Renka, R. J.; Cline, A. K. *Rocky Mt. J. Math.* **1984**, *14*, 223.
- (71) Fernández-Ramos, A.; Truhlar, D. G. *J. Chem. Theory Comput* **2005**, *1*, 1063.

- (72) Fernández-Ramos, A.; Truhlar, D. G.; Corchado, J.; Espinosa-García, J. *J. Phys. Chem. A* **2002**, *106*, 4957.
- (73) Renka, R. J. *SIAM J. Sci. Stat. Comput.* **1987**, *8*, 393.
- (74) Renka, R. J. *ACM Trans. Math. Software* **1993**, *19*, 81.
- (75) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes*, 3rd ed.; Cambridge University Press: Cambridge, UK, 2007; p 492.
- (76) Truong, T. N.; Lu, D.-h.; Lynch, G. C.; Liu, Y.-P.; Melissas, V. S.; Stewart, J. J. P.; Steckler, R.; Garrett, B. C.; Isaacson, A. D.; González-Lafont, A.; Rai, S. N.; Hancock, G. C.; Joseph, T.; Truhlar, D. G. *Comput. Phys. Commun.* **1993**, *75*, 143.
- (77) Zheng, J.; Zhang, S.; Lynch, B. J.; Corchado, J. C.; Chuang, Y.-Y.; Fast, P. L.; Hu, W.-P.; Liu, Y.-P.; Lynch, G. C.; Nguyen, K. A.; Jackels, C. F.; Fernández Ramos, A.; Ellingson, B. A.; Melissas, V. S.; Villa, J.; Rossi, I.; Coitiño, E. L.; Pu, J.; Albu, T. V.; Steckler, R.; Garrett, B. C.; Isaacson, A. D.; Truhlar, D. G. *POLYRATE*, 2008; University of Minnesota: Minneapolis, MN, 2008.
- (78) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem A* **2004**, *108*, 6908.
- (79) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.
- (80) Sharp, T. E.; Johnston, H. S. *J. Chem. Phys.* **1962**, *37*, 1541.
- (81) Carmichael, H.; Johnston, H. S. *J. Chem. Phys.* **1964**, *41*, 1975.
- (82) Sansón, J. A.; Sánchez, M. L.; Corchado, J. *J. Phys. Chem. A* **2005**, *110*, 589.

CT100285A

Adaptive Steered Molecular Dynamics of the Long-Distance Unfolding of Neuropeptide Y

Gungor Ozer,[†] Edward F. Valeev,[‡] Stephen Quirk,[§] and Rigoberto Hernandez^{*‡}

Center for Computational and Molecular Science and Technology, School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia 30332-0400, Department of Chemistry, Virginia Tech, Blacksburg, Virginia 24061, and Kimberly-Clark Corporation, Atlanta, Georgia 30076-2199

Received June 11, 2010

Abstract: Neuropeptide Y (NPY) has been found to adopt two stable conformations in vivo: (1) a monomeric form called the PP-fold in which a polyproline tail is folded onto an α -helix via a β -turn and (2) a dimeric form of the unfolded proteins in which the α -helices interact with each other via side chains. The transition pathway and rates between the two conformations remain unknown and are important to the nature of the binding of the protein. Toward addressing this question, the present work suggests that the unfolding of the PP-fold is too slow to play a role in NPY monomeric binding unless the receptor catalyzes it to do so. Specifically, the dynamics and structural changes of the unfolding of a monomeric NPY protein have been investigated in this work. Temperature accelerated molecular dynamics (MD) simulations at 500 K under constant (N,V,E) conditions suggests a hinge-like unraveling of the tail rather than a random unfolding. The free energetics of the proposed unfolding pathway have been described using an adaptive steered MD (SMD) approach at various temperatures. This approach generalizes the use of Jarzynski's equality through a series of stages that allows for better convergence along nonlinear and long-distance pathways. Results acquired using this approach provide a potential of mean force (PMF) with narrower error bars and are consistent with some of the earlier reports on the qualitative behavior of NPY binding.

1. Introduction

The true nature of conformational changes undertaken by a given bioactive ligand during its binding to a receptor remains an elusive and important target for the development of novel drugs. The binding of a small ligand to a large membrane receptor is a dynamic process and is therefore difficult to observe using classical experimental approaches. Although atomic force microscopy (AFM) presents insightful information regarding the force required to unfold a particular protein, the detailed unfolding process is generally not observed in AFM experiments. Computer simulation tech-

niques—to the limit of the accuracy of the model potentials and the integration of the equations of motion—provide a useful approach for elucidating the complete unfolding pathway of the protein of interest.^{1–5} For example, in silico simulation using forced molecular dynamics agrees well with the corresponding AFM force pulling experiments.^{6–9}

The neuropeptide Y (NPY) ligand has been a primary target of many recent pharmacological studies because of its implicated function in the brain.^{10–14} Consisting of 36 amino acids, NPY is the most abundant neuropeptide in the mammalian central nervous system¹⁰ and widely expressed in the peripheral nervous system.¹¹ Several important physiological activities such as induction and control of food intake, inhibition of anxiety, increase in memory retention, presynaptic inhibition of neurotransmitter release, vasoconstriction, and regulation of ethanol consumption have been attributed to NPY.¹² The multifunctionality of NPY is the

* To whom correspondence should be addressed E-mail: hernandez@chemistry.gatech.edu.

[†] Georgia Institute of Technology.

[‡] Virginia Polytechnic Institute and State University.

[§] Kimberly-Clark Corporation.



Figure 1. The structure of NPY (with the water solvent hidden) shown at 5 points along a single steered MD unfolding pathway at 500 K. The NPY backbone is illustrated using a brown ribbon. The first frame shows a stable folded NPY protein derived and equilibrated over the coordinates from the PDB file, 1PPT, and the last frame is an illustration of the unfolded protein (1RON). Both structures have the same 36 amino acid sequence: Y–P–S–K–P–D–N–P–G–E–D–A–P–A–E–D–M–A–R–Y–Y–S–A–L–R–H–Y–I–N–L–I–T–R–Q–R–Y. The first eight of these residues comprise the tail. The remaining frames illustrate the unhooking of the tail from the α helix as obtained from a particular unfolding trajectory in this work. The three residues most clearly marking the unhooking of the tail are shown in atomistic detail: LEU24 (in black on the helix), ALA12 (in black on the turn), and PRO5 (in red on the tail).

result of its affinity to bind to at least six receptor subtypes—enumerated as Y1 through Y6—belonging to the rhodopsin-like superfamily of G protein-coupled receptors. It has been shown that receptors Y1, Y4, and Y6 are closely related to each other.¹³ A recent study on the evolution of neuropeptide Y receptors (Y3 was not investigated) has led to a partitioning into three subfamilies of receptors: Y1/Y4/Y6, Y2, and Y5.¹⁴

NPY is a member of the pancreatic polypeptide (PP) hormone family that includes also pancreatic polypeptide (PP) and peptide YY (PYY).¹⁵ All three ligands share a common hairpin-like structure in tertiary form called the PP-fold. Therein, the N-terminal residues (1–8) adopt a polyproline type II helical conformation (tail). Residues 9–13 form a loop that allows the tail to fold onto an α -helix (residues 14–31), and the C-terminal residues (32–36) are so flexible that they do not participate in the α -helical conformation (14–31).^{16–18} NMR studies have shown that NPY adopts a different conformation in the dimeric form^{19–21} or when bound to membrane mimetic, dodecylphosphocholine (DPC) micelles.^{22,23} In this particular state, the NPY tail is observed to be destabilized and positioned away from the α -helix. Recently, Bettio et al. reported, in contrast to earlier reports,^{16–18} that at low concentrations monomeric NPY favors a less ordered structure in which the β -turn of NPY is more destabilized.²⁴

The numerical study described herein aims to provide a dynamical explanation for the mechanism performed by an NPY molecule during its structural transition between the reported open (PDB²⁵ ID: 1PPT²⁶) and closed (PDB ID: 1RON²¹) conformations. Figure 1 demonstrates a reduced representation of the unfolding of NPY using only five ribbon diagrams in order: the folded form (pp-fold), three intermediate structures, and an unfolded form. Knowledge of the pathway may be of use in the design of ligands to stimulate NPY toward the desired fold in vivo, regulators for the binding of NPY to lipid membranes, and alternative receptors. The present work, in particular, provides some insight into the likely form—PP-fold or free tail—adopted by NPY as it binds to a receptor. This article is structured as follows: High temperature MD simulations are used to accelerate the unfolding process and to observe a possible unfolding pathway for said process. The proposed unfolding pathway is investigated using steered molecular dynamics (SMD)

simulations. The free energy along this path is generally obtained from the SMD trajectories through the use of Jarzynski's nonequilibrium work relation. Unfortunately, the standard application of this approach did not converge within available computational resources. An auxiliary central result of this work is the development of a stepwise adaptive SMD scheme for the calculation of the free energy along a nonlinear and large-distance pathway, in section 2.3. The time scale of the structural stability of NPY is obtained by way of a determination of the transition state theory rates on the computed surfaces. We observe that the NPY tail follows a hinge-like motion while folding/unfolding. We have also confirmed that the PP-fold conformation of NPY is favored in monomeric form, which was previously proposed by Nordmann et al.^{16–18}

2. Methods

2.1. Accelerated MD Simulations at Elevated Temperature. The relative dynamics of the α -helix and tail in NPY immersed in a periodic box of water molecules have been simulated using several computational protocols to overcome the long times needed to follow simulations of the folding process. The focus of the simulations is the unfolding of NPY, as it is faster than the folding process while still revealing the folding pathway(s). The initial state of the unfolding process—namely, the protein's crystal structure—is also more clearly defined than the structures of the unfolded protein basin, and this offers additional numerical advantages in attempts to map out the pathway.²⁷

Molecular dynamics simulations have been carried out using the NAMD²⁸ molecular dynamics integrator with the forces in NPY specified through the CHARMM force field.²⁹ The water molecules are treated using the TIP3P³⁰ model, and 13 178 water molecules are included in the cube. A time step of 1 fs has been employed in all simulations. Electrostatic interactions have been calculated through the particle mesh Ewald (PME) method.³¹ Solvated structures are initialized by inserting NPY into an appropriately sized cavity created within an equilibrated neat water box. These are equilibrated at 50 K for 5 ps and subsequently heated gradually to the temperature of interest. An NPT equilibration run (at the desired final temperature) is then performed to ensure that the cubic box has a density consistent with 1.0

atm of pressure. Temperature control is realized within the NAMD program by integrating the Langevin equation with the Brunger–Brooks–Karplus (BBK) method, which is a natural extension of Verlet integration. This results in an ensemble of structures in which NPY is constrained to its folded state within an equilibrated solvent inside of a cubic box with sides roughly between 70 and 75 Å.

Each member of the ensemble of solvated folded-NPY structures is allowed to freely propagate for 5 ps under constant (N, V, E) conditions. It is common practice to run such simulations under constant (N, V, T) conditions using thermostats on all the atoms in the system. However, this has the possible negative side effect of suppressing fluctuations in energy that lead to correlated energy flows between molecules and therefore overly lose correlation as the system evolves in the heat bath.³² For example, several popular MD packages, including NAMD, have recently been shown to suffer from a serious problem associated with the random number generators implemented in their thermostat algorithms.³³ In order to differentiate the unfolding mode from any other mode in the system, the alternative is to run the simulation under constant (N, V, E) conditions at an energy that is thermodynamically consistent with the temperature. This has the disadvantage that the total energy of the box is constant, but with a sufficiently large water box the effective dynamics of the NPY protein will still be that of an open system at constant temperature. The results from a small number of (N, V, T) and (N, V, E) MD simulations are described later, but the conclusion is that all of the remaining simulations could be performed using constant (N, V, E) conditions without losing the notion of temperature along the unfolding path.

Although we are primarily interested in the unfolding dynamics of NPY at 310 K, the duration of such trajectories is so long that it would entail simulations that are cost-prohibitive. Among several accelerated dynamics approaches now available in the literature, we chose to overcome this obstacle using temperature acceleration,³⁴ as it has been previously reported to accelerate the unfolding process without altering the pathway.³⁵ This is valid assuming that thermal unfolding of proteins demonstrates Arrhenius behavior.⁵ However, some reports claim that protein unfolding can show non-Arrhenius behavior,^{36,37} and therefore temperature acceleration may result in losing some intermediate states.^{38,39} In the present context of NPY unfolding, the resulting potentials of mean force (shown below) contain only a single barrier at both low and high temperatures and hence are consistent with the requirement of Arrhenius behavior. Preliminary runs were tested at $T = 300$ K, 367 K, 433 K, and 500 K in a cubic box with sides of 75 Å solvated with equilibrated water (TIP3P) molecules. As will be shown below, NPY unfolded only at 500 K within 100 ps, and hence it became the temperature of choice for the temperature accelerated MD simulations in this work. A temperature of 500 K is well above any natural biological temperature and is also above the protein melting temperature, T_m . An experimental system under these conditions would exhibit different dynamics than the biological case. The water system in the computer model, however, remains as a

metastable and superheated liquid because neither chemical bond breaking-and-making or evaporation pathways are available to it. The key assumption is that the dynamical pathways also remain in the same universality class, and thus we require additional tests to confirm the predictions of correlation functions using temperature acceleration. As will be shown below, the model system exhibits the appropriate chemical structures (in the same universality class) as those of the lower temperature.

2.2. Measurables and Correlation Functions of the Trajectories. Analysis of the trajectories was carried out by several methods. Both pepstat, which is our own code, and the NAMD/VMD package were used for trajectory analysis, with the latter focusing on the graphical representations of the trajectories.

Although the tail section exhibits the most dramatic dynamical changes, structural metrics were collected throughout the protein simulations. Within the polyproline tail (residues 1–12), the time dependence of the end-to-end distance and radius of gyration, $R_g^2 = 1/N \sum_{k=1}^N (r_k - r_{\text{mean}})^2$, are measured. The time dependence in the tail-to-helix distance is inferred by way of the pairwise distances between residue pairs, 1–31, 4–27, 5–24, 7–20, and 8–16.

The results shown below [cf. Figure 3b] suggest that the unfolding pathway involves the unhinging of the tail away from the α -helix instead of sliding. This unhinging occurs about the pivot represented by the ALA12 residue and is measurable through a so-called tail-turn-helix angle. While the α -helix is relatively stiff through this unfolding, the N-terminal of the polyproline tail—and particularly TYR1 to LYS4—is much floppier. The remaining residues (PRO5 to ASP11) on the tail follow a smoother unhinging and can be used to define the tail-turn-helix angle.

2.3. The Unfolding Path and the Potential of Mean Force (PMF). The domain of the energy landscape of even a small protein such as NPY has a high dimensionality. The identification of an unfolding pathway is therefore useful because it greatly reduces this dimensionality. Once identified, the energetics along this pathway are determined by the potential of mean force (PMF). [See, e.g., ref 40.] The importance of the PMF as well as the difficulty in calculating it has led to the development of far too many approaches to list here. Instead, we focus on those approaches which rely on sampling the states directly from trajectories. Unfortunately, the use of unconstrained trajectories is cost prohibitive when the processes of interest are very slow and dominated by deep minima. Instead, SMD can accelerate such processes by applying steering forces along the chosen unfolding pathway. Such a nonequilibrium process would not seem to provide the unconstrained structures required to obtain the equilibrium PMF. This problem was resolved by Jarzynski when he showed that an appropriately weighted average of the nonequilibrium work over many such SMD trajectories leads to the PMF.^{41,42} Jarzynski's equality has been validated numerically on several systems such as deca-alanine stretching by Park and Schulten,⁴³ Ace–ALA₈–NMe unfolding and ligand diffusion in globins by Xionget et al.,⁴⁴ and Angeli's salt decomposition by Torras et al.⁴⁵ It has been compared to existing biased MD techniques, such as to umbrella

sampling⁴⁶ and to targeted MD,⁴⁷ yielding comparable results. It has also been verified in the context of experimental results such as RNA unfolding by Liphardt⁴⁸ and a mechanical oscillator.⁴⁹ Below, we provide a review of Jarzynski's inequality and its implementation within SMD to obtain the PMF along a selected pathway. In so doing, we also introduce a generalization of this approach to account for long-range nonlinear unfolding pathways.

2.3.1. Review of Jarzynski's Equality. Jarzynski's equality was originally expressed in terms of classical Hamiltonian systems.^{41,42} It was extended to thermostatted stochastic systems by Crooks.⁵⁰ Crooks' introduction of a heat bath ensures that after sufficient time upon reaching a given nonequilibrium state, the system will reach an equilibrium with the environment at no additional cost of work. Jarzynski's equality for *dissipated* Hamiltonian systems can be stated as follows. Suppose a classical mechanical system consists of N particles, denoted by the phase space variables z , which are surrounded by a large enough heat bath. A constraint on the configuration space z_x is imposed through the projection $\xi_x = \xi_x(z_x)$ acting in configuration space alone. The constrained Hamiltonian may be written as

$$H_\xi^{\text{SEB}}(\Gamma, \Theta) = H^{\text{SE}}(z; \Theta_x) + H^{\text{B}}(\Theta) \quad (1a)$$

$$= T^{\text{S}}(\xi) + H_{\xi_x}^{\text{E}}(\Gamma; \Theta_x) + H^{\text{B}}(\Theta) \quad (1b)$$

where S, E and B denote the constrained system, environment, and bath, respectively; the subscript x (p) refers to the position (momentum) components; and T^{S} is the kinetic energy for the constrained system variables. The system variables not constrained by ξ —viz., the environment—comprise a space of dimension lower than $6N$, and its phase space variables are represented through Γ . The phase space variables Θ comprise the positions Θ_x and momenta Θ_p of the bath, and their dynamics are weakly coupled to Γ in the $H_{\xi_x}^{\text{E}}$ term. The constraint ξ_x is typically one-dimensional and serves as an order parameter or reaction path that defines a state of the system. The space defined by Γ_x is orthogonal to ξ_x and denotes the environment exclusive of the bath Θ . The nonequilibrium process between two points in the constrained space is driven by the addition of a time-dependent Hamiltonian

$$H' = H'(\xi_x, t) \quad (2)$$

that acts only on ξ_x . That is, the total time-dependent Hamiltonian is $H^{\text{T}} = H_\xi^{\text{SEB}} + H'$. In what follows, we will not generally distinguish between the phase space ξ and configuration space ξ_x variables, for simplicity.

The change in the energy as the system is carried from an initial state ξ_0 to a final state ξ_t corresponds to the work done by $H'(\xi, t)$ through this $\xi_t \leftarrow \xi_0$ process,

$$W_{\xi_t \leftarrow \xi_0}^{\xi}(\Gamma_t, \Theta_t, \Gamma_0, \Theta_0) = H_{\xi_t}^{\text{E}}(\Gamma_t; \Theta_t) - H_{\xi_0}^{\text{E}}(\Gamma_0; \Theta_0) \quad (3a)$$

$$= H'(\xi_t, t) - H'(\xi_0, 0) \quad (3b)$$

where Γ_t and Θ_t are connected to Γ_0 and Θ_0 through the propagator during the $\xi_t \leftarrow \xi_0$ process for a time t .

The equilibrium partition functions associated with the initial and final points associated with the $\xi \leftarrow 0$ process can be rewritten in terms of the original system variables as^{51,52}

$$Z_\xi^{\text{SE}} = \int dz e^{-\beta H^{\text{SE}}(z)} \delta(\xi(z) - \xi) \quad (4a)$$

$$= \int dz d\Theta e^{-\beta \{H_\xi^{\text{SEB}}(\Gamma, \Theta) + H'(\xi, t)\}} \delta(\xi(z) - \xi) \quad (4b)$$

which is related to the potential of mean force, $G(\xi)$, through the reversible work theorem, $\ln Z_\xi^{\text{S}} = -\beta G(\xi)$. In terms of these free energies, Jarzynski's equality,^{41,42,50} is

$$G(\xi_t) = G(\xi_0) - \frac{1}{\beta} \ln \langle e^{-\beta W_{\xi_t \leftarrow \xi_0}^{\xi}} \rangle_0 \quad (5)$$

where the ensemble average is taken over the initial variables (z, Θ) satisfying the constraint, $\xi(z_x) = \xi_0$. Note that, similar to the ground-state dominance in the calculation of a partition function, the Jarzynski average is dominated by the trajectories with the lowest work change.

Jarzynski's inequality follows from eq 5 through the use of Jensen's inequality:

$$G(\xi_t) - G(\xi_0) \leq \langle W_{\xi_t \leftarrow \xi_0}^{\xi} \rangle_0 \quad (6)$$

Alternatively, the use of a cumulant expression provides the second-order cumulant (SOC) expression

$$G(\xi_t) - G(\xi_0) \approx \langle W_{\xi_t \leftarrow \xi_0}^{\xi} \rangle_0 - \frac{1}{2} \beta \langle \langle [W_{\xi_t \leftarrow \xi_0}^{\xi}]^2 \rangle \rangle_0 - \langle W_{\xi_t \leftarrow \xi_0}^{\xi} \rangle_0^2 \quad (7)$$

which is surprisingly accurate for small nonequilibrium processes or environments with a Gaussian response.^{43,53,54}

2.3.2. Adaptive Scheme for Jarzynski's Equality. As will be seen below, the application of the Jarzynski equality for the extended motion of a finite number of NPY unfolding trajectories provides a very weak upper bound to the PMF. In fact, it is so weak that the cumulant expansion of eq 5 presents a dramatically large deviation between the second-order cumulant and the exponential average, as demonstrated in Figure 8. In order to treat such extended systems, we have developed an adaptive version of Schulten's algorithm⁴³ in which the Jarzynski equality is applied through a series of shorter steps. It is *adaptive* in the sense that the initial configuration for a given step is obtained (or adapted) from the trajectories of the previous step.

The overall unfolding path is initially partitioned into N steps marked by its end points, $\xi_0, \xi_1, \dots, \xi_N$. The i th iteration is initiated at ξ_{i-1} and Γ_{i-1} while the bath Θ_{i-1} is sampled from the appropriate canonical ensemble. Each such bath, $\Theta_\alpha^{\xi_t \leftarrow \xi_{i-1}}(t_{i-1})$, leads to M trajectories labeled by α for the $\xi_i \leftarrow \xi_{i-1}$ process. This, in turn, leads to a distribution of values in the work $W_\alpha^{\xi_t \leftarrow \xi_{i-1}}(t)$, environment $\Gamma_\alpha^{\xi_t \leftarrow \xi_{i-1}}(t)$, and bath $\Theta_\alpha^{\xi_t \leftarrow \xi_{i-1}}(t)$ for times t within the i th step. At the end of the iteration, the average work $W_{\xi_t \leftarrow \xi_{i-1}}^{\xi}(t_i)$ is computed according to the Jarzynski equality (eq 5). There then exists a trajectory α' for which its work $W_{\alpha'}^{\xi_t \leftarrow \xi_{i-1}}(t_i)$ is closest to the average work $W_{\xi_t \leftarrow \xi_{i-1}}^{\xi}(t_i)$. The initial value of the environment Γ_i for the $(i + 1)$ th iteration is then taken to be the corresponding $\Gamma_{\alpha'}^{\xi_t \leftarrow \xi_{i-1}}(t_i)$. Meanwhile the algorithm is initiated with values

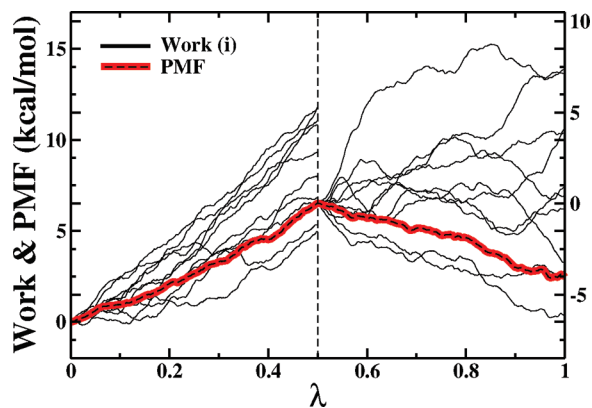


Figure 2. Illustration of the adaptive scheme applied to a system where the unfolding path is divided into two steps (with 10 trajectories α each). Black solid curves are the work for each of the 10 trajectories at each step. The PMF along a given substep is shown with a thick-red highlighted black-dashed curve. The right y axis tick marks are labeled for the second-step work trajectories with the 0 position located at the final average work value of the first substep. The left y axis tick marks are labeled for both the first step work trajectories and the overall PMF. (This figure is drawn for illustration purposes only and is not based on real physical data.)

(ξ_0, Γ_0) matching the initial structure of the system and environment. In the case of NPY, this amounts to the crystal structure of the entire protein, while ξ refers only to the constrained angle spanned by the helix and tail.

A proof of this algorithm begins by considering the application of the adaptive procedure to divide a single step into two substeps, as illustrated in Figure 2, along a specific unfolding path λ for the corresponding system variables ξ . For simplicity, but without a loss of generality, we suppose that the system is carried along by a nonequilibrium process from state $\xi = 0$ at initial time 0 to a final state $\xi = 1$ at a final time t . For each of M realizations labeled by α of the $1 \leftarrow 0$ process, the trajectories of the environment $\Gamma_\alpha^{1 \leftarrow 0}(t)$ and the bath $\Theta_\alpha^{1 \leftarrow 0}(t)$ can be formally constructed. The work done along each of these trajectories is $W_\alpha^{1 \leftarrow 0}[\Gamma_\alpha^{1 \leftarrow 0}(t), \Theta_\alpha^{1 \leftarrow 0}(t), \Gamma_\alpha^{1 \leftarrow 0}(0), \Theta_\alpha^{1 \leftarrow 0}(0)]$ as specified by eq 3. The PMF of this process is

$$\Delta G^{1 \leftarrow 0} = -\frac{1}{\beta} \ln \langle e^{-\beta W_\alpha^{1 \leftarrow 0}} \rangle_\alpha \quad (8)$$

where the average is taken over the M realizations starting with the same initial ξ_0 and Γ_0 and various initial bath configurations $\Theta_\alpha(0^{1 \leftarrow 0})$.

The single step can now be partitioned into two steps in which the system is stopped at an intermediate time t' and the corresponding position ξ' . For each of the original M trajectories in the $1 \leftarrow 0$ process, this partitions the work into two components:

$$W_\alpha^{\xi' \leftarrow 0} = H_{\xi'}^{\text{SB}}[\Gamma_\alpha^{1 \leftarrow 0}(t'), \Theta_\alpha^{1 \leftarrow 0}(t')] - H_{\xi'}^{\text{SB}}[\Gamma_\alpha^{1 \leftarrow 0}(0), \Theta_\alpha^{1 \leftarrow 0}(0)] \quad (9a)$$

$$W_\alpha^{1 \leftarrow \xi'} = H_{\xi'}^{\text{SB}}[\Gamma_\alpha^{1 \leftarrow 0}(t), \Theta_\alpha^{1 \leftarrow 0}(t)] - H_{\xi'}^{\text{SB}}[\Gamma_\alpha^{1 \leftarrow 0}(t'), \Theta_\alpha^{1 \leftarrow 0}(t')] \quad (9b)$$

from which the free energy change for the first step can be easily obtained using Jarzynski's equality

$$\Delta G_{\xi' \leftarrow 0} = -\frac{1}{\beta} \ln \langle e^{-\beta W_\alpha^{\xi' \leftarrow 0}} \rangle_\alpha \quad (10)$$

For the second substep, however, each trajectory specified by eq 9b starts at a different value of the environment, $\Gamma_\alpha^{1 \leftarrow 0}(t')$. We now introduce a $\xi' \leftarrow \xi'$ process during which ξ' is held fixed and the environment $\Theta_\alpha^{\xi' \leftarrow \xi'}(t')$ relaxes in time τ from 0 to τ_α for some arbitrary final time τ_α , which is likely different for each trajectory α . The work to move the system from the state at the end of the process described in eq 9a along this $\xi' \leftarrow \xi'$ process is

$$\Delta W_\alpha^{\xi' \leftarrow \xi'} = H_{\xi'}^{\text{SB}}[\Gamma_\alpha^{\tau_\alpha}(t'), \Theta_\alpha^{\tau_\alpha}(t')] - H_{\xi'}^{\text{SB}}[\Gamma_\alpha^{\xi' \leftarrow 0}(t'), \Theta_\alpha^{\xi' \leftarrow 0}(t')] \quad (11)$$

and the work to return to the final point of the $1 \leftarrow 0$ process is

$$W_\alpha^{1 \leftarrow \xi'} = H_{\xi'}^{\text{SB}}[\Gamma_\alpha^{1 \leftarrow \xi'}(t), \Theta_\alpha^{1 \leftarrow \xi'}(t)] - H_{\xi'}^{\text{SB}}[\Gamma_\alpha^{\tau_\alpha}(t'), \Theta_\alpha^{\tau_\alpha}(t')] \quad (12)$$

The $\xi' \leftarrow \xi'$ process can be allowed to propagate for as long as it takes for $\Gamma_\alpha^{\tau_\alpha}(t')$ to be equal to some $\Gamma_\alpha^{1 \leftarrow \xi'}(t')$ which is independent of α . The existence of such a common end point is assured if the process is ergodic and the system is found in a single local basin of attraction. The requirement of ergodicity is a weak constraint given that the environment is coupled to a bath. The requirement for a single basin is also weak because the environment must access all possible such basins with zero-work paths. This motivates a new path for a *restricted* $1 \leftarrow \xi'$ process starting at the fixed end point $\Gamma_\alpha^{1 \leftarrow \xi'}(t')$, and its work is given by

$$W_\alpha^{1 \leftarrow \xi'} = H_{\xi'}^{\text{SB}}[\Gamma_\alpha^{1 \leftarrow \xi'}(t), \Theta_\alpha^{1 \leftarrow \xi'}(t)] - H_{\xi'}^{\text{SB}}[\Gamma_\alpha^{1 \leftarrow \xi'}(t'), \Theta_\alpha^{1 \leftarrow \xi'}(t')] \quad (13)$$

where the stochastic $\Theta_\alpha^{1 \leftarrow \xi'}(t')$ has replaced the formally propagated $\Theta_\alpha^{\tau_\alpha}(t')$. That is, the bath decoherence time is sufficiently fast so that the detailed propagation can be ignored while the initial bath $\Theta_\alpha^{1 \leftarrow \xi'}(t')$ in the $1 \leftarrow \xi'$ process is Gaussian random. The PMF of the restricted $1 \leftarrow \xi'$ process is

$$\Delta G^{1 \leftarrow \xi'} = -\frac{1}{\beta} \ln \langle e^{-\beta W_\alpha^{1 \leftarrow \xi'}} \rangle_\alpha \quad (14)$$

The average in eq 8 can thus be written as

$$\langle e^{-\beta W_\alpha^{1 \leftarrow 0}} \rangle_\alpha = \langle e^{-\beta \{W_\alpha^{1 \leftarrow \xi'} + \Delta W_\alpha^{\xi' \leftarrow \xi'} + W_\alpha^{\xi' \leftarrow 0}\}} \rangle_\alpha \quad (15)$$

where it should be noted that the sum in the exponent in the RHS is not equal to $W_\alpha^{1 \leftarrow 0}(t)$, nor is the trajectory the same after t' . However, the averages are equal because they are both nonequilibrium $1 \leftarrow 0$ processes between the same initial and final points satisfying Jarzynski's equality. Meanwhile, the work in the $\xi' \leftarrow \xi'$ process is

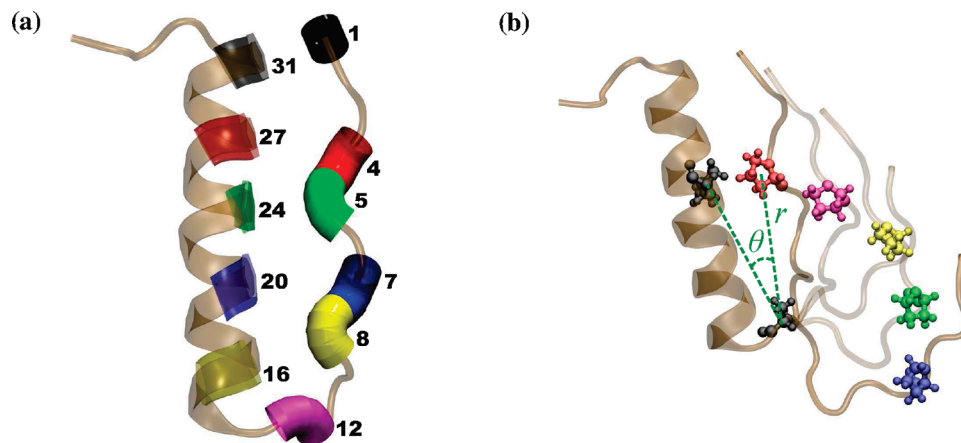


Figure 3. (a) A backbone ribbon diagram of NPY shown in brown with the helix emphasized by the thick ribbons as usual. The residue, ALA12, at the turn is shown in magenta and acts as the hinge. Pairs of residues that are in contact in the folded NPY and whose relative distances and angles are tracked in the following are color-coded as in the following scheme: black for TYR1 and ALA31, red for LYS4 and TYR27, green for PRO5 and LEU24, blue for ASN7 and TYR20, and yellow for PRO8 and ASP16. Note that residue positions 14–31 correspond to the helix. (b) The unfolding path is illustrated on the right, wherein the helix and hinge regions are held fixed while five images of the tail are overlaid. The PRO5 residue, which is explicitly used for steering relative to the fixed residues LEU24 and ALA12 (shown in black), is shown in five different colors along the unfolding path: red → magenta → yellow → green → blue.

zero because the system was allowed to relax freely. Hence,

$$\begin{aligned} \langle e^{-\beta W_{\alpha}^{1 \leftarrow 0}} \rangle_{\alpha} &= \langle e^{-\beta (W_{\alpha}^{\Gamma^{1 \leftarrow \xi'}} + W_{\alpha}^{\xi' \leftarrow 0})} \rangle_{\alpha} \\ &= \langle e^{-\beta W_{\alpha}^{\Gamma^{1 \leftarrow \xi'}}} \rangle_{\alpha} \times \langle e^{-\beta W_{\alpha}^{\xi' \leftarrow 0}} \rangle_{\alpha} \end{aligned} \quad (16)$$

The second equality follows from the fact that the trajectories in the $1 \leftarrow \xi'$ and $\xi' \leftarrow 0$ processes are uncoupled and independently sampled. Combining eqs 8, 10, 14, and 16, we obtain the desired result:

$$\Delta G^{1 \leftarrow 0} = \Delta G^{1 \leftarrow \xi'} + \Delta G^{\xi' \leftarrow 0} \quad (17)$$

where the initial value of the environment $\Gamma^{1 \leftarrow \xi'}(t')$ at the beginning of the $1 \leftarrow \xi'$ process, in principle, can be chosen to be any arbitrary (but the same) state that is accessible to a $\xi' \leftarrow \xi'$ process. However, the choice of that intermediate state will affect the accuracy and convergence of the approach insofar as better choices would be more easily accessible and thus require less numerical relaxation in the evolution of the $1 \leftarrow \xi'$ process. The best such choice is one that corresponds to a typical structure (not the minimum energy state) associated with the nonequilibrium process. To this end, we choose $\Gamma^{1 \leftarrow \xi'}(t')$ according to the $\Gamma_{\alpha}^{1 \leftarrow \xi'}(t')$ corresponding to the trajectory α , which minimizes the work difference, $|\Delta G^{\xi' \leftarrow 0} - W_{\alpha}^{\xi' \leftarrow 0}(t')|$.

Repeated application of eq 17 and the associated prescription for the choice of intermediate environment variables Γ for N steps gives rise to the desired final expression for the adaptive free energy difference:

$$\Delta G = \sum_{i=1}^N \Delta G^{i \leftarrow (i-1)} \quad (18)$$

where i labels the corresponding steps. In the limit that the “environment variables” are empty—i.e., that the dimensionality of the Γ space is zero—the adaptive procedure reduces

to the use of the Jarzynski equality with the additivity trivially arising from the fact that the free energy is a state function.

In so far as the bath has been assumed to be Gaussian, the adaptive procedure should fail if the second-order cumulants in the work of a given set of trajectories begin to be nontrivial. As is shown below, the adaptive procedure does indeed satisfy this requirement.

2.3.3. Implementation of the Method. In this work, steered MD simulation is performed by pulling PRO5 at a constant velocity relative to the α helix on NPY. The choice of PRO5 is motivated both by experiment and computation. It has been previously reported that amino acids 1–4 of NPY (TYR1 to LYS4) form salt bridges with corresponding receptors.⁵⁵ Recent studies have indicated that binding hot spots at protein–protein interfaces exhibit high frequency fluctuation.⁵⁶ This suggests that the four residues from TYR1 to LYS4 of the NPY tail fluctuate faster than the other tail residues. Therefore, the choice of PRO5, rather than one of these other residues, allows us to drive the unfolding of the semirigid tail (including residues PRO5 to ASP11) while allowing the residues from TYR1 to LYS4 to fluctuate freely. Meanwhile the α helix must be represented by at least two fixed points so as to define the requisite hinge motion. These residues are LEU24 on the α helix and ALA12 on the hinge connecting the helix to the polyproline tail. The constrained system can therefore be designated through two variables: the LEU24–ALA12–PRO5 angle and the ALA12–PRO5 distance.

The external forces that carry the system along the unfolding path, $\xi(z_x)$, are imposed by way of a predefined potential $H'(\xi(z_x); \lambda)$. With the addition of this new potential, the extended time-dependent Hamiltonian, H^{ext} , becomes

$$H^{\text{ext}}(z, \Theta; \lambda) = H_{\xi}^{\text{SEB}}(\Gamma, \Theta) + H'(\xi(z_x); \lambda) \quad (19)$$

where

$$H(\xi(z_x); \lambda) = \frac{1}{2}k[\xi_x(z_x) - \lambda]^2 \quad (20)$$

for a specified time-dependent process $\lambda(t)$. In the case of NPY, the pulling process is staged in N linear steps so as to approximate the circular unbinding process in a coordinate frame in which the center of mass of the hinge (ALA12) is the origin. The position of the center of mass of PRO5 \vec{r} encodes the radius and angle of the system $\xi_x(z_x)$. Thus, for each step i , the auxiliary potential in eq 20 becomes

$$U_i(\vec{r}) = \frac{1}{2}k[\vec{r}(t) - (\vec{r}_i + v_i\vec{n}_i t)]^2 \quad (21)$$

where \vec{r}_i is the position of PRO5 at the beginning of the interval, v_i is the velocity to move the particle to the end in the fixed time step, and \vec{n}_i is the direction between the initial and final positions of PRO5. The position $\lambda_i \equiv (\vec{r}_i + v_i\vec{n}_i t)$ can be associated with an auxiliary particle (or dummy atom) that follows smoothly the prescribed unfolding path. As it does so, it exerts a work on the system ξ that is given by

$$\Delta W_i(t) = \int_{t_{i-1}}^t \vec{F}_i \vec{n}_i v dt \quad (22)$$

where the force $\vec{F}_i = -\nabla U_i(\xi_x(z_x))$ is related to the corresponding potential of eq 21. The corresponding free energy change, ΔG^{t-0} , at time t within the i th interval can now be calculated using the adaptive work expression in eq 18, i.e.,

$$e^{-\beta\Delta G_{t-0}} = \langle e^{-\beta\Delta W_i(t)} \rangle_i \times \prod_{j=1}^{i-1} \langle e^{-\beta\Delta W_j(t_j)} \rangle_j \quad (23)$$

where the subscript on the angle brackets denotes the averaging over the trajectories in the corresponding interval.

2.4. Transition State Theory and Rates. The experimental results, unfortunately, do not provide a potential of mean force that can be used to compare directly to the computational work. Instead, we use the relative stability of the folded and unfolded states (as suggested by the calculated ΔG^{u-f}) to compare to the experimentally known stable structures. In addition, the rates of the unfolding and folding processes can be determined using transition state theory for the PMF determined along the unfolding path. These will be compared to the findings from both the molecular dynamics trajectories and experiment.

The simple transition state rate is

$$K = \frac{k_B T}{h} e^{-\frac{\Delta G^\ddagger}{k_B T}} \quad (24)$$

where ΔG^\ddagger is the free energy barrier of the transition. Although much work has been done to go beyond this simple estimate,^{57–59} it is reasonably accurate for the order of magnitude of the rate.

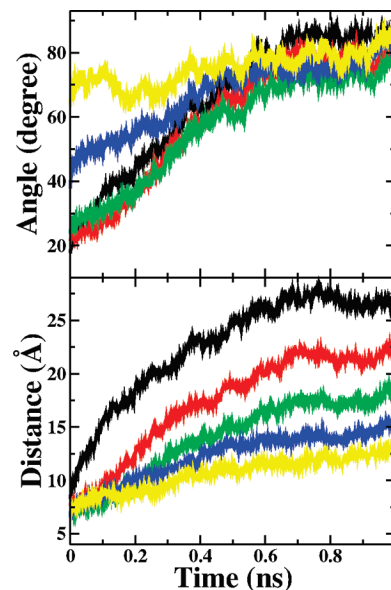


Figure 4. (Bottom) Average time-dependent displacements as NPY unfolds at 500 K, shown for several tail–helix residue pairs defined in Figure 3 using the same labeling scheme. (Top) Corresponding time-dependent angles spanned by a given pair of residues with respect to ALA12.

3. Results and Discussion

3.1. Identifying the Unfolding Path of NPY. Unfolding of NPY was first investigated through unconstrained MD simulations. MD trajectories were propagated using NAMD with the CHARMM force field in an explicit water solvent (TIP3P). At each of several temperatures, 300, 367, 433, and 500 K, 50 independent free MD simulations were integrated for 1 ns. At low temperatures, no unfolding was observed within the 1 ns observation window of the trajectories (not shown). At 500 K, all of the 50 generated trajectories unfolded in less than 1 ns.

Detailed analysis of the time dependence of the helix–tail separation in the 500 K unfolding trajectories reveals a hinge-like motion. The distance between the five pairs of residues initially in contact within the folded NPY are shown in the bottom panel of Figure 4. Pairs of residues farther from the turn (ALA12) move to more distant positions as the protein unfolds. All but the farthest residue pairs sweep a similar angle relative to the turn (ALA12), as shown in the top panel of Figure 4. This suggests that the tail hinges away from the helix about the turn during the unfolding process. It does not, however, follow this path linearly. The farthest residue pairs violate the quantitative agreement because the residues at the end of the tail are much more mobile and exhibit large fluctuations in position.

Both experimental^{60,61} and computational⁵⁵ studies have suggested that residues 1–4 of the NPY tail form a pharmacophore that plays an active role during NPY binding to receptors. As postulated by Ertekin et al.,⁵⁶ interface residues that are in close contact with binding protein residues have a higher packing density and exhibit high frequency fluctuation.⁵⁶ The dynamics of the tail shown in Figure 5 are in good agreement with these previous reports. The part of the tail that is proximal to the hinge (including

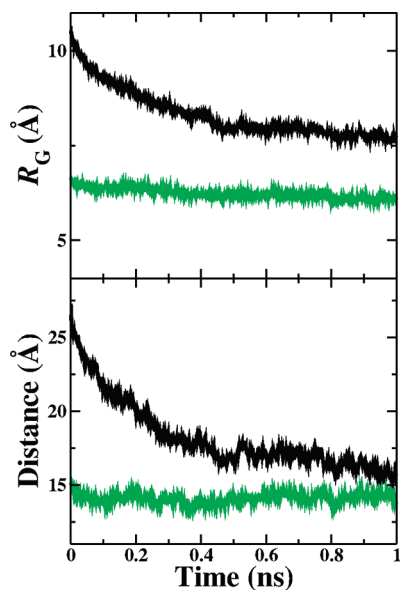


Figure 5. (Top and bottom) Radius of gyration and end-to-end displacement as a function of simulation time of two different segments of the NPY tail: ALA12 to TYR1 (black) and ALA12 to PRO5 (green). [cf. Figure 3b for the identification of the residues.]

all the residues up to PRO5) have nearly similar geometric properties (in terms of the length and radius of gyration) through the entire unfolding process. The rest of the tail, however, exhibits a significant geometric change through the unfolding process. It appears to be relaxation of the tail end toward a more compact structure in the vicinity of the proximal part of the tail.

The unfolding path thus appears to be primarily following the unhinging of the proximal part of the tail about the ALA12 hinge. Through this process, the tail appears to be nearly rigid up to PRO5, while the more distant residues are much more mobile. Hence, PRO5 is associated in the remainder of this work with the unfolding (reaction) path illustrated in Figure 4. Following Daggett and co-workers,² we therefore suppose that this unfolding path is followed not just at the elevated temperature of 500 K but also at experimentally accessible temperatures.

3.2. The PMF along the Unfolding Path. Our objective is to learn about the dynamics of NPY at temperatures relevant to the experimental systems. The temperature accelerated MD simulations provided us rates only at the locally stable temperature of 500 K. They also suggested an unfolding path along which we can calculate the PMF at lower temperatures for the purpose of obtaining relative rate information as will be done in the next subsection. The PMF must be calculated at 500 K for comparison with the MD simulations. For the lower temperature, we chose 310 K, as it is the body or *in vivo* temperature and is the temperature at which several experimental studies have explored the NPY dynamics.^{17,18,22} The determination of the PMF at these two temperatures is nontrivial because the models are quite large (consisting of 40 123 atoms), for which a single nanosecond trajectory takes approximately 100 h on one computer core. Nevertheless, the nonequilibrium SMD approaches described in the previous section were used to obtain the PMFs. The

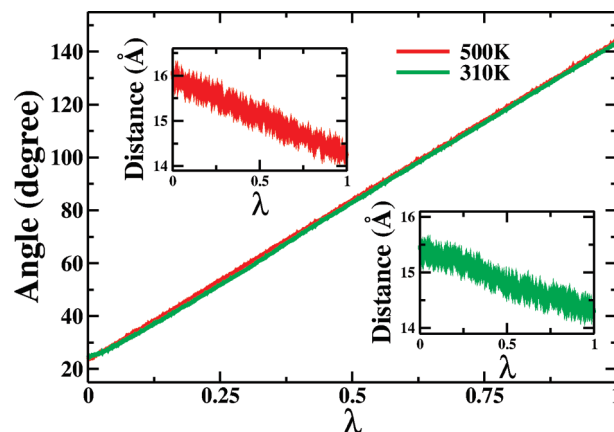


Figure 6. The average displacements in the system shown along the parametrized path λ for the adaptive SMD at 310 K and 500 K. The displacements $\xi_x = (r, \theta)$ fixed by the nonequilibrium process correspond to the radial distance r of PRO5 from the hinge (ALA12) and the angle θ spanned by the PRO5–ALA12 and LEU24–ALA12 vectors.

nonequilibrium simulations were realized using NAMD with the CHARMM force field for NPY in an explicit water solvent (TIP3P). All standard configuration parameters were the same as in the unconstrained MD simulations. The PMFs determined by either SMD approach required 110 h running on 48 2.33 GHz Intel 64 CPUs for 144 1-ns trajectories at a cost of 5280 CPU hours.

Steered MD trajectories have been obtained at a high temperature (500 K) as well as at body temperature (310 K). The unhinging of the tail was steered by pulling PRO5 (coupled to a dummy atom through a spring constant as per eq 20) relative to the virtually fixed residues ALA12 at the turn of the loop and LEU24 on the α helix. The unfolding path, which the dummy atom follows, is a discretization of the pseudocircular path shown in Figure 3b with each of the N finite steps taken to be linear. Specifically, the external force was applied on PRO5 to steer it from an initial configuration of the PRO5–ALA12–LEU24 angle θ_{initial} and radius r_{initial} to the final values, θ_{final} and r_{final} . At 500 K, $\theta_{\text{initial}} = 24.36^\circ$ and $r_{\text{initial}} = 16.09 \text{ \AA}$. At 310 K, $\theta_{\text{initial}} = 24.41^\circ$ and $r_{\text{initial}} = 15.49 \text{ \AA}$. At both temperatures, the final configuration is $\theta_{\text{final}} = 144.4^\circ$ and $r_{\text{final}} = 14.3 \text{ \AA}$. The initial configurations for the two temperatures differ because each was prepared from equilibration runs at the respective temperatures. All control parameters, such as the pulling velocity ($v = 33 \text{ \AA/ns}$) and the spring constant ($k = 7.2 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$), were kept identical to each other so as to render comparable results. The degree to which the PRO5 residue followed the unfolding path through the SMD simulations is shown in Figure 6. On average, both θ and r follow the linear displacement well, as expected for a constant velocity pulling SMD simulation. The fluctuations around the average are small and also consistent with this conclusion.

At each temperature, 144 independent SMD trajectories were generated. (The number is 144, not 100, because of technical reasons related to the architecture of the particular computer cluster and the number of simultaneous trajectories—three—that could be run per core without increasing the wall clock time.) This number was sufficient to converge the

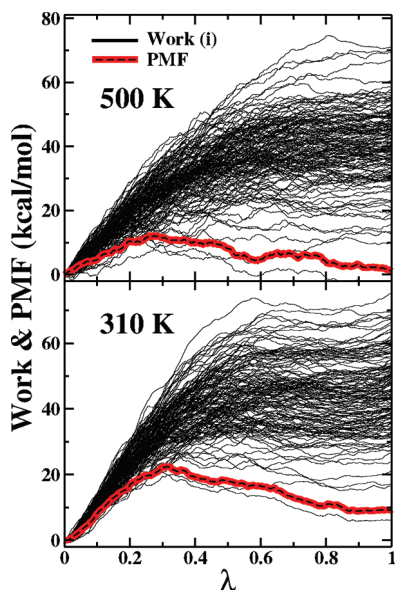


Figure 7. The work for 144 individual trajectories α (in black) and the PMF (in thick-red highlighting of a black-dashed curve) obtained using the Jarzynski equality displayed as a function of the parametrized unfolding path at 500 K (top panel) and 310 K (bottom panel).

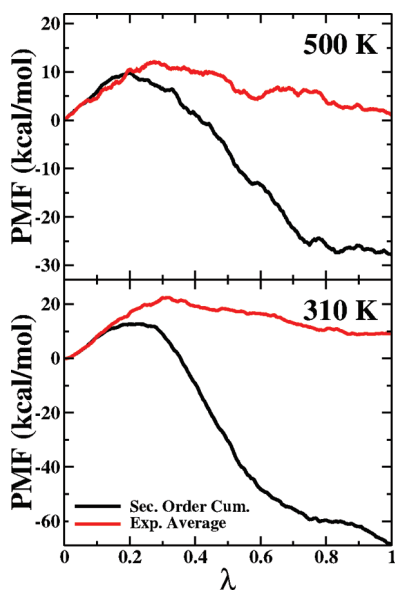


Figure 8. The PMF obtained using Jarzynski's equality (in red, cf. eq 5) and second-order cumulant expression (in black, cf. eq 7) obtained from a standard SMD calculation with 144 trajectories displayed as a function of the parametrized unfolding path at 500 K (top panel) and 310 K (bottom panel).

adaptive SMD trajectories and therefore serves as a good foil for the comparison of the two methods utilizing a similar amount of computational resources. Figure 7 shows the work and the averaged PMF using Jarzynski's relation at both 500 K (top) and 310 K (bottom). There are only a limited number of trajectories contributing to the PMF of the system at each temperature. This suggests a need for many more trajectories in order to converge the Jarzynski average. Indeed, the original deca-alanine in vacuum SMD PMFs calculated by

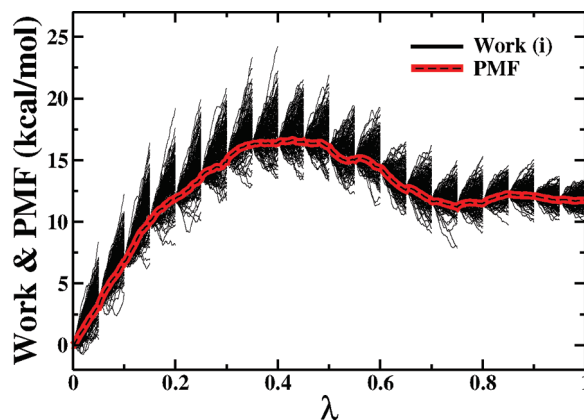


Figure 9. The work for 144 individual trajectories α (in black) and the PMF (in thick-red highlighting of a black-dashed curve) obtained using *adaptive* SMD displayed as a function of the parametrized unfolding path at 500 K.

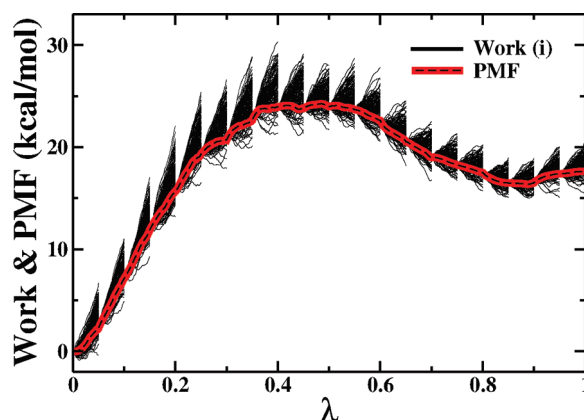


Figure 10. The work for 144 individual trajectories α (in black) and the PMF (in thick-red highlighting of a black-dashed curve) obtained using *adaptive* SMD displayed as a function of the parametrized unfolding path at 310 K.

the Schulten group⁴³ required over 10 000 trajectories on this much smaller system.

The lack of convergence of this approach (using a limited number of trajectories) is also illustrated by the comparison of the PMF between Jarzynski's average and the second-order cumulant expression shown in Figure 8. The two expressions are equal in the limit that the work distribution is Gaussian because of the well-known Marcinkiewicz's theorem.⁶² The lack of agreement between the two expressions is due both to the use of too few trajectories and also the fact that the observed trajectories were able to stray far from the relevant configurations. The consequence of the latter is that the statistics of the work contributions are far from Gaussian, and hence the second-order cumulant expression deviates greatly from Jarzynski's average.

The adaptive SMD method described in section 2.3 pre-empted the work distribution of such high barrier PMFs from losing their Gaussian nature by partitioning the unfolding path into several steps over which the PMF undergoes smaller changes. For the curved unfolding path illustrated in Figure 3b, we found convergence when we used 20 steps and a mere 144 trajectories per step. As noted earlier, the total computational cost is almost the same, excluding the

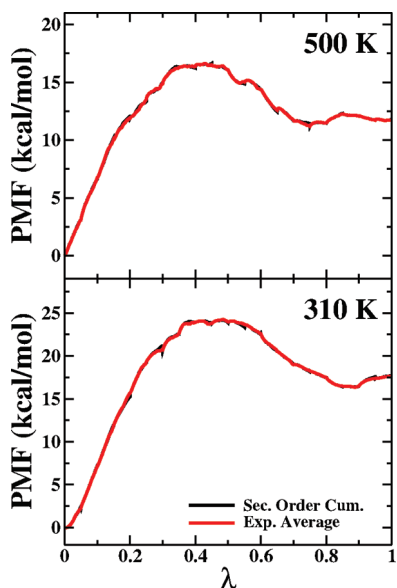


Figure 11. The PMF obtained using Jarzynski's equality (in red, cf. eq 5) and second-order cumulant expression (in black, cf. eq 7) obtained from an *adaptive* SMD calculation with 144 trajectories displayed as a function of the parametrized unfolding path at 500 K (top panel) and 310 K (bottom panel). Note that the black curves are nearly entirely covered by the red curves and hence not very visible.

negligible cost required for trajectory comparison at the end of each step. As before, 144 independent *adaptive* SMD trajectories were generated for each of the two temperatures, 310 and 500 K.

The work and the averaged PMF using *adaptive* SMD (eq 18) are shown in Figures 9 (500 K) and 10 (310 K). Unlike in the results for the standard SMD simulations shown above, the PMFs are not dominated by the lowest energy trajectories. On the contrary, the PMF for each step has contributions from several trajectories. The results obtained for the PMF using the *adaptive* SMD method (cf. eq 18) with Jarzynski's equality (cf. eq 5) shown in Figures 9 and 10 are reproduced in Figure 11. Therein, the PMFs obtained with the second-order cumulant expression (cf. eq 7) are also shown. The agreement is remarkable, as the differences are not visible at this level of resolution. Though not shown, the number of sampled trajectories was doubled, leading to no significant change in the converged PMFs. Thus, the *adaptive* nonequilibrium process appears to result in a better estimate of the PMF with a limited number of trajectories, i.e., computational resources.

The PMFs in Figure 11 also provide information about the energetics of the unfolding process of NPY at the two temperatures, 310 and 500 K. The barrier heights to unfolding at 310 and 510 K are 24 and 17 kcal/mol, respectively. The 7 kcal/mol lowering of the barrier height at the higher temperature presumably results from the fact that the orthogonal degrees of freedom have a higher frequency at the folded state than at the barrier to unfolding. The difference is even more dramatic when one compares the ratio of the barrier height in units of $k_B T$. At 310 and 510 K, these ratios are 40 and 17, respectively, which suggests that the rates at 310 K are several orders of magnitude slower

than at 510 K. As NPY had exhibited only partial unfolding at the high temperature, it is therefore not surprising that the low temperature MD simulations did not unfold within the 1 ns observation window. In addition, the folded state has a lower PMF and is therefore predicted to be the more stable form for monomeric NPY at 310 K.

3.3. The Folding and Unfolding Rates of NPY. The barrier height for the transition from the folded to unfolded conformations of NPY has been found to be 24 and 17 kcal/mol for 310 and 500 K, respectively. From these activation energy values, the rates have been calculated as $5.1 \times 10^{-5} \text{ s}^{-1}$ and $5.5 \times 10^5 \text{ s}^{-1}$ again for 310 and 500 K, respectively. The inverse of these rates corresponds to a lifetime for the NPY unfolding transition. At the elevated temperature (500 K), this lifetime is 1.8 μs and is consistent with the fact that the NPY trajectories would explore the unfolded space within 1 ns, as seen in the MD simulations. At body temperature (310 K), this would suggest a lifetime of over 5 h, which is consistent with the fact that none of the low temperature NPY proteins unfolded during the MD simulations.

4. Conclusions

We have developed an *adaptive* algorithm extending the Schulten–Jarzynski SMD method for the calculation of PMFs when the subsystem is dragged across long nonlinear paths. In such cases, the PMF can span many $k_B T$'s, leading to the sampling of nonequilibrium trajectories with work functions that fluctuate over a very large energy range. Consequently, only a small fraction of the trajectories generated from the SMD contribute nontrivially to the Jarzynski average. In order to numerically converge this average, one then needs to generate a large number of trajectories, which can be cost prohibitive. The *adaptive* algorithm allows one to break up the SMD calculation in a series of steps. The free energy difference across each such step is much smaller, and thereby allows convergence of the Jarzynski average with significantly fewer trajectories. In this sense, the *adaptive* algorithm is not formally better than the standard approach, but it is significantly more numerically efficient.

The SMD approach using the Jarzynski equality has been implemented experimentally in molecular force pulling experiments by several groups^{48,49} with the underlying theory having been recently clarified by Zimanyi and Silbey.⁵² The numerical success of the *adaptive* technique developed here could also be extended to such molecular force pulling experiments. Instead of using single constant velocity force pulling, the *adaptive* procedure would suggest the use of staged (or stepped) force pulling events. The pauses between the stages need only be held long enough so that the environment to the constrained system can relax (while applying zero work.)

The unfolding path of NPY has been suggested by temperature accelerated MD simulations to be the unhooking of the polyproline tail away from the α helix about the turn (near ALA12.) The NPY tail maintains its overall shape between PRO5 and ASP11 while unhooking away from the NPY helix. As the NPY unfolds along the path, the first four N-terminal residues (TYR1 to LYS4) fluctuate freely when

no biasing force is applied on them. This observation is consistent with earlier reports which hypothesized that these four residues on the polyproline tail of NPY form a pharmacophore at the NPY–receptor interface during NPY bioactivity.⁵⁵ This was earlier justified by the fact that protein–protein interfaces have been seen to be enriched by the presence of high frequency fluctuating residues.⁵⁶ The observation that this unfolding process can be reduced to a dominant one-dimensional pathway is not uncommon. Several groups^{63,64} have reported the possibility of such a simplification if the dynamics are dominated by passage across a single barrier, as we have seen here for NPY.

The potentials of mean force along the folding path provide a more detailed view of the dynamics. This was possible because of a generalization of SMD (also known as force-biased simulations) using the adaptive scheme introduced in this work. The barrier heights and associated rates of the NPY unfolding transition at an elevated temperature (500 K) and the *in vivo* temperature (310 K) agree well with the numerical MD simulations (reported here) and those authors^{16–18} which have proposed the stability of PP-fold on the basis of their experimental findings. At the *in vivo* temperature, we have determined an unfolding rate for NPY on a time scale longer than 5 h. The typical single-domain protein folding/unfolding time scale is a few microseconds at the fastest and a couple hundred microseconds at the slowest.⁶⁵ We thus conclude that at 310 K monomeric NPY does not unfold. This conclusion is consistent with our preliminary unconstrained MD simulations in which NPY did not unfold at temperatures up to 433 K. The fact that the unfolded NPY state has a higher free energy than the folded structure also suggests that the NPY monomer in solution is folded in the pancreatic–polypeptide (PP) fold. This result is also consistent with the experimental hypothesis that the NPY dimer is biologically inactive in solution because the tail moves away from the PP-fold.¹⁸ This indirectly suggests that the biological activity of the NPY monomer results from the stability of the folded structure in agreement with the energetic stability found in this work. Recently, Bader et al.²² reported that the micelle-bound form of NPY demonstrates a less ordered conformation than the PP-fold. In this less-ordered conformation, the NPY tail is observed to be fluctuating (Figure 3 in Bader et al.) while the α -helix remains stable. Our results suggest that this is due to the specific contacts, formed between micelle and side chains of the NPY α -helix, replacing the favorable polyproline tail and α -helix contacts observed in the PP-fold.

Acknowledgment. This work has been partially supported by the National Science Foundation (NSF) through Grant No. CHE 0749580. The computing resources necessary for this research were provided in part by the National Science Foundation through TeraGrid resources provided by the Purdue Dell PowerEdge Linux Cluster (Steele) under grant number TG-CTS090079, and by the Center for Computational Molecular Science & Technology through Grant No. CHE 0946869.

References

- (1) Galzitskaya, O. V.; Finkelstein, A. V. A theoretical search for folding/unfolding nuclei in three-dimensional protein structure. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11299–11304.
- (2) Best, R. B.; Li, B.; Steward, A.; Daggett, V.; Clarke, J. Can non-mechanical proteins withstand force? Stretching barnase by atomic force microscopy and molecular dynamics simulation. *Biophys. J.* **2001**, *81*, 2344–2356.
- (3) Ng, S. P.; Rounsevell, R. W. S.; Steward, A.; Geierhaas, C. D.; Williams, P. M.; Paci, E.; Clarke, J. Mechanical unfolding of TNfn3: The unfolding pathway of a fnIII domain probed by protein engineering, AFM and MD simulation. *J. Mol. Biol.* **2005**, *350*, 776–789.
- (4) Hisatomi, Y.; Katagiri, D.; Neya, S.; Hara, M.; Hoshino, T. Analysis of the unfolding process of green fluorescent protein by molecular dynamics simulation. *J. Phys. Chem. B* **2008**, *112* (29), 8672–8680.
- (5) Mayor, U.; Johnson, C. M.; Daggett, V.; Fersht, A. R. Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 13518–13522.
- (6) Lu, H.; Isralewitz, B.; Krammer, A.; Vogel, V.; Schulten, K. Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophys. J.* **1998**, *75*, 662–671.
- (7) Grater, F.; Shen, J.; Jiang, H.; Gautel, M.; Grubmuller, H. Mechanically induced titin kinase activation studied by force-probe molecular dynamics simulations. *Biophys. J.* **2005**, *88*, 790–804.
- (8) Gao, M.; Craig, D.; Vogel, V.; Schulten, K. Identifying unfolding intermediates of FN-III10 by steered molecular dynamics. *J. Mol. Biol.* **2002**, *323*, 939–950.
- (9) Lu, H.; Schulten, K. Steered molecular dynamics simulations of force-induced protein domain unfolding. *Proteins* **1999**, *35*, 453–463.
- (10) Gray, T.; Morley, J. Neuropeptide Y: Anatomical distribution and possible function in mammalian nervous system. *Life Sci.* **1986**, *38*, 389–401.
- (11) Dumont, Y.; Fournier, A.; Quirion, R. Neuropeptide Y and neuropeptide Y receptor subtypes in brain and peripheral tissues. *Progr. Neurobiol.* **1992**, *38*, 125–167.
- (12) Turton, M.; O’Shea, D.; Bloom, S. *Central effects of neuropeptide Y with emphasis on its role in obesity and diabetes*; Academic Press: San Diego, CA, 1997.
- (13) Larhammar, D. Structural diversity of receptors for neuropeptide Y, peptide YY and pancreatic polypeptide. *Regul. Pept.* **1996**, *65*, 165–174.
- (14) Wraith, A.; Tornsten, A.; Chardon, P.; Harbitz, I.; Chowdhary, B. P.; Andersson, L.; Lundin, L.-G.; Larhammar, D. Evolution of the neuropeptide Y receptor family: Gene and Chromosome duplications deduced from the cloning and mapping of the five receptor subtype genes in pig. *Genome Res.* **2000**, *10*, 302–310.
- (15) Larhammar, D. Evolution of neuropeptide Y, peptide YY, and pancreatic polypeptide. *Regul. Pept.* **1996**, *62*, 1–11.
- (16) Li, X.; Sutcliffe, M. J.; Schwartz, T. W.; Dobson, C. M. Sequence-specific proton NMR assignments and solution structure of bovine pancreatic polypeptide. *Biochemistry* **1992**, *31*, 1245–1253.

- (17) Darbon, H.; Bernassau, J.; Deleuze, C.; Chenu, J.; Roussel, A.; Cambillau, C. Solution conformation of human neuropeptide Y by ¹H nuclear magnetic resonance and restrained molecular dynamics. *Eur. J. Biochem.* **1992**, *209*, 765–771.
- (18) Nordmann, A.; Blommers, M.; Fretz, H.; Arvinte, T.; Drake, F. Aspects of the molecular structure and dynamics of neuropeptide Y. *Eur. J. Biochem.* **1999**, *261*, 216–226.
- (19) Cowley, D.; Hoflack, J.; Pelton, J.; Saudek, V. Structure of neuropeptide Y dimer in solution. *Eur. J. Biochem.* **1992**, *205*, 1099–1106.
- (20) Mierke, D.; Durr, H.; Kessler, H.; Jung, G. Neuropeptide Y: Optimized solid-phase synthesis and conformational analysis in trifluoroethanol. *Eur. J. Biochem.* **1992**, *206*, 39–48.
- (21) Monks, S.; Karagianis, G.; Howlett, G.; Norton, G. Solution structure of human neuropeptide Y. *J. Biomol. NMR* **1996**, *8*, 379–390.
- (22) Bader, R.; Bettio, A.; Beck-Sickinger, A. G.; Zerbe, O. Structure and dynamics of micelle-bound neuropeptide Y: Comparison with unligated NPY and implications for receptor selection. *Genome Res.* **2000**, *10*, 302–310.
- (23) Lerch, M.; Mayrhofer, M.; Zerbe, O. Structural similarities of micelle-bound peptide YY (PYY) and neuropeptide Y (NPY) are related to their affinity profiles at the Y receptors. *J. Mol. Biol.* **2004**, *339*, 1153–1168.
- (24) Bettio, A.; Dinger, M. C.; Beck-Sickinger, A. G. The neuropeptide Y monomer in solution is not folded in the pancreatic-polypeptide fold. *Protein Sci.* **2002**, *11*, 1834–1844.
- (25) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (26) Blundell, T. L.; Pitts, J. E.; Tickle, I. J.; Wood, S. P.; Wu, C.-W. X-ray analysis (1.4-Å resolution) of avian pancreatic polypeptide: Small globular protein hormone. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 4175–4179.
- (27) Daggett, V.; Fersht, A. R. Is there a unifying mechanism for protein folding. *Trends Biochem. Sci.* **2003**, *28*, 18–25.
- (28) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *28*, 1781–1802.
- (29) Brooks, B.; Brucoleri, R.; Olafson, R.; States, D.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (30) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (31) Park, P. J.; Lee, S. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (32) Nose, S. Constant-temperature molecular dynamics. *J. Phys.: Condens. Matter* **1990**, *2*, SA115–SA119.
- (33) Cerutti, D. S.; Duke, R.; Freddolino, P. L.; Fan, H.; Lybrand, T. P. A vulnerability in Popular Molecular Dynamics Packages Concerning Langevin and Andersen Dynamics. *J. Chem. Theory Comput.* **1999**, *4*, 1669–1680.
- (34) Okur, A.; Roe, D. R.; Cui, G.; Hornak, V.; Simmerling, C. Improving Convergence of Replica-Exchange Simulations through Coupling to a High-Temperature Structure Reservoir. *J. Chem. Theory Comput.* **2007**, *3*, 557–568.
- (35) Day, R.; Daggett, V. Increasing Temperature Accelerates Protein Unfolding Without Changing the Pathway of Unfolding. *J. Mol. Biol.* **2002**, *322*, 189–203.
- (36) Chan, H. S.; Dill, K. A. Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics. *Proteins* **1998**, *30*, 2–33.
- (37) Matagne, A.; Jamin, M.; Chung, E. E.; Robinson, C. V.; Radford, S. E.; Dobson, C. M. Thermal unfolding of an intermediate is associated with non-Arrhenius. *J. Mol. Biol.* **2000**, *297*, 193–210.
- (38) Khan, F.; Chuang, J. I.; Gianni, S.; Fersht, A. R. The kinetic pathway of folding of barnase. *J. Mol. Biol.* **2003**, *333*, 169–186.
- (39) Nguyen, H.; Jaeger, M.; Moretto, A.; Gruebele, M.; Kelly, J. W. Tuning the free energy landscape of a WW domain by temperature, mutation, and truncation. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *352*, 370–381.
- (40) Chandler, D. *Introduction to modern statistical mechanics*; Oxford: New York, 1987.
- (41) Jarzynski, C. Equilibrium free-energy differences from non-equilibrium measurements: A master-equation approach. *Phys. Rev. E* **1997**, *56*, 5018–5035.
- (42) Jarzynski, C. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* **1997**, *78*, 2690–2693.
- (43) Park, S.; Schulten, K. Calculating potentials of mean force from steered molecular dynamics simulations. *J. Chem. Phys.* **2004**, *120*, 5946–5961.
- (44) Xiong, H.; Crespo, A.; Marti, M.; Estrin, D.; Roitberg, A. E. Free energy calculations with non-equilibrium methods: applications of the Jarzynski relationship. *Theor. Chem. Acta* **2006**, *116*, 338–346.
- (45) Torras, J.; de M. Seabra, G.; Roitberg, A. E. A Multiscale Treatment of Angeli's Salt Decomposition. *J. Chem. Theory Comput.* **2009**, *5*, 37–46.
- (46) Piccinini, E.; Ceccarelli, M.; Affinito, F.; Brunetti, R.; Jacoboni, C. Biased Molecular Simulations for Free-Energy Mapping: A Comparison on the KcsA Channel as a Test Case. *J. Chem. Theory Comput.* **2008**, *4*, 173–183.
- (47) Huang, H.; Ozkirimli, E.; Post, C. B. Comparison of Three Perturbation Molecular Dynamics Methods for Modeling Conformational Transitions. *J. Chem. Theory Comput.* **2009**, *5*, 1304–1314.
- (48) Liphardt, J.; Dumont, S.; Smith, S. B.; Bustamante, C. Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski's equality. *Science* **2002**, *296*, 1832–1835.
- (49) Douarache, F.; Ciliberto, S.; Petrosyan, A.; Rabbiosi, L. An experimental test of the Jarzynski equality in a mechanical experiment. *Europhys. Lett.* **2005**, *70*, 593–599.
- (50) Crooks, G. E. Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *J. Stat. Phys.* **1998**, *90*, 1481–1487.
- (51) Hummer, G.; Szabo, A. Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 3658–3661.

- (52) Zimanyi, E. N.; Silbey, R. J. The work-Hamiltonian connection and the usefulness of the Jarzynski equality for free energy calculations. *J. Chem. Phys.* **2009**, *130*, 171102.
- (53) Park, S.; Khalili-Araghi, F.; Tajkhorshid, E.; Schulten, K. Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. *J. Chem. Phys.* **2003**, *119*, 3559–3566.
- (54) Amaro, R.; Tajkhorshid, E.; Luthey-Schulten, Z. Developing an energy landscape for the novel function of a (beta/alpha)₈ barrel: Ammonia conduction through HisF. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7599–7604.
- (55) Sylte, I.; Andrianjara, C.; Calvet, A.; Pascal, Y.; Dahl, S. Molecular dynamics of NPY Y1 receptor activation. *Bioorg. Med. Chem.* **1999**, *7* (12), 2737–2748.
- (56) Ertekin, A.; Nussinov, R.; Haliloglu, T. Association of putative concave protein-binding sites with the fluctuation behavior of residues. *Protein Sci.* **2006**, *15*, 2265–2277.
- (57) Hnggi, P.; Talkner, P.; Borkovec, M. Reaction-rate theory: Fifty years after Kramers. *Rev. Mod. Phys.* **1990**, *62*, 251–341, and references therein.
- (58) Pollak, E.; Talkner, P. Reaction rate theory: What it was, where it is today, and where is it going? *Chaos* **2005**, *15*, 026116–1–11.
- (59) Hernandez, R.; Bartsch, T.; Uzer, T. Transition state theory in liquids beyond planar dividing surfaces. *Chem. Phys.* **2010**, *370*, 270–276.
- (60) Beck-Sickinger, A. G.; Wieland, H. A.; Wittneben, H.; Willim, K.-D.; Rudolf, K.; Jung, G. Complete L-alanine scan of neuropeptide Y reveals ligands binding to Y1 and Y2 receptors with distinguished conformations. *Eur. J. Biochem.* **1994**, *225* (3), 947–958.
- (61) Fournier, A.; Gagnon, D.; Quirion, R.; Dumont, Y.; Pheng, L.-H.; St-Pierre, S. Conformational and biological studies of neuropeptide Y analogs containing structural alterations. *Mol. Pharmacol.* **1994**, *45*, 93–101.
- (62) Marcinkiewicz, J. Sur une propriete de la loi de Gauss. *Math. Z* **1939**, *44*, 612–618.
- (63) Berezhkovskii, A.; Szabo, A. One-dimensional reaction coordinates for diffusive activated rate processes in many dimensions. *J. Chem. Phys.* **2005**, *122*, 014503–014506.
- (64) Rhee, Y. M.; Pande, V. S. One-dimensional reaction coordinate and the corresponding potential of mean force from commitment probability distribution. *J. Phys. Chem. B* **2005**, *109*, 6780–6786.
- (65) Kubelka, J.; Hofrichter, J.; Eaton, W. A. The protein folding speed limit. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76–88.

CT100320G

JCTC

Journal of Chemical Theory and Computation

Efficient Multistate Reactive Molecular Dynamics Approach Based on Short-Range Effective Potentials

Hanning Chen,^{†,‡} Pu Liu,^{†,§} and Gregory A. Voth^{*,||}

Department of Chemistry, James Franck Institute, and Computation Institute, University of Chicago, 5735 South Ellis Avenue, Chicago, Illinois 60637, Department of Chemistry, Northwestern University, 2145 Sheridan Road, Evanston, Illinois 60208-3113, and Johnson & Johnson Pharmaceutical Research & Development, 665 Stockton Drive, Exton, Pennsylvania 19341

Received June 11, 2010

Abstract: Nonbonded interactions between molecules usually include the van der Waals force and computationally expensive long-range electrostatic interactions. This article develops a more efficient approach: the effective-interaction multistate empirical-valence-bond (EI-MS-EVB) model. The EI-MS-EVB method relies on a mapping of all interactions onto a short-range and thus, computationally efficient effective potential. The effective potential is tabulated by matching its force to known trajectories obtained from the full-potential empirical multistate empirical-valence-bond (MS-EVB) model. The effective pairwise interaction depends on and is uniquely determined by the atomic configuration of the system, varying only with respect to the hydrogen-bonding topology. By comparing the EI-MS-EVB and full MS-EVB calculations of several equilibrium and dynamic properties important to hydrated excess proton solvation and transport, we show that the EI-MS-EVB model produces very accurate results for the specific system in which the tabulated potentials were generated. The EI-MS-EVB potential also transfers reasonably well to similar systems with different temperatures and box sizes. The EI-MS-EVB method also reduces the computational cost of the nonbonded interactions by about 1 order of magnitude in comparison with the full algorithm.

1. Introduction

An empirical force field must either explicitly or implicitly account for all of the electrostatic interactions between charged particles. The long-range electrostatic interaction,^{1–5} which decays as an inverse function of the interparticle distance r , is a crucial element of many molecular simulations—especially for highly charged biological systems such as DNA.^{6–8} Reliable and accurate calculations of the Coulomb force are possible using lattice summation methods, including the original Ewald summation,⁹ particle–particle particle–mesh (P3M),^{10,11} and particle–mesh Ewald (PME)

algorithms.¹² Numerous simulations based on these techniques have produced results in good agreement with experimental data. Even with carefully chosen parameters, however, the computational cost of the original Ewald summation is $O(N^{3/2})$,¹³ with N being the number of charged particles in the system. By employing the fast Fourier transform, P3M and PME significantly accelerate long-range electrostatic calculations and reduce the cost to $O(N \log N)$.^{10,11} Although all three methods are much faster than a direct evaluation of all possible pairs, they are still very computationally expensive for large systems,¹⁴ and thus may not be well suited for highly scalable computations.

Fast multipole methods (FMMs), an alternative approach to the Ewald algorithms, are based on the multipole expansion of the electrostatic potential.^{15–18} Although their asymptotic computational scaling is claimed to be $O(N)$, much better than that of particle–mesh methods, in a trial

* Corresponding author e-mail: gavoth@uchicago.edu.

[†] These authors contributed equally to this work.

[‡] Northwestern University.

[§] Johnson & Johnson Pharmaceutical Research & Development.

^{||} University of Chicago.

conducted by Pollock and Glosli, the latter actually outperformed FMMs for all systems studied.¹⁷ Furthermore, it is not easy to obtain a reasonably good conservation of the total system energy using FMMs. A very high-order expansion must be used, at the cost of significantly more CPU (central processing unit) cycles. However, as pointed out by one of the reviewers, some recent progress has substantially reduced the prefactor implicit in the $O(N)$ notation,^{19,20} making the improved FMMs quite competitive.

Among the various ways to improve the efficiency of long-range force calculations, the most straightforward and convenient one is to truncate the interaction beyond a certain (usually fairly short) distance. In other words, in this approach, any nonbonded forces between particles separated by more than a certain cutoff distance are simply neglected.^{21,22} At the cutoff distances typically used in simulations, however, the long-range force is not yet sufficiently small to be ignored. The resulting discontinuity is known to produce severe artifacts in various properties of the simulated systems.^{23–36} Although the discontinuity can be diminished by smoothing or shifting functions, such methods cannot completely remove the artifacts.^{37–40}

Alternatively, interactions between particles at large length scales can be treated as a homogeneous dielectric medium in the mean-field sense. The contribution from all molecules beyond the cutoff distance can then be approximated using a Barker–Watts reaction-field (RF) correction.^{24,29,41–45} The problem with this approach is that many systems of interest (e.g., aqueous solutions of large proteins) exhibit intrinsically heterogeneous dielectric properties. In such cases, RF corrections with a short cutoff will induce significant artifacts. These errors can be reduced by setting a large cutoff, but this solution increases the computational expense, thereby limiting the applicability of the RF method.

The above summary is consistent with the concept that expensive, explicit computations are both necessary and inevitable for an accurate description of long-range interactions. However, recent thinking suggests that the effective electrostatic interaction in condensed systems might decay substantially faster than an inverse power of r . Indeed, several short-range potentials have been constructed that appear to account for the long-range electrostatic interaction.^{46–48} The damped, force-shifted (DFS) potential⁴⁶ is one good example; not only is it computationally efficient, with $O(N)$ computational cost, but it satisfactorily reproduces some thermodynamic and dynamic properties for a variety of systems, including argon in water, liquid water, crystalline water, NaCl crystals, and NaCl aqueous solutions. A weakness of the DFS potential is that some of its key variables (such as the damping parameter itself) can be optimized only by trial and error. Moreover, the fact that it relies on a predefined analytical function might limit the DFS potential's ability to represent the "best" effective short-range interaction in a wide range of systems.

Very recently, a different and systematic method for determining the optimal effective short-range potential was developed.^{48,49} This method is called coarse-graining in interaction space (CGIS). The goal is to minimize differences between the forces generated by a candidate effective short-

range potential and those generated by the full Coulomb potential. This innovative approach represents a "bottom-up" strategy, fundamentally different from the more common "top-down" strategies described previously. This force-matching (FM) algorithm^{50,51} takes into account the effects of the full Coulomb interaction, generating a statistically accurate effective^{48,49} potential. Furthermore, the new potential need not be a predefined analytical function. This method has successfully reproduced several important structural, energetic, and dynamical properties of bulk water systems.⁴⁸ The subsequent work by Shi et al.⁴⁹ took the theory a step further by analytically demonstrating that the effective short-range force approaches zero naturally at the FM cutoff and by deriving its corresponding analytic approximation.

Turning now to a more specific focus, the multistate empirical-valence-bond (MS-EVB)^{52–57} method represents a significant theoretical and computational advance toward understanding proton solvation and transport (PS&T). This phenomenon is critical to many chemical, biochemical, and biophysical systems. The MS-EVB method has provided statistically reliable and accurate descriptions of excess hydrated proton behavior in a wide variety of environments, including small water clusters⁵⁸ and the bulk and interfacial water phase,^{54,55,59} the aquaporin channels,^{60,61} the influenza A virus M2 channel,⁶² cytochrome *c* oxidase,⁶³ and liquid-phase imidazole.⁶⁴ The MS-EVB approach describes proton transport as a multipathway and multistep reaction, wherein distinct reactant-like (or product-like) intermediates with different chemical and hydrogen-bonding topologies are represented as individual states of an MS-EVB Hamiltonian matrix. The diagonal elements of the matrix represent the diabatic energies of the MS-EVB states, and the off-diagonal elements provide the coupling between any pair of states. The MS-EVB method is a sort of multistate or multiconfiguration molecular dynamics (MD) approach that allows chemically reactive processes such as Grotthuss proton shuttling^{56,57} to be modeled.

After diagonalization of the MS-EVB Hamiltonian and identification of the eigenvector with the lowest eigenvalue, the system nuclei are propagated with MD in accordance with the Hellmann–Feynman theorem. Empirical potentials and atomistic forces (hereafter denoted E&F) need to be calculated for each element in the EVB matrix, except for those off-diagonal elements that correspond to noncoupling MS-EVB states.^{52,54,55} In aqueous solutions, for example, approximately 30 MS-EVB states and roughly the same number of MS-EVB coupling pairs are typically required to describe the delocalized excess proton charge defect and to satisfactorily conserve total system energy. Thus, one should expect about 60 computationally expensive long-range electrostatic lattice summations to be carried out for each time step. This calculation is the dominant source of CPU cycles in an MS-EVB calculation. A new approach for accurately and efficiently evaluating the long-range electrostatic interactions in the MS-EVB method is therefore highly desirable for many interesting applications, especially those involving an implementation within a highly scalable computing environment.

The CGIS algorithm folds long-range electrostatic interactions into short-range effective potentials.^{48,49} In the present study, we demonstrate that this algorithm can be generalized to greatly accelerate MS-EVB calculations by simplifying most of the nonbonded interactions. This demonstration is important because proton transport as described by the MS-EVB model has various exponentially sensitive properties (such as diffusion barriers) that might not be properly modeled by more ad hoc effective interaction schemes. For a given atomistic configuration, distinct MS-EVB states differ only in their hydrogen-bonding topology. Thus, once the E&F have been determined for one EVB state, those of other states can be calculated conveniently and rapidly by identifying the few atom pairs that differ between the states. It is therefore not surprising that such short-range effective potentials are a particularly efficient tool for MS-EVB calculations. As will be demonstrated later, this “effective-interaction” MS-EVB (EI-MS-EVB) approach can substantially reduce the cost of nonbonded calculations in the MS-EVB calculations while producing structural, energetic, and dynamical properties comparable to those of the original full-potential (FP) MS-EVB model. More importantly, the improvement in computational efficiency becomes pronounced in larger systems.

The remaining sections of this article are organized as follows: Section 2 provides an overview of the CGIS FM algorithm and the MS-EVB methodology, followed by a detailed explanation of the construction of effective short-range potentials. Section 3 compares some physical properties of systems simulated using the EI-MS-EVB method and the regular (FP) MS-EVB method and discuss the efficiency of the algorithms. Conclusions and future directions are given in section 4.

2. Methods

2.1. Force-Matching Algorithm. The FM algorithm^{50,51} has previously been used to construct short-range effective potentials via the CGIS methodology.^{48,49} Given that the system itself is not coarse-grained and that nonbonded atomistic potentials are usually pairwise and additive, this procedure can produce very accurate effective potentials. Indeed, this accuracy has been demonstrated for various systems.⁴⁸

Briefly, the optimal parameters $\{g_m\}$ of an effective short-range interaction can be systematically derived from a known trajectory by variationally minimizing the residual function

$$\chi^2(\{g_m\}) = \frac{1}{N} \sum_{i,l} \|\vec{F}_{i,l}^{\text{ref}} - \vec{F}_{i,l}^{\text{eff}}(g_m, r_c)\|^2$$

In this formula, N is the number of atoms in the system, $\vec{F}_{i,l}^{\text{ref}}$ is the known force acting on atom i in frame l of the reference trajectory, and

$$\vec{F}_{i,l}^{\text{eff}}(g_m, r_c) = \sum_{d=1}^{N_p} \phi_{i,l}^d(r^n) g_d$$

is the effective force calculated from the configuration r^n given the parameter set $\{g_m\}$. In the effective force formula,

N_p is the number of elements in $\{g_m\}$, r^n represents the coordinates of the atoms, and $\phi_{i,l}^d(r^n)$ is a vector whose elements depend on the system configuration. Beyond the cutoff distance r_c , the effective force $\vec{F}_{i,l}^{\text{eff}}(g_m, r_c)$ is defined to be zero. A more detailed description of the force-matching procedure can be found elsewhere.^{48,50} It is worth emphasizing that, because $\vec{F}_{i,l}^{\text{ref}}$ is deduced from a real MD trajectory, it includes contributions from the whole system. Even though the effective force $\vec{F}_{i,l}^{\text{eff}}(g_m, r_c)$ includes contributions only from particles within a spherical region around atom i , it is constructed to imitate $\vec{F}_{i,l}^{\text{ref}}$ as closely as possible and therefore also includes information on the long-range interactions.

Mathematically, the residual function can be rewritten as a multidimensional quadratic function of the parameter set $\{g_m\}$ ^{65–67}

$$\chi^2 = \sum_{d,d'} G_{dd'} g_d g_{d'} - 2 \sum_d b_d g_d + \chi_0^2 \quad (1)$$

where

$$G_{dd'} = \frac{1}{N} \left\langle \sum_i \sum_l \phi_{i,l}^d(r^n) \cdot \phi_{i,l}^{d'}(r^n) \right\rangle \quad (2)$$

$$b_d = \frac{1}{N} \left\langle \sum_i \sum_l \phi_{i,l}^d(r^n) \cdot \vec{F}_{i,l}^{\text{ref}} \right\rangle$$

$$\chi_0^2 = \frac{1}{N} \left\langle \sum_i \sum_l \vec{F}_{i,l}^{\text{ref}} \cdot \vec{F}_{i,l}^{\text{ref}} \right\rangle \quad (3)$$

Differentiating the residual function with respect to the parameter set yields a normal system of equations from which the effective short-range force parameters can be determined as

$$\sum_{d'=1}^{N_p} G_{dd'} g_{d'} = b_d \quad (4)$$

for $d = 1, \dots, N_p$.

Note that $\vec{F}_{i,l}^{\text{ref}}$ and $\vec{F}_{i,l}^{\text{eff}}(g_m, r_c)$ need not include all of the force components. As the present work will demonstrate, the formalism is applicable even when only part of the force is considered.

2.2. MS-EVB Multistate MD Method. Because a hydrated excess proton is shared among several solvating water molecules (charge defect delocalization), it is not appropriate to simulate PS&T using conventional empirical force fields with MD. In a manner similar to quantum mechanics, the MS-EVB framework describes delocalized hydrated excess protons as a linear combination of basis states $|i\rangle$ corresponding to distinct hydrogen-bonding topologies.^{53–55} Given a configuration of system nuclei \mathbf{x} , the state function $|\Psi\rangle$ of a delocalized excess proton is written as

$$|\Psi\rangle = \sum_i^N c_i(\mathbf{x}) |i\rangle \quad (5)$$

where N is the total number of basis states and c_i represents the normalized state coefficients. The MS-EVB state am-

plitude squared, c_i^2 , represents the probability of finding the system in state $|i\rangle$. The state function $|\Psi\rangle$ at a given instant is determined by a Hamiltonian matrix \mathbf{H} whose elements h_{ij} are typically described in terms of an empirical force field. Given \mathbf{H} , the ground state $|\Psi\rangle$ and lowest-energy $\mathbf{E}(\mathbf{x})$ eigenstate can be obtained by solving the eigenvalue problem

$$\mathbf{c}^T \mathbf{H} \mathbf{c} = \mathbf{E}(\mathbf{x}) \quad (6)$$

where \mathbf{c} is the N -dimensional eigenvector with elements c_i , $i = 1, 2, \dots, N$, as described in eq 5. According to the Hellmann–Feynman theorem, the force exerted on atom i for a given nuclear configuration can be expressed as

$$\mathbf{F}_i(\mathbf{x}) = -\left\langle \Psi_0 \left| \frac{\partial \mathbf{H}}{\partial \mathbf{x}_i} \right| \Psi_0 \right\rangle = -\sum_{m,n} c_m c_n \frac{\partial h_{mn}(\mathbf{x})}{\partial \mathbf{x}_i} \quad (7)$$

2.3. Effective Short-Range Forces for the EI-MS-EVB Model. Under a full-potential model (denoted here simply as the MS-EVB model), the diagonal elements $h_{ii}(\mathbf{x})$ of the Hamiltonian matrix \mathbf{H} are described by the potential energy function⁵⁵

$$h_{ii} = V_{\text{H}_3\text{O}^+}^{\text{intra}} + \sum_k^{N_{\text{H}_2\text{O}}} V_{\text{H}_2\text{O}}^{\text{intra},k} + \sum_k^{N_{\text{H}_2\text{O}}} V_{\text{H}_3\text{O}^+, \text{H}_2\text{O}}^{\text{inter},k} + \sum_{k < k'}^{N_{\text{H}_2\text{O}}} V_{\text{H}_2\text{O}}^{\text{inter},kk'} \quad (8)$$

The term $V_{\text{H}_3\text{O}^+}^{\text{intra}}$ represents the intramolecular potentials for hydronium, and $\sum_k^{N_{\text{H}_2\text{O}}} V_{\text{H}_2\text{O}}^{\text{intra},k}$ is the intramolecular potential of the flexible, simple point-charge underlying water (SPC/Fw) model developed by Wu et al.⁵⁵ The term $\sum_{k < k'}^{N_{\text{H}_2\text{O}}} V_{\text{H}_2\text{O}}^{\text{inter},kk'}$ represents all nonbonded interactions between water molecules, and the term $V_{\text{H}_3\text{O}^+, \text{H}_2\text{O}}^{\text{inter},k}$ represents the intermolecular potential between hydronium and water molecules and can be written as⁵⁵

$$V_{\text{H}_3\text{O}^+, \text{H}_2\text{O}}^{\text{inter},k} = 4\epsilon_{\text{OO}_w} \left[\left(\frac{\sigma_{\text{OO}_w}}{R_{\text{OO}_k}} \right)^{12} - \left(\frac{\sigma_{\text{OO}_w}}{R_{\text{OO}_k}} \right)^6 \right] + 4\epsilon_{\text{HO}_w} \left[\left(\frac{\sigma_{\text{HO}_w}}{R_{\text{HO}_k}} \right)^{12} - \left(\frac{\sigma_{\text{HO}_w}}{R_{\text{HO}_k}} \right)^6 \right] + \sum_m^4 \sum_{n_k}^3 \frac{q_m^{\text{H}_3\text{O}^+} q_{n_k}^{\text{H}_2\text{O}}}{R_{mn_k}} + V_{\text{OO}_k}^{\text{rep}} + V_{\text{HO}_k}^{\text{rep}} \quad (9)$$

In addition to the standard Lennard-Jones (LJ) and Coulomb potentials, two repulsive terms ($V_{\text{OO}_k}^{\text{rep}}$ and $V_{\text{HO}_k}^{\text{rep}}$) are required to correctly describe interactions between hydronium ion and the water molecules in its first solvation shell. In practical terms, the repulsive terms improve consistency with the high-level ab initio potential energy surface. For the sake of simplicity, only $\sum_{k < k'}^{N_{\text{H}_2\text{O}}} V_{\text{H}_2\text{O}}^{\text{inter},kk'}$ and the first three terms of $V_{\text{H}_3\text{O}^+, \text{H}_2\text{O}}^{\text{inter},k}$ (eq 9) are included in the effective interaction FM procedure. The intramolecular terms $V_{\text{H}_3\text{O}^+}^{\text{intra}}$ and $V_{\text{H}_2\text{O}}^{\text{intra},k}$, as well as the repulsive terms $V_{\text{OO}_k}^{\text{rep}}$ and $V_{\text{HO}_k}^{\text{rep}}$, are many-body, already short-range, and computationally efficient. They can therefore be calculated directly from the empirical functions, by means described in another work.⁵⁵

The off-diagonal (coupling) elements $h_{ij}(\mathbf{x})$ ($i \neq j$) in eq 6 are defined to be nonzero only when the hydronium ions of states $|i\rangle$ and $|j\rangle$ share a transferring proton. These elements can be expressed as⁵⁵

$$h_{ij} = (V_{\text{const}}^{ij} + V_{\text{ex}}^{ij}) A(R_{\text{OO}}, \mathbf{q}) \quad (10)$$

where V_{const}^{ij} is a constant and V_{ex}^{ij} is the electrostatic potential between the H_5O_2^+ Zundel complex and the remaining water molecules.⁵⁵ The latter term is given by

$$V_{\text{ex}}^{ij} = \sum_m^7 \sum_k^{N_{\text{H}_2\text{O}}-1} \sum_{n_k}^3 \frac{q_{n_k}^{\text{H}_2\text{O}} q_m^{\text{ex}}}{R_{mn_k}} \quad (11)$$

where q_m^{ex} represents exchange charges of the H_5O_2^+ complex and $q_{n_k}^{\text{H}_2\text{O}}$ is the atomic charge of water molecule k . The term $A(R_{\text{OO}}, \mathbf{q})$ in eq 10 is a geometric scale factor dependent on the positions of the atoms forming the H_5O_2^+ complex. Thus, only the exchange charge term V_{ex}^{ij} is involved in the construction of an effective short-range force for off-diagonal terms.

Unless otherwise specified, the construction of effective short-range potentials and all simulations (using both the EI-MS-EVB and MS-EVB models) were carried out in a cubic volume with 216 water molecules and one excess proton at 298.15 K. The box size was 18.621 Å, yielding an average density of 1.0 g/cm³. For thermostatic calculations, a constant- NVT ensemble of simulations with a Nosé–Hoover thermostat⁶⁸ was used. For dynamical properties, the constant- NVE ensemble was employed.

In the MS-EVB simulations, long-range electrostatic forces were treated by Ewald summation with a relative tolerance of 10^{-6} . A spherical cutoff of 9.0 Å was chosen for the LJ interactions. In the EI-MS-EVB simulations, the short-range effective potentials employed a slightly longer spherical cutoff radius of 9.24 Å, which is one-half of the simulation box size of the atomistic system for which the effective potential is derived (i.e., the largest possible cutoff that could be used). A leapfrog algorithm was applied to integrate the equations of motion, with a time step of 1 fs. In total, 48 ns of simulation data were produced, consisting of six independent 4-ns simulations for both MS-EVB and EI-MS-EVB cases. All simulations were performed using the DL_EVB program,⁶⁹ derived from the DL_POLY package.⁷⁰

3. Results

3.1. Instantaneous Atomistic Forces. Because the present study aims to improve the efficiency of MS-EVB simulations without sacrificing accuracy, the short-range effective potentials should be assessed in terms of force deviations between the EI-MS-EVB and MS-EVB models for a given configuration of the system nuclei. The force deviation ΔF_{ix} for an atom i along the x axis can be expressed as

$$\Delta F_{ix} = \text{FM}_{ix} - \text{FP}_{ix} \quad (12)$$

where the forces FM_i and FP_i are calculated under the EI-MS-EVB and MS-EVB models, respectively. As all systems in the present study are isotropic, the observed distributions of ΔF_i are indistinguishable along the three Cartesian axes (data not shown). Consequently, all distributions of ΔF_i presented in this article have been averaged over the three axes. The standard deviation $\delta(\Delta F)$ is calculated by the equation

$$\delta(\Delta F) = \sqrt{\frac{1}{MN} \sum_{j=1}^M \sum_{i=1}^N (\Delta F_{ij} - \overline{\Delta F_{ij}})^2} \quad (13)$$

where M is the number of atomistic configurations and N is the number of atoms in the system. To ensure statistical reliability, 6000 configurations were randomly selected from each of the six EI-MS-EVB trajectories, which were obtained at temperatures of 260, 270, 280, 298.15, 310, and 320 K. The resulting ΔF_i distributions and the corresponding $\delta(\Delta F)$ values are shown in Figure 1. Although the wings of the force deviation distributions are slightly wider than the standard Gaussian, the distributions have a zero mean for all six temperatures, demonstrating the absence of systematic error in the effective short-range potentials. Moreover, the standard deviations indicate that about 90% of the individual ΔF_i values are less than $0.5 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-1}$ in absolute value. (The probability of obtaining a ΔF_i magnitude larger than $1.0 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-1}$ is only about 1%.) The average value of $\delta(\Delta F)$ is approximately $0.3 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-1}$, only 2% of the average force ($14 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-1}$) exerted on the atoms.

To further validate the effective short-range potentials, several systems were simulated with larger box sizes at 298.15 K (but otherwise similar to those described above). The greater volume substantially increases the number of atoms beyond the cutoff radius. As depicted in Figure 2, the variance of ΔF_{ix} generally increases with box size, except for the largest system with a box size of 37.106 Å, which exhibits a $\delta(\Delta F_{ix})$ value similar to that of the medium-sized systems. The moderate increase is expected because the effective short-range potentials were based solely on information available within the original-size system, which was relatively small. Nevertheless, $\delta(\Delta F)$ increased by only 0.11 $\text{kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-1}$ (Figure 2) when the box size nearly doubled from 18.621 to 37.106 Å. Given the intrinsic limitations of the accuracy of the underlying empirical force field, errors in the range of a few percent can be viewed as being essentially negligible. The effective potentials generated by the present force-matching scheme should therefore be reasonably transferable to other temperatures and box sizes.

3.2. Radial Distribution Functions. The solvation structure of a hydrated excess proton essentially governs the extent of its charge defect delocalization, which can be characterized by a radial distribution function (RDF). Figure 3 compares the RDFs obtained in EI-MS-EVB and MS-EVB simulations. With the exception of RDF(O^*-H), where the EI-MS-EVB and MS-EVB models deviate in the rarely sampled core region ($<3.0 \text{ \AA}$), the two approaches produce almost indistinguishable descriptions of excess proton solvation in water. The EI-MS-EVB method generates a somewhat lower probability of finding water hydrogen atoms close to the hydronium oxygen at short distances (Figure 3b), but this feature is not directly related to the process of proton transport.

3.3. Free Energy Profile of Proton Transfer. The free energy profile of proton transfer between two water molecules can be calculated by the equation

$$\Delta E(c_1^2 - c_2^2) = -RT \ln[P(c_1^2 - c_2^2)] - C \quad (14)$$

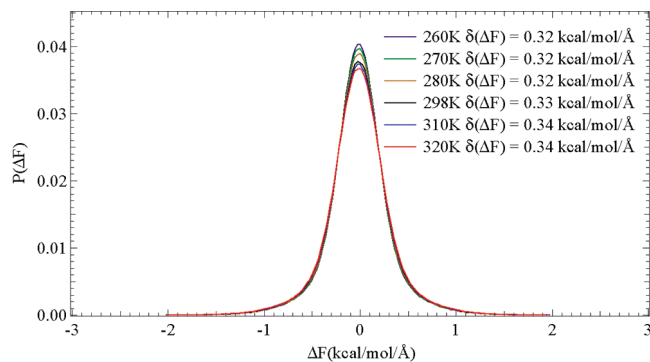


Figure 1. Distribution and standard deviation of the force difference ΔF_{ix} as a function of temperature.

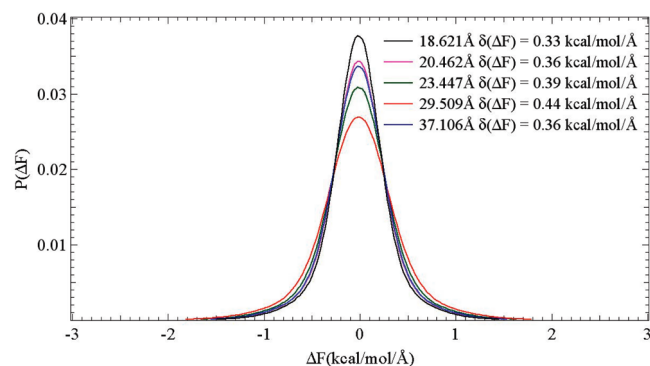


Figure 2. Distribution and standard deviation of the force difference ΔF_{ix} as a function of box size.

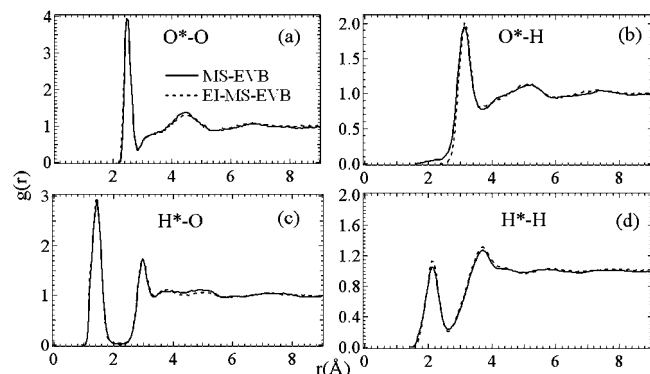


Figure 3. Comparison of the radial distribution functions (RDFs) calculated by the EI-MS-EVB and the full-potential MS-EVB methods. (a) RDFs of water oxygen atoms O around the hydronium oxygen atom O^* , (b) RDFs of water hydrogen atoms H around the hydronium oxygen atom O^* , (c) RDFs of water oxygen atoms O around the hydronium hydrogen atoms H^* , (d) RDFs of water hydrogen atoms H around the hydronium hydrogen atoms H^* .

where a “coordinate” relevant to the proton transfer process can be defined as the difference between the largest and second-largest MS-EVB amplitudes, c_1^2 and c_2^2 . The function $P(c_1^2 - c_2^2)$ is thus the probability distribution of that coordinate. In this expression, R is the molar gas constant ($8.314 \text{ J mol}^{-1} \text{ K}^{-1}$), T is the system temperature (298.15 K), and C is an arbitrary constant that can be adjusted to define the point of zero free energy. Typical values for reference^{52,54–57} are $c_1^2 - c_2^2 \approx 0.45$ for the Eigen cation (H_3O_4^+) and $c_1^2 - c_2^2 \approx 0.0$ for the Zundel cation (H_5O_2^+);

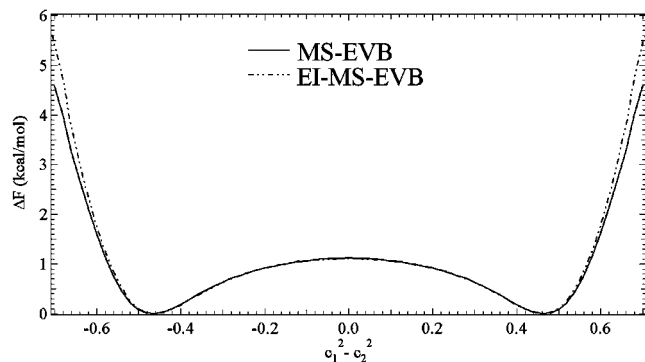


Figure 4. Free energy profile of the proton-transfer coordinate for the EI-MS-EVB model versus the full-potential MS-EVB calculation.

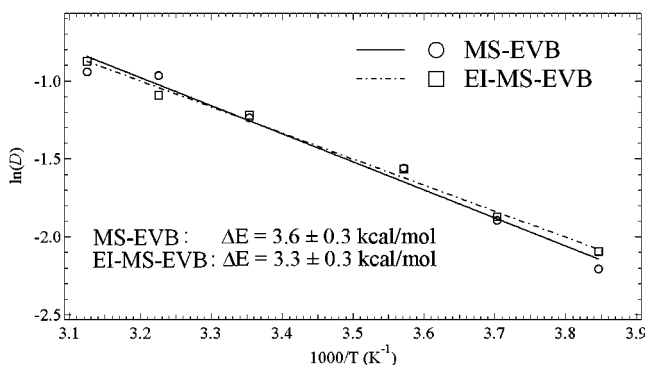


Figure 5. Temperature dependence of the proton diffusion coefficient for the EI-MS-EVB model versus the full-potential MS-EVB calculation.

the charge defect is more localized in the former case. As demonstrated in Figure 4, the EI-MS-EVB simulations nearly perfectly reproduce not only the free energy barrier of proton transfer, but also the probability distribution of the reactive coordinate. The central peak is slightly lower compared to the MS-EVB case, by only about 0.02 kcal/mol.

3.4. Temperature Dependence of the Diffusion Coefficient. The activation energy for proton transport, ΔE_a , can be described according to the classical Arrhenius equation

$$D_{H^+} = A \exp(-\Delta E_a/RT) \quad (15)$$

where A is a temperature-independent constant. By plotting $\ln(D_{H^+})$ against $1/T$, the value of ΔE_a can be determined from a simple linear fit. Simulations were carried out at 260, 270, 280, 310, and 320 K to obtain the temperature dependence of D_{H^+} , as shown in Figure 5. On average, the deviation between D_{H^+} values obtained in EI-MS-EVB and MS-EVB simulations is about $0.03 \text{ \AA}^2/\text{ps}$ for all temperatures. The two approaches are essentially indistinguishable with respect to this property, as the error bar for D_{H^+} is $\pm 0.03 \text{ \AA}^2/\text{ps}$. The ΔE_a value for the EI-MS-EVB model is $3.6 \pm 0.3 \text{ kcal/mol}$, whereas that for the MS-EVB model is $3.3 \pm 0.3 \text{ kcal/mol}$. Again, the two values are statistically consistent.

3.5. Effect of Box Size on the Proton Diffusion Coefficient. This section reports the transferability of the effective short-range potentials in the EI-MS-EVB method to other box sizes. In particular, we investigate the variation in D_{H^+} with system size. EI-MS-EVB and MS-EVB simulations

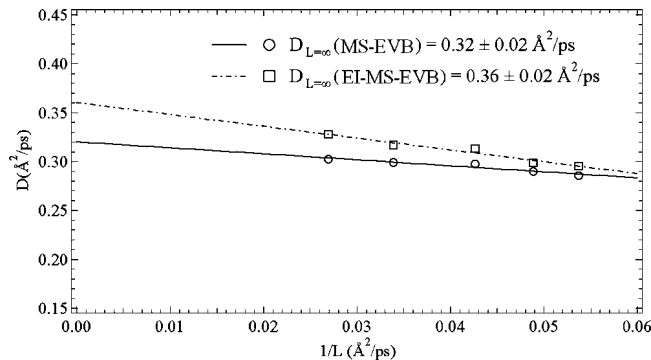


Figure 6. Effect of box size on the proton diffusion coefficient for the EI-MS-EVB model versus the full-potential MS-EVB calculation.

were performed on cubic systems with box sizes L of 20.462, 23.447, 29.509, and 37.106 \AA in addition to the original system. Figure 6 demonstrates that the error ranges of all corresponding D_{H^+} pairs overlap, except for the largest box size. The actual D_{H^+} values calculated with the EI-MS-EVB method are consistently somewhat greater than those calculated with the MS-EVB method. Following the procedure of Yeh and Hummer,⁷¹ the D_{H^+} value for an infinite volume can be estimated by linearly extrapolating D_{H^+} to the point $1/L = 0$. The asymptotic value of D_{H^+} with the FP EI-MS-EVB model turns out to be $0.36 \pm 0.02 \text{ \AA}^2/\text{ps}$, only $0.04 \text{ \AA}^2/\text{ps}$ larger than that determined with the MS-EVB approach. The trend toward increasing deviations with box size reflects the slight inconsistency between instantaneous atomistic forces and the effective potential, as discussed in section 3.1. Despite this small discrepancy, the EI-MS-EVB simulations correctly predict the nature of the relationship between box size and D_{H^+} . Moreover, the projected values of D_{H^+} remain in good agreement even for the larger systems.

3.6. Computational Efficiency. This section summarizes the most important result from the present article—that of increased computational efficiency with the EI-MS-EVB model. Two simulations were performed to compare the efficiency of nonbonded interaction calculations. In the first simulation, electrostatic forces were calculated using the smooth particle–mesh Ewald (SPME)¹² algorithm with a tolerance of 10^{-6} , and LJ interactions were calculated using a simple spherical cutoff of radius 9.0 \AA . In the second simulation, effective short-range potentials were applied with a simple spherical cutoff at 9.26 \AA . The two approaches are denoted FP (full-potential MS-EVB model) and EI (EI-MS-EVB model), respectively.

Both runs were performed on a single core of a computer equipped with IBM PowerPC970 quad-core processors and 8 Gb of memory. To avoid introducing additional complexities related to parallel computing, both runs were carried out on a single CPU. The results for four different system sizes are summarized in Table 1, which reports the total CPU time spent on calculations of nonbonded interactions over 10000 time steps. At all system sizes, EI is significantly faster than FP, being nearly 4 times faster than FP even for a relatively small system with 5185 atoms. Very encouragingly, but not surprisingly, the efficiency improvement increases with

Table 1. Computational Efficiency of Nonbonded Interaction Calculations as a Function of System Size, Comparing the Full-Potential MS-EVB (FP) and EI-MS-EVB (EI) Approaches

number of atoms	FP time (s)	EI time (s)	ratio (FP/EI)
649	9912	3414	2.9
5185	41229	8267	5.0
12427	104801	13889	7.5
34252	313810	27267	11.5

system size. For the system with 34252 atoms, the EI approach is roughly 10 times faster than the FP approach.

Although the present results are for only a single processor, when considering the difficulty in constructing an efficient global parallelization of the fast Fourier transform (used by the SPME calculation), the computational efficiency of the EI-MS-EVB approach will become even more pronounced when the scalability of the algorithm is considered. Molecular simulations of large-scale biological systems are typically performed on many CPUs, requiring such parallelization. The EI-MS-EVB approach might therefore be better suited to such applications than the full-potential MS-EVB method, where calculations of long-range, nonbonded interactions (particularly electrostatics) are computationally less scalable. It has recently been demonstrated,⁷² for example, that highly scalable biomolecular MD simulations are within reach when implementing short-range (cutoff) electrostatic potentials similar in spirit to those in the EI-MS-EVB approach.

4. Conclusions

Depending on the system of interest, the time and length scales of importance to reactive events such as those involved in PS&T can vary widely: from picoseconds to microseconds and from nanometers to micrometers, respectively. Because the system coordinates are usually propagated on a time scale of femtoseconds in an MD algorithm, MS-EVB applications can be extremely computationally challenging. The EI-MS-EVB approach utilizes an effective short-range potential^{48,49} and accurately reproduces reference trajectories from the full MS-EVB method. By comparing the instantaneous forces on single atoms from the same configuration of nuclei, we have shown that EI-MS-EVB model is not only accurate but also transferable over a range of system temperatures and box sizes. The accuracy of the EI-MS-EVB method was further confirmed by evaluating several key properties: RDFs, free energy profiles, activation energies of proton transport, and proton diffusion coefficients. Most importantly, the EI-MS-EVB model was found to outperform the full-potential MS-EVB model in terms of computational efficiency. For larger systems, the EI-MS-EVB approach is approximately 10 times faster than the SPME-based FP MS-EVB calculation on a single processor. In addition, the EI-MS-EVB method of calculating short-range interactions should be much easier to decompose and parallelize than the fast Fourier transform of the SPME method, which is intrinsically a global operation. This is an important consideration for complex applications and large systems, which typically require many thousands of CPUs.

Very recently, the FM algorithm has even been used to construct effective short-range forces from explicit evaluations of the long-range Coulomb interaction.⁴⁸ The resulting effective interaction between unit charges successfully reproduces many structural, dynamic, and thermodynamic properties of various systems, including liquid water, solvated ions, and hydrophobic solutes. This methodology⁴⁸ and its analytic approximation⁴⁹ is closely related to the present approach. Provided the van der Waals interaction is correctly accounted for, the effective nonbonded interaction obtained by force matching should be recoverable from the “universal” effective short-range electrostatic interaction identified previously.⁴⁸ Furthermore, integrating all the nonbonded interactions in this manner can reduce the number of operations per particle pair to just one table lookup in both the EI and charge-scaled schemes.

In future work, one goal will be to extend the EI-MS-EVB approach to the self-consistent, iterative MS-EVB (SCI-MS-EVB)⁷³ method. This approach determines the EVB amplitude separately for each protonated complex, enforcing consistency with the EVB amplitudes of other protonated complexes in an iterative manner until the total potential energy of the multiproton system has converged. The SCI-MS-EVB method demands significantly more CPU cycles for nonbonded interactions than does the single-proton MS-EVB. The SCI-MS-EVB approach also typically requires five or more iterations for the EVB amplitudes to converge. Incorporating the EI-MS-EVB model into the SCI-MS-EVB framework should therefore greatly facilitate the investigation of PS&T behavior in highly acidic systems such as the proton-exchange membranes used in fuel cells.⁷⁴

Acknowledgment. This research was supported by the National Science Foundation (CHE-0719522) and the Department of Energy (DE-FG02-05ER15724). The computational resources utilized in this research were provided by the following NSF programs: Partnerships for Advanced Computational Infrastructure, Distributed Terascale Facility (DFT), and Terascale Extensions: Enhancements to the Extensible Terascale Facility.

Supporting Information Available: Atom types in the EI-MS-EVB model and a table of numerical pairwise nonbonded forces fitted by the force-matching method. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

References

- (1) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–218.
- (2) Sagui, C.; Darden, T. A. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 155–179.
- (3) Tobias, D. J. *Curr. Opin. Struct. Biol.* **2001**, *11*, 253–261.
- (4) Harvey, S. C. *Proteins* **1989**, *5*, 78–92.
- (5) Smith, P. E.; van Gunsteren, W. F. In *Computer Simulations of Biomolecular Systems: Theoretical and Experimental Applications*; ESCOM: Leiden, The Netherlands, 1993; pp 182–212.
- (6) Norberg, J.; Nilsson, L. *Biophys. J.* **2000**, *79*, 1537–1553.

- (7) Cheathan, T. E., III; Brooks, B. R. *Theor. Chem. Acc.* **1999**, *99*, 279–288.
- (8) Auffinger, P.; Westhof, E. In *Encyclopedia of Computational Chemistry*; John Wiley & Sons: New York, 1998; pp 1628–1639.
- (9) Ewald, P. P. *Ann. Phys.* **1921**, *64*, 253–287.
- (10) Hockney, R. W.; Eastwood, J. W. *Computer Simulation Using Particles*; McGraw-Hill: New York, 1981.
- (11) Darden, T. A.; York, D. M.; Pedersen, L. G. *J. Chem. Phys.* **1993**, *98*, 10089–92.
- (12) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (13) Perram, J. W.; Petersen, H. G.; DeLeeuw, S. W. *Mol. Phys.* **1988**, *65*, 875–893.
- (14) Toukmaji, A.; Board, J. A. *Comput. Phys. Commun.* **1996**, *95*, 78–92.
- (15) Greengard, L.; Rokhlin, V. *J. Comput. Phys.* **1987**, *73*, 325–348.
- (16) Board, J. A.; Causey, J. W.; Leathrum, J. F.; Windemuth, A.; Shulten, K. *Chem. Phys. Lett.* **1992**, *198*, 89–94.
- (17) Pollock, E.; Glosli, J. *Comput. Phys. Commun.* **1996**, *95*, 93–110.
- (18) Figuerido, F.; Levy, R.; Zhou, R.; Berne, B. J. *J. Chem. Phys.* **1997**, *106*, 9835–9849.
- (19) Cheng, H.; Greengard, L.; Rokhlin, V. *J. Comput. Phys.* **1999**, *155*, 468–98.
- (20) Greengard, L. F.; Huang, J. *J. Comput. Phys.* **2002**, *180*, 642–58.
- (21) McCammon, J. A.; Harvey, S. C. *Dynamics of Proteins and Nucleic Acids*; Cambridge University Press: Cambridge, U.K., 1987.
- (22) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: New York, 1987.
- (23) Brooks, C. L., III; Pettitt, B. M.; Karplus, M. *J. Chem. Phys.* **1985**, *83*, 5897–5908.
- (24) Hunenberger, P. H.; van Gunsteren, W. F. *J. Chem. Phys.* **1998**, *108*, 6117–6134.
- (25) Madura, J. D.; Pettitt, B. M. *Chem. Phys. Lett.* **1988**, *150*, 105–108.
- (26) Schreiber, H.; Steinhauser, O. *Chem. Phys.* **1992**, *168*, 75–89.
- (27) Neumann, M.; Steinhauser, O.; Pawley, G. S. *Mol. Phys.* **1984**, *52*, 97–113.
- (28) Baker, N. A.; Hünenberger, P. H.; McCammon, J. A. *J. Chem. Phys.* **1999**, *110*, 10679–10692.
- (29) Wood, R. H. *J. Chem. Phys.* **1995**, *103*, 6177–6187.
- (30) Brooks, C. L., III. *J. Chem. Phys.* **1987**, *86*, 5156–5162.
- (31) Straatsma, T. P.; Berendsen, H. J. C. *J. Chem. Phys.* **1988**, *89*, 5876–5886.
- (32) Kalko, S. G.; Sese, G.; Padro, G. A. *J. Chem. Phys.* **1996**, *104*, 9578–9585.
- (33) Resat, H.; McCammon, J. A. *J. Chem. Phys.* **1996**, *104*, 7645–7651.
- (34) Dang, L. X.; Pettitt, B. M.; Rossky, P. J. *J. Chem. Phys.* **1992**, *96*, 4046–4047.
- (35) Badert, J. S.; Chandler, D. *J. Chem. Phys.* **1992**, *96*, 6423–4427.
- (36) Del Buono, G. S.; Figueirido, F. E.; Levy, R. M. *Chem. Phys. Lett.* **1996**, *263*, 521–529.
- (37) Loncharich, R. J.; Brooks, B. R. *Proteins* **1989**, *6*, 32–45.
- (38) Lau, K. F.; Alper, H. E.; Thacher, T. S.; Stouch, T. R. *J. Phys. Chem.* **1994**, *98*, 8185–8792.
- (39) Prevost, M.; van Belle, D.; Lippens, G.; Wodak, S. *Mol. Phys.* **1990**, *76*, 587–603.
- (40) Steinbach, P. J.; Brooks, B. R. *J. Comput. Chem.* **1994**, *15*, 667–683.
- (41) Barker, J. A.; Watts, R. O. *Mol. Phys.* **1973**, *26*, 789–792.
- (42) Hummer, G.; Soumpasis, D. M.; Neumann, M. *Mol. Phys.* **1992**, *77*, 769–785.
- (43) Chipot, C.; Millot, C.; Maigret, B.; Kollman, P. A. *J. Chem. Phys.* **1994**, *101*, 7953–7962.
- (44) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451–5459.
- (45) Daura, X.; Hünenberger, P. H.; Mark, A. E.; Querol, E.; Avilés, F. X.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **1996**, *118*, 6285–6294.
- (46) Fennell, C. J.; Gezelter, J. D. *J. Chem. Phys.* **2006**, *124*, 234104.
- (47) Wolf, D.; Keblinski, P.; Phillpot, S. R.; Eggebrecht, J. *J. Chem. Phys.* **1999**, *110*, 8254–8282.
- (48) Izvekov, S.; Swanson, J. M.; Voth, G. A. *J. Phys. Chem. B* **2008**, *112*, 4711–4724.
- (49) Shi, Q.; Liu, P.; Voth, G. A. *J. Phys. Chem. B* **2008**, *112*, 16230–7.
- (50) Izvekov, S.; Parrinello, M.; Burnham, C. J.; Voth, G. A. *J. Chem. Phys.* **2004**, *120*, 10896–10913.
- (51) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 6573–86.
- (52) Schmitt, U. W.; Voth, G. A. *J. Phys. Chem. B* **1998**, *102*, 5547–5551.
- (53) Schmitt, U. W.; Voth, G. A. *J. Chem. Phys.* **1999**, *111*, 9361–9381.
- (54) Day, T. J. F.; Soudackov, A. V.; Cuma, M.; Schmitt, U. W.; Voth, G. A. *J. Chem. Phys.* **2002**, *117*, 5839–5849.
- (55) Wu, Y.; Chen, H.; Wang, F.; Paesani, F.; Voth, G. A. *J. Phys. Chem. B* **2007**, *112*, 467–482.
- (56) Voth, G. A. *Acc. Chem. Res.* **2006**, *39*, 143–50.
- (57) Swanson, J. M. J.; Maupin, C. M.; Chen, H.; Petersen, M. K.; Xu, J.; Wu, Y.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 4300–14.
- (58) Iyengar, S. S.; Petersen, M. K.; Day, T. J. F.; Burnham, C. J.; Teige, V. E.; Voth, G. A. *J. Chem. Phys.* **2005**, *123*, 084309.
- (59) Day, T.; Schmitt, U.; Voth, G. *J. Am. Chem. Soc.* **2000**, *122*, 12027–12028.
- (60) Chen, H.; Wu, Y.; Voth, G. A. *Biophys. J.* **2006**, *90*, L73–75.
- (61) Chen, H.; Ilan, B.; Wu, Y.; Zhu, F.; Schulten, K.; Voth, G. A. *Biophys. J.* **2007**, *92*, 46–60.
- (62) Chen, H.; Wu, Y.; Voth, G. A. *Biophys. J.* **2007**, *93*, 3470–3479.

- (63) Xu, J.; Voth, G. A. *Proc. Natl. Acad. Sci.* **2005**, *102*, 6795–6800.
- (64) Chen, H.; Yan, T.; Voth, G. A. *J. Phys. Chem. A* **2009**, *113*, 4507–17.
- (65) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 4116–4127.
- (66) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. *J. Chem. Phys.* **2008**, *128*, 244114.
- (67) Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. *J. Chem. Phys.* **2008**, *128*, 244115.
- (68) Nosé, S. *Mol. Phys.* **1984**, *52*, 255–268.
- (69) Smondyrev, A. M.; Voth, G. A. *Biophys. J.* **2002**, *83*, 1987–1996.
- (70) Smith, W.; Forester, T. R. *J. Mol. Graphics* **1996**, *14*, 136–141.
- (71) Yeh, I. C.; Hummer, G. *J. Phys. Chem. B* **2004**, *108*, 15873–15879.
- (72) Schulz, R.; Lindner, B.; Petridis, L.; Smith, J. C. *J. Chem. Theory Comput.* **2009**, *5*, 2798–808.
- (73) Wang, F.; Voth, G. A. *J. Chem. Phys.* **2005**, *122*, 144105–9.
- (74) Petersen, M. K.; Voth, G. A. *J. Phys. Chem. B* **2006**, *110*, 18594–18600.

CT100318F

Automated Sampling Assessment for Molecular Simulations Using the Effective Sample Size

Xin Zhang, Divesh Bhatt, and Daniel M. Zuckerman*

Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15213

Received May 6, 2010

Abstract: To quantify the progress in the development of algorithms and force fields used in molecular simulations, a general method for the assessment of the sampling quality is needed. Statistical mechanics principles suggest the populations of physical states characterize equilibrium sampling in a fundamental way. We therefore develop an approach for analyzing the variances in state populations, which quantifies the degree of sampling in terms of the effective sample size (ESS). The ESS estimates the number of statistically independent configurations contained in a simulated ensemble. The method is applicable to both traditional dynamics simulations as well as more modern (e.g., multicanonical) approaches. Our procedure is tested in a variety of systems from toy models to atomistic protein simulations. We also introduce a simple automated procedure to obtain approximate physical states from dynamic trajectories: this allows sample-size estimation in systems for which physical states are not known in advance.

1. Introduction

The field of molecular simulations has expanded rapidly in the last two decades and continues to do so with progressively faster computers. Furthermore, significant effort has been devoted to the development of more sophisticated algorithms^{1–5} and force fields^{6–10} for use in both physical and biological sciences. To quantify progress - and indeed to be sure progress is occurring - it is critical to assess the efficiency of the algorithms. Moreover, if the quality of sampling is unknown, we cannot expect to appreciate fully the predictions of molecular mechanics force fields: after all, statistical ensembles, whether equilibrium or dynamical, are the essential output of force fields. These issues demand a gauge to assess the quality of the generated ensembles¹¹ - one which is automated, nonsubjective, and applicable regardless of the method used to generate the ensembles.

Ensembles are of fundamental importance in the statistical mechanical description of physical systems: beyond the description of fluctuations intrinsic to the ensembles, all thermodynamic properties are obtained from them.¹² The quality of simulated ensembles is governed by the number of uncorrelated samples present in the ensemble. Due to

significant correlations between successive frames in, say, a dynamics trajectory, the number of uncorrelated samples cannot be directly gauged from the total number of frames. Rather, the number of statistically independent configurations in the ensemble (or the effective sample size, ESS) is required.^{13–16} This effective sample size has remained difficult to assess for reasons described below. In this work, we present a straightforward method to determine the ESS of an ensemble - regardless of the method used to generate the ensemble - by quantifying variances in populations of physical states.

A conventional view of sample size based on a dynamical simulation is given by the following equation

$$\text{ESS} = \frac{t_{\text{sim}}}{t_{\text{corr}}[f]} \quad (1)$$

where t_{sim} is the simulation time, and $t_{\text{corr}}[f]$ is the correlation time¹⁶ for the observable f , which is presumed to relax most slowly. However, the estimation of the correlation time is data intensive and potentially very sensitive to noise in the tail of the correlation function.¹⁷ Other approaches for assessing correlations have, therefore, been proposed. For example, Mountain and Thirumalai^{18,19} introduced the “ergodic measure”, which quantifies the time required for the

* Corresponding author e-mail: ddmmzz@pitt.edu.

observable to appear ergodic. Flyvbjerg and Petersen¹⁷ developed a block averaging method which can be adapted to yield a correlation time and ESS.²⁰

The key challenge in applying eq 1, however, is the choice of an observable f which consistently embodies the slowest motions across the incredible variety of molecular systems. Indeed, it is well appreciated that different observables exhibit different correlation times (e.g., ref 21). For example, in a typical molecule, bond lengths become decorrelated faster than dihedral angles. Nevertheless, apparently fast observable rarely are fully decoupled from the rest of the system: slower motions ultimately couple to the fast motions and influence their distributions in typical cases.²¹ On the whole, there is significant ambiguity in the use of a hand-picked observable to estimate “the” correlation time - not to mention, subjectivity. Moreover, the ultimate goal of simulation, arguably, is not to compute a particular ensemble average but to generate a truly representative ensemble of configurations, from which any observable can be averaged.

Several years ago, Lyman and Zuckerman proposed that the configuration-space distribution itself could be used as a fundamental observable.²² In particular, it was pointed out that if configuration space is divided into different regions or bins, then the resulting “structural histogram” of bin populations could be a critical tool in assessing sampling. The idea was subsequently used to quantify sample size in at least two studies: Lyman and Zuckerman developed a scheme to quantify ESS for trajectories with purely sequential correlations based on variances in the bins of the structural histogram;¹⁶ Grossfield and co-workers suggested a bootstrapping approach for estimating ESS based on structural histograms.¹⁵ The present work expands on ideas from these studies. There has been related work for sequentially correlated Markov chains.^{23,24}

This study extends the earlier structural-histogram approaches by focusing on physical or metastable states. Qualitatively, a physical state can be defined as a region of configuration space for which the internal time scales are much shorter than those for transitions between different states.²⁵ The populations of physical states seem an intuitive choice for quantifying sampling quality, since they reflect slow time scales by construction. Indeed, the state populations along with state definitions (addressed in section 2.1) can be said to embody the equilibrium ensemble. This type of argument can be made semiquantitative by noting that any ensemble average $\langle f \rangle$ can be expressed in terms of state populations p_i and state-specific averages $\langle f \rangle_i$ for state i , because $\langle f \rangle = \sum_i p_i \langle f \rangle_i$. Thus, the goal of sampling can be described as obtaining both (i) state populations and (ii) well-sampled ensembles within each state.

Statistical mechanics principles strongly suggest, moreover, that state populations should be viewed as the most critical slow observables. To see why, consider states i and j defined by regions of configuration space V_i and V_j . The ratio of state populations is given by the ratio of state partition functions

$$\frac{p_i}{p_j} = \frac{Z_i}{Z_j} = \frac{\int_{V_i} d\mathbf{r} \exp(-U(\mathbf{r})/k_B T)}{\int_{V_j} d\mathbf{r} \exp(-U(\mathbf{r})/k_B T)} \quad (2)$$

where Z_i is the partition function for state i , U is the potential energy of the system, T is the temperature, and \mathbf{r} represents all configuration-space coordinates. Equation 2 indicates that state populations cannot be determined without good sampling within each state. In other words, it would seem impossible for an algorithm (which is correct for arbitrary systems) to predict state populations without having already sampled correctly within states (see Figure 1). For this reason, the state populations can be considered the fundamental set of slow observables - a physically motivated choice of structural histogram. We will use variances in state populations to estimate ESS, an approach which applies to both dynamic and nondynamic (e.g., exchange) simulations.

Accordingly, an important prerequisite for the estimation of ESS is the determination of physical states. In this work, we use a particularly simple method for the approximation of physical states that uses information present in a dynamics trajectory regarding the transition rates between different regions. Regions showing high transition rates with each other belong in the same physical state. Further, this procedure also highlights the hierarchical nature of the energy landscape. Our state approximation scheme is based on ideas of Chodera et al.²⁵ who developed approximated physical states by determining a division of the total configuration space that maximizes the self-transition probabilities (i.e., the divisions represent metastable states). See also ref 26. Our state-approximation method can also be used with short dynamics trajectories initiated from configurations obtained from nondynamic simulations.

We emphasize, nevertheless, that our procedure for ESS estimation can be used with states discovered by different means.

The manuscript is organized as follows. First, we describe in detail the procedure we use to estimate the effective sample size. Then, we present results for several models with different levels of complexity - a two-state toy model, butane, calmodulin, dileucine, and Met-enkephalin. Our ESS results are compared with the previous “decorrelation time” approach.²⁷ We also analyzed multi- μ s atomistic simulations for the membrane protein rhodopsin.¹⁵ We then discuss the practical aspects of the procedure and present conclusions. Further, in the Appendix, we describe the simple, automated procedure used to determine approximate physical states.

2. Methods and Systems

We have argued above that the populations of physical states are fundamental observables for assaying the equilibrium ensemble. We therefore propose that the statistical quality of an equilibrium ensemble be quantified using variances in state populations. As usual, the variances will decrease with better sampling. Importantly, however, simple binomial statistics permit a fairly precise quantification of the ESS - i.e., the number of statistically independent configurations to which an ensemble is equivalent - regardless of the number

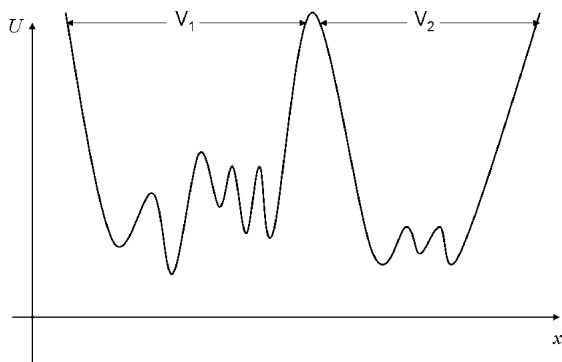


Figure 1. A schematic two-state potential energy landscape illustrating eq 2. The states are defined by the “volumes” V_1 and V_2 . The distributions of configurations within each state help to determine the overall ratio of state populations in eq 2.

of configurations in the original ensemble. Below, we will address the issues of computing variances from dynamical and nondynamical simulations as well as methods for approximating physical states.

The key technical idea in connecting the variance in a state’s population to the ESS follows work presented in ref 16: an analytic estimate for the variance can be computed based on a known number of independent samples. If one “turns around” this idea, given the observed variance, an estimate for the number of independent samples can be immediately obtained. In particular, given a region j of configuration space with fractional population p_j , the variance in p_j based on N independent samples is $\sigma_j^2 = p_j(1 - p_j)/N$. In practice, this variance is estimated from repeated independent simulations, each yielding a value for p_j . The ESS based on populations recorded for region j can therefore be estimated via

$$N_j^{\text{eff}} = \frac{\bar{p}_j(1 - \bar{p}_j)}{\sigma_j^2} \quad (3)$$

where \bar{p}_j is the observed average population in region j . Equation 3 gives the ESS for one simulation characterized by σ_j^2 . For N_{obs} simulations, the total ESS is N_j^{eff} .

Both \bar{p}_j and σ_j^2 can be computed from N_{obs} repeated simulations

$$\bar{p}_j = \frac{1}{N_{\text{obs}}} \sum_i p_j^{(i)}, \sigma_j^2 = \frac{1}{N_{\text{obs}}} \sum_i (p_j^{(i)} - \bar{p}_j)^2 \quad (4)$$

where $p_j^{(i)}$ is the population of state j from simulation i .

Two important points are implicit in these estimators (both of which are discussed further, below). First, our analysis assumes reasonable \bar{p}_j values have been obtained in the simulations to be analyzed - although a low value of N^{eff} can suggest additional sampling is advisable. Second, our effective sample size will have the lower bound $N^{\text{eff}} > N_{\text{obs}}$, so in practice estimates such that $N^{\text{eff}} \approx N_{\text{obs}}$ suggest poor sampling.

As noted in ref 16, eq 3 is actually a limiting form appropriate for large N . Although it is straightforward to include corrections accounting for the fact that only $N - 1$ observations are independent (because \bar{p}_j is the observed

average among the p_j values used in estimating the variance), the effect is unimportant compared to the intrinsic fluctuations in N^{eff} .

Each region or state will yield its own estimate for the ESS via eq 3, but we are interested in the smallest ESS reflecting the slowest time scales. As described below, in this report, we use a hierarchical decomposition of configuration space which leads to only two states at the top level by construction. In turn, these two states yield identical ESS values by eq 3. Alternatively, if a full hierarchy is not constructed, one can simply select the lowest ESS value as the best quantification of sampling, reflecting the worst bottleneck encountered; in such cases, it may be of interest to set a minimum \bar{p}_j value for the governing state (e.g., 0.01–0.05) to avoid the ESS being dominated by a relatively minor state.

We note that, based on eq 3, the minimal value which can be determined for a single simulation is one, and generally $N_j^{\text{eff}} > 1$. Thus, a value of $O(1)$ is strongly suggestive of inadequate sampling.

The current approach, in essence, uses a block-averaging strategy¹⁷ and can be contrasted with the previous work of Lyman and Zuckerman for dynamical trajectories.¹⁶ The present work computes block-style variances of quantities (the state populations) whose statistical behavior is straightforward to analyze - e.g., via eq 3. The earlier approach, in contrast, directly exploited sequential correlations to do “hypothesis testing”: Do the snapshots of a trajectory separated by a fixed time interval behave as though independent?¹⁶ The earlier work also used population variances and an analog of eq 3; however, physical states were not required because individual configurations were used, rather than block averages as in the present work - which tend to convolute time scales (sections 3.5 and 4.3).

2.1. Hierarchical Approximation of Physical States.

The approximation of physical states has previously been addressed in some detail, particularly in the context of developing Markov models.²⁵ Below, and in the Appendix, we describe a simpler approach used in this work. As we elaborate in the Discussion, it appears that our ESS analysis does not require a particularly precise specification of physical states. Because our prescription is to find the slowest time scale (i.e., smallest ESS) among the many which may be present, and because our physical states are reasonable, the approach works reliably. On the other hand, although eq 3 can be applied to an arbitrary region in principle, it can “get fooled” into overestimating the ESS if only a small part of a state is considered: see section 4 for details.

We emphasize that our ESS analysis described above is distinct from the states analyzed, and other reasonable state decomposition procedures can be used.

The Appendix details the hierarchical state approximation scheme adopted here, which is closely related to the work of Chodera et al.²⁵ In brief, given the best data available, we first divide configuration space into small regions or bins (following refs 16 and 28), which do not necessarily correspond to energy basins. Based on one or more dynamical trajectories (perhaps those being analyzed for ESS), we estimate rates among each pair of regions. Starting from the

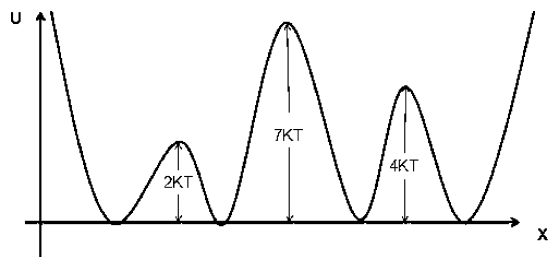


Figure 2. A one-dimensional potential energy landscape with four basins separated by three barriers.

fastest pairwise rates, the bins are combined into statelike aggregated regions. By construction, all pairwise rates within each aggregate are faster than rates between aggregates. The process is continued to construct a full hierarchy until all aggregates are combined (see Figures 3 and 4). The approximate states used to estimate the ESS are based on the top (i.e., two-state) level of the hierarchy, which reflects the slowest time scales as desired.

Our rate estimation procedure is well-suited to our purpose of ESS estimation. First, it is fairly simple and typically requires a small fraction of the computational cost of the simulation being analyzed. More importantly, as noted in the Discussion, it performs as well as a somewhat more complex approach we implemented (data not shown). Although our procedure (and others²⁵) requires dynamical trajectories to estimate interbin transition rates, this does not mean prohibitively expensive dynamics simulations must be performed, as we now discuss.

2.1.1. State Approximation from Noncontinuous Dynamical Trajectories. Because our state approximation scheme depends on continuous dynamical trajectories, the question arises as to how states can be obtained when sampling has been performed using a nondynamical method such as replica exchange.^{3,29,30} Although exchange simulations use continuous trajectories which contain the necessary information for estimating rates among local regions,³¹ other sampling methods may not employ dynamical trajectories at all (e.g., see ref 28).

In fact, states can be approximated based on a set of short dynamics trajectories run after a possibly costly nondynamical trajectory. In particular, a set of M trajectories (we use $M = 20$ below) can be initiated from random configurations selected from the nondynamical simulation. These short trajectories need only be long enough to permit exploration *within* states. There is no need for transitions between states. The only modification to the state approximation scheme described previously is that it may not be possible to iterate the combination procedure until all states are combined. Rather, the process will terminate after regions with measurable transition rates are combined. A set of approximate states will remain for which no interstate transitions have been recorded. For each of these remaining states, an ESS estimate can be obtained via eq 3. Because of our interest in the slowest time scales, the overall ESS will be taken as the minimum among the various state-specific values.

The scheme just described is tested below and compared with the use of longer trajectories for state approximation.

We, again, emphasize that short dynamic trajectories are only used to approximate states, whereas ESS is, subsequently, computed from the much longer nondynamic trajectories. The nondynamic trajectories are presumed to sample all the relevant states.

2.2. A Caveat: Self-Consistent but Not Absolute ESS. Without prior knowledge or assumptions about a landscape, it would appear impossible to know whether every important state has been visited in a given simulation. This is not a limitation of our analysis per se but of any attempt to estimate ESS based on simulation data. Nevertheless, it is important to make this caveat clear.

Therefore, the goal of the present analysis is not to assess the coverage of configuration space but to self-consistently assess sampling quality given the states visited in the simulation. In other words, we answer, “What is the statistical quality of the sampling based on the configurational states visited in a given set of simulations?” Our ESS estimation can therefore be viewed as an upper bound to the true ESS based on the full configuration space. ESS estimation, nevertheless, is essential for assessing efficiency in algorithms and precisely specifying the predictions of modern force fields.

On the other hand, so long as a state has been visited in a simulation, it can greatly affect the sample size. For instance, if a state has been visited only once, the estimate of its population variance will be large and lead (correctly) to a small ESS.

2.3. Estimating Variances in State Populations. The heart of our approach is to estimate ESS based on variances in state populations using eq 3. Clearly, then, without reliable variance estimates, we cannot expect ESS values to be reliable.

For dynamical simulations - i.e., simulations yielding trajectories in which correlations are purely sequential, such as MD and “ordinary” (Markov chain) MC - there is more than one way to estimate a variance suitable for ESS calculation via eq 3. Ideally, a number of independent dynamics runs would be started from significantly different initial conditions. Nevertheless, multiple simulations started from the same configuration will also reveal the variance associated with the duration of each run: for instance, if only one simulation makes a transition to an alternative basin, a large variance and small ESS estimate will result, appropriately. It is important to note that the ESS thus calculated is characteristic of one of the simulation trajectories, so that N_{obs} independent trajectories imply an ESS which is N_{obs} times as large. This discussion also indicates that a single long trajectory can be divided into segments (“blocks”) which can be used for variance estimation.

More complex simulation methods, such as replica exchange,^{3,29,30} may require multiple independent runs for careful variance estimation. To see why in the case of replica exchange, note that continuous trajectories will traverse a ladder of different “conditions” (e.g., temperatures or force fields), but often only a single condition is of interest. By the construction of such an algorithm, configurations appearing at one time at the condition of interest may be strongly correlated with configurations occurring later on -

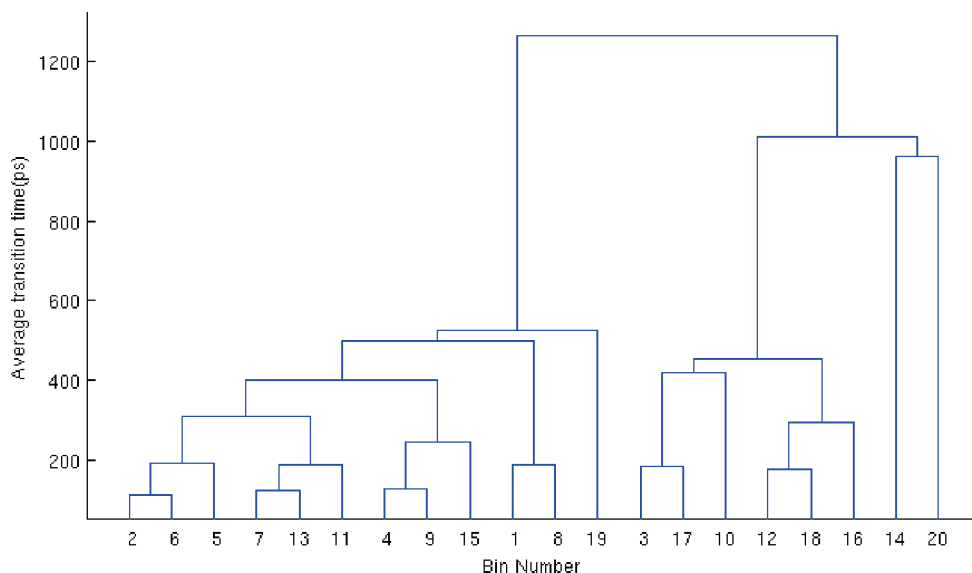


Figure 3. Hierarchical physical states for dileucine shown via the average transition time required for transition among bin pairs. Bin pairs that combine “faster” (i.e., have shorter transition time) are combined at a lower level of the hierarchy.

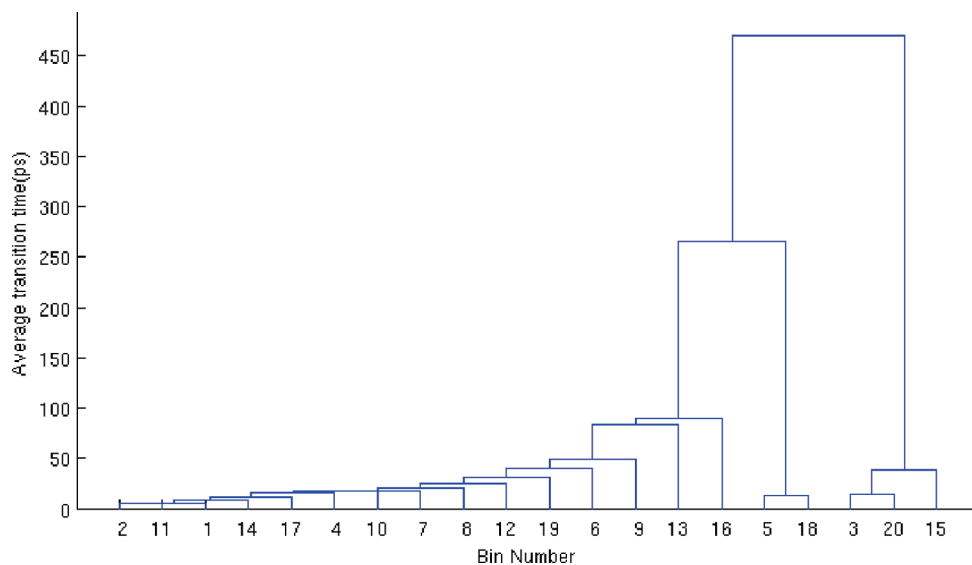


Figure 4. Hierarchical physical states for butane shown via the average transition time ($1/k_{ij}^{\text{eff}}$) required for transition among bin pairs. Bin pairs that combine “faster” (i.e., have shorter transition time) are combined at a lower level of the hierarchy.

but not with configurations in between, when a different trajectory may have occupied the condition of interest. In sharp contrast to dynamics simulation, correlations may be nonsequential. This absolutely precludes estimating the variance by simply cutting up the equilibrium ensemble into blocks or segments. Such a variance may not reflect sampling quality and could misleadingly reflect only diffusivity among ladder levels.²¹ We note that subtleties in estimating uncertainties in replica exchange simulations have been noted previously.^{32–34} Further, it may be possible to use the independent continuous trajectories from one replica exchange run to provide \bar{p}_j values in eq 4; see also ref 16.

For a nondynamical simulation method, the only sure way to estimate a variance which reflects the underlying ESS is by multiple independent runs. The extra cost could be modest if each run is sufficiently short and such runs would, of course, enhance sampling, i.e., they would “pay for them-

selves”. In any case, the cost seems worthwhile when it permits careful quantification of the results.

2.4. Systems Studied. We study several systems using the ESS procedure described above to establish correctness and robustness of the procedure. The systems range from toy models and small molecules to coarse-grained and atomistic proteins.

Toy Models with Known Sample Size. First, we study simple toy models for which the correct sample size is known in advance, to establish the correctness of the procedure. The toy system has n idealized “states” that correspond to preset values of independently drawn random numbers. The sample size in such toy models is simply the number of random numbers drawn by construction. We use two such toy models: $n = 2$ (and both states with equal population), and $n = 5$ (with state probabilities 0.1, 0.15, 0.2, 0.25, 0.3). An application of eq 3 to the two-state system yields, by

construction, the same sample size in both the states. On the other hand, the effective sample sizes obtained may, in general, be different when the number of states is greater than 2. Thus, the five-state toy model is useful in determining the consistency in the sample sizes obtained in the different states.

The sampling in these toy models is nondynamic and uncorrelated. Thus, the use of such models illustrate the applicability of the effective sample size determined by eq 3 to nondynamic sampling. Results for the toy models and all other systems are given in section 3.

Systems with a Priori Known Physical States. In contrast to independent sampling in the toy models, dynamics-based sampling in molecular systems is not typically independent and the sample size is not known in advance. Nevertheless, a knowledge of physical states allows for an independent estimate of the ESS by computing the variances in the known physical states and comparing with the estimate obtained via approximate hierarchical states. Thus, the robustness of the procedure described in section 2 with regard to definitions of physical states can be checked. We study two such systems with *a priori* known states: butane and calmodulin. A second, independent ESS estimate for these systems is derived from a time correlation analysis.¹⁸

We study a standard all-atom butane model using the OPLSAA force field.⁷ This system has three well-known states: trans, gauche+, and gauche-. The 1 μ s dynamical trajectory is generated at 298 K using Langevin dynamics (as implemented in Tinker v. 4.2.2) in vacuum with friction constant 91/ps.

We also study the N-terminal domain of calmodulin, which has the two known physical states: the apo form (PDB id -1CFD) and the holo form (PDB id -1CLL). A long trajectory (5.5×10^7 MC sweeps) was generated by using “dynamic” Monte Carlo (small, single-atom moves only) as previously described.³⁵ To permit transitions, we use a simple alpha-carbon model with a double-G \ddot{o} potential to stabilize the two physical states. Full details of this model are given elsewhere.³⁵

Systems with Unknown Physical States. For most biomolecular systems, the physical states are not known in advance. For this reason, we test our method on several such systems, starting with two peptides: leucine dipeptide (acetaldehyde-(Leu)₂-n-methylamide) and Met-enkephalin (NH₃⁺-Tyr-[Gly]₂-Phe-Met-COO⁻). We use the Charmm27 force field for leucine dipeptide and OPLSAA force field for Met-enkephalin and generate trajectories using overdamped Langevin dynamics (in Tinker v 4.2.2) at 298 K with a friction constant of 5/ps for both. For leucine dipeptide we use a uniform dielectric of 60 and the GB/SA solvation for Met-enkephalin.^{36,37} For each system, a 1 μ s simulation is performed with frames stored every 1 ps for Met-enkephalin and every 10 ps for leucine dipeptide.

We then study a much more complex system - rhodopsin.^{15,38} We analyze 26 independent 100 ns molecular dynamics simulations of rhodopsin in a membrane containing 50 1-stearoyl-2-docosahexaenoyl-phosphatidylethanolamine (SDPE) molecules, 49 1-stearoyl-2-docosahexaenoyl-phosphatidylcholine (SDPC) molecules, and 24 cholesterol. There is an

explicit water environment embedded in a periodic box. The all-atom CHARMM27 force field was used. We analyze only protein coordinates under the assumption that these will include the slowest time scales.

2.5. Independent ESS Estimates. We would like to compare ESS estimates obtained from our new procedure to independent “reference” results. Independent ESS estimates can be obtained in several ways, depending on the system and simulation method to be analyzed.

For uncorrelated sampling in the toy models, the ESS is known in advance: it is simply the number of samples used to obtain the state variance. In this case, we merely check that knowledge of the variances along is sufficient to recover the number of samples.

In some molecular systems, such as butane and calmodulin in this study, physical states are known in advance. Independent variance (and hence ESS) estimates are then obtained using the “exact states”. These are compared to ESS estimates obtained fully automatically based on states approximated from trajectories. In systems with a small number of states, additional ESS estimates can be approximately obtained simply by counting transitions.

Whether or not physical states are known, if a dynamics (or Markov Chain MC) trajectory is analyzed, independent ESS estimates can be obtained using our previously developed structural decorrelation time analysis¹⁶ and eq 1. This approach uses a t_{corr} reflecting the time to sample the whole distribution. In work with model one-dimensional systems (data not shown), we have found that the ESS is estimated within a factor of 2 using the method of ref 16; therefore, ESS estimates based on decorrelation time are shown as ranges.

3. Results

3.1. Nondynamic Toy Systems. First, we establish the formal correctness of our method for estimating N^{eff} . For this purpose, we study the toy models described in section 2.4 for which the sample size is known in advance. For each toy model, we draw N independent samples and estimate the sample size using the procedure described in section 2.

To determine whether an accurate estimate of N^{eff} ($\equiv N$) is obtained, we also compute both the mean value and standard deviation of N^{eff} based on many repeats. As suggested by eq 3, the N^{eff} variation depends on variances of both the mean population and the population variance (these quantities are equal across the states for a two-state system). Further, care must be taken to account for the nonlinear dependence of N^{eff} on the state variance in eq 3.

For the two-state model, with $N = 2000$, we obtain a mean value of $\langle N^{\text{eff}} \rangle = 2004$, with a standard deviation of 57.4. Similarly, for $N = 4000$, we obtain a mean $\langle N^{\text{eff}} \rangle = 4041$ with a standard deviation 117.6. This confirms our basic premise of using eq 3 based on the binomial distribution. The intrinsic fluctuations in the estimates, about 3% in both cases, presumably do not decrease with increasing N due to the nonlinearity of eq 3.

In the five-state model estimates of the sample sizes in each state are different (see section 2), and such a model is

Table 1. Automated and Independent Effective Sample Sizes for Butane and Calmodulin^a

	approximate states			known states	time correlation	counting
	1	2	3			
butane	6064	6236	6200	5865	5000–10000	6000
calmodulin	93	90	92	91	80–160	80

^a ESS estimates obtained from eq 3 using three different sets of approximate physical sets are shown in columns 2–4. Also shown are ESS estimates from eq 3 and the known physical states (column 5), the structural decorrelation time analysis¹⁶ (column 6), and from counting the number of transitions (column 7).

a further step in confirming eq 3 in a more heterogeneous case. Using $N = 2000$, and states with fractional populations 0.1, 0.15, 0.2, 0.25, and 0.3, the mean sample sizes (standard deviations) are obtained as 2007 (70), 1998 (57), 1974 (35), 1966 (79), and 1986 (63), respectively. There is a good agreement across the states as well as with the correct sample size $N = 2000$.

3.2. Systems with A Priori Known Physical States. We turn next to molecular systems with known physical states for which long dynamics trajectories are available. This is essentially the simplest case for a molecular system, because two independent estimates of ESS can be obtained, as described below. Comparison of our blind, automated procedure to these independent estimates further establishes the correctness and robustness of the procedure. Additionally, because our automated state-construction procedure is somewhat stochastic (see Appendix), we repeat the procedure to understand the fluctuations in our ESS estimates.

We obtained multiple estimates of ESS as described above using a single long trajectory for each of the two systems with known physical states - butane and calmodulin. Table 1 shows results for N^{eff} for the two systems, including three different estimates of N^{eff} from eq 3 based on different sets of approximate states. Comparison is also made to the use of eq 3 based on known physical states and to the range of effective sample sizes obtained using time correlation analysis. For both butane and calmodulin, the procedure is very “robust” in estimating N^{eff} , as different binning procedures give similar estimates. These estimates also agree with the range of sample sizes suggested by the correlation time analysis and with counts of transitions. For butane, the total number of transitions among the three state is about 6000. For calmodulin, the total number of transitions is 80. These results also agree with the estimates in Table 1.

3.3. Systems with Unknown Physical States. Exact physical states are not known in advance for most biomolecular systems. Thus, we test the approach described in section 2 to determine ESS in three such systems - dileucine, Met-enkephalin, and rhodopsin. Because the physical states are not well-defined, we can only obtain independent estimates from the time correlation analysis. A single 1 μs trajectory is analyzed for each of the peptides, whereas 26 trajectories of 100 ns each are studied for rhodopsin.

Table 2 shows repeated ESS estimates using our approximate states with eq 3 as well as the time-correlation analysis for both dileucine and Met-enkephalin. There is good agreement between our variance-based estimates and those from time correlation analysis for both systems.

Table 2. Effective Sample Sizes for Dileucine and Met-Enkephalin^a

	approximate states			time correlation
	1	2	3	
dileucine	1982	1878	1904	1100–2200
Met-enkephalin	416	362	365	250–500

^a Equation 3 is used on the final two states in the hierarchical picture obtained by three different repetitions of the binning procedure (columns 2–4), and the ESS is independently estimated from the structural decorrelation time correlation (column 5).

We proceed to analyze the sample size of 26 rhodopsin trajectories based on our approximate states with eq 3. Our analysis gives three physical states, with sample sizes 1.93, 1.99, and 2.73, respectively, per 100 ns trajectory. The three states are never further connected in full hierarchy, since transitions are not observed between some bin pairs. The three N^{eff} estimates, nevertheless, are quite similar and all are $O(1)$. However, eq 3 always yields a value ≥ 1 , indicating that the 100 ns rhodopsin values are effectively minimal and reflect inadequate sampling. In ref 15, Grossfield and co-workers examined the same trajectories with principal components and cluster populations. They concluded, similarly, that rhodopsin’s fluctuations are not well described by 100 ns of dynamics and that the sampling is not fully converged even for individual loops.

3.4. Application to Discontinuous Trajectories. Although sample size estimation using eq 3 is applicable to nondynamical simulation methods, the underlying physical states, approximated from transition rates between regions of configuration space (see Appendix) may not be easy to calculate from nondynamical trajectories. We therefore investigate the feasibility of running short dynamics trajectories starting from configurations previously obtained from nondynamic simulations and then estimating ESS based on states from the short dynamics simulations.

For this purpose, we ran a series of 20 short Langevin simulations for both dileucine and Met-enkephalin, starting from configurations obtained in the original long trajectories which serve as proxies for well-sampled ensembles by an arbitrary method. For both systems, we approximated states as described in section 2.1 and estimated the ESS as $\min\{N_j^{\text{eff}}\}$. For simulation segments as short as 200 ps we could obtain the correct ESS within a factor of 2 (dileucine) or 3 (Met-enkephalin), whereas the longest time scales (i.e., decorrelation time) in these systems exceed a nsec.¹⁶ However, a precise estimate of the ESS required 1–3 ns segments.

We note that Chodera et al.²⁵ also used discontinuous trajectories in their state approximation scheme. As noted in the Appendix, our scheme is a simplified version of theirs.

3.5. Spurious Results from Unphysical States. Thus far, we have focused on using physical states with eq 3, based on the arguments presented in the Introduction. In principle, however, eq 3 can be applied to an arbitrary region. To confirm the need for using states, here we investigate what happens when only part of a state is used. We will see that spurious ESS estimates results.

Table 3. Spurious ESS Estimates When Physical States Are Not Used^a

bin number	ESS
1	12567
2	61380
3	82080
4	91820
5	292640
6	71180
7	240200
8	5600
9	162720
10	310260

^a Butane sample size is estimated in each of 10 arbitrary regions of configuration space. The actual sample size is ~ 6000 , based on a 1 μ s Langevin dynamics trajectory.

The system we examine is butane. We divide the configuration space into 10 “bins” using Voronoi cells³⁹ and perform *no* combination into physical states. We estimate the effective sample size using eq 3 for each bin. We examine a 300 ns trajectory, for which $N^{\text{eff}} \approx 2000$.

Table 3 shows estimates of ESS obtained for each of the 10 arbitrary bins, which are not approximate states. The estimates show a dramatic bin dependence.

The problem with using bins rather than states results for simulations which use dynamics. In fact, arbitrary bins can be used in eq 3 if sampling is fully uncorrelated; we verified this using a fixed number of butane configurations which were essentially uncorrelated. However, when dynamics are present, the variance of one bin is a convolution of state-population variances and fast intra-state processes. We discuss this in more detail below.

4. Discussion

4.1. Is the ESS Measure Too Strict? It certainly can be argued that many observables of interest will “converge” to satisfactory accuracy and precision even with small sample sizes. However, the ESS measure should be valuable in two important regards: (i) as an objective measure of sampling quality that can be applied to any method to enable unbiased comparison and (ii) as a measure of the quality of ensemble generated, which can be expected to embody structural details of interest in biomolecular simulation.

4.2. Diagnosing Poor Sampling. A key outstanding issue is how to know when sampling is inadequate, at least in the self-consistent sense of section 2.2. The “diagnosis” of poor sampling is intimately connected with the idea of estimating ESS by subdividing a dynamics trajectory into smaller, equal segments.

First, consider subdividing a dynamics trajectory into smaller, equal segments to estimate the population mean and variances. If the trajectory is very long compared to all correlation times, no serious problems will arise. If the sample size estimate for each of these segments is $O(1)$, however, then the method does not reliably give the estimate of the sample size of the total trajectory and likely overestimates it. For example, if the correct total number of independent configurations in the full trajectory is 10, and we subdivide it into 20 equal segments, then each of the segment will give a sample size of 1, which is the minimum number possible

using eq 3. This leads to an overestimate of the sample size. But the problem is easily diagnosed by $ESS \sim 1$ for each segment. If division into fewer segments still leads to $ESS \sim 1$, sampling is likely inadequate.

It is of interest to consider a special case of poor sampling, where trajectories started from different initial conditions visit mutually exclusive states - i.e., have no overlap. In this case, the \bar{p}_j values in eq 3 will not be known correctly. Nevertheless, because some $p_j^{(i)}$ values in eq 4 will be zero, the analysis will correctly “sense” a maximal variance with $ESS \sim O(1)$ for each simulation. In other words, poor sampling can still be diagnosed.

4.3. The Inadequacy of Arbitrary Regions for ESS Estimation. It is somewhat difficult to understand the reason for spurious results for ESS obtained using a correlated dynamics trajectory from bins that are a small part of a physical state as in section 3.5. A two-state thought experiment is instructive. Consider a system with two basins, A and B, separated by a barrier. Imagine that we divide the full space into many bins, of which the seventh is a small part of state A and has the (true) probability of p_7 . In ideal uncorrelated sampling, the observed outcomes should be in the bin with probability p_7 and out of the bin with probability $1-p_7$. However, in dynamical sampling, if the system is trapped in state A (with a fractional population p_A) for the observation time, the observed probability in the bin turns out to be p_7/p_A instead of p_7 . Conversely, if a trajectory segment is trapped in state B, the observed population of bin 7 is zero. The variance of this observed distribution when $p_7 \ll p_A$ is much lower than the binomial case; physically, the fast time scales within state A act to “smooth out” population variation within a small part of the state. The estimated ESS obtained using a correlated (i.e., dynamical) trajectory thus will typically appear to be larger based on such a bin. This is seen in Table 3, except for one bin which corresponds, roughly, to a physical state.

5. Conclusions

We have developed a new method to assess the effective sample size, which quantifies the degree of sampling in molecular simulations. Our approach is based on the fundamental role of physical states and hence of variances in their populations. A major feature of the method is that it is applicable both to dynamical and nondynamical simulation methods and gives a tool to compare sampling and efficiencies of different molecular simulation algorithms. Our previous approach was applicable only to dynamical (sequentially correlated) molecular simulation algorithms.¹⁶ Another feature of the new procedure is that it is applicable to discontinuous trajectories as well. We also demonstrated that our procedure is fairly insensitive to the precise definitions of physical states - a fact that is expected to be of importance for systems for which actual physical states are not known in advance. We applied the approach to systems ranging from discrete toy models to an all-atom treatment of rhodopsin.

To supplement the estimation of the effective sample size, we also developed a simple procedure for the automated determination of physical states, which is based on previous

work.²⁵ This procedure yields, in a natural way, a hierarchical picture of the configurational space, based on transition rates between regions of configurational space.

Acknowledgment. We are grateful to the authors of ref 15 for providing their trajectories for analysis, and we received helpful input from David Jasnow, Derek Cashman, Ying Ding, Artem Mamonov, and Bin Zhang. Funding for this work was provided by the NIH (grant GM076569) and NSF (grant MCB-0643456).

Appendix: A Simple Hierarchical Scheme for Approximating Physical States from Dynamical Trajectories. In this Appendix, we describe our physical state discovery method and its results. In this method, bins or regions in configurational space are combined to give the physical states, as discussed below in more detail. Our method is based on the work of Chodera et al.²⁵ but is simpler. There is no Markovian requirement on the selection of bins. Indeed, a typical bin in a configurational space for a large multidimensional system may itself encompass several separate minima. We emphasize that our procedure is designed solely for the purpose of estimating sample size and is not claimed to be an extremely precise description of states.

Our approach explicitly shows the hierarchical nature of the configurational space^{40,41} and focuses on the slowest time scale - which is of paramount importance for the estimation of the effective sample size in the main text.

Method

Use of Rates To Describe Conformational Dynamics.

Our approximate states are constructed based on rates between regions of configuration space, which are a fundamental property that emerges uniquely from the natural system dynamics. Following ref 25, we first decompose the conformational space into multiple bins as detailed below. Subsequently, we combine bins that have the highest transition rates between them, iterating to create a hierarchical description. This procedure is based on the physical idea of separation of time scales: there are faster time scales (high transition rates) associated with regions within a single physical state and slower time scales for transitions between states. Furthermore, “fast” and “slow” time scales are not absolute, necessitating a hierarchical description following precedents.^{40,41}

Binning Decomposition of the Configurational Space.

We divide the whole configuration space into m bins and determine the physical states by combination of these regions. All data reported here used $m = 20$. The value $m = 20$ was motivated by our intuition that regions with less than 5% population should not be allowed to dominate ESS. However, we obtained very similar results using larger m values of 40 and 60.

The procedure to decompose the whole configurational space (with N configurations) into m bins is as follows:²⁸

- A reference configuration i is picked at random from the trajectory.
- The distance of the configuration i to all remaining configurations in the trajectory is then computed, based on an appropriate metric discussed later.
- The configurations are sorted according to distance, and the closest N/m configurations are removed.

- Steps 1–3 are repeated $m - 1$ times on the progressively smaller set of remaining configurations, resulting in a total of m reference configurations.

For the distance metric, we select the root-mean squared deviation (rmsd) of the full molecule, estimated after alignment. We note that using just the backbone rmsd may be a poor distance metric for peptides as it ignores side chain kinetics. However, other metrics may prove useful.

After reference structures are selected, we decompose the whole space into bins based on a Voronoi construction. That is, for each configuration, we calculate the rmsd of this configuration to each of the m reference structures. We assign the configuration to the bin associated with the reference structure, with which the configuration has the smallest rmsd.

Calculation of Rates among Bins and Bin Combination.

We compute the mean first passage time (MFPT) from each bin, i , to every other bin, j , using a continuous dynamical trajectory or a set of trajectories. The rate from bin i to bin j is the inverse of that MFPT. In general, the rate from bin i to bin j is not the same as the rate from bin j to bin i - and we take a linear average of these two rates to define an effective rate between bin i and bin j , k_{ij}^{eff} . The effective rates are then used to construct a hierarchy of states. Rates may also be computed via alternate methods such as via transition matrices, and these different definitions may lead to somewhat different approximate physical states; however, the estimates of the effective sample size should be fairly robust, based on our experience varying other parameters.

Hierarchy. In essence, we perform hierarchical clustering.⁴² We construct a hierarchy of states by combining bins together if all pairs of rates k_{ij}^{eff} exceed a cutoff, k_c . The cutoff is then decreased. We start with $k_c = 1/\text{min}(\text{MFPT})$ and progressively decrease k_c (or, equivalently, increase the transition time cutoff) until the next smallest k_{ij}^{eff} value is reached. The new set of k_{ij}^{eff} is, then, calculated between the new set of bins. With a decrease in k_c , more bins are combined resulting fewer states. Ultimately all bins are combined if transitions among all bin pairs are present in the trajectories which are analyzed. See Results below.

The rule of unanimity - the requirement for fast transitions among *all* bin pairs in a state - is important for ESS estimation. In physical terms, it prevents a bin which “straddles” two states from combining with bins on both “sides” of the straddled barrier (until a suitably low k_c is employed). In turn, this absence of straddling prevents anomalous ESS estimates.

We note that the hierarchical picture can be significantly affected by the time interval between snapshots underlying the MFPT calculations. For example, although a trajectory may have a low likelihood (hence a low rate) to cross over the $2k_B T$ barrier in Figure 2 in time τ_1 , it may easily cross that barrier for a long enough time interval, τ_2 . Thus, a hierarchical picture at the lowest level can differentiate the two left states of Figure 2 if the rates are computed from the dynamic trajectory with snapshots at every τ_1 interval. On the other hand, if the rates are computed using the τ_2 interval, $2k_B T$ barrier cannot be resolved at the lowest hierarchical level. As an extreme case, if the interval between snapshots is longer than the largest correlation

time in the system, then the rates to bin i from any other bin is simply proportional to the equilibrium population of bin i - and the application of the procedure described above is not appropriate.

Figures 3 and 4 show the hierarchical physical for dileucine and butane, respectively. Both start with $m = 20$ initial bins and combine all the way to a single state. The effective sample size is calculated from the two-state level of the hierarchy as described in section 2.

References

- (1) Frenkel, D.; Smit, B. *Understanding Molecular Simulations*; Academic Press: San Diego, 2002.
- (2) Berg, B. A.; Neuhaus, T. *Phys. Rev. Lett.* **1992**, *68*, 9–12.
- (3) Swendsen, R. H.; Wang, J.-S. *Phys. Rev. Lett.* **1986**, *57*, 2607–2609.
- (4) Okamoto., Y. *J. Mol. Graphics Modell.* **2004**, *22*, 425–439.
- (5) Abrams, J. B.; Tuckerman, M. E. *J. Phys. Chem. B* **2008**, *112*, 15742–15757.
- (6) 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.
- (7) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
- (8) 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.
- (9) Ren, P.; Ponder, J. J. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.
- (10) Lamoureux, G.; Mackerell, A.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185–5197.
- (11) Keller, B.; Daura, X.; van Gunsteren, W. F. *J. Chem. Phys.* **2010**, *132*, 074110.
- (12) Reich, L. E. *A Modern Course in Statistical Physics*; Wiley-VCH: Berlin, 2009.
- (13) Wenzel, S.; Janke, W. *Phys. Rev. B* **2009**, *79*, 014410.
- (14) Binder, K.; Heermann, D. W. *Monte Carlo Simulation in Statistical Physics*; Springer: Berlin, 1997.
- (15) Grossfield, A.; Feller, S. E.; Pitman, M. C. *Proteins: Struct., Funct., Bioinf.* **2007**, *67*, 31–40.
- (16) Lyman, E.; Zuckerman, D. M. *J. Phys. Chem. B* **2007**, *111*, 12876–12882.
- (17) Flyvbjerg, H.; Petersen, H. G. *J. Chem. Phys.* **1989**, *91*, 461–466.
- (18) Mountain, R. D.; Thirumalai, D. *J. Phys. Chem.* **1989**, *93*, 6975–6979.
- (19) Mountain, R. D.; Thirumalai, D. *Int. J. Mod. Phys. C* **1990**, *1*, 77–89.
- (20) Ding, Y.; Mamonov, A. B.; Zuckerman, D. M. *J. Phys. Chem. B* **2010**, *114*, 5870–5877.
- (21) Grossfield, A.; Zuckerman, D. M. *Annu. Rep. Comput. Chem.* **2009**, *5*, 23–46.
- (22) Lyman, E.; Zuckerman, D. M. *Biophys. J.* **2006**, *91*, 164–172.
- (23) Diaconis, P.; Holmes, S.; Neal, R. M. *Ann. Appl. Probab.* **2000**, *10*, 720–752.
- (24) Diaconis, P.; Saloff-Coste, L. *J. Comput Syst. Sci.* **1998**, *57*, 20–36.
- (25) Chodera, J. D.; Singhal, N.; Swope, W. C.; Pande, V. S.; Dill, K. A. *J. Chem. Phys.* **2007**, *126*, 155101.
- (26) Noe, F.; Horenko, I.; Schutte, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.
- (27) Lyman, E.; Zuckerman, D. M. *J. Chem. Phys.* **2007**, *127*, 065101.
- (28) Zhang, X.; Mamonov, A. B.; Zuckerman, D. M. *J. Comput. Chem.* **2009**, *30*, 1680–1691.
- (29) Earl, D. J.; Deem, M. W. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910–3916.
- (30) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140–150.
- (31) Buchete, N.-V.; Hummer, G. *Phys. Rev. E* **2008**, *77*, 030902.
- (32) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, *3*, 26–41.
- (33) Huang, X.; Bowman, G. R.; Pande, V. S. *J. Chem. Phys.* **2008**, *128*, 205106.
- (34) Rosta, E.; Hummer, G. *J. Chem. Phys.* **2009**, *131*, 134104.
- (35) Zuckerman, D. M. *J. Phys. Chem. B* **2004**, *108*, 5127–5137.
- (36) Michel, J.; Taylor, R. D.; Essex, J. J. *J. Chem. Theory Comput.* **2006**, *2*, 732–739.
- (37) Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156–169.
- (38) Grossfield, A.; Feller, S.; Pitman, M. *Proc. Natl. Acad. Sci.* **2006**, *103*, 4888–4893.
- (39) Voronoi, G. *J. Reine. Angew. Math.* **1907**, *133*, 97–178.
- (40) Fraunfelder, H.; Parak, F.; Young, R. D. *Annu. Rev. Biophys. Chem.* **1988**, *17*, 451–479.
- (41) Wales, D. J. *J. Chem. Phys.* **2009**, *130*, 204111.
- (42) Ward, J. H. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.

JCTC

Journal of Chemical Theory and Computation

A Graphics Processing Unit Implementation of Coulomb Interaction in Molecular Dynamics

Prateek K. Jha,[†] Rastko Sknepnek,[‡] Guillermo Iván Guerrero-García,[‡] and
Monica Olvera de la Cruz^{*,‡,†,§}

*Department of Chemical and Biological Engineering, Department of Materials Science
and Engineering, and Department of Chemistry, Northwestern University,
Evanston Illinois 60201*

Received June 29, 2010

Abstract: We report a GPU implementation in HOOMD Blue of long-range electrostatic interactions based on the orientation-averaged Ewald sum scheme, introduced by Yakub and Ronchi (*J. Chem. Phys.* **2003**, *119*, 11556). The performance of the method is compared to an optimized CPU version of the traditional Ewald sum available in LAMMPS, in the molecular dynamics of electrolytes. Our GPU implementation is significantly faster than the CPU implementation of the Ewald method for small to a sizable number of particles ($\sim 10^5$). Thermodynamic and structural properties of monovalent and divalent hydrated salts in the bulk are calculated for a wide range of ionic concentrations. An excellent agreement between the two methods was found at the level of electrostatic energy, heat capacity, radial distribution functions, and integrated charge of the electrolytes.

1. Introduction

The introduction of highly optimized, specialized hardware, that is, the graphics processing unit (GPU), has allowed for rendering high definition, nearly photorealistic 3D scenes in real time on a standard personal computer. Ever increasing market demand for fast and realistic graphics has driven a rapid development of inexpensive GPU devices, with a doubling of computational power every 12 months. A modern GPU is a highly parallel, multithreaded device with floating point speed close to 1 TFLOPS and a bandwidth in the 100 GB/s range. The GPU derives its superb computational power from its design, specialized in performing intensive computations on large sets of data in parallel. In recent years, the GPU hardware has become available to nongraphical applications through the advent of general-purpose programmability of the device. Problems that can take advantage of the high-throughput parallel computations can greatly benefit from the GPU architecture and easily reach a 100-fold

increase in performance over equivalent implementation on a CPU.^{1,2} A notable example is molecular dynamics (MD) with reports of GPU implementations achieving speed-ups in excess of 100 times compared to the standard MD codes. However, the high level of data parallelization comes at the expense of limited caching and flow control compared to the CPU.^{1,3–5} Thus, in most cases, it is not possible to simply recompile existing CPU codes on the GPU, and it is often required to substantially redesign existing methods and to develop new algorithms.

In order to reduce finite-size effects, periodic boundary conditions are imposed in a typical MD simulation. That is, if a particle crosses the simulation box boundary, it immediately reappears from the opposite side. Equivalently, this can be seen as if the system has been replicated infinitely many times in each direction and each particle has infinitely many images. In principle, one has to include contributions from all the images of all the particles in order to compute the total energy of the system. In practice, this is seldom necessary, and it is sufficient to cut off interactions at a certain distance r_c and evaluate only the interaction between particles that are within r_c from each other. Formally, if we assume that a system containing N particles is homogeneous

* To whom correspondence should be addressed. E-mail: m-olvera@northwestern.edu.

[†] Department of Chemical and Biological Engineering.

[‡] Department of Materials Science and Engineering.

[§] Department of Chemistry.

and isotropic with density ρ , then the error introduced in the total energy by truncating the potential at r_c is⁶

$$U_{\text{error}} = \frac{N\rho}{2} \int_{r_c}^{\infty} u(r) 4\pi r^2 dr \quad (1)$$

where $u(r)$ is the true, nontruncated potential and we explicitly used the fact that the system is isotropic to write the integral in spherical coordinates. If $u(r) \propto r^{-\alpha}$ with $\alpha > 3$, the correction $U_{\text{tail}} \propto r_c^{3-\alpha}$ can be made arbitrarily small by increasing the cutoff distance r_c . However, if $u(r)$ falls off slower than r^{-3} , any such cutoff will result in a divergent correction to the total energy. Most intermolecular potentials fall off faster than r^{-3} and can be considered short-range. Practically, we can safely truncate them at a suitable cutoff distance, typically chosen to be less than half the diameter of the simulation box, an approximation commonly known as the nearest image convention. Important exceptions are Coulomb and dipolar interaction potentials that fall off with distance as r^{-1} and r^{-3} , respectively. These electrostatic potentials describe interaction between point charges and dipoles ubiquitous in nature, most notably in biological systems. It has been shown in the past that a truncation of long-range interactions can lead to artifacts like the formation of nonphysical structures in ionic liquids.^{7,8}

A proper treatment of the electrostatic interaction is a necessary feature in a general-purpose molecular dynamics code. The Ewald summation method⁹ (henceforth referred to as the ES method) and its derivatives are most commonly used, though several alternatives exist.^{10,11} The trick behind the ES method is to separate the electrostatic energy into a short-range and a long-range contribution, with the long-range contribution computed efficiently in reciprocal space. The numerical effort needed to calculate the total electrostatic energy using ES method scales as $O(N^{3/2})$ with the system size.⁶ The computational expense can be reduced to $O(N \log N)$ by interpolating charges to a lattice and using fast Fourier transform to compute the reciprocal space sum. This is the basis of the smoothed particle mesh Ewald (SPME) method,¹² used in several MD packages. However, an implementation of these methods on the GPU is a challenging task since the long-range contribution has to be treated carefully in order to harvest the full benefit of the massive data parallelization. A successful implementation of the SPME method on the GPU has been recently reported.² Also, an alternative algorithm based on the multipole expansion has been proposed.¹³ Unfortunately, although very efficient, both schemes are complex, and full apprehension of these algorithms requires intimate knowledge of the GPU architecture.

In this paper, we take a different approach and present results of the GPU implementation of a treatment of the electrostatic interaction recently introduced by Yakub and Ronchi (henceforth referred to as the YR method).^{14–16} This approximation is particularly suitable for isotropic ionic fluids. The expressions for the electrostatic energy and the interparticle force are remarkably simple and can be easily implemented into an existing MD code.

2. Methodology

The total electrostatic energy of a system of N charges placed in a cubic box of length L with periodic boundary conditions is

$$E_{\text{el}} = \frac{1}{2} \frac{1}{4\pi\epsilon_0\epsilon_r} \sum_{\vec{n}} \sum'_{i,j=1}^N \frac{q_i q_j}{|\vec{r}_i - \vec{r}_j + \vec{n}L|} \quad (2)$$

where ϵ_0 is the vacuum permittivity, ϵ_r is the relative static permittivity, \vec{r}_i (\vec{r}_j) is the position of charge q_i (q_j). $\vec{n} = (n_x, n_y, n_z)$, where n_x , n_y , and n_z are arbitrary integers, counts all periodic images. The prime in the second sum indicates that the $i = j$ term should be omitted for $\vec{n} = 0$, and the $1/2$ prefactor accounts for double counting. The sum in eq 2 is only conditionally convergent and cannot be directly used in simulations. The idea behind the Ewald method is to separate eq 2 into short- and long-range parts, each expressed as a rapidly converging sum. The total electrostatic energy can be written as the sum of these two contributions plus a constant self-energy contribution⁶

$$E_{\text{el}} = E_{\text{short}} + E_{\text{long}} + E_{\text{self}} \quad (3)$$

The short-range contribution is calculated in the real space as

$$E_{\text{short}} = \frac{1}{2} \frac{1}{4\pi\epsilon_0\epsilon_r} \sum_{i,j=1}^N \sum_{i \neq j} \frac{q_i q_j}{|\vec{r}_i - \vec{r}_j|} \text{erfc}(\sqrt{\alpha}|\vec{r}_i - \vec{r}_j|) \quad (4)$$

where $\text{erfc}(x) = 1 - \text{erf}(x)$ is the complementary error function and α is the Ewald parameter. The long-range sum (E_{long}) is evaluated in the reciprocal space as

$$E_{\text{long}} = \frac{1}{2L^3} \frac{1}{\epsilon_0\epsilon_r} \sum_{\vec{k} \neq 0} \frac{\exp(-k^2/4\alpha)}{k^2} |S(\vec{k})| \quad (5)$$

where $\vec{k} = (2\pi)/(L)\vec{n}$ are the reciprocal lattice vectors, and

$$S(\vec{k}) = \sum_i q_i e^{i\vec{k} \cdot \vec{r}_i} \quad (6)$$

is the charge structure factor. In addition, a self-energy term

$$E_{\text{self}} = -\frac{1}{4\pi\epsilon_0\epsilon_r} \sqrt{\frac{\alpha}{\pi}} \sum_{i=1}^N q_i^2 \quad (7)$$

arises, and it has to be added to the sum of short- and long-range terms. Note that the Ewald parameter α is related to the position of splitting between short- and long-range parts in the Ewald sum. In a simulation, α has to be carefully tuned to ensure the most optimal performance.

We briefly summarize the YR method. A detailed derivation is presented in the original paper.¹⁴ In an ordered phase, the crystal lattice sets a natural direction for the simulation box. This is not the case in fluids where all directions are equivalent; that is, there is no preferred orientation of the simulation box. Thus, Yakub and Ronchi proposed to average eq 3 over all directions of the reciprocal lattice vector \vec{k} ; that is, $E_{\text{el}} = \langle E_{\text{el}} \rangle$, where

$$\langle \dots \rangle = \frac{1}{4\pi} \int_0^\pi d\theta \sin \theta \int_0^{2\pi} d\phi \dots \quad (8)$$

is the average over the polar angle θ and the azimuthal angle ϕ . Note that the averaging is performed over all possible orientations of \vec{k} while \vec{r}_i is kept fixed; thus it is only necessary to average the E_{long} term since E_{short} and E_{self} terms have no θ or ϕ dependence. If we impose electroneutrality, $\sum_{i=1}^N q_i = 0$, the expression for the angularly averaged total electrostatic energy takes a surprisingly simple form¹⁴

$$E_{\text{el}} = \frac{1}{2} \sum_{\substack{i,j \\ i \neq j}} \phi^{(C)}(r_{ij}) \quad (9)$$

with the pair potential

$$\phi^{(C)}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}} \left[1 + \frac{1}{2} \left(\frac{r_{ij}}{r_m} \right)^3 \right] \quad (10)$$

where $r_{ij} = |\vec{r}_i - \vec{r}_j|$, and $r_m = (3/4\pi)^{1/3}L$ is the radius of a sphere of volume L^3 . Note that unlike the adjustable parameter α in the Ewald method, r_m is fixed by size of the simulation box.¹⁴ One counts only the interactions between particles at distances $0 \leq r_{ij} \leq r_m$. A drawback of eq 10 is that the pair potential $\phi^{(C)}(r_{ij})$ is nonzero at its minimum at $r_{ij} = r_m$, $\phi^{(C)}(r_m) = 3q_i q_j / 8\pi\epsilon_0\epsilon_r r_m$, that results in a jump in the cutoff scheme. It is therefore convenient to shift this potential by $-\phi^{(C)}(r_m)$ to bring the boundary values to zero. That is, a modified interionic potential is defined

$$\tilde{\phi}^{(C)}(r_{ij}) = \begin{cases} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}} \left[1 + \frac{1}{2} \left(\frac{r_{ij}}{r_m} \right) \left(\left(\frac{r_{ij}}{r_m} \right)^2 - 3 \right) \right] & r_{ij} < r_m \\ 0 & r_{ij} \geq r_m \end{cases} \quad (11)$$

such that $\tilde{\phi}^{(C)}(r_{ij}) \rightarrow 0$ as $r_{ij} \rightarrow r_m$. By using the electroneutrality condition, $\sum_{i=1}^N q_i = 0$, once again, the expression for the total electrostatic energy (eq 9) in terms of the modified interionic potential can be written as

$$E_{\text{el}} = - \sum_{i=1}^N \frac{3q_i^2}{16\pi\epsilon_0\epsilon_r r_m} + \frac{1}{2} \sum_{\substack{i,j \\ i \neq j}} \tilde{\phi}^{(C)}(r_{ij}) \quad (12)$$

However, the expression for the interparticle force $\vec{f}_{ij} = -\nabla\phi_{ij}^{(C)} = -\nabla\tilde{\phi}_{ij}^{(C)}$ is not affected. Equation 11 is the effective electrostatic pair potential $\tilde{\phi}^{(C)}$ associated with the Coulombic system and is the central result of the YR method. Thus, for an isotropic electroneutral system subject to periodic boundary conditions, long-range effects of the electrostatic interaction can be expressed in terms of a finite range potential.

It is worth mentioning that the cutoff radius $r_m = (3/4\pi)^{1/3}L \approx 0.62L$ is larger than $L/2$. This fact has to be accounted for when calculating the electrostatic potential

$$\Phi^{(C)}(\vec{r}_i) = \sum_{j \neq i} \tilde{\phi}^{(C)}(r_{ij}) \quad (13)$$

on a charge q_i located at \vec{r}_i . Namely, the electrostatic contribution of some charges has to be included twice, that is, as the original charges and as their ‘‘phantom’’ images.¹⁴ These ions are contained in the shaded region in Figure 1 and are obtained by the overlap of a sphere of radius r_m

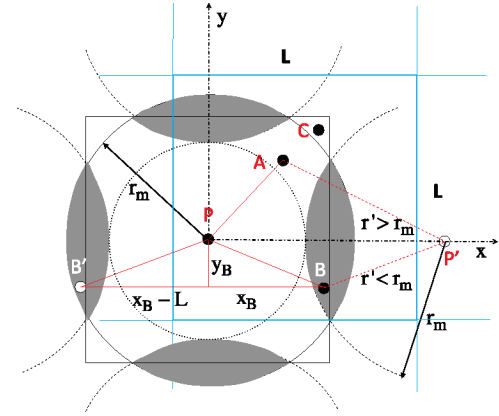


Figure 1. Main unit cell and spheres of radius r_m centered on an ion P and its nearest images P' . Shaded regions indicate the overlap of these spheres. The effective interaction of ion B with ion P is counted twice, both as the original ion and its ‘‘phantom’’ image B' . Ion A is counted only once since it is in a nonoverlap region, and the effective interaction between ions P and C is zero. The blue line indicates the boundaries of the cubic simulation box. See section 3 for description of the dotted circle.

centered on an ion and six spheres of the same radius centered on the images of the ion. To illustrate the calculation of the effective pair potential in the YR method, we show four particles in the xy plane ($z = 0$) with coordinates $P(0, 0, 0)$, $A(x_A, y_A, 0)$, $B(x_B, y_B, 0)$, and $C(x_C, y_C, 0)$. Particle A is in the nonoverlap region of the sphere centered at P , and thus its interaction with P needs to be counted once. That is, the effective pair potential between particles P and A is

$$\tilde{\phi}_{PA}^{(C)} = \tilde{\phi}^{(C)}(\sqrt{x_A^2 + y_A^2}) \quad (14)$$

Particle B is in the overlap region of the sphere centered at P , and thus its interaction with P needs to be counted twice, both with particle B and its ‘‘phantom’’ image B' . The effective pair potential between P and B is

$$\tilde{\phi}_{PB}^{(C)} = \tilde{\phi}^{(C)}(\sqrt{x_B^2 + y_B^2}) + \tilde{\phi}^{(C)}(\sqrt{(x_B - L)^2 + y_B^2}) \quad (15)$$

Since particle C is outside the sphere centered at P , it needs not be counted, and the effective pair potential between P and C , $\tilde{\phi}_{PC}$, is zero.

3. Implementation Details

A direct consequence of the cutoff radius $r_m = (3/4\pi)^{1/3}L$ being larger than $L/2$ is that each ion has more than N neighbors, rendering the use of a neighbor list impractical, from the point of view of both the memory required to store it and the overhead to update it. Instead, one simply loops over all ions and decides which ones contribute once, which twice, and which do not contribute at all to the sum in eq 13. This implies that when implementing the YR method into the HOOMD Blue package, it is not possible to use the sophisticated EvaluatorPair class template specifically designed for the ease of implementing additional short-range potentials. Instead, we implemented a specialization of the

```

1:  $i \leftarrow \text{blockIdx.x} \times \text{blockDim.x} + \text{threadIdx.x}$ 
2:  $\phi_i \leftarrow 0, \vec{f}_i \leftarrow 0$ 
3: for  $j = 1, N_{\text{ions}}$  do
4:   if  $i \neq j$  then
5:      $r_{ij} \leftarrow |\vec{r}_i - \vec{r}_j|$ 
6:     if  $r_{ij} \leq r_m$  then
7:       Compute potential  $\tilde{\phi}^{(C)}(r_{ij})$  using Eq. (11)
8:       Compute force  $\tilde{f}_{ij} = -\nabla \tilde{\phi}_{ij}^{(C)}(r_{ij})$ 
9:        $\phi_i \leftarrow \phi_i + \tilde{\phi}^{(C)}(r_{ij}), \vec{f}_i \leftarrow \vec{f}_i + \tilde{f}_{ij}$ 
10:      if  $r_{ij} > L - r_m$  then
11:        for  $\vec{e} = (-1, -1, -1), (-1, -1, 0), (-1, -1, 1), \dots, (1, 1, 1)$  do
12:           $\tilde{r}_{ij} \leftarrow \vec{r}_i - \vec{r}_j + L\vec{e}$ 
13:          if  $\tilde{r}_{ij}$  is inside the shaded regions in Figure 1 then
14:            Compute potential  $\tilde{\phi}^{(C)}(|\tilde{r}_{ij}|)$  using Eq. (11)
15:            Compute force  $\tilde{f}_{ij} = -\nabla \tilde{\phi}_{ij}^{(C)}(|\tilde{r}_{ij}|)$ 
16:             $\phi_i \leftarrow \phi_i + \tilde{\phi}^{(C)}(|\tilde{r}_{ij}|), \vec{f}_i \leftarrow \vec{f}_i + \tilde{f}_{ij}$ 
17:          end if
18:        end for
19:      end if
20:    end if
21:  end if
22: end for

```

Figure 2. Pseudocode for computing pair Coulomb interactions in the YR method. *blockIdx* and *threadIdx* are standard CUDA structures that contain information about the current execution block and thread, respectively.

PotentialPair template with a custom EvaluatorPairCoulomb class designed to avoid costly use of the HOOMD's neighbor list system.

The CUDA kernel for computing the pair Coulomb interaction in the YR approximation is described in Figure 2. Each thread handles one ion i of the main cell, and one loops over all ions j different from i . If the interionic distance $r_{ij} \leq L - r_m$, that is, if both ions are inside the dotted circle in Figure 1, their contribution to the Coulomb energy is counted once. On the other hand, if $L - r_m < r_{ij} \leq r_m$, one needs to include the contribution of the image ion $\tilde{r} = \vec{r}_i + L\vec{e}$ as well, if $\tilde{r}_{ij} = \vec{r}_i - \vec{r}_j$ is inside one of the shaded regions in Figure 1. \vec{e} is one of the vectors $(-1, -1, -1), (-1, -1, 0), (-1, -1, 1), \dots, (1, 1, 1)$, excluding $(0, 0, 0)$. Finally, if $r_{ij} > r_m$, the ion pair (i, j) is ignored. Since the order in which the contributions from different ions are added to the force and potential sums is irrelevant, a fully coalesced memory read is trivially achievable.

In order to compare the performance and accuracy of the YR method against the ES method in electrolyte systems, we performed MD simulations of hydrated monovalent and divalent electrolytes, with valence $z_+ = -z_- = 1$ and $z_+ = -z_- = 2$, respectively. We use the restricted primitive model (RPM), where an ion is modeled as a hard sphere with a point charge embedded in its center immersed in a continuum dielectric medium. The excluded volume of ions is modeled by the repulsive part of the shifted Lennard-Jones (LJ) potential^{17,18}

$$U_{\text{LJ}}(r) = \begin{cases} 4\epsilon_{\text{LJ}} \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right] + \epsilon_{\text{LJ}} & r < 2^{1/6}\sigma \\ 0 & r \geq 2^{1/6}\sigma \end{cases} \quad (16)$$

where σ is the diameter of bulk hydrated ions, taken as 6.6 Å and 8.25 Å for monovalent and divalent ions respectively.^{19,20} $\epsilon_{\text{LJ}} = 1k_{\text{B}}T$ is the LJ interaction strength, where k_{B} is the Boltzmann constant and T is the temperature. RPM has been quite successful in the prediction of thermodynamic properties of bulk ionic solutions, and in describing several interesting phenomena associated with charged colloidal

systems—such as charge inversion and charge reversal.^{21,22}

In charge inversion, co-ions and counterions switch their roles near an electrified surface, and in charge reversal, the native surface charge of a colloid is overcompensated by counterions. These effects are due to the ion-size correlations, treated in a coarse-grained description by associating an excluded volume to the hydrated ions.

MD simulations were performed in an NVT ensemble at a reduced temperature $T^* = k_{\text{B}}T/\epsilon = 1$ with a time step of 0.005τ , where $\epsilon = 1k_{\text{B}}T$ and $\tau = (m\sigma^2/\epsilon)^{1/2}$ are the reduced LJ units²³ of energy and time, respectively, and m is the mass of ions, set to unity. The relative static permittivity of the solvent is $\epsilon_r = 78.4$, corresponding to an aqueous solution. The average electrostatic energy per ion is defined as

$$E^* = \frac{\langle E_{\text{el}} \rangle}{Nk_{\text{B}}T} \quad (17)$$

where E_{el} is defined in eq 12 and $\langle \dots \rangle$ stands for the time average. The heat capacity per ion is defined as

$$C^* = \frac{C_v}{Nk_{\text{B}}} = \frac{\langle E_{\text{el}}^2 \rangle - \langle E_{\text{el}} \rangle^2}{N(k_{\text{B}}T)^2} \quad (18)$$

where C_v is the heat capacity in real units. These averages and corresponding standard deviations were calculated from the snapshots collected every 100 time steps, which is well beyond the sample correlation time determined from the associated autocorrelation function.⁶ A total of 100 000 to 1 million MD time steps were performed, where the longer runs correspond to the more dilute systems. The time averages are calculated from the second half of each run, well beyond the equilibration time.

The YR method is implemented in the development version of the HOOMD Blue package,²⁴ revision 3109. HOOMD Blue currently supports only single precision arithmetic. Simulations were performed on NVIDIA GTX 295 and GTX 480 GPUs installed in a custom built workstation with an Intel Core i7 920 CPU, 12 GB of RAM, running the Fedora 12 Linux operating system, CUDA 2.2, and NVIDIA Linux driver version 195.36.24. In all runs, only one of the two GPUs on the GTX 295 card was used while the other was kept idle. No monitors were attached to either GTX 295 or GTX 480 cards. On GTX 295, maximum performance is achieved with 64 CUDA threads per block, while on GTX 480, the most optimal thread per block count was 160. Simulations with the ES method were performed in LAMMPS^{25,26} on 32 CPU cores, that is, on four IBM iDataplex blades with two quad-core 2.4 GHz Intel Xeon E5520 processors, 48 GB of memory, and interconnected through a DDR InfiniBand network.

4. Results and Discussion

The YR method requires $O(N^2)$ computations to evaluate the total electrostatic energy, as opposed to $O(N^{3/2})$ computations in the ES method. However, due to its simplicity, computation of the electrostatic interaction in the YR method requires a relatively small number of simple arithmetic

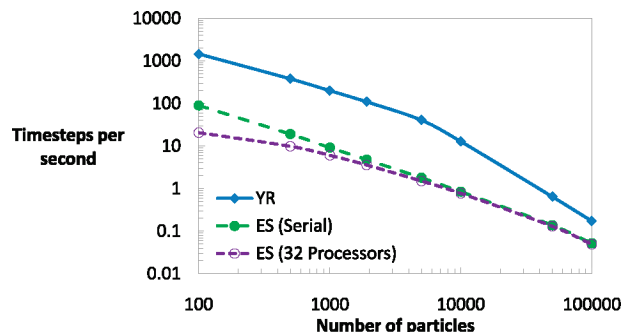


Figure 3. Time steps per second against the number of particles for the YR method on an NVIDIA GTX 480 GPU, and the serial and parallel executions of the ES method on an Intel Xeon computer cluster, for 0.1 M concentration of monovalent salt. The time steps per second for the ES method are defined per unit processor.

operations compared to significantly more complex calculations needed for the same evaluation in the ES method. Therefore, the YR method is significantly faster than the ES method even for 10^5 particles, as shown in Figure 3. The performance gain is even higher when compared to the parallel execution of the ES method, since a significant amount of time is spent in communication between processors for this range of simulation sizes. Note that while the YR method is free of adjustable parameters, the performance of the ES method is sensitive to changes in the real-space cutoff and the reciprocal space precision. We use a real space cutoff of approximately one-fifth of the simulation box size and the reciprocal space calculations were performed with a precision of 10^{-5} . Simulations on the NVIDIA GTX 480 were approximately twice as faster as that on the NVIDIA GTX 295.

Next, we evaluate the thermodynamic predictions of the YR and ES methods for a range of concentrations of monovalent and divalent electrolytes. The number of ions in the simulation box can be chosen arbitrarily as long as the electroneutrality condition is preserved. In this study, we use 1912 ions, a number chosen to balance a full utilization of the GPU with reasonably short execution times. The calculated values of electrostatic energy and heat capacity from the ES and YR methods are shown in Tables 1 and 2 for monovalent and divalent electrolytes, respectively. We observed an excellent agreement between the two methods for a wide range of concentrations—even for divalent ions, where the electrostatic correlations are stronger. Note that there is an appreciable difference in the heat capacities obtained by the ES and the YR methods for dilute systems, in particular for the monovalent case. We attribute this to the inability of the MD to successfully equilibrate a dilute system of charges,^{27,28} and we believe this is not a drawback of the YR method.

To evaluate the concordance of the YR and ES methods in reproducing the structural details of the electrical double layer, we calculate the radial distribution functions, $g_{++}(r^*)$ and $g_{+-}(r^*)$, of the like- and unlike-charged ions, respectively, where $r^* = r/\sigma$ is the reduced distance from the center of the reference ion. These quantities are averaged from snapshots taken every 10 time steps. Radial distribution

Table 1. Average Electrostatic Energy per Ion, $E^* = \langle E_{ei} \rangle / Nk_B T$ and Heat Capacity per Ion, $C^* = C_i / Nk_B$ Calculated by the ES and YR Methods for 1:1 Bulk Hydrated Electrolyte at Different Salt Concentrations (ρ)^a

ρ [M]	E_{ES}^*	E_{YR}^*	C_{ES}^*	C_{YR}^*
1.0	-0.4470(4)	-0.4472(3)	0.095(1)	0.093(1)
0.75	-0.4144(3)	-0.4145(4)	0.091(1)	0.094(2)
0.5	-0.3722(3)	-0.3721(3)	0.095(1)	0.092(1)
0.25	-0.3066(3)	-0.3067(3)	0.085(1)	0.084(1)
0.1	-0.2316(2)	-0.2319(3)	0.074(1)	0.077(1)
0.075	-0.2105(1)	-0.2106(3)	0.069(1)	0.072(1)
0.05	-0.1828(2)	-0.1828(2)	0.062(1)	0.065(1)
0.025	-0.1415(2)	-0.1419(2)	0.055(1)	0.055(1)
0.01	-0.09829(5)	-0.09842(3)	0.0424(2)	0.0428(2)
0.0075	-0.08712(6)	-0.08713(4)	0.0366(2)	0.0375(1)
0.005	-0.07329(4)	-0.07331(2)	0.0314(2)	0.0321(1)
0.0025	-0.05403(7)	-0.05415(3)	0.0212(1)	0.0239(2)
0.001	-0.03530(2)	-0.03529(1)	0.0170(1)	0.0150(1)

^a Uncertainties in the last digit are indicated in parentheses.

Table 2. Average Electrostatic Energy per Ion, $E^* = \langle E_{ei} \rangle / Nk_B T$ and Heat Capacity per Ion, $C^* = C_i / Nk_B$ Calculated by the ES and YR Methods for 2:2 Bulk Hydrated Electrolyte at Different Salt Concentrations (ρ)^a

ρ [M]	E_{ES}^*	E_{YR}^*	C_{ES}^*	C_{YR}^*
1.0	-2.056(3)	-2.056(4)	0.27(1)	0.28(1)
0.75	-1.931(2)	-1.931(3)	0.28(1)	0.28(1)
0.5	-1.773(2)	-1.773(2)	0.28(1)	0.30(1)
0.25	-1.533(2)	-1.533(2)	0.31(1)	0.31(1)
0.1	-1.254(3)	-1.254(2)	0.32(1)	0.31(1)
0.075	-1.173(2)	-1.173(3)	0.31(1)	0.31(2)
0.05	-1.063(2)	-1.063(3)	0.31(2)	0.31(1)
0.025	-0.888(3)	-0.887(2)	0.30(1)	0.29(1)
0.01	-0.676(2)	-0.675(3)	0.29(2)	0.26(2)
0.0075	-0.616(2)	-0.617(2)	0.24(1)	0.25(1)
0.005	-0.535(3)	-0.535(3)	0.27(2)	0.25(2)
0.0025	-0.413(2)	-0.414(2)	0.20(1)	0.21(1)
0.001	-0.283(1)	-0.285(1)	0.159(3)	0.164(9)

^a Uncertainties in the last digit are indicated in parentheses.

functions calculated by the two methods show excellent agreement, as is clear from the curves being virtually indistinguishable in Figures 4 and 5. Further, in both methods, $g_{++}(r^*)$ and $g_{+-}(r^*)$ approach one far from the central ion and at the border of the simulation box, as shown in the insets of Figures 4 and 5. This condition is necessary to ascertain that the system is free of finite size effects, and it is often not met in truncation schemes for handling electrostatic interactions, even for significantly large simulation box sizes.²⁹ For the monovalent ions at 0.01 M (Figure 4a), the contact values show an attraction and a repulsion between unlike- and like-charged ions, respectively, as is expected of bare Coulomb interactions. Interestingly, for the 1 M concentration (Figure 4b), the excluded volume of hydrated monovalent ions leads to a slight attraction between like-charged ions. For the divalent case at 0.005 M (Figure 5a), we observe repulsion and attraction of like- and unlike-charged ions, respectively, that increased in comparison to the monovalent instance at 0.01 M (Figure 4a). In addition, at a 0.5 M concentration (Figure 5b) of divalent ions, we observe a region of charge inversion between $r^* \approx 1.7$ and $r^* \approx 2.6$.

A more stringent test is the calculation of the integrated charge of ions,³⁰

$$P_i(r) = z_i + \int_0^r \left[\sum_{j=+,-} z_j \rho_j g_{ij}(r') \right] 4\pi r'^2 dr' \quad (19)$$

where ρ_j is the bulk density of ion species j in the simulation box. $P_i(r)$ corresponds to the net charge inside a sphere of radius r centered at an ion of species i and hence measures the neutralization of such an ion by the surrounding ionic cloud. At the surface of an ion of species i ($r = 0$), the integrated charge is equal to its valence z_i , whereas sufficiently far from the ion ($r \rightarrow \infty$), $P_i(r)$ approaches zero due to the electroneutrality condition. $P_+(r)$ is identical to $-P_-(r)$ for electrolytes symmetric in valence and size.

The integrated charge of a positive ion, $P_+(r^*)$, for the monovalent and divalent electrolytes are displayed in Figures 6 and 7, respectively. As expected from the radial distribution functions, the concordance of the YR and ES methods is very good for both monovalent and divalent salts, especially near the ionic surface. The fluctuations in the integrated charge near the border of the simulation box are displayed in the insets of Figures 6 and 7. As a check of the global electroneutrality condition, we require that the integrated charge $P_+(r^*)$ approaches zero near the boundary, which is indeed met by the two methods disregarding minor statistical fluctuations. For the monovalent electrolyte at 0.01 M concentration (Figure 6a), the profile of $P_+(r^*)$ shows a monotonic neutralization of the ionic charge. In contrast, for 1 M concentration (Figure 6b), a nonmonotonic neutralization is observed. In fact, there is a region near the ionic surface where the sign of the integrated charge is opposite the sign

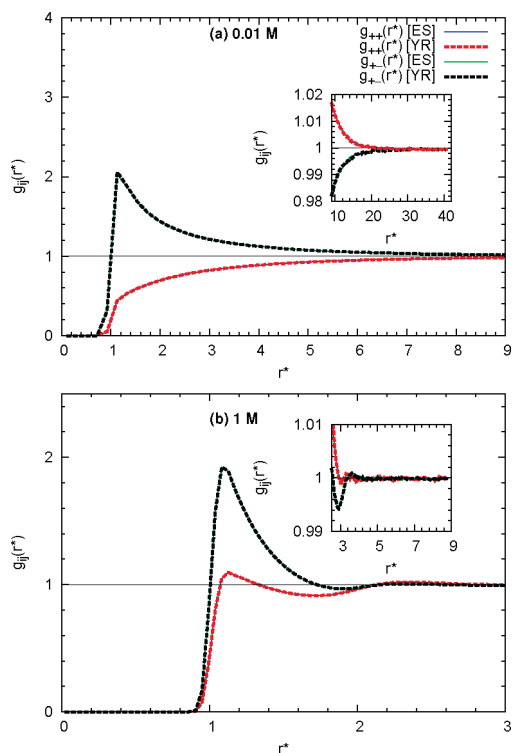


Figure 4. Pair distribution functions, $g_{++}(r^*)$ and $g_{+-}(r^*)$, for 1:1 electrolyte at different concentrations. Bold and dashed lines indicate the ES and YR methods, respectively. Notice that the profiles obtained by the ES and YR methods are virtually indistinguishable. Behavior near the box boundary is shown in the insets.

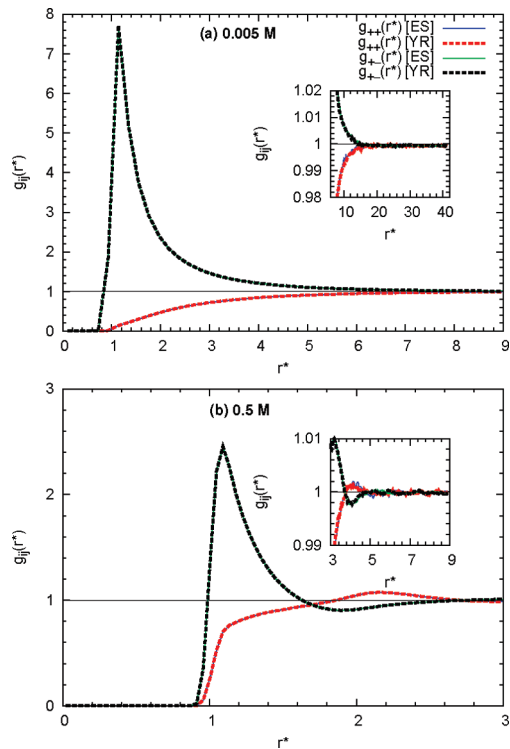


Figure 5. Pair distribution functions, $g_{++}(r^*)$ and $g_{+-}(r^*)$, for 2:2 electrolyte at different concentrations. Bold and dashed lines indicate the ES and YR methods, respectively. Notice that the profiles obtained by the ES and YR methods are virtually indistinguishable. Behavior near the box boundary is shown in the insets.

of the central ion, indicating charge reversal. This behavior is caused by the large excluded volume associated with the hydrated monovalent ions. For the 0.005 M concentration of divalent electrolyte (Figure 7a), a monotonic ionic neutralization behavior akin to that of the 0.01 M monovalent case (Figure 6a) is observed. However, for the 0.5 M concentration of the divalent electrolyte (Figure 7b), the magnitude of maximum charge reversal near the ionic surface increased compared to the 1 M monovalent instance (Figure 6b), and several oscillations in the integrated charge are observed.

5. Conclusion

We have implemented an efficient method for long-range electrostatic interactions in the molecular dynamics on graphics processing units (GPU) based on the scheme originally proposed by Yakub and Ronchi.¹⁴ The method is implemented in the MD package HOOMD Blue.²⁴ In order to test the accuracy of this method applied to the electrolyte systems, thermodynamic and structural properties of bulk hydrated monovalent and divalent salts were calculated. An excellent agreement was found with respect to the conventional Ewald summation method, available in LAMMPS. The current implementation of the YR method is particularly suited for moderate to high concentrations of charges. Its limited applicability to dilute systems is not a flaw of the method but, we believe, is an artifact of MD simulations related to their inability to reach thermodynamic equilibrium in a reasonable time. Additionally, the GPU implementation

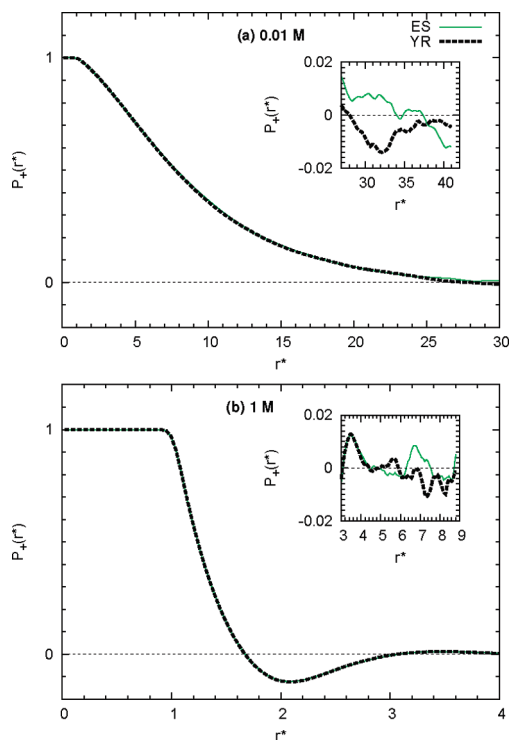


Figure 6. Integrated charge, $P_+(r^*)$, of the 1:1 electrolyte at different concentrations. Bold and dashed lines indicate the ES and YR methods, respectively. Notice that the profiles obtained by the ES and YR methods are virtually indistinguishable. Fluctuations of the integrated charge near the boundary of the box are shown in the insets.

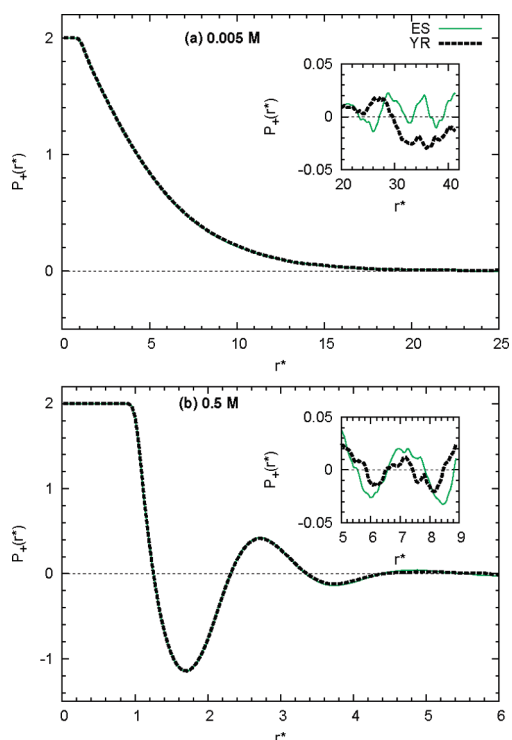


Figure 7. Integrated charge, $P_+(r^*)$, of the 2:2 electrolyte at different concentrations. Bold and dashed lines indicate the ES and YR methods, respectively. Notice that the profiles obtained by the ES and YR methods are virtually indistinguishable. Fluctuations of the integrated charge near the boundary of the box are shown in the insets.

is significantly faster than the fully optimized Ewald summation method for the simulation sizes commonly used in simulations of electrolytes (10^2 to 10^5).

We would like to mention that there is another class of finite or short-range methods for electrostatic interactions such as the Wolf method and its variations^{31–33} that can potentially also benefit from the GPU's high FLOPS count. However, in such schemes, the cutoff and the damping constant must be calibrated for each particular system, whereas the YR method is free of adjustable parameters. The present implementation can be easily extended to study more complicated systems including charged spherocylinders,^{34,35} nanoparticles and colloids,^{33,36–40} asymmetric ionic liquids,^{41–44} and polyelectrolyte solutions and networks,^{45–48} with the incorporation of the corresponding short-range interactions that are already available in the GPU codes. Efforts in these directions are currently underway.

Acknowledgment. G.I.G.-G. thanks Eugene Yakub for his invaluable help in clarifying certain numerical and theoretical aspects of the YR method. The authors thank Graziano Vernizzi for stimulating discussions and William Kung for suggestions on the manuscript. We thank anonymous reviewers for useful suggestions. Molecular dynamics simulations were in part performed on Quest cluster at Northwestern University. A part of GPU simulations presented in this work were executed on GTX 480 GPUs provided by NVIDIA through their professor partnership program. P.K.J. and M.O.d.I.C. are supported by NSF under Award number DMR-0907781. R.S. is supported by the NSEC program on integrated nanopatterning (EEC-0647560) at Northwestern University. G.I.G.-G. is supported by the MRSEC program of the NSF (DMR-0520513) at the Materials Research Center at Northwestern University.

References

- (1) Anderson, J. A.; Lorenz, C. D.; Travestet, A. *J. Comput. Phys.* **2008**, *227*, 5342–5359.
- (2) Harvey, M. J.; De Fabritiis, G. *J. Chem. Theory Comput.* **2009**, *5*, 2371–2377.
- (3) Stone, J. E.; Phillips, J. C.; Freddolino, P. L.; Hardy, D. J.; Trabuco, L. G.; Schulten, K. *J. Comput. Chem.* **2007**, *28*, 2618–2640.
- (4) Owens, J. D.; Houston, M.; Luebke, D.; Green, S.; Stone, J. E.; Phillips, J. C. *Proc. IEEE* **2008**, *96*, 879–899.
- (5) 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*, 864–872.
- (6) Frenkel, D.; Smit, B. *Understanding Molecular Simulation, Second ed.: From Algorithms to Applications* (Computational Science Series, Vol 1), 2nd ed.; Academic Press: New York, 2001.
- (7) Adams, D. *J. Chem. Phys. Lett.* **1979**, *62*, 329–332.
- (8) Brush, S. G.; Sahlin, H. L.; Teller, E. *J. Chem. Phys.* **1966**, *45*, 2102.
- (9) de Leeuw, S. W.; Perram, J. W.; Smith, E. R. *Philos. Trans. R. Soc. London, Ser. A* **1980**, *373*, 27–56.
- (10) Koehl, P. *Curr. Opin. Str. Bio.* **2006**, *16*, 142–151.

- (11) Karttunen, M.; Rottler, J.; Vattulainen, I.; Sagui, C. In *Computational Modeling of Membrane Bilayers*; Feller, S. E., Ed.; Academic Press: New York, 2008; Vol. 60, pp 49–89.
- (12) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pederson, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (13) Hardy, D. J.; Stone, J. E.; Schulten, K. *Parallel Comput.* **2009**, *35*, 164–177.
- (14) Yakub, E.; Ronchi, C. *J. Chem. Phys.* **2003**, *119*, 11556–11560.
- (15) Yakub, E.; Ronchi, C. *J. Low Temp. Phys.* **2005**, *139*, 633–643.
- (16) Yakub, E. *J. Phys. A—Math. Gen.* **2006**, *39*, 4643.
- (17) Messina, R.; González-Tovar, E.; Lozada-Cassou, M.; Holm, C. *Europhys. Lett.* **2002**, *60*, 383.
- (18) Jiménez-Ángeles, F.; Messina, R.; Holm, C.; Lozada-Cassou, N. *J. Chem. Phys.* **2003**, *119*, 4842–4856.
- (19) Nightingale, E. R. *J. Phys. Chem.* **1959**, *63*, 1381–1387.
- (20) Israelachvili, J. *Intermolecular and Surface Forces*, 2nd ed.; Academic Press: London, 1992.
- (21) Quesada-Pérez, M.; González-Tovar, E.; Martín-Molina, A.; Lozada-Cassou, M.; Hidalgo-Álvarez, R. *ChemPhysChem* **2003**, *4*, 234–248.
- (22) Messina, R. *J. Phys.: Condens. Matter* **2009**, *21*, 113102.
- (23) Rapaport, D. C. *The Art of Molecular Dynamics Simulation*, 2nd ed.; Cambridge University Press: Cambridge, U. K., 2004.
- (24) HOOMD Blue. <http://codeblue.umich.edu/hoomd-blue/> (August 20, 2010).
- (25) Plimpton, S. *J. Comput. Phys.* **1995**, *117*, 1–19.
- (26) LAMMPS Molecular Dynamics Simulator. <http://lammps.sandia.gov/> (August 20, 2010).
- (27) Oran, E. S.; Boris, J. P. *J. Phys. IV France* **1995**, *5*, 609.
- (28) Zhang, Z.; Duan, Z. *Chem. Phys.* **2004**, *297*, 221–233.
- (29) Linse, P.; Andersen, H. C. *J. Chem. Phys.* **1986**, *85*, 3027–3041.
- (30) Guerrero-García, G. I.; González-Tovar, E.; Chávez-Páez, M. *Phys. Rev. E* **2009**, *80*, 021501.
- (31) Wolf, D.; Keblinski, P.; Phillpot, S. R.; Eggebrecht, J. *J. Chem. Phys.* **1999**, *110*, 8254–8282.
- (32) Fennell, C. J.; Gezelter, J. D. *J. Chem. Phys.* **2006**, *124*, 234104.
- (33) Avendaño, C.; Gil-Villegas, A. *Mol. Phys.* **2006**, *104*, 1475–1486.
- (34) Avendaño, C.; Gil-Villegas, A.; González-Tovar, E. *J. Chem. Phys.* **2008**, *128*, 044506.
- (35) Avendaño, C.; Gil-Villegas, A.; González-Tovar, E. *Chem. Phys. Lett.* **2009**, *470*, 67–71.
- (36) Linse, P.; Lobaskin, V. *Phys. Rev. Lett.* **1999**, *83*, 4208–4211.
- (37) Guerrero-García, G. I.; González-Tovar, E.; Lozada-Cassou, M.; Guevara-Rodríguez, F. D. *J. Chem. Phys.* **2005**, *123*, 034703.
- (38) Guerrero-García, G. I.; González-Tovar, E.; de la Cruz, M. O. *Soft Matter* **2010**, *6*, 2056–2065.
- (39) Guerrero-García, G. I.; González-Tovar, E.; Chávez-Páez, M.; Lozada-Cassou, M. *J. Chem. Phys.* **2010**, *132*, 054903.
- (40) dos Santos, A. P.; Diehl, A.; Levin, Y. *J. Chem. Phys.* **2010**, *132*, 104105.
- (41) Yan, Q. L.; de Pablo, J. J. *Phys. Rev. Lett.* **2001**, *86*, 2054–2057.
- (42) Yan, Q. L.; de Pablo, J. J. *Phys. Rev. Lett.* **2002**, *88*, 095504.
- (43) Hynninen, A. P.; Dijkstra, M.; Panagiotopoulos, A. Z. *J. Chem. Phys.* **2005**, *123*, 084903.
- (44) Panagiotopoulos, A. Z. *J. Phys.: Condens. Matter* **2005**, *17*, S3205–S3213.
- (45) Liao, Q.; Dobrynin, A. V.; Rubinstein, M. *Macromolecules* **2003**, *36*, 3399–3410.
- (46) Yin, D.-W.; de la Cruz, M. O.; de Pablo, J. J. *J. Chem. Phys.* **2009**, *131*, 194907.
- (47) Narambuena, C.; Leiva, E.; Chávez-Páez, M.; E. Pérez, E. *Polymer* **2010**, *51*, 3293–3302.
- (48) Hsiao, P.-Y.; Luijten, E. *Phys. Rev. Lett.* **2006**, *97*, 148301.

JCTC

Journal of Chemical Theory and Computation

Lowest-Lying Conformers of Alanine: Pushing Theory to Ascertain Precise Energetics and Semiexperimental R_e Structures

Heather M. Jaeger,[†] Henry F. Schaefer III,[†] Jean Demaison,[‡] Attila G. Császár,^{*,§} and Wesley D. Allen^{*,†}

Center for Computational Chemistry and Department of Chemistry, University of Georgia, Athens, Georgia 30602, Laboratoire de Physique des Lasers, Atomes et Molécules, UMR CNRS 8523, Université de Lille I, 59655 Villeneuve d'Ascq Cedex, France, and Laboratory of Molecular Spectroscopy, Institute of Chemistry, Eötvös University, H-1518 Budapest 112, P.O. Box 32, Hungary

Received January 12, 2010

Abstract: The two lowest-energy gas-phase conformers, **Ala-I** and **Ala-IIA**, of the natural amino acid L-alanine (Ala) have been investigated by means of rigorous ab initio computations. Born–Oppenheimer (BO) equilibrium structures (r_e^{BO}) were fully optimized at the coupled-cluster [CCSD(T)/cc-pVTZ] level of electronic structure theory. Corresponding semiexperimental (SE) equilibrium structures (r_e^{SE}) of each conformer were determined for the first time by least-squares refinement of 11–15 structural parameters on modified, experimental rotational constant data from 10 isotopologues. The SE equilibrium rotational constants were obtained by, first, refitting Fourier transform microwave spectra using the method of predicate observations and, second, correcting the resulting effective rotational constants with theoretical vibration–rotation interaction constants (α_i). Careful analysis is made of the procedures to account for vibrational distortion, which proves essential to defining precise structures in flexible molecules such as Ala. Because Ala possesses no symmetry, has several large-amplitude nuclear motions, and exhibits conformers with different hydrogen bonding patterns, it is one of the most difficult cases where reliable equilibrium structures have now been determined. The relative energy of the alanine conformers was pinpointed using first-principles composite focal point analyses (FPA), which employed extrapolations using basis sets as large as aug-cc-pV5Z and electron correlation treatments as extensive as CCSD(T). The FPA computations place the **Ala-IIA** equilibrium structure higher in energy than that of **Ala-I** by a mere 0.45 kJ mol⁻¹ (38 cm⁻¹), showing that the two lowest-lying conformers of alanine are nearly isoenergetic; inclusion of zero-point vibrational energy increases the relative energy to 2.11 kJ mol⁻¹ (176 cm⁻¹). The yet unobserved **Ala-IIB** conformer is found to be separated from **Ala-IIA** by a vibrationally adiabatic isomerization barrier of only 16 cm⁻¹.

I. Introduction

Flexible molecules have potential energy surfaces (PESs) characterized by flat regions and low barriers for conforma-

tional isomerization.^{1–3} L-Alanine (Ala) and all other natural amino acids exhibit such characteristics and have a sizable number of low-energy conformers.^{4–19} The primary differences between the conformers of gas-phase amino acids, which exist exclusively in neutral form, are the number and types of intramolecular hydrogen bonds occurring for various configurations of the amino, carboxylic acid, and any polar side-chain groups. In the case of Ala, the dihedral angle about the central carbon–carbon bond, $\tau(\text{O}=\text{C}-\text{C}_\alpha-\text{N})$, remains

* To whom correspondence should be addressed. E-mail: wdallen@uga.edu (W.D.A.); csaszar@chem.elte.hu (A.G.C.).

[†] University of Georgia.

[‡] Université de Lille.

[§] Eötvös University.

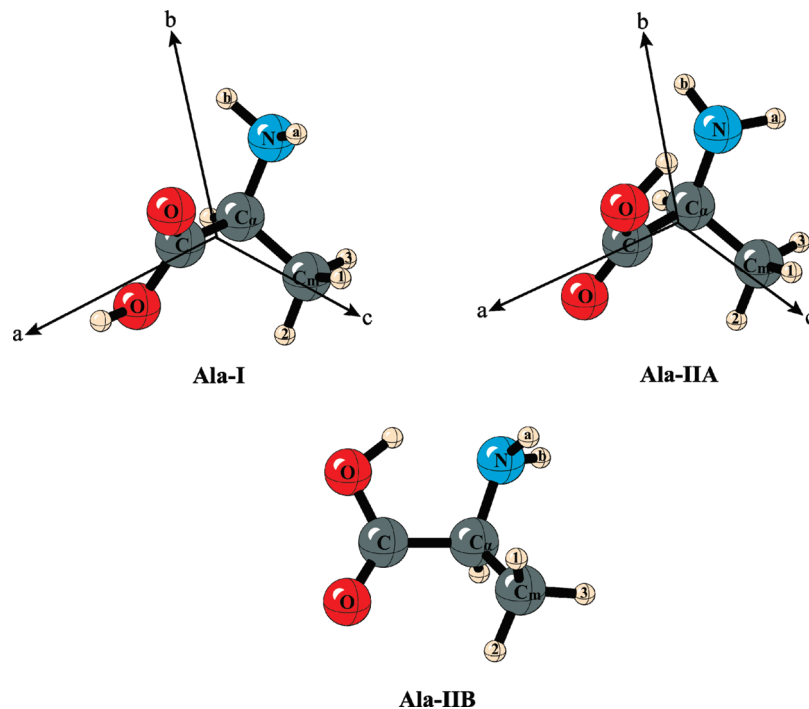


Figure 1. L-Alanine conformers **I**, **IIA**, and **IIB**.

consistently near 0° or 180° for all conformers, despite the variety of possible hydrogen bonds.

Experiments on the structure(s) of free Ala have included gas-phase electron diffraction (GED),^{11,12} jet-cooled millimeterwave (MMW) and Fourier transform microwave (FTMW) spectroscopy in molecular beams,^{5,9,10} and matrix-isolation infrared spectroscopy.¹³ The GED results for Ala cannot provide a clear distinction between the multiple conformers present at the elevated temperature of the experiments. Moreover, the derived r_g , r_α , and r_α^0 structural parameters differ substantially from the corresponding equilibrium (r_e) values because of temperature-dependent rotational–vibrational effects, which can be as large as those induced by conformational changes. The low-temperature MMW and FTMW molecular beam experiments^{5,9,10} clearly identified two gas-phase conformers, **Ala-I** and **Ala-IIA** (Figure 1). The failure to observe other low-energy conformers given by electronic structure computations^{4,6–8} has been attributed to vibrational relaxation in the free-jet expansions.²⁰ The matrix-isolation infrared experiments¹³ also observed two conformers of alanine.

Select r_0 and r_s parameters for **Ala-I** and **Ala-IIA** have been determined⁵ from two sets of FTMW rotational constants involving 10 isotopologues of each conformer. Unfortunately, this approach is not sufficient to obtain an accurate, well-defined empirical structure. Equilibrium structures, free from undesirable isotopic, rotational–vibrational, and temperature effects, are often difficult, if not impossible, to obtain experimentally, especially for flexible molecules. Vibrational distortion, arising from flat, anharmonic regions on the PES, can greatly influence the effective, experimental rotational constants, leading to sizable isotopic effects even at low temperature. Consequently, for conformers of flexible molecules, only equilibrium structures can be compared to one another with any degree of validity. For example,

differences between the backbone structures of glycine and alanine should be ascertained from r_e parameters (see Section III.F).

A protocol has been established whereby a semiexperimental equilibrium structure (r_e^{SE}) can be determined by first correcting empirical, effective ground-state rotational constants with ab initio vibration–rotation interaction constants (α_i) and then performing a structural refinement on the resulting “experimental” equilibrium rotational constants (B_e^{SE}).²¹ This combined experimental and theoretical approach has been successfully applied in many studies,^{21–34} including work that has given r_e^{SE} structures for the lowest-energy conformers of the neutral amino acids glycine (Gly)²⁴ and proline (Pro).²³ In this investigation, accurate r_e^{SE} structures of **Ala-I** and **Ala-IIA** are determined after refitting spectroscopic constants to the observed rotational transitions,⁵ deriving B_e^{SE} constants for 10 isotopologues of each conformer, and imposing geometric constraints from Born–Oppenheimer equilibrium structures (r_e^{BO}) obtained at the highest feasible level of ab initio electronic structure theory [CCSD(T)/cc-pVTZ, vide infra]. This is the first study to conjoin theory and experiment to derive reliable equilibrium structures, including detailed error analyses for both theoretical and experimental procedures, for a molecule as large and flexible as Ala.

All low-energy conformers of Ala possess intramolecular hydrogen bonds that significantly stabilize these structures, increase their rigidity, and provide challenges for electronic structure theory, as shown in numerous previous ab initio studies.^{1,7,14,35–37} The sensitivity of Ala conformational energies to the level of electronic structure theory has been demonstrated by Császár,^{1,7} and Figure 2 vividly displays the energetic variations observed⁷ for **Ala-I**, **Ala-II(A/B)**, and **Ala-III(A/B)**. The **Ala-III** conformers are derived from the **Ala-I** structure in Figure 1 by a 180° rotation of the

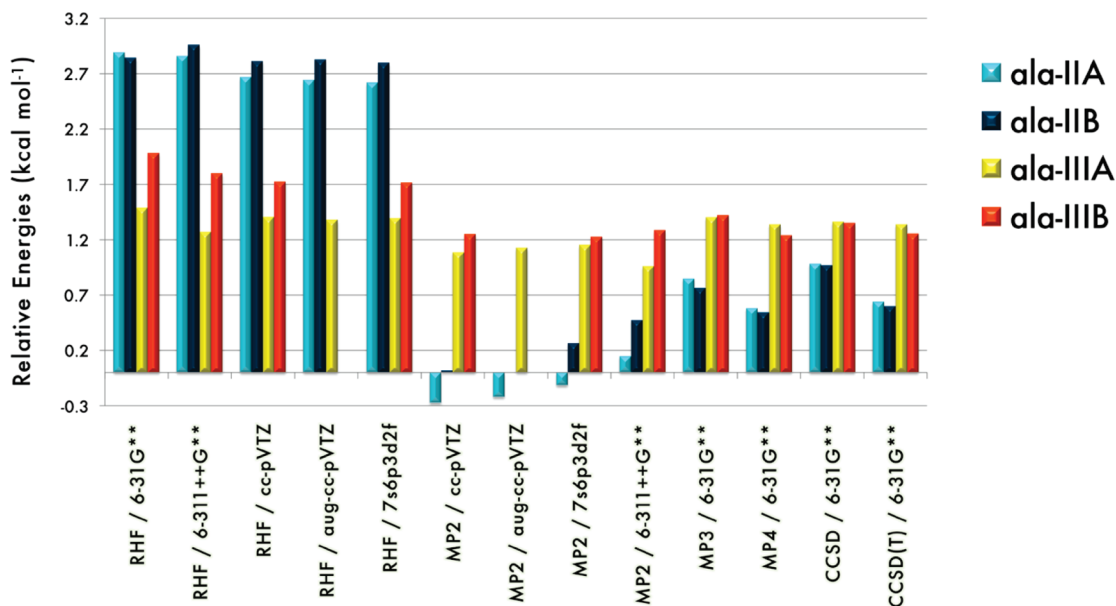


Figure 2. Equilibrium energies relative to **Ala-I** at various levels of electronic structure theory.

–COOH moiety about the C–C_α bond, the two variants differing in whether the carboxyl O–H bond is oriented toward (**IIIA**) or away from (**IIIB**) the methyl substituent at C_α.⁷ Electron correlation reverses the energy ordering of **Ala-II(A/B)** and **Ala-III(A/B)**, in accord with known deficiencies of Hartree–Fock theory in predicting conformational energies of amino acids.^{1,14,38} Strong basis set dependence of conformational energies is also exhibited; large basis sets with diffuse functions are necessary to fully capture the differences in intramolecular hydrogen bonding interactions. In this work, the relative energy of **Ala-IIA** with respect to **Ala-I** in the nonrelativistic, ab initio limit is determined by applying the composite focal point analysis (FPA) approach^{39–44} that has been used successfully in previous studies on amino acids^{1,7,8,14,23,24} and many other species.^{40,43,45–50}

II. Computational Methods

II.A. Semiexperimental Equilibrium Structures. The derivation of r_c^{SE} structures involves three main steps: optimization of reliable r_c^{BO} structures, computation of an ab initio cubic force field with subsequent evaluation of α_i constants to extract equilibrium B_c^{SE} parameters from the experimental rotational constants, and a tight least-squares structural fit to selected B_c^{SE} values for several isotopologues, incorporating r_c^{BO} constraints as necessary. In this study, the r_c^{BO} geometries of **Ala-I** and **Ala-IIA** were fully optimized using frozen-core (FC) CCSD(T) coupled-cluster theory^{51–53} paired with the correlation-consistent cc-pVTZ basis set of [4s3p2d1f] and [3s2p1d] quality for {C, N, O} and H, respectively.⁵⁴ While the inclusion of core electron correlation during these demanding geometry optimizations was not feasible, the corresponding effects⁵⁵ on the r_c^{BO} parameters are expected to lie within the uncertainties of most r_c^{SE} parameters and are partially corrected during the least-squares fit. Geometry optimizations were carried out in natural internal coordinates^{56,57} using a quasi-Newton–Raphson method implemented in the PSI3 package.⁵⁸ The optimization

of highly flexible coordinates was facilitated with a fixed Hessian matrix evaluated at the MP2 level with a (9s5p) double- ζ valence basis set⁵⁹ (DZ) at a point near the target minimum. Energy gradients were computed by finite differences of energies provided by the MOLPRO⁶⁰ package using a five-point central difference formula to ensure numerical accuracy for both high- and low-frequency modes. Finally, minima were verified by evaluating the molecular gradients analytically using the MAB-ACESII⁶¹ program. Cartesian coordinates of the CCSD(T)/cc-pVTZ r_c^{BO} structures of **Ala-I** and **Ala-IIA** are provided in Supporting Information (Table S1).

For both conformers of Ala, anharmonic force fields were determined at the all-electron MP2/6-31G(d)⁶² level at the corresponding minima to avoid the nonzero force dilemma.⁶³ The cubic and semidiagonal quartic force constants in normal coordinates were evaluated by numerical differentiation of analytically computed second derivatives. Built-in features⁶⁴ of MAB-ACESII then gave the vibration–rotation interaction constants for all isotopologues according to the second-order vibrational perturbation theory⁶⁵ (VPT2) formula

$$\alpha_i^B = -\frac{2B_c^2}{\omega_i} \left[\sum_{\xi=a,b,c} \frac{3(a_i^{b\xi})^2}{4I_\xi} + \sum_{j(\neq i)} (\zeta_{ij}^b)^2 \frac{(3\omega_i^2 + \omega_j^2)}{\omega_i^2 - \omega_j^2} + \pi \left(\frac{c}{h}\right)^{1/2} \sum_j \phi_{ij} a_j^{bb} \left(\frac{\omega_i}{\omega_j^{3/2}}\right) \right] \quad (1)$$

in which indices (i, j) denote normal coordinates (Q_i, Q_j) with harmonic vibrational frequencies (ω_i, ω_j), I_ξ is a principal moment of inertia, $a_i^{b\xi}$ is a first derivative of inertial tensor element $I_{b\xi}$ with respect to Q_i , ζ_{ij}^b is a Coriolis coupling constant, and ϕ_{ij} is a cubic force constant in the reduced normal coordinate space. In lowest-order and without centrifugal distortion corrections, the effective ground-state rotational constants (B_0) are related to their equilibrium counterparts by the expression

$$B_e - B_0 = \frac{1}{2} \sum_i \alpha_i^B = -B_e^2 \left[\frac{3}{4} \sum_{i\xi} \frac{(a_i^{b\xi})^2}{\omega_i^2 I_\xi} - \sum_{i<j} \frac{(\zeta_{ij}^b)^2 (\omega_i - \omega_j)^2}{\omega_i \omega_j (\omega_i + \omega_j)} + \pi \left(\frac{c}{h}\right)^{1/2} \sum_{ij} \phi_{ij} a_j^{bb} \omega_j^{-3/2} \right] \quad (2)$$

which was employed in this study to obtain semiexperimental B_e^{SE} constants. Note that all Coriolis resonance terms appearing in eq 1 are canceled in the reduced form on the right side of eq 2, an important point often not fully appreciated.

The weighted least-squares refinement^{66–68} for the r_e^{SE} structures employed linear combinations of simple valence internal coordinates and was carried out with the MolStruct⁶⁹ code. The weights were chosen as the reciprocal statistical uncertainties in the experimentally derived rotational constants. For both **Ala-I** and **Ala-IIA**, experimental data is available for ten isotopologues, yielding 30 B_e^{SE} constants each (Supporting Information, Table S2). However, not all of these constants proved suitable for the r_e^{SE} refinements. Because **Ala-I** and **Ala-IIA** possess no symmetry, the number of independent geometric parameters (33) is greater than the experimental data set, necessitating the use of r_e^{BO} structural constraints. The least-squares refinements were performed on select sets of internal coordinates and rotational constants. Within least-squares fits, the standard errors intrinsic to each variable and the deviations for the rotational constants were monitored carefully.

The success of the r_e^{SE} procedure depends on the number of isotopologues with accurate experimental rotational constants that can be used to determine meaningful structural parameters, the accuracy of the anharmonic force fields and theoretical α_i constants, the quality of the r_e^{BO} least-squares constraints, and the validity of modeling vibrational effects via first-order vibration–rotation interaction (eq 2). The utility of α_i constants suffers more from the inherent approximations within VPT2 for large, flexible molecules with highly anharmonic vibrational modes than for small, rigid molecules exhibiting predominantly harmonic motions and small rovibrational couplings. By employing eq 2, higher-order vibration–rotation interactions and centrifugal distortion are neglected, despite their enhanced significance for flexible molecules. Centrifugal distortion contamination appears in both the experimental rotational constants and the theoretical correction of B_0 to extract B_e^{SE} .⁷⁰ Thus, while the effective rotational constants of certain isotopologues may describe the observables accurately, caution must be exercised in using these constants to refine the semiexperimental structure.

II.B. Conformational Energetics. The method of focal point analysis (FPA)^{39–44} provides a means of systematically approaching and monitoring convergence of ab initio computations toward the one-particle complete basis set (CBS) limit and the fully correlated many-electron wave function (full configuration interaction, FCI). In this study, an FPA investigation of the **Ala-IIA–Ala-I** relative energy was executed with correlation-consistent basis sets augmented with diffuse functions,^{54,71} aug-cc-pVXZ (X = D, T, Q, 5). Hartree–Fock (X = T, Q, 5) and MP2 (X = Q, 5) energies

were extrapolated to the CBS limit using standard exponential and inverse cubic formulas, respectively.^{72,73} Higher-order correlation effects were incorporated by means of additive CCSD/aug-cc-pVQZ and CCSD(T)/aug-cc-pVTZ increments. Core correlation was included by appending the difference between all-electron and frozen-core CCSD(T)/cc-pCVTZ results to the valence FPA limit. The frozen-core CCSD(T)/cc-pVTZ r_e^{BO} geometries were adopted as reference structures in the FPA computations.

The zero-point vibrational energies (ZPVEs) of **Ala-I** and **Ala-IIA** were first computed from the MP2/6-31G(d) anharmonic force fields via the expression

$$\text{ZPVE} = \frac{1}{2} \sum_i \omega_i + \frac{1}{4} \sum_{i \leq j} \chi_{ij} \quad (3)$$

where χ_{ij} denotes the second-order vibrational anharmonicity constants derived from VPT2.⁶⁵ The effect of anharmonicity on the ZPVE correction (Δ_{ZPVE}) to the **Ala-IIA–Ala-I** energy separation was less than 0.02 kJ mol⁻¹. Therefore, our final Δ_{ZPVE} value (+1.66 kJ mol⁻¹) was evaluated from harmonic vibrational frequencies computed at the highest feasible level of theory, all-electron MP2 with a pared aug-cc-pVTZ basis set.⁷⁴

III. Results and Discussion

III.A. Lowest-Energy Conformers of Ala. Extensive conformational searching for Gly¹⁴ and Ala,^{4,6,7} the two smallest amino acids, has revealed 8 and 13 distinct conformers, respectively. An unmistakable correspondence exists between the Gly and Ala conformers because both have inert side groups (–H for Gly, –CH₃ for Ala) leading to the same intramolecular hydrogen bonding possibilities. A bifurcated hydrogen bond forms between the carbonyl oxygen atom and the amino hydrogen atoms in the global minima **Gly-I** and **Ala-I**. Upon ~180° rotation of the –COOH plane, hydrogen bonding occurs with –OH as the proton donor and –NH₂ as the acceptor, resulting in the **Gly-IIa** and the **Ala-II** conformers. The suffix in the **Gly-IIa** designation indicates a *non*-planar backbone, although accurate FPA computations find a barrier to planarity of only 21 ± 5 cm⁻¹.²⁴ The two Ala structures corresponding to **Gly-IIa** exist as a nearly isoenergetic pair, **Ala-IIA** and **Ala-IIB**, having the same H-bonding arrangement but different orientations of the methyl group (Figure 1). **Ala-I** has repeatedly been observed as the predominant conformer in the rotational spectra of alanine,^{5,9,10} in accord with high-level theoretical results. In fact, large basis CCSD, CCSD(T), and MP4 single-point energy computations at MP2/6-311++G** optimum geometries determine **Ala-I** to be more stable than the **Ala-II** conformers by 200–300 cm⁻¹.⁷ The same levels of theory predict that the next conformers (**Ala-III**) are, again, 200–300 cm⁻¹ higher in energy than the **Ala-II** conformers. While the **Ala-I** and **Ala-IIA** conformers were identified^{5,9} in the observed rotational spectra by ¹⁴N nuclear quadrupole coupling, **Ala-IIB** and higher-energy conformers were never observed.

Prior speculation on the absence of **Ala-IIB** in the observed rotational spectra was based on a presumably low

interconversion barrier for **Ala-IIB** \rightarrow **Ala-IIA**. To elucidate this issue, we computed CCSD(T) energy points with the aug-cc-pVTZ basis set at the MP2/6-311++G(d,p) stationary structures of **Ala-IIA**, **Ala-IIB**, and the connecting transition state optimized in this work. The resulting well depth of **Ala-IIB** with respect to the interconversion barrier is only 34 cm^{-1} , which is reduced to a minuscule 16 cm^{-1} upon vibrational correction. To obtain this vibrationally adiabatic barrier, MP2/cc-pVTZ harmonic frequencies were computed, and ZPVEs were evaluated by excluding at each stationary point the contribution from the normal mode connecting **Ala-IIA** to **Ala-IIB**. In summary, the small amount of energy required to interconvert the **Ala-II** conformers is indeed representative of the conformational flexibility of Ala and may rationalize the absence of **Ala-IIB** in the molecular beam experiments.^{20,75,76}

III.B. Refitting the Rotational Spectra of Ala. Before determining r_c^{SE} structures, we refit the existing rotational spectra of alanine to more firmly establish the rotational constants and their uncertainties for the structural analysis. In the original spectroscopic study,⁵ the rotational, centrifugal distortion, and nuclear quadrupole hyperfine constants of alanine were simultaneously determined from a global fit of a chosen Hamiltonian to the measured transitions. From a statistical point of view, this method is meritorious and has the advantage of being simple. However, overly optimistic uncertainties are produced when the data set for the global fit is small, as is the case here. Moreover, from a numerical perspective, correlations are induced between the centrifugal distortion constants and the remaining rotational parameters, worsening the condition number.⁷⁷ Finally, “masked” errors that do not yield outlying residuals become more prevalent.⁷⁸ For these reasons, we first corrected the transitions for the nuclear quadrupole hyperfine structure and then fit the hypothetical unperturbed rotational transitions to a standard Watson Hamiltonian.⁷⁹ It could be argued that this approach might give biased rotational parameters containing systematic errors due to inaccuracies in the nuclear quadrupole hyperfine constants. However, when several hyperfine components of the same rotational transition are measured, as for a great majority of the reported transitions,⁵ the hypothetical, unperturbed frequencies may be calculated using the intensity-weighted mean of the multiplets.⁸⁰ As a consequence, accurate knowledge of the nuclear quadrupole hyperfine constants is unnecessary and the possible contribution of the spin-rotation interaction is canceled. Furthermore, our approach permits the elimination of outliers, the estimation of the uncertainty of the measurements, and an increase in the reliability of the rotational frequencies.

Another issue in the original fits⁵ is that the full set of quartic centrifugal distortion constants was not determinable for many isotopologues, and hence these constants were fixed to values for the parent (or ^{15}N) species. Our computations revealed significant variations in the centrifugal distortion constants from one isotopologue to another (*vide infra*). Therefore, we used the method of predicate observations⁸¹ in our refitting, in which the *ab initio* “scaled” centrifugal distortion constants (or the constants of another isotopologue) are input data in a weighted least-squares fit. Though this

method permits the approximate determination of the centrifugal distortion constants, it must be used with care, and it is essential to check that the derived constants are really compatible with the experimental data. In our fits the weights of the predicate observations were varied to keep the corresponding “jackknifed” residuals, $t(i)$, small (typically less than 3), where $t(i)$ is the i th residual divided by its standard deviation calculated by omitting the i th transition.⁷⁸

Parent values and isotopic shifts for the effective rotational constants and quartic centrifugal distortion constants of the **Ala-I** and **Ala-IIA** isotopologues are reported in Tables 1 and 2, respectively. Three sets of data are tabulated: our results from refitting the observed lines (refit), our CCSD(T)/cc-pVTZ theoretical values (theor), and the original experimental constants (expt).⁵ Rotational constant shifts associated with heavy-atom (non-hydrogen) isotopic substitution exhibit modest differences (2–5 kHz) between original and refit values and are relatively independent of the method used to fit the rotational spectra. However, much larger deviations are found between the original and refit values for many of the D-substituted isotopologues. The largest discrepancies (in kHz) are 868 for B_0 of O–D (**Ala-I**), 476 for A_0 of $\text{C}_m\text{-3D}$ (**Ala-IIA**), 84 for A_0 of N–D_a (**Ala-I**), 47 for A_0 of N–D_a (**Ala-IIA**), and 46 for B_0 of $\text{C}_m\text{-3D}$ (**Ala-IIA**). Comparing these discrepancies to the average residual of the structural fits for Ala and Gly, around 20 kHz (ref 24 and below), it becomes clear that these rotational constants should not be given much weight in the determination of r_c^{SE} structures.

The theoretical isotopic shifts (Tables 1 and 2) are based on (A_0, B_0, C_0) constants, which are determined by conjoining our MP2/6-31G(d) vibration–rotation interaction constants and CCSD(T)/cc-pVTZ equilibrium rotational constants (B_e^{BO}). The theoretical and experimental heavy-atom isotopic shifts of the (A_0, B_0, C_0) constants are generally in remarkable agreement. The mean absolute deviations between refit and theor isotopic shifts among the (^{13}C , $^{13}\text{C}_\alpha$, $^{13}\text{C}_m$, ^{15}N) isotopologues are 0.2 and 0.4 MHz in the **Ala-I** and **Ala-IIA** cases, respectively. On the other hand, most isotopic-shift disparities for the deuterated isotopologues are greater than 1 MHz. The proximity of the D_b position to the methyl group seems to enhance the error in the vibrationally corrected rotational constants of the N–D_b isotopologue, especially in comparison to N–D_a. The two largest absolute discrepancies are (8.7, 3.5) MHz for [A_0 (**Ala-IIA**), B_0 (**Ala-I**)] of the triply deuterated methyl isotopologues, $\text{C}_m\text{-3D}$. Nevertheless, on a percentage basis, the discord between the refit and theor isotopic shifts is less than 2% even in these instances. For the centrifugal distortion constants, the refit and theor isotopic shifts agree quite well for **Ala-I**, similarly to other molecules.^{82,83} However, considerable differences occur for the Δ_{JK} , Δ_K , and δ_K isotopic shifts of **Ala-IIA**. The underlying cause is not transparent and is not specific to the deuterated isotopologues.

Relevant to the structure refinements, the number of fitted transitions for the ^{15}N isotopologues of **Ala-I** and **Ala-IIA** is relatively small. This is particularly true in the **Ala-IIA** case, where 17 lines were used to determine 8 parameters (3 rotational and 5 quartic centrifugal distortion constants).

Table 1. Isotopic Shifts of Effective Rotational Constants (A_0 , B_0 , C_0) and A-Reduced Quartic Centrifugal Distortion Constants (Δ_J , Δ_{JK} , Δ_K , δ_J , δ_K) for **Ala-I**: Original Experimental Constants from Ref 5 (expt), Current Refitting of Observed Lines (refit), and CCSD(T)/cc-pVTZ Theoretical Values (theor)^a

constant	parent	¹³ C	¹³ C _α	¹³ C _m	¹⁵ N	C _α -D	C _m -3D	O-D	N-D _a	N-D _b	MAD	
A ₀ expt	5066.1456(4)	-0.941	-8.821	-95.254	-48.851	-113.253	-462.026	-13.968	-104.850	-176.847		
	refit	5066.1455(7)	-0.941	-8.821	-95.254	-48.851	-113.250	-462.025	-13.965	-104.934	0.777	
	theor	5031.4685	-0.939	-8.995	-94.714	-48.400	-112.952	-459.265	-12.775	-104.774	-175.431	
B ₀ expt	3100.9506(3)	-9.668	-12.419	-33.767	-50.457	-51.813	-175.397	-109.991	-57.095	-49.805		
	refit	3100.9507(5)	-9.667	-12.421	-33.767	-50.458	-51.818	-175.404	-110.859	-57.103	-49.804	1.347
	theor	3067.4442	-9.505	-12.209	-33.324	-50.093	-50.240	-171.944	-109.247	-54.348	-48.269	
C ₀ expt	2264.0134(2)	-5.178	-5.343	-31.628	-34.806	-5.462	-141.514	-61.745	-41.615	-45.195		
	refit	2264.0131(4)	-5.178	-5.342	-31.628	-34.806	-5.458	-141.510	-61.749	-41.608	-45.197	0.307
	theor	2258.8416	-5.199	-5.370	-31.518	-34.921	-6.150	-141.609	-61.727	-40.245	-44.884	
Δ _J expt	2.452	-0.035	-0.029	-0.069	-0.121	-0.026	0	0	0	0		
	refit	2.445(7)	-0.024	-0.086	-0.072	-0.113	-0.138	-0.312	-0.262	-0.281	0.007	0.033
	theor	2.409	-0.027	-0.040	-0.042	-0.100	-0.195	-0.325	-0.245	-0.262	0.108	
Δ _{JK} expt	-6.391	0.052	0	0.324	0.112	0	0	0	0	0		
	refit	-6.38(1)	0.03	0.45	0.29	0.11	0.74	1.61	0.69	0.77	-0.74	0.08
	theor	-6.373	0.104	0.127	0.292	0.126	0.590	1.613	0.634	0.694	-0.741	
Δ _K expt	5.410	0.022	0	-0.21	0.124	0	0	0	0	0		
	refit	5.37(5)	0.07	-0.09	-0.18	0.15	-0.69	-1.55	-0.28	-0.36	0.53	0.08
	theor	5.424	-0.077	-0.114	-0.298	-0.032	-0.568	-1.570	-0.324	-0.413	0.478	
δ _J expt	0.5696	0.0013	-0.0216	-0.0109	-0.0294	0	0	0	0	0		
	refit	0.574(2)	0.008	-0.028	-0.010	-0.033	-0.061	-0.094	-0.066	-0.084	0.058	0.009
	theor	0.570	-0.009	-0.014	-0.011	-0.023	-0.070	-0.095	-0.053	-0.077	0.064	
δ _K expt	10.3777	0.0083	0.2823	-0.4577	-0.3647	0.5423	0	0	0	0		
	refit	10.37(3)	0.07	-0.31	-0.36	-0.34	-0.71	-1.80	-0.11	-1.36	0.22	0.13
	theor	9.656	-0.025	-0.058	-0.297	-0.273	-0.775	-1.734	-0.303	-1.442	-0.042	

^a Units: MHz for (A_0 , B_0 , C_0) and kHz for (Δ_J , Δ_{JK} , Δ_K , δ_J , δ_K); MAD = mean absolute deviation between refit and theor isotopic shifts. Large deviations of theoretical and experimental rotational constants are italicized. The CCSD(T)/cc-pVTZ rotational constants include MP2/6-31G(d) vibrational corrections.

Table 2. Isotopic Shifts of Effective Rotational Constants (A_0 , B_0 , C_0) and A-Reduced Quartic Centrifugal Distortion Constants (Δ_J , Δ_{JK} , Δ_K , δ_J , δ_K) for **Ala-IIA**: Original Experimental Constants from Ref 5 (expt), Current Refitting of Observed Lines (refit), and CCSD(T)/cc-pVTZ Theoretical Values (theor)^a

constant	parent	¹³ C	¹³ C _α	¹³ C _m	¹⁵ N	C _α -D	C _m -3D	O-D	N-D _a	N-D _b	MAD	
A ₀ expt	4973.0558(6)	-0.138	-10.402	-88.351	-54.385	-115.510	-441.476	-138.202	-35.075	-164.893		
	refit	4973.0546(35)	-0.136	-10.399	-88.356	-54.384	-115.537	-441.952	-138.198	-35.028	-164.887	2.566
	theor	4950.175	-0.177	-11.079	-88.826	-53.339	-117.566	-433.218	-134.674	-31.794	-168.219	
B ₀ expt	3228.3379(5)	-12.458	-13.622	-39.366	-42.997	-54.464	-198.172	-8.122	-114.542	-77.573		
	refit	3228.3375(56)	-12.456	-13.622	-39.366	-42.997	-54.462	-198.218	-8.124	-114.543	-77.576	1.910
	theor	3183.801	-12.216	-13.218	-38.315	-42.527	-51.598	-195.056	-7.333	-113.036	-70.879	
C ₀ expt	2307.8090(3)	-6.190	-5.382	-33.334	-31.976	-8.888	-148.755	-27.187	-61.942	-44.052		
	refit	2307.8090(42)	-6.191	-5.382	-33.333	-31.977	-8.890	-148.704	-27.182	-61.935	-44.047	0.858
	theor	2316.254	-6.326	-5.539	-33.356	-32.264	-10.313	-148.500	-26.600	-65.042	-45.800	
Δ _J expt	2.13(1)	0.038	-0.036	0.036	-0.004	-0.055	-0.004	-0.004	-0.004	-0.004		
	refit	2.11(6)	0.10	-0.02	0.09	0.01	-0.06	-0.06	-0.01	0.16	-0.01	0.09
	theor	1.397	-0.016	-0.022	-0.016	-0.055	-0.107	-0.152	0.016	-0.149	-0.086	
Δ _{JK} expt	-4.84(6)	-0.175	0.207	0.199	-0.358	-0.358	-0.358	-0.358	-0.358	-0.358		
	refit	-4.8(3)	-0.37	0.13	-0.03	-0.43	-0.18	-0.18	-0.27	-0.17	-0.25	0.37
	theor	-2.611	0.047	0.055	0.067	0.098	0.232	0.459	-0.100	0.306	0.279	
Δ _K expt	4.98(2)	0	0	0	0	-0.002	-0.002	-0.002	-0.002	-0.002		
	refit	4.6(5)	0.67	0.44	-0.63	0.43	0.42	0.42	0.44	0.43	0.43	0.64
	theor	2.772	-0.032	-0.069	-0.092	-0.058	-0.328	-0.574	0.009	-0.124	-0.390	
δ _J expt	0.41(1)	-0.010	-0.025	-0.015	-0.0087	-0.009	-0.009	-0.009	-0.009	-0.009		
	refit	0.41(2)	-0.01	-0.03	-0.04	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	0.02
	theor	0.257	-0.004	-0.006	-0.0002	-0.012	-0.033	-0.025	0.007	-0.023	-0.047	
δ _K expt	7.35(1)	0	0	0	0	0	0	0	0	0		
	refit	7.2(7)	0.59	0.11	0.26	0.15	0.97	0.97	-1.79	-1.68	-1.78	1.02
	theor	4.686	-0.008	-0.026	-0.172	-0.127	-0.388	-0.940	-0.320	-0.099	-0.317	

^a Units: MHz for (A_0 , B_0 , C_0) and kHz for (Δ_J , Δ_{JK} , Δ_K , δ_J , δ_K); MAD = mean absolute deviation between refit and theor isotopic shifts. Large deviations of theoretical and experimental rotational constants are italicized. The CCSD(T)/cc-pVTZ rotational constants include MP2/6-31G(d) vibrational corrections.

Accordingly, the standard deviation of the ¹⁵N(**Ala-IIA**) fit is only 0.2 kHz, less than 10% of the estimated experimental accuracy (3 kHz). For this reason, the standard deviations of the ¹⁵N parameters are perhaps one order of magnitude too small.

Highly anharmonic vibrational motions, such as the internal rotation of the methyl group, twisting along the backbone, or fluid rocking motion of the amino group, complicate the determination of vibrational corrections to

the effective experimental rotational constants. Fortunately, for isotopologues that exhibit similar vibrational effects as the parent, the error in the vibrational corrections is systematic and the least-squares refinement can still produce equilibrium structures with small standard errors. The statistical outliers (among the rotational constants) stem from isotopologues for which the substituted atom undergoes large-amplitude, anharmonic motion or yields large isotopic shifts. As such, isotopic substitutions at peripheral hydrogen

Table 3. Equilibrium Structures of **Ala-I**^a

parameters ^b	semiexperimental r_e			
	r_e^{BO} CCSD(T)/cc-pVTZ	Fit 1	Fit 2	Fit 3
$r(\text{C}-\text{C}_\alpha)$	1.5236	1.519(2)	1.518(2)	1.520(3)
$r(\text{C}_\alpha-\text{C}_m)$	1.5316	1.524(4)	1.526(3)	1.522(4)
$r(\text{C}_\alpha-\text{N})$	1.4570	1.446(4)	1.445(3)	1.448(4)
$r(\text{C}=\text{O})$	1.2085	1.208(6)	1.208(5)	1.207(7)
$r(\text{C}-\text{O})$	1.3551	1.347(5)	1.348(4)	1.349(6)
$r(\text{N}-\text{H})_{\text{Avg}}$	1.0156	1.0156	1.0156	1.014(4)
$\angle(\text{C}-\text{C}_\alpha-\text{C}_m)$	108.70	109.0(2)	108.8(2)	109.0(3)
$\angle(\text{C}-\text{C}_\alpha-\text{N})$	113.24	113.4(3)	113.5(3)	113.3(4)
$\angle(\text{C}_\alpha-\text{C}=\text{O})$	125.36	124.8(4)	126.1(4)	125.1(4)
$\angle(\text{O}-\text{C}=\text{O})$	122.81	122.7(1)	124.9(3)	122.7(2)
$\angle(\text{C}-\text{O}-\text{H})$	105.82	105.82	105.82	105.9(6)
$\angle(\text{C}-\text{C}_\alpha-\text{H})$	107.39	107.39	107.39	106.5(5)
$\tau(\text{O}-\text{C}-\text{C}_\alpha-\text{C}_m)$	-73.65	-71.9(3)	-71.8(3)	-71.9(4)
$\tau(\text{O}-\text{C}-\text{C}_\alpha-\text{H})$	43.98	43.98	43.98	47.5(9)
$\tau(\text{O}=\text{C}-\text{C}_\alpha-\text{N})$	-17.27	-16.2(5)	-16.1(4)	-16.2(8)

CCSD(T)/cc-pVTZ r_e^{BO} constraints			
$r(\text{O}-\text{H})$	0.9680	$\angle(\text{C}_\alpha-\text{C}_m-\text{H})_{\text{Diff1}}$	-0.54
$r(\text{C}_\alpha-\text{H})$	1.0925	$\angle(\text{C}_\alpha-\text{C}_m-\text{H})_{\text{Diff2}}$	2.02
$r(\text{N}-\text{H})_{\text{Diff}}$	0.0012	$\tau(\text{O}=\text{C}-\text{O}-\text{H})$	-0.90
$r(\text{C}_m-\text{H})_{\text{Avg}}$	1.0911	$\tau(\text{C}_\alpha-\text{C}-\text{O}-\text{H})$	177.80
$r(\text{C}_m-\text{H})_{\text{Diff1}}$	0.0075	$\tau(\text{C}-\text{C}_\alpha-\text{N}-\text{H}_a)$	54.01
$r(\text{C}_m-\text{H})_{\text{Diff2}}$	0.0007	$\tau(\text{C}-\text{C}_\alpha-\text{N}-\text{H}_b)$	-59.37
$\angle(\text{C}_\alpha-\text{N}-\text{H})_{\text{Avg}}$	108.66	$\tau(\text{H}-\text{C}_\alpha-\text{C}_m-\text{H}_1)$	180.86
$\angle(\text{H}_a-\text{N}-\text{H}_b)$	104.72	$\tau(\text{H}-\text{C}_\alpha-\text{C}_m-\text{H}_2)$	-58.90
$\angle(\text{C}_\alpha-\text{C}_m-\text{H})_{\text{Avg}}$	110.08	$\tau(\text{H}-\text{C}_\alpha-\text{C}_m-\text{H}_3)$	62.11

coordinate definitions	
$r(\text{N}-\text{H})_{\text{Avg}} = [r(\text{N}-\text{H}_a) + r(\text{N}-\text{H}_b)]/2$	
$r(\text{N}-\text{H})_{\text{Diff}} = [r(\text{N}-\text{H}_a) - r(\text{N}-\text{H}_b)]$	
$\angle(\text{C}_\alpha-\text{N}-\text{H})_{\text{Avg}} = [\angle(\text{C}_\alpha-\text{N}-\text{H}_a) + \angle(\text{C}_\alpha-\text{N}-\text{H}_b)]/2$	
$r(\text{C}_m-\text{H})_{\text{Avg}} = [r(\text{C}_m-\text{H}_1) + r(\text{C}_m-\text{H}_2) + r(\text{C}_m-\text{H}_3)]/3$	
$r(\text{C}_m-\text{H})_{\text{Diff1}} = 2[r(\text{C}_m-\text{H}_1)] - r(\text{C}_m-\text{H}_2) - r(\text{C}_m-\text{H}_3)$	
$r(\text{C}_m-\text{H})_{\text{Diff2}} = r(\text{C}_m-\text{H}_2) - r(\text{C}_m-\text{H}_3)$	
$\angle(\text{C}_\alpha-\text{C}_m-\text{H})_{\text{Avg}} = [\angle(\text{C}_\alpha-\text{C}_m-\text{H}_1) + \angle(\text{C}_\alpha-\text{C}_m-\text{H}_2) + \angle(\text{C}_\alpha-\text{C}_m-\text{H}_3)]/3$	
$\angle(\text{C}_\alpha-\text{C}_m-\text{H})_{\text{Diff1}} = 2[\angle(\text{C}_\alpha-\text{C}_m-\text{H}_1)] - \angle(\text{C}_\alpha-\text{C}_m-\text{H}_2) - \angle(\text{C}_\alpha-\text{C}_m-\text{H}_3)$	
$\angle(\text{C}_\alpha-\text{C}_m-\text{H})_{\text{Diff2}} = \angle(\text{C}_\alpha-\text{C}_m-\text{H}_2) - \angle(\text{C}_\alpha-\text{C}_m-\text{H}_3)$	

^a Distances in Å, angles in deg. Boldface denotes parameters included in the least-squares fits. Note that in Fits 1 and 2, all hydrogen atoms were fully positioned by the constraints, whereas Fit 3 provided four internal coordinates involving hydrogen atoms. ^b Refer to Figure 1 for atom labels.

atoms may be difficult to fit and should be treated judiciously to avoid vibrational contamination of r_e^{SE} structures.

III.C. r_e^{SE} Structure of Ala-I. The semiexperimental structures of **Ala-I** resulting from three different least-squares refinements, labeled Fit 1 through Fit 3, are reported in Table 3, along with the associated frozen-core CCSD(T)/cc-pVTZ r_e^{BO} parameters/constraints. The values refined in each fit are highlighted in boldface type with standard errors in parentheses, whereas all other parameters necessary to define the molecular structure were constrained to the CCSD(T)/cc-pVTZ values listed in normal type. In Fit 1 only the rotational constants of the parent and isotopologues involving heavy-atom substitution were used; accordingly, all hydrogen atoms were fully positioned by the r_e^{BO} constraints. By omitting the deuterated species, the errors in the B_e^{SE} data arising from large-amplitude vibrational effects are reduced and become more systematic. In Fit 1, the weighted root-mean-square (rms) residual of the 15 chosen rotational constants is only 16 kHz, and the standard errors of the fit for bond distances and angles are no greater than 0.007 Å and 0.6°, respectively. Clearly, the rotational constants of the parent, the ¹⁵N, and

the three unique ¹³C isotopologues provide enough information to determine the positions of all heavy atoms in **Ala-I**. Nonetheless, even in the highly constrained Fit 1 some of the optimized parameters are strongly correlated, hindering their explicit determination. The introduction of further constraints, for example fixing either $r(\text{C}-\text{O})$ or $r(\text{C}=\text{O})$, gave similar results. We note that the standard deviations of the r_e^{SE} (Fit 1) parameters are underestimations because the uncertainty of the several fixed parameters is not taken into account.

Fit 2 employed the same structural variables and constraints as Fit 1 but added selected rotational constants from the deuterated isotopologues to the B_e^{SE} data set. In particular, Fit 2 included $A_e^{\text{SE}}(\text{C}_\alpha-\text{D})$, $B_e^{\text{SE}}(\text{C}_m-3\text{D})$, $B_e^{\text{SE}}(\text{O}-\text{D})$, $C_e^{\text{SE}}(\text{O}-\text{D})$, $A_e^{\text{SE}}(\text{N}-\text{D}_b)$, and $C_e^{\text{SE}}(\text{N}-\text{D}_b)$, all of which had a residual <0.5 MHz in Fit 1. Fit 2 reduces the standard error of each structural parameter (Table 3) vis-à-vis Fit 1, while maintaining a reasonably small residual (24 kHz).

Fit 3 incorporated all of the observed rotational constants except those for the C_m-3D and $\text{N}-\text{D}_b$ isotopologues; in addition, $B_e^{\text{SE}}(\text{N}-\text{D}_a)$ was excluded and the weight of

$r_c^{\text{SE}}(\text{N}-\text{D}_a)$ was decreased, making the $\text{N}-\text{D}_a$ isotopologue less influential in the fit. Problems with nonsystematic errors necessitating exclusion of rotational constant data for the C_m-3D , $\text{N}-\text{D}_b$, and $\text{N}-\text{D}_a$ isotopologues were identified in section III.B above. The expanded data set for Fit 3 allowed some hydrogen-atom coordinates to be refined, the tightest fit to the data (weighted rms of 25 kHz) being obtained by releasing $r(\text{N}-\text{H})_{\text{Avg}}$, $\angle(\text{C}-\text{C}_\alpha-\text{H})$, $\angle(\text{C}-\text{O}-\text{H})$, and $\tau(\text{O}-\text{C}-\text{C}_\alpha-\text{H})$. Additional parameters could not be refined without introducing large deviations in both hydrogen- and heavy-atom positions. In this regard the structures of the methyl and carboxyl groups are under-determined by the experimental data, due to the lack of isotopic substitution on the oxygen atoms and on individual methyl hydrogen atoms. In summary, Fit 3 provides the best currently possible r_c^{SE} structure of **Ala-I** by refining 15 of the 33 geometric degrees of freedom on 23 vibrationally corrected, semiexperimental equilibrium rotational constants.

III.D. r_c^{SE} Structure of Ala-IIA. One satisfactory fit was achieved for the semiexperimental structure of **Ala-IIA**. Initially, incorporating only rotational constants of heavy-atom isotopologues and structural parameters involving heavy-atom positions, as in Fit 1 of **Ala-I**, resulted in an r_c^{SE} structure that had surprisingly large standard errors and poor agreement with the CCSD(T)/cc-pVTZ r_c^{BO} parameters. Most notably, the semiexperimental C–O distance had a standard error of 0.02 Å and was 0.04 Å shorter than the CCSD(T)/cc-pVTZ value, while the fit to the ^{15}N data was poor. As mentioned above, the ^{15}N assignments and the fitted data appear to be correct, but the originally reported and refitted uncertainties are probably too optimistic. Therefore, the structure of **Ala-IIA** was determined again after increasing the experimental uncertainties (reciprocal weights in the least-squares fit) of the three ^{15}N rotational constants by a factor of 20. This modification, labeled Fit 1', was validated by an approximate one-half reduction in the standard errors of the structural parameters. The C–O bond distance displayed an error of ± 0.01 Å and became a much more reasonable 0.01 Å shorter than the r_c^{BO} value. The final structural parameters of **Ala-IIA** refined in Fit 1' are reported in boldface in Table 4, along with the CCSD(T)/cc-pVTZ constraints invoked. While the success of the **Ala-IIA** fit is gratifying, the statistical errors are significantly larger than observed for Fit 1 of **Ala-I**.

Attempts to incorporate rotational constants of the deuterated isotopologues in the structural refinement of **Ala-IIA** did not reduce the statistical errors. Including the only two rotational constants that had residuals under 0.5 MHz in Fit 1' did not improve the heavy-atom structural parameters, unlike Fit 2 of **Ala-I**. Fitting the data for the $\text{C}_\alpha-3\text{D}$ and $\text{N}-\text{D}_a$ isotopologues once again proved problematic. Adding the O–D rotational constants and releasing the $\angle(\text{C}-\text{O}-\text{H})$ parameter distorted the **Ala-IIA** structure considerably, causing the carbonyl bond distance to deviate from the r_c^{BO} value by an unacceptable 0.1 Å. Other parameters for the hydroxyl hydrogen atom were released with similar or more pronounced distortion of the overall structure. Therefore, the B_c^{SE} data for the O–D isotopologue are disappointingly unable to yield the structure within the

Table 4. Equilibrium Structures of **Ala-IIA**^a

parameters ^b	CCSD(T)/cc-pVTZ	semiexperimental r_e (Fit 1')	
$r(\text{C}-\text{C}_\alpha)$	1.5347	1.530(3)	
$r(\text{C}_\alpha-\text{C}_m)$	1.5282	1.529(4)	
$r(\text{C}_\alpha-\text{N})$	1.4726	1.460(4)	
$r(\text{C}=\text{O})$	1.2052	1.205(9)	
$r(\text{C}-\text{O})$	1.3431	1.33(1)	
$\angle(\text{C}-\text{C}_\alpha-\text{C}_m)$	108.08	108.2(3)	
$\angle(\text{C}-\text{C}_\alpha-\text{N})$	109.44	109.8(4)	
$\angle(\text{C}_\alpha-\text{C}=\text{O})$	122.66	122.0(8)	
$\angle(\text{O}-\text{C}=\text{O})$	123.27	123.2(4)	
$\tau(\text{O}-\text{C}-\text{C}_\alpha-\text{C}_m)$	257.77	254.5(4)	
$\tau(\text{O}=\text{C}-\text{C}_\alpha-\text{N})$	195.28	192.5(5)	

CCSD(T)/cc-pVTZ r_e^{BO} constraints			
$r(\text{O}-\text{H})$	0.9789	$\angle(\text{C}_\alpha-\text{C}_m-\text{H})_{\text{Avg}}$	110.18
$r(\text{C}_\alpha-\text{H})$	1.0926	$\angle(\text{C}_\alpha-\text{C}_m-\text{H})_{\text{Diff1}}$	-0.60
$r(\text{N}-\text{H})_{\text{Avg}}$	1.0132	$\angle(\text{C}_\alpha-\text{C}_m-\text{H})_{\text{Diff2}}$	-0.33
$r(\text{N}-\text{H})_{\text{Diff}}$	0.0009	$\tau(\text{O}-\text{C}-\text{C}_\alpha-\text{H})$	140.31
$r(\text{C}_m-\text{H})_{\text{Avg}}$	1.0915	$\tau(\text{O}=\text{C}-\text{O}-\text{H})$	178.18
$r(\text{C}_m-\text{H})_{\text{Diff1}}$	-0.0067	$\tau(\text{C}_\alpha-\text{C}-\text{O}-\text{H})$	-4.10
$r(\text{C}_m-\text{H})_{\text{Diff2}}$	0.0005	$\tau(\text{C}-\text{C}_\alpha-\text{N}-\text{H}_a)$	89.86
$\angle(\text{C}-\text{O}-\text{H})$	104.26	$\tau(\text{C}-\text{C}_\alpha-\text{N}-\text{H}_b)$	208.10
$\angle(\text{C}-\text{C}_\alpha-\text{H})$	106.67	$\tau(\text{H}-\text{C}_\alpha-\text{C}_m-\text{H}_1)$	60.01
$\angle(\text{H}_a-\text{N}-\text{H}_b)$	106.90	$\tau(\text{H}-\text{C}_\alpha-\text{C}_m-\text{H}_2)$	179.80
$\angle(\text{C}-\text{N}-\text{H})_{\text{Avg}}$	110.61	$\tau(\text{H}-\text{C}_\alpha-\text{C}_m-\text{H}_3)$	-60.14

^a Distances in Å, angles in deg. Boldface denotes parameters included in the least-squares fits. Note that in Fit 1' all hydrogen atoms were fully positioned by the constraints. ^b Refer to Figure 1 for atom labels and Table 3 for coordinate definitions.

strong $\text{OH}\cdots\text{N}$ hydrogen bond. In summary, only the experimental rotational constants of the heavy-atom isotopologues yield useful information, and thus the r_c^{SE} structure of **Ala-IIA** is considerably less well determined than that of **Ala-I**.

III.E. Discussion of the Ala Structures. A comparison of prior experimental r_g , r_α , r_z , r_0 , and r_s parameters with the current r_c^{SE} and r_c^{BO} results is made in Tables 5 and 6 for **Ala-I** and **Ala-IIA**, respectively. Considerable vibrational effects are present in all previous experimental structures, and several structural parameters exhibit disturbing differences. The disparities are more prominent for the bond distances than for the bond angles. Our equilibrium r_c^{SE} and r_c^{BO} results allow unphysical or misleading values to be identified among the vibrationally averaged parameters. The most important defects are $r_g(\text{C}_\alpha-\text{C}_m) = 1.509(16)$ Å and $r_s(\text{C}_\alpha-\text{C}_m) = 1.57(1)$ Å compared to $r_c^{\text{SE}}(\text{C}_\alpha-\text{C}_m) = 1.522(4)$ Å for **Ala-I**; $r_0(\text{C}-\text{O}) = 1.37(2)$ Å compared to $r_c^{\text{SE}}(\text{C}-\text{O}) = 1.33(1)$ Å for **Ala-IIA**; and $r_s(\text{C}_\alpha-\text{N}) = 1.430(9)$ Å compared to $r_c^{\text{SE}}(\text{C}_\alpha-\text{N}) = 1.460(4)$ Å for **Ala-IIA**. Excessive deviations from the r_c^{SE} and r_c^{BO} values and underestimated experimental uncertainties are exhibited in several cases, such as $r_s(\text{C}_\alpha-\text{C}_m)$ of **Ala-IIA**, while anomalous vibrationally averaged distances smaller than the corresponding equilibrium bond length occur in other instances such as $r_z(\text{C}-\text{O})$ of **Ala-I**.

Several systematic studies^{84–87} have established the expected accuracy of CCSD(T)/cc-pVTZ geometric parameters, allowing a reliable assessment of our r_c^{BO} and r_c^{SE} structures of alanine. For 19 small (H, C, N, O, F) molecules, all-electron CCSD(T)/cc-pVTZ equilibrium bond distances have a mean error (std. dev.) of +0.0002 (0.0023) Å, whereas

Table 5. Selected **Ala-I** Structural Parameters (Å and deg) from Different Methodologies

	r_g/r_α ref 11 ^a	r_z ref 12	r_0 ref 5	r_s ref 5	r_e^{SE} Fit 3	r_e^{BO} CCSD(T)/cc-VTZ
$r(\text{C}-\text{C}_\alpha)$	1.544(10)	1.527(11)	1.51(1)	1.48(1)	1.520(3)	1.5236
$r(\text{C}_\alpha-\text{C}_m)$	1.509(16)	1.536(11)	1.53(2)	1.57(1)	1.522(4)	1.5316
$r(\text{C}_\alpha-\text{N})$	1.471(7)	1.453(2)	1.45(1)	1.438(9)	1.448(4)	1.4570
$r(\text{C}=\text{O})$	1.192(2)	1.197(1)	1.24(2)		1.207(7)	1.2085
$r(\text{C}-\text{O})$	1.347(3)	1.341(2)	1.33(2)		1.349(6)	1.3551
$\angle(\text{C}-\text{C}_\alpha-\text{C}_m)$	111.6(11)	111.9(2)	108.3(6)	109(1)	109.0(3)	108.70
$\angle(\text{C}-\text{C}_\alpha-\text{N})$	110.1(8)	112.9(3)	115(1)	117(1)	113.3(4)	113.24
$\angle(\text{C}_\alpha-\text{C}=\text{O})$	125.6(7)	125.7(3)	125(2)		125.1(4)	125.36
$\angle(\text{C}_\alpha-\text{C}-\text{O})$	110.3(7)	110.3(2)	113(2)			111.82
$\tau(\text{O}=\text{C}-\text{C}_\alpha-\text{N})$	-17.2(18)	-16.6(4)			-16.2(8)	-17.27

^a r_g for distances, r_α for angles.

Table 6. Selected **Ala-IIA** Structural Parameters (Å and deg) from Different Methodologies

	r_0 ref 5	r_s ref 5	r_e^{SE} Fit 1'	r_e^{BO} CCSD(T)/cc-VTZ
$r(\text{C}-\text{C}_\alpha)$	1.524(7)	1.517(7)	1.530(3)	1.5347
$r(\text{C}_\alpha-\text{C}_m)$	1.543(8)	1.571(9)	1.529(4)	1.5282
$r(\text{C}_\alpha-\text{N})$	1.458(9)	1.430(9)	1.460(4)	1.4726
$r(\text{C}=\text{O})$	1.20(2)		1.205(9)	1.2052
$r(\text{C}-\text{O})$	1.37(2)		1.33(1)	1.3431
$\angle(\text{C}-\text{C}_\alpha-\text{C}_m)$	107.1(3)	107.6(8)	108.2(3)	108.08
$\angle(\text{C}-\text{C}_\alpha-\text{N})$	111.7(7)	111.8(7)	109.8(4)	109.44
$\angle(\text{C}_\alpha-\text{C}=\text{O})$	125(1)		122.0(8)	122.66
$\angle(\text{C}_\alpha-\text{C}-\text{O})$	113(2)			114.02
$\tau(\text{O}=\text{C}-\text{C}_\alpha-\text{N})$	167(1)		192.5(5)	195.28

bond angles have a mean absolute error (MAE) of about 0.5° .⁸⁵ A very favorable cancellation of basis set incompleteness and electron correlation errors is responsible for such high accuracy. Statistics are not available for dihedral angles, but a larger MAE of perhaps $1-2^\circ$ is probable. Because 1s electron correlation contracts bond lengths in first-row diatomics by $0.0005-0.0025$ Å,^{55,88} frozen-core CCSD(T)/cc-pVTZ r_e^{BO} distances are expected to be too large by at least $0.001-0.003$ Å. Therefore, the general $r_e^{\text{BO}} > r_e^{\text{SE}}$ trend for bond distances in Tables 3 and 4 is nicely explained.

An investigation of 18 small, rigid molecules⁸⁹ showed that the MAE in the relative magnitude of the sum of theoretical α_i constants, $\sum_i \alpha_i^{\text{B}}/B_0$, was only 0.225% at the MP2/cc-pVDZ level of theory with respect to CCSD(T)/cc-pVQZ benchmarks. The resulting MAE for r_e^{SE} distances was a mere 0.0005 Å. Because our MP2/6-31G(d) α_i constants were computed with a basis set comparable to cc-pVDZ, electronic structure errors in the $B_e - B_0$ VPT2 vibrational corrections are not expected to have an appreciable effect on our r_e^{SE} results for Ala. An important caveat to this conclusion is that the test molecules of ref 89 did not have the troublesome, large-amplitude vibrational modes present in the alanine conformers. Nonetheless, the largest sources of error in the r_e^{SE} parameters are the modeling of vibrational effects via VPT2 theory, the phenomenological nature of the underlying empirical rotational constants, and the gaps in the isotopologic data. Taking into account all sources of error in both the theoretical and semiexperimental methods, the agreement in Tables 3 and 4 between the r_e^{BO} and r_e^{SE} structures of **Ala-I** and **Ala-IIA** is quite satisfactory. The dihedral angle $\tau(\text{O}=\text{C}-\text{C}_\alpha-\text{N})$ in **Ala-I** is a notable point of accord.

The conformational change from **Ala-I** to **Ala-IIA** yields considerable shifts in a few bond distances and angles.

Particularly prominent is the shift of the semiexperimental $\angle(\text{C}-\text{C}_\alpha-\text{N})$ angle from $113.3(4)^\circ$ in **Ala-I** to $109.8(4)^\circ$ in **Ala-IIA**, consistent with the trans-angle rule⁹⁰ of hyperconjugative and steric effects. In the r_s structures,⁵ there is also a large **Ala-I**–**Ala-IIA** difference in this angle, but the shift is overestimated, and $\angle(\text{C}-\text{C}_\alpha-\text{N})$ is much too large for both conformers. The carbonyl oxygen is involved in a bifurcated hydrogen bond in **Ala-I** but is uncomplexed in **Ala-IIA**. In both the r_e^{SE} and r_e^{BO} structures, the hydrogen bond formation is accompanied by an expected lengthening of $r(\text{C}=\text{O})$ by $0.002-0.003$ Å. While the r_0 structures⁵ exhibit C=O bond elongation, the magnitude of the effect is $0.04(3)$ Å, a severe overestimation.

A key measure of the intramolecular hydrogen bonding in the Ala conformers is the associated heavy-atom distance $R(\text{N}\cdots\text{O})$. In the **Ala-I** [r_e^{SE} , r_e^{BO}] structures, $R(\text{N}\cdots\text{O}) = [2.825(12), 2.841]$ Å, while the corresponding values for **Ala-IIA** are $R(\text{N}\cdots\text{O}) = [2.605(18), 2.607]$ Å. Values for $R(\text{N}\cdots\text{O})$ hydrogen-bond distances computed at several levels of electronic structure theory are presented in Table S3 of the Supporting Information. The variations among the results demonstrate that our CCSD(T)/cc-pVTZ r_e^{BO} values for $R(\text{N}\cdots\text{O})$ should be accurate to ± 0.01 Å or better. **Ala-IIA** exhibits a larger, 0.034 Å discrepancy between the r_e^{SE} and r_e^{BO} H-bond lengths because of the aforementioned difficulty in determining the nitrogen-atom position. Likewise, both $\tau(\text{O}=\text{C}-\text{C}_\alpha-\text{N})$ and $r(\text{C}_\alpha-\text{N})$ of **Ala-IIA** significantly stray from the respective r_e^{BO} values. The much shorter $R(\text{N}\cdots\text{O})$ distance in **Ala-IIA** correctly reflects the greater strength of the $\text{OH}\cdots\text{N}$ hydrogen bond in this conformer compared to the $\text{NH}\cdots\text{O}$ bifurcated hydrogen bonds in **Ala-I**. Despite these relative hydrogen bond strengths, **Ala-I** is lower in energy than **Ala-IIA**, as definitively shown in section III.G below. The compensating energetic factor is the ~ 5 kcal mol⁻¹ more favorable (cis) arrangement of the carboxyl group in **Ala-I**.

III.F. Comparison of Ala and Gly Structures. A profitable comparison of the structures of the two simplest amino acids is afforded by our determination of the first r_e^{SE} parameters for **Ala-I** and **Ala-IIA** combined with analogous r_e^{SE} results for **Gly-Ip** and **Gly-IIn** from our earlier work.²⁴ The **Ala-I**–**GlyIp** differences in the heavy-atom bond distances are $\Delta r(\text{C}-\text{C}_\alpha) = +0.009$, $\Delta r(\text{C}_\alpha-\text{N}) = +0.007$, $\Delta r(\text{C}=\text{O}) = 0.000$, and $\Delta r(\text{C}-\text{O}) = -0.004$ Å, while the corresponding **Ala-IIA**–**GlyIIn** differences are $\Delta r(\text{C}-\text{C}_\alpha) = +0.006$, $\Delta r(\text{C}_\alpha-\text{N}) = -0.002$, $\Delta r(\text{C}=\text{O}) = +0.003$, and

Table 7. Focal Point Analysis of the **Ala-IIA–Ala-I** Energy Difference (kJ mol⁻¹)^a

	$\Delta E_e(\text{RHF})$	$\delta[\text{MP2}]$	$\delta[\text{CCSD}]$	$\delta[\text{CCSD(T)}]$	$\Delta E_e[\text{CCSD(T)}]$
aug-cc-pVDZ	10.81	-10.73	+2.63	-1.57	+1.14
aug-cc-pVTZ	10.50	-11.26	+2.86	-1.74	+0.35
aug-cc-pVQZ	10.51	-11.28	+2.97	[-1.74]	+0.46
aug-cc-pV5Z	10.54	-11.25	[+2.97]	[-1.74]	+0.52
CBS	[10.56]	[-11.21]	[+2.97]	[-1.74]	[+0.58]
extrapolation	$a + be^{-cX}$ ($X = 3, 4, 5$)	$a + bX^{-3}$ ($X = 4, 5$)	additive	additive	

$$\Delta E_0(\text{final}) = \Delta E_e[\text{CCSD(T)/CBS}] + \Delta_{\text{core}}[\text{CCSD(T)/cc-pCVTZ}] + \Delta_{\text{ZPVE}}[\text{MP2/aug-cc-pVTZ (pared)}] = +0.58 - 0.13 + 1.66 = \mathbf{2.11} \text{ kJ mol}^{-1}$$

^a The symbol δ denotes the increment in the relative energy (ΔE_e) with respect to the preceding level of theory in the hierarchy RHF \rightarrow MP2 \rightarrow CCSD \rightarrow CCSD(T). Square brackets signify results obtained from basis set extrapolations or additivity assumptions. Final predictions are boldfaced.

$\Delta r(\text{C}-\text{O}) = -0.003 \text{ \AA}$. Among these small changes, only the $\Delta r(\text{C}-\text{C}_\alpha)$ shifts are clearly significant compared to the uncertainty of the r_c^{SE} parameters. Likewise, the only significant change among the bond angles of the Gly and Ala heavy-atom frameworks occurs for $\angle(\text{C}-\text{C}_\alpha-\text{N})$, whose **Ala-I–GlyIp** and **Ala-IIA–GlyIn** shifts are -2.1° and -1.6° , respectively. Therefore, the main differences between the bond distances and angles in Gly and Ala are highly localized at the site of the methyl substitution.

The torsion angle $\tau(\text{O}=\text{C}-\text{C}_\alpha-\text{N})$ characterizes the deviation of the amino acid backbone from planarity. In **Gly-Ip** this angle is zero because the molecule has a symmetrical bifurcated hydrogen bond and adopts C_s point-group symmetry. Substitution of the methyl group in Ala breaks this symmetry significantly and leads to a torsion angle of $16.2(8)^\circ$ in **Ala-I**. In contrast, the backbones of **Gly-In** and **Ala-IIA** exhibit $\tau(\text{O}=\text{C}-\text{C}_\alpha-\text{N})$ angles of $11(2)^\circ$ and $12.5(5)^\circ$, respectively, which are essentially equivalent within the given uncertainties.

III.G. Relative Energy of Ala Conformers. The focal-point analysis of the energy of **Ala-IIA** relative to **Ala-I** (ΔE_e) is presented in Table 7. Showing rapid convergence to the CBS limit, the RHF relative energy and the MP2 correlation increment are converged to better than 0.1 kJ mol^{-1} using the aug-cc-pVTZ basis set. Basis sets with diffuse functions were employed specifically to treat the hydrogen bonding interactions.

The electron correlation sequence for ΔE_e shows less rapid convergence than the atomic-orbital basis set series. As seen in earlier studies,^{1,7,14} Hartree–Fock theory proves unreliable for conformational energetics of amino acids, placing **Ala-IIA** above **Ala-I** by a substantial $10.56 \text{ kJ mol}^{-1}$. The MP2 correlation energy largely rectifies this overestimation, but in the CBS limit, MP2 erroneously predicts that **Ala-IIA** is 0.65 kJ mol^{-1} lower in energy than **Ala-I**. With more sophisticated treatments of electron correlation, **Ala-I** is restored as the lowest energy conformer. The final frozen-core result is $\Delta E_e[\text{CCSD(T)/CBS}] = +0.58 \text{ kJ mol}^{-1}$, and appending the effect of core electron correlation (Δ_{core}), we obtain $\Delta E_e = +0.45 \text{ kJ mol}^{-1}$. The incorporation of connected quadruple excitations in coupled-cluster wave functions is not currently feasible for alanine, but several benchmark studies^{48,91–97} have shown that $\delta[\text{CCSDT(Q)}]$ relative-energy increments are typically about an order of magnitude smaller than $\delta[\text{CCSD(T)}]$ values. Therefore, considering all sources of error, our final equilibrium energy

difference is $\Delta E_e = +0.5(3) \text{ kJ mol}^{-1}$, in which the uncertainty estimate represents a 95% confidence interval.

Zero-point vibrational energy (ZPVE) increases the **Ala-IIA–Ala-I** energy separation by 1.66 kJ mol^{-1} , yielding $\Delta E_0 = +2.1(3) \text{ kJ mol}^{-1}$. Thus, ZPVE effects constitute almost 80% of the energy difference at 0 K. The low-frequency vibrational modes that were problematic in the r_c^{SE} analysis do not appear to add significant uncertainty to the ΔE_0 determination, as less than 1% of the ZPVE effect arises from anharmonic corrections.

IV. Summary

This investigation is the first to conjoin theory and experiment to not only determine reliable semiexperimental r_e structures (r_e^{SE}) for conformers of a molecule as large and flexible as alanine (Ala) but also to analyze in detail the factors contributing to the accuracy of such parameters. It is shown convincingly that an accurate r_e^{SE} structure for a flexible molecule can indeed be determined if procedures developed for (semi)rigid systems are carefully employed. For alanine, we find that the outcome of the r_e^{SE} least-squares refinement depends critically on the accuracy of the equilibrium rotational constants, as expected, as well as the attendant uncertainties, which is less expected. Therefore, our study commenced by refitting all the spectroscopic constants of **Ala-I** and **Ala-IIA** to the experimentally measured rotational transitions to ensure a dependable reference data set. A predicate observations scheme using ab initio quartic centrifugal distortion information appears to work well even for such a flexible molecule. In refining r_e^{SE} structures for Ala, we discovered that not all effective rotational constants can be utilized, even if their apparent uncertainty is small. The problem results mostly from the effective nature of the empirical rotational constants and, to a lesser extent, from limitations of the theoretical vibration–rotation interaction treatment. It is essential to constrain the r_e^{SE} fit using accurate Born–Oppenheimer equilibrium (r_e^{BO}) parameters, obtained here at the frozen-core CCSD(T)/cc-pVTZ level of electronic structure theory. A proper choice of the fitted and constrained parameters is paramount to obtaining good r_e^{SE} results. In general, the heavy-atom positions are well determined by the fits, whereas the hydrogen atoms must be constrained. Avoiding overfitting requires particular attention to statistical details.

The r_e^{SE} parameters determined in this study demonstrate that vibrational effects *must* be removed to get meaningful

structures for large and flexible systems from rotational constants. Specifically, previous vibrationally averaged r_g / r_a , r_z , r_0 , and r_s structures for Ala are shown to be defective, exhibiting errors as large as 0.04 Å for bond distances, 3° for bond angles, and 25° for torsion angles. Therefore, small and intrinsic conformation-induced changes are reliably discerned only when precise r_e structures are known, because vibrational effects can mask the true variations. Our r_e^{SE} results are significant in this regard because they provide the first sound comparison of empirically based structures for the two simplest amino acids, Gly and Ala.

Through convergent focal-point analysis (FPA) ab initio computations, the energy difference between the lowest conformers of alanine has been pinpointed for the first time, proving that **Ala-I** and **Ala-IIA** are almost isoenergetic. The **Ala-IIA** equilibrium structure is higher in energy than that of **Ala-I** by a mere 0.5(3) kJ mol⁻¹, and with inclusion of zero-point vibrational energy (ZPVE), this relative energy is still only 2.1(3) kJ mol⁻¹. Our high-level computations also reveal that the unobserved **Ala-IIB** conformer has a tenuous existence as a distinct species, being separated from **Ala-IIA** by a vibrationally adiabatic isomerization barrier less than 0.2 kJ mol⁻¹.

Much attention has been afforded glycine and alanine as essential origin-of-life molecules, and as such, their existence in interstellar space has been actively researched. Until now, only a few molecules of possible biochemical interest have been detected with certainty in interstellar environments: glycolaldehyde, a small “sugar”;⁹⁸ acetamide, a molecule with a peptide bond;⁹⁹ and aminoacetonitrile, a precursor of glycine.¹⁰⁰ Glycine has been detected only tentatively.^{101,102} The difficulties of detecting glycine may be explained partly by the small dipole-moment components of its most stable conformer (**Gly-Ip**), for example, $\mu_a = 0.91$ D.¹⁰³ In contrast, for **Ala-I** the μ_b dipole component has been measured to be 1.6 D.⁹ The present study confirms unequivocally that **Ala-I** is the most stable form of α -alanine and supports the somewhat imprecise dipole moment measurements of Godfrey et al.⁹ Because b -type transitions have larger line strengths than a -type transitions, the μ_b component of **Ala-I** might be large enough to permit the interstellar detection of α -alanine, provided it is sufficiently abundant in the source. The interplay of theory and experiment could prove very productive toward this goal.

Acknowledgment. Dr. Steven Wheeler is thanked for helpful discussions. The research in Athens at the University of Georgia was supported by the U.S. National Science Foundation, Grant CHE-0749868. The work performed in Hungary and the Athens/Budapest collaboration received support from the Hungarian Scientific Research Fund, OTKA K72885 and IN77954, respectively. The joint work between Lille and Budapest was partially supported by EGIDE. The high-accuracy ab initio computations used resources of the National Energy Research Scientific Computing Center (NERSC), which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Supporting Information Available: Cartesian coordinates of CCSD(T)/cc-pVTZ equilibrium structures; refit (A_0 , B_0 , C_0) and semiexperimental (A_e , B_e , C_e) rotational constants for **Ala-I** and **Ala-IIA** isotopologues; hydrogen-bond distances $R(\text{N}\cdots\text{O})$ for various levels of theory, and complete refs 60 and 61. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Császár, A. G.; Perczel, A. *Prog. Biophys. Mol. Biol.* **1999**, *71*, 243.
- (2) Dian, B. C.; Clarkson, J. R.; Zwier, T. S. *Science* **2004**, *303*, 1169.
- (3) Robertson, E. G.; Simons, J. P. *Phys. Chem. Chem. Phys.* **2000**, *3*, 1.
- (4) Gronert, S.; O'Hair, R. A. J. *J. Am. Chem. Soc.* **1995**, *117*, 2071.
- (5) Blanco, S.; Lesarri, A.; Lopez, J. C.; Alonso, J. L. *J. Am. Chem. Soc.* **2004**, *126*, 11675.
- (6) Cao, M.; Newton, S. Q.; Pranata, J.; Schäfer, L. *J. Mol. Struct.* **1995**, *332*, 251.
- (7) Császár, A. G. *J. Phys. Chem.* **1996**, *100*, 3541.
- (8) Császár, A. G. *J. Mol. Struct.* **1994**, *346*, 141.
- (9) Godfrey, P. D.; Firth, S.; Hatherley, L. D.; Brown, R. D.; Pierlot, A. P. *J. Am. Chem. Soc.* **1993**, *115*, 9687.
- (10) Hirata, Y.; Kubota, S.; Watanabe, S.; Momose, T.; Kawaguchi, K. *J. Mol. Spectrosc.* **2008**, *251*, 314.
- (11) Iijima, K.; Beagley, B. *J. Mol. Struct.* **1991**, *248*, 133.
- (12) Iijima, K.; Nakano, M. *J. Mol. Struct.* **1999**, *486*, 255.
- (13) Stepanian, S. G.; Reva, I. D.; Radchenko, E. D.; Adamowicz, L. *J. Phys. Chem. A* **1998**, *102*, 4623.
- (14) Császár, A. G. *J. Am. Chem. Soc.* **1992**, *114*, 9568.
- (15) Maul, R.; Ortmann, F.; Preuss, M.; Hannewald, K.; Bechstedt, F. *J. Comput. Chem.* **2007**, *28*, 1817.
- (16) Rak, J.; Skurski, P.; Simons, J. P.; Gutowski, M. *J. Am. Chem. Soc.* **2001**, *123*, 11695.
- (17) Snoek, L. C.; Robertson, E. G.; Kroemer, R. T.; Simons, J. P. *Chem. Phys. Lett.* **2000**, *321*, 49.
- (18) Szidarovszky, T.; Czakó, G.; Császár, A. G. *Mol. Phys.* **2009**, *107*, 761.
- (19) Wilke, J. J.; Lind, M. C.; Schaefer, H. F.; Császár, A. G.; Allen, W. D. *J. Chem. Theory Comput.* **2009**, *5*, 1511.
- (20) Godfrey, P. D.; Brown, R. D.; Rodgers, F. M. *J. Mol. Struct.* **1996**, *376*, 65.
- (21) Pulay, P.; Meyer, W.; Boggs, J. E. *J. Chem. Phys.* **1978**, *68*, 5077.
- (22) Botschwina, P.; Oswald, M.; Flugge, J.; Heyl, A.; Oswald, R. *Chem. Phys. Lett.* **1993**, *209*, 117.
- (23) Allen, W. D.; Czinki, E.; Császár, A. G. *Chem.—Eur. J.* **2004**, *10*, 4512.
- (24) Kasalová, V.; Allen, W. D.; Schaefer, H. F.; Czinki, E.; Császár, A. G. *J. Comput. Chem.* **2007**, *28*, 1373.
- (25) Thiel, W.; Scuseria, G.; Schaefer, H. F.; Allen, W. D. *J. Chem. Phys.* **1988**, *89*, 4965.

- (26) Breidung, J.; Cosleou, J.; Demaison, J.; Sarka, K.; Thiel, W. *Mol. Phys.* **2004**, *102*, 1827.
- (27) Cazzoli, G.; Cludi, L.; Contento, M.; Puzzarini, C. *J. Mol. Spectrosc.* **2008**, *251*, 229.
- (28) Demaison, J.; Császár, A. G.; Dehayem-Kamadjeu, A. *J. Phys. Chem. A* **2006**, *110*, 13609.
- (29) Demaison, J.; Lievin, J.; Császár, A. G. *J. Phys. Chem. A* **2008**, *112*, 4477.
- (30) Demaison, J.; Margules, L.; Maeder, H.; Sheng, M.; Rudolph, H. D. *J. Mol. Spectrosc.* **2008**, *252*, 169.
- (31) East, A. L. L.; Allen, W. D.; Klippenstein, S. J. *J. Chem. Phys.* **1995**, *102*, 8506.
- (32) East, A. L. L.; Johnson, C. S.; Allen, W. D. *J. Chem. Phys.* **1993**, *98*, 1299.
- (33) Gauss, J.; Stanton, J. F. *J. Phys. Chem. A* **2000**, *104*, 2865.
- (34) Gordon, V. D.; Nathan, E. S.; Apponi, A. J.; McCarthy, M. C.; Thaddeus, P.; Botschwina, P. *J. Chem. Phys.* **2000**, *113*, 5311.
- (35) Boese, A. D.; Martin, J. M. L.; Klopper, W. *J. Phys. Chem. A* **2007**, *111*, 11122.
- (36) Lane, J. R.; Kjaergaard, H. G. *J. Chem. Phys.* **2009**, *131*, 034307.
- (37) Paizs, B.; Salvador, P.; Császár, A. G.; Duran, M.; Suhai, S. *J. Comput. Chem.* **2001**, *22*, 196.
- (38) Sellers, H. L.; Schafer, L. *Chem. Phys. Lett.* **1979**, *63*, 609.
- (39) Allen, W. D.; East, A. L. L.; Császár, A. G., In *Structures and Conformations of Non-Rigid Molecules*; Laane, J., Dakkouri, M., van der Veeken, B., Oberhammer, H., Eds.; Kluwer: Dordrecht, The Netherlands, 1993; pp 343–373.
- (40) Császár, A. G.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **1998**, *108*, 9751.
- (41) Császár, A. G.; Tarczay, G.; Leininger, M. L.; Polyansky, O. L.; Tennyson, J.; Allen, W. D. In *Spectroscopy from Space*, Demaison, J., Sarka, K., Eds.; Kluwer: Dordrecht, The Netherlands, 2001; pp 317–339.
- (42) East, A. L. L.; Allen, W. D. *J. Chem. Phys.* **1993**, *99*, 4638.
- (43) Gonzales, J. M.; Pak, C.; Cox, R. S.; Allen, W. D.; Tarczay, G.; Császár, A. G. *Chem.—Eur. J.* **2003**, *9*, 2173.
- (44) Schuurman, M. S.; Allen, W. D.; Schleyer, P. v. R.; Schaefer, H. F. *J. Chem. Phys.* **2005**, *122*, 104302.
- (45) Czakó, G.; Braams, B. J.; Bowman, J. M. *J. Phys. Chem. A* **2008**, *112*, 7466.
- (46) Schuurman, M.; Muir, S.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **2004**, *120*, 11586.
- (47) Czakó, G.; Nagy, B.; Tasi, G.; Somogyi, A.; Šimunek, J.; Noga, J.; Braams, B. J.; Bowman, J. M.; Császár, A. G. *Int. J. Quantum Chem.* **2009**, *109*, 2393.
- (48) Simmonett, A. C.; Schaefer, H. F.; Allen, W. D. *J. Chem. Phys.* **2009**, *130*, 044301.
- (49) Wheeler, S. E.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **2004**, *121*, 8800.
- (50) Wheeler, S. E.; Robertson, K. A.; Allen, W. D.; Schaefer, H. F.; Bomble, Y. T.; Stanton, J. F. *J. Phys. Chem. A* **2007**, *111*, 3819.
- (51) Bartlett, R. J.; Watts, J. D.; Kucharski, S. A.; Noga, J. *Chem. Phys. Lett.* **1990**, *165*, 513.
- (52) Bartlett, R. J.; Watts, J. D.; Kucharski, S. A.; Noga, J. *Chem. Phys. Lett.* **1990**, *167*, 609.
- (53) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.
- (54) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- (55) Császár, A. G.; Allen, W. D. *J. Chem. Phys.* **1996**, *104*, 2746.
- (56) Fogarasi, G.; Zhou, X.; Taylor, P. W.; Pulay, P. *J. Am. Chem. Soc.* **1992**, *114*, 8191.
- (57) Pulay, P.; Fogarasi, G.; Pang, F.; Boggs, J. E. *J. Am. Chem. Soc.* **1979**, *101*, 2550.
- (58) 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.
- (59) Dunning, T. H., Jr. *J. Chem. Phys.* **1970**, *53*, 2823.
- (60) Werner, H.-J. et al. *Molpro, A Package of Ab Initio Programs*, version 2006.1.
- (61) ACESII, J. F. Stanton et al. For current version, see <http://www.aces2.de>. See also Stanton, J. F.; Gauss, J.; Watts, J. D.; Lauderdale, W. J.; Bartlett, R. J. *Int. J. Quantum Chem. Symp.* **1992**, *26*, 879.
- (62) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; Defrees, D. J.; Pople, J. A. *J. Chem. Phys.* **1982**, *77*, 3654.
- (63) Allen, W. D.; Császár, A. G. *J. Chem. Phys.* **1993**, *98*, 2983.
- (64) Stanton, J. F.; Lopreore, C. L.; Gauss, J. *J. Chem. Phys.* **1998**, *108*, 7190.
- (65) Mills, I. M., In *Molecular Spectroscopy: Modern Research*; Rao, K. N., Mathews, C. W., Eds.; Academic Press: New York, 1972; Vol. 1, pp 115–140.
- (66) Curl, R. F. *J. Comput. Phys.* **1970**, *6*, 367.
- (67) Gordy, W.; Cook, R. L. In *Microwave Molecular Spectra*; John Wiley & Sons: New York, 1984; Vol. 18, pp 647–724.
- (68) Lees, R. M. *J. Mol. Spectrosc.* **1970**, *33*, 124.
- (69) MolStruct is an abstract program developed by Wesley D. Allen for use within Mathematica (Wolfram Research Inc., Champaign, IL) to perform diverse fits of molecular structures to sets of isotopologic rotational constants.
- (70) Demaison, J. *Mol. Phys.* **2007**, *105*, 3109.
- (71) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.
- (72) Feller, D. *J. Chem. Phys.* **1993**, *98*, 7059.
- (73) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639.
- (74) Pared aug-cc-pVTZ basis set specifications: for the C, C α , N, and O atoms, aug-cc-pVTZ sans the diffuse f functions; for the amino and hydroxyl hydrogens, aug-cc-pVTZ sans the diffuse d functions; for C $_m$, as well as the methyl and C α hydrogens, cc-pVTZ.
- (75) Felder, P.; Günthard, H. H. *Chem. Phys.* **1982**, *71*, 9.
- (76) Ruoff, R. S.; Klots, T. D.; Emilsson, T.; Gutowsky, H. S. *J. Chem. Phys.* **1990**, *93*, 3142.
- (77) Belsley, D. A. *Conditioning Diagnostics: Colinearity and Weak Data in Regression*; Wiley: Chichester, U.K., 1991.

- (78) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression in Outlier Detection*; Wiley: New York, 1987.
- (79) Watson, J. K. G. In *Vibrational Spectra and Structure*; Durig, J. R., Ed. Elsevier: Amsterdam, 1977; pp 1–89.
- (80) Rudolph, H. D. Z. *Naturforsch.* **1968**, *23a*, 540.
- (81) Bartell, L. S.; Romanesko, D. J.; Wong, T. C. In *Molecular Structure by Diffraction Methods*; Sims, G. A., Sutton, L. E., Eds.; The Chemical Society: London, 1975; Vol. 3, p 72.
- (82) Császár, A. G.; Fogarasi, G. *J. Chem. Phys.* **1989**, *89*, 7647.
- (83) Wlodarczak, G.; Burie, J.; Demaison, J.; Vormann, K.; Császár, A. G. *J. Mol. Spectrosc.* **1989**, *134*, 297.
- (84) Demaison, J.; Herman, M.; Liévin, J. *Int. Rev. Phys. Chem.* **2007**, *26*, 391.
- (85) Bak, K. L.; Gauss, J.; Jørgensen, P.; Olsen, J.; Helgaker, T.; Stanton, J. F. *J. Chem. Phys.* **2001**, *114*, 6548.
- (86) Demaison, J.; Margulès, L.; Boggs, J. E. *Chem. Phys.* **2000**, *260*, 65.
- (87) Margulès, L.; Demaison, J.; Boggs, J. E. *J. Mol. Struct.* **2000**, *500*, 245.
- (88) Martin, J. M. L. *Chem. Phys. Lett.* **1995**, *242*, 343.
- (89) Pawlowski, F.; Jørgensen, P.; Olsen, J.; Hegelund, F.; Helgaker, T.; Gauss, J.; Bak, K. L.; Stanton, J. F. *J. Chem. Phys.* **2002**, *116*, 6482.
- (90) Rasanen, M.; Aspiala, A.; Homanen, L.; Murto, J. *J. Mol. Struct.* **1982**, *96*, 81.
- (91) Bomble, Y. J.; Vazquez, J.; Kállay, M.; Michauk, C.; Szalay, P. G.; Császár, A. G.; Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **2006**, *125*, 064108.
- (92) Czakó, G.; Mátyus, E.; Simmonett, A. C.; Császár, A. G.; Schaefer, H. F.; Allen, W. D. *J. Chem. Theory Comput.* **2008**, *4*, 1220.
- (93) Feng, H.; Allen, W. D. *J. Chem. Phys.* **2010**, *132*, 094304.
- (94) Simmonett, A. C.; Stibrich, N. J.; Papas, B. N.; Schaefer, H. F.; Allen, W. D. *J. Phys. Chem. A* **2009**, *113*, 11643.
- (95) Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vazquez, J.; Stanton, J. F. *J. Chem. Phys.* **2004**, *121*, 11599.
- (96) Wheeler, S. E.; Houk, K. N.; Schleyer, P. v. R.; Allen, W. D. *J. Am. Chem. Soc.* **2009**, *131*, 2547.
- (97) Wheeler, S. E.; Robertson, K. A.; Allen, W. D.; Schaefer, H. F.; Bomble, Y. J.; Stanton, J. F. *J. Phys. Chem. A* **2007**, *111*, 3819.
- (98) Hollis, J. M.; Jewell, P. R.; Lovas, F. J.; Remijan, A. *Astrophys. J. Lett.* **2004**, *613*, L45.
- (99) Hollis, J. M.; Lovas, F. J.; Remijan, A. J.; Jewell, P. R.; Ilyushin, V. V.; Kleiner, I. *Astrophys. J. Lett.* **2006**, *643*, L25.
- (100) Belloche, A.; Menten, K. M.; Comito, C.; Müller, H. S. P.; Schilke, P.; Ott, J.; Thorwirth, S.; Hieret, C. *Astron. Astrophys.* **2008**, *482*, 179.
- (101) Kuan, Y. J.; Charnley, S. B.; Huang, H. C.; Kisiel, Z. *Astrophys. J.* **2003**, *593*, 848.
- (102) Snyder, L. E.; Lovas, F. J.; Hollis, J. M.; Friedel, D. N.; Jewell, P. R.; Remijan, A.; Ilyushin, V. V.; Alekseev, E. A.; Dyubko, S. F. *Astrophys. J.* **2005**, *619*, 914.
- (103) Lovas, F. J.; Zobov, N.; Fraser, G. T.; Suenram, R. D. *J. Mol. Spectrosc.* **1995**, *171*, 189.

CT1000236

JCTC

Journal of Chemical Theory and Computation

Mixed Quantum Mechanics/Molecular Mechanics Scoring Function To Predict Protein–Ligand Binding Affinity

Seth A. Hayik,^{†,‡} Roland Dunbrack, Jr.,[†] and Kenneth M. Merz, Jr.*[‡]

Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, Pennsylvania 19111, and Department of Chemistry, Quantum Theory Project, University of Florida, P.O. Box 118435, Gainesville, Florida 32611

Received June 8, 2010

Abstract: Computational methods for predicting protein–ligand binding free energy continue to be popular as a potential cost-cutting method in the drug discovery process. However, accurate predictions are often difficult to make as estimates must be made for certain electronic and entropic terms in conventional force field based scoring functions. Mixed quantum mechanics/molecular mechanics (QM/MM) methods allow electronic effects for a small region of the protein to be calculated, treating the remaining atoms as a fixed charge background for the active site. Such a semiempirical QM/MM scoring function has been implemented in AMBER using the DivCon program and tested on a set of 23 metalloprotein–ligand complexes, where QM/MM methods provide a particular advantage in the modeling of the metal ion. The binding affinity of this set of proteins can be calculated with an R^2 of 0.64 and a standard deviation of 1.88 kcal/mol without fitting and an R^2 of 0.71 and a standard deviation of 1.69 kcal/mol with fitted weighting of the individual scoring terms. In this study we explore the use of various methods to calculate terms in the binding free energy equation, including entropy estimates and minimization standards. From these studies we found that using the rotational bond estimate of ligand entropy results in a reasonable R^2 of 0.63 without fitting. We also found that using the ESCF energy of the proteins without minimization resulted in an R^2 of 0.57, when using the rotatable bond entropy estimate.

Introduction

It is important to accurately model ligand interactions so that computational screening can be effectively applied to lower the cost and time involved in drug discovery.^{1,2} Protein–ligand interactions involve a complicated mixture of electrostatic, dispersion, and other interactions and therefore represent a challenge to model and predict accurately. Methods for predicting ligand affinity vary in the scoring function and structure prediction methods used, and each has different advantages and disadvantages.^{3–5} Often, these schemes sacrifice accuracy to gain speed or improve accuracy at greater computational expense but then are too complicated and slow to be used for large-scale applications.^{6,7} For these

reasons research in this area continues to thrive, searching for a model that achieves a good compromise between speed and accuracy to predict binding affinities and score binding poses.

Empirical scoring functions are composed of several terms trained on experimental binding data to generate general parameters.^{8–10} These functions may perform well and are quick, but have limitations due to the way these scoring functions are derived and the training sets used to construct them. The accuracy of empirical and knowledge-based scoring functions is dependent on the size and variety of the training set used. These functions may fail if the systems being examined are too different from any ligands in the training set.¹¹ This leaves the scoring function without knowledge of the systems being examined, and therefore, it essentially must draw conclusions from estimates of other ligands.

* Corresponding author phone: (352) 392-6973; fax: (352) 392-8722; e-mail: merz@qtp.ufl.edu.

[†] Fox Chase Cancer Center.

[‡] University of Florida.

Another class of scoring functions is constructed using a potential based on classical molecular mechanics force fields.^{12–16} These methods either exclude electronic effects or account for them with an empirical parameter that may not always apply. These methods also have problems accurately predicting properties for nonstandard residues and metalloenzymes, which are difficult to describe with classical methods,¹⁷ although some successful examples have been reported.^{18,19}

Quantum mechanics (QM) methods have begun to demonstrate their usefulness as scoring functions for calculating ligand binding free energies as computer power has increased. Until recently, QM methods were primarily used only for smaller systems because the cost associated with these methods was untenable for large-scale screening. Efforts have been made to decrease the time required for QM calculations of proteins using various methods, making them more viable in scoring functions.^{20–24} We recently developed QMScore,^{25–27} a scoring function using a full QM potential that uses a linear scaling semiempirical method within the program DivCon^{22,23} to calculate properties of the protein, ligand, and complex and to combine them into a scoring function to give a binding free energy.

QM methods have also been incorporated into molecular mechanics programs²⁸ and have seen a resurgence recently as computing power has increased to create mixed quantum mechanics/molecular mechanics (QM/MM) methods.²⁹ These methods have proven quite successful for molecular dynamics simulations and reaction mechanism studies in biological systems.^{30–33} QM/MM methods allow a small region of interest to be explored in more detail while treating the surroundings with a faster method to save computational time. Some recent structure-based drug design methods have used quantum mechanics,^{34,35} QM/MM,^{36–40} or QM/QM⁴¹ methods to calculate protein–ligand binding. These methods take advantage of the mixed method's ability to specify a region of interest, usually the ligand, with an expensive Hamiltonian while treating the remaining system with a cheaper Hamiltonian to get more accuracy for the ligand and its surroundings while not becoming prohibitively expensive.

It has been shown that polarization^{42–45} and charge transfer effects^{26,27} can play an important role in docking a ligand to a target, effects that classical methods often cannot take into account. In particular, QM methods have shown their potential in predicting binding affinities for metalloenzymes. Difficulties with metalloenzymes arise from the expanded valence and the high charge of the metal atom and the charge transfer associated with ligand binding.¹⁹ Our previously developed method, QMScore,^{25–27} allows the entire protein to be calculated using a semiempirical QM potential, with the drawback that it is time-consuming to calculate the entire protein at a QM level. However, the benefit is that this method allows electronic effects such as charge transfer to be captured throughout the protein.

QM/MM methods allow polarization effects near the ligand to be taken into account, as QMScore does, without the need to include the entire protein in the quantum region. While long-range polarization may impact ligand binding slightly, Illingworth et al. found that a majority of the

polarization energy is within 5 Å of the ligand and polarization's effect on the charges was fairly short-ranged when examining MM charge polarization.⁴⁶ Including the first shell of residues allows this polarization to be fully incorporated into the energy terms of the scoring function. These QM/MM methods also have an advantage over many other methods in that specific parameters are not necessary for the ligand or many common metals that may be present.

In this paper, we develop a scoring function comprising several terms: an interaction energy derived from QM/MM energies calculated with AMBER 10 with DivCon, a solvation term based on the change in surface area upon binding, and entropic terms based on the QM/MM frequencies of the system or the rotatable bonds in the ligand to determine the overall binding free energy. We compare the results of the new scoring function with the full QM calculations of Raha and Merz, focusing on a set of zinc metalloenzymes previously modeled with QMScore in a full QM treatment.²⁵ Our results show that the QM/MM-based scoring function can be optimized to perform nearly as well as the full QM calculation. In addition, we examine a number of factors of the calculation that can be altered to explore accuracy/speed trade-offs for this QM/MM method, such as the number of minimizations and the entropy calculation method.

Methods

QM/MM Implementation. The linear scaling program DivCon,^{22,23} integrated into AMBER 10, was used to model the QM region around the ligand. This combined (active site/ligand) region was included in the semiempirical QM calculation, while atoms outside of this region were treated with the classical AMBER ff99SB⁴⁷ force field. This gives the effective Hamiltonian of the system, which is a sum of the MM (\hat{H}_{MM}), QM (\hat{H}_{QM}), and QM/MM interaction ($\hat{H}_{\text{QM/MM}}$) Hamiltonians:

$$\hat{H}_{\text{eff}} = \hat{H}_{\text{MM}} + \hat{H}_{\text{QM}} + \hat{H}_{\text{QM/MM}} \quad (1)$$

Splitting the system into regions like this leaves many covalent bonds between the two regions severed, resulting in dangling bonds and charge imbalances in both regions. This is addressed by adding a hydrogen link atom to the QM atom in the broken bond that is formed, similar to the implementation in Dynamo.⁴⁸ The link atom is forced to lie along the bond vector so that no extra degrees of freedom are introduced to the system, and it is treated like a regular QM atom throughout the calculation. The forces on the link atom are then distributed to the QM and MM atoms that make up the bonding pair, and any interactions involving the MM atom are treated classically.⁴⁹

The electrostatics of the link atom must be carefully considered to avoid false polarization of the QM boundary atom and to maintain a constant charge for the QM region. This is accomplished by spreading the charge of the QM region removed from \hat{H}_{MM} across all of the MM atoms in the system at the beginning of a run. This initial setup step adds a slight cost to the QM/MM calculation, but is not prohibitively expensive as it is a one-time calculation. To

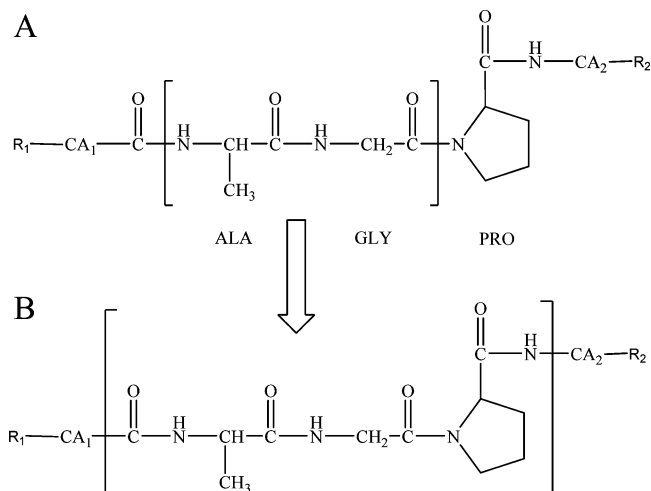


Figure 1. Graphical representation of the QM/MM partitioning scheme. The area in brackets represents the QM region. (A) The rough QM region with a PRO residue at the boundary cuts the two peptide bonds between the C and N. (B) The refined structure extends the QM boundary to include the PRO residue and moves the boundary to cut the C–CA bond to avoid spurious polarization at the boundary.

avoid overpolarization of the QM/MM boundary, the classical atom involved in the link atom bond is ignored by the link atom, and the van der Waals interactions of the bonding pair are treated classically. This results in more stable charge distributions for the atoms in the QM region⁴⁹ and eliminates any false, highly repulsive forces between the link atom and the MM bonding atom.

The QM/MM method in AMBER allows for the minimization of a system of interest using the conjugate gradient or steepest descent methods. In a QM/MM calculation the forces of the QM region are calculated by the QM program being used and are then transferred and added to the MM forces on the QM atoms, while the MM forces are calculated by AMBER. This allows a fairly simple minimization to be undertaken for a QM/MM calculation while still removing the need to calculate parameters for an organic molecule or metal ion and taking advantage of the parameters available in the force field for defined atom types.

Preparation of QM/MM Calculations. In a QM/MM calculation, the boundary must be carefully chosen so that the approximations and assumptions made in a QM/MM system do not significantly alter the accuracy of the calculation. The QM region should not be too large to realize the cost savings with the MM portion of the energy function. In this study the QM region was selected to include not only the ligand, but also the first shell of the active site to capture electronic changes from binding. Any protein residue with at least one atom within 5 Å of the zinc ion was included in the QM region. The QM region was expanded so that highly polarizable peptide bonds were not cut by the boundary as shown in Figure 1. If a proline was present, the entire residue was included along with its peptide bond. This accommodates the QM cut so that it is not adjacent to the carbonyl group of the peptide bond, providing more distance between the polar region and the QM/MM boundary. If any disulfide bond was included in the QM region, the region was

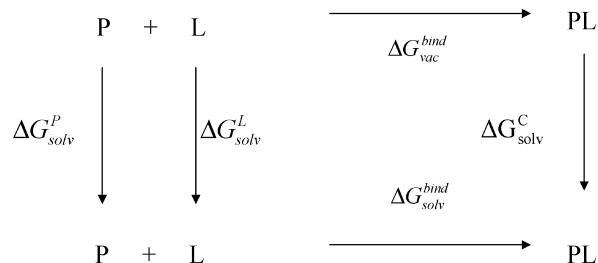


Figure 2. Schematic view of the thermodynamic cycle to calculate binding affinity. The cycle calculates the protein (P), ligand (L), and complex (C) in vacuum and then transfers them to solvent to find the solvation free energy.

expanded to include both partners. Once the QM region was properly defined, the charge of the QM region was calculated on the basis of what residues were present and the charge of the ligand. The active site was visually inspected to confirm the accuracy of the charge.

Preparation of Proteins. The structures used in this study were downloaded from the Protein Data Bank (PDB).⁵⁰ Protons were added to the heavy atoms using the LEAP module in AMBER 5.0 based on standard geometries and physiological pH, and energy minimization was performed to relax the added protons. Each of these complexes contained a zinc ion in the active site, and a zinc-bound water molecule was added to the uncomplexed proteins to fill the exposed valence. These structures were used as the starting point in this study.

From this point the structures were split into the QM and MM regions for the QM/MM calculations as described above and minimized for 500 cycles using steepest descent and a maximum of another 1500 cycles using the conjugate gradient method in version 10 of AMBER.⁵¹ The long-range cutoff for both the QM and MM regions was set to 100 Å to include long-range effects in the binding affinity calculations, unless otherwise stated. After the initial structures were minimized, the solvation free energies of the protein, ligand, and complex were calculated in DivCon. In addition to the solvation free energy, the vibrational frequencies of the ligand and protein alone and in complex were calculated using DivCon, along with a count of the rotatable bonds on the ligand, to provide an estimate of the vibrational entropy change associated with binding. The convergence criterion for the vibrational entropy step was increased to 1×10^{-6} for the changes in the energy and density matrix, and the long-range cutoff was again set to 100 Å.

Binding Affinity Calculation. The binding affinity of a protein–ligand complex can be determined using the thermodynamic cycle shown in Figure 2. The quantity of interest is $\Delta G_{\text{bind}}^{\text{sol}}$, the free energy change of binding in solvent, which can be calculated via the alternative path of first desolvating the protein and the ligand, binding of the protein and ligand into a complex, and then solvating the complex. If the free energy change on solvation for each entity X is expressed as ΔG_{sol}^X , then the desolvation contributes $-\Delta G_{\text{sol}}^{\text{protein}}$ and $-\Delta G_{\text{sol}}^{\text{ligand}}$ to $\Delta G_{\text{bind}}^{\text{sol}}$, while the solvation of the complex contributes $+\Delta G_{\text{sol}}^{\text{complex}}$. The free energy of binding in solution can therefore be broken down into the gas-phase binding free energy and the solvation free energy:

$$\Delta G_{\text{bind}}^{\text{sol}} = \Delta G_{\text{bind}}^{\text{gas}} + \Delta \Delta G_{\text{solv}} \quad (2)$$

where

$$\Delta \Delta G_{\text{solv}} = \Delta G_{\text{solv}}^{\text{complex}} - (\Delta G_{\text{solv}}^{\text{protein}} + \Delta G_{\text{solv}}^{\text{ligand}}) \quad (3)$$

The gas-phase free energy of binding contains an enthalpic term from the electrostatic and nonpolar interaction energies and an entropic term from the degrees of freedom of the individual systems:

$$\Delta G_{\text{bind}}^{\text{gas}} = \Delta H_{\text{bind}}^{\text{gas}} - T \Delta S_{\text{bind}}^{\text{gas}} \quad (4)$$

$$\Delta S_{\text{bind}}^{\text{gas}} = \Delta S_{\text{SASA}} + \Delta S_{\text{vib}} \quad (5)$$

where ΔS_{SASA} is the entropy change from the solvent-accessible surface area (SASA) and is discussed in further detail below.

The enthalpy term is calculated from the SCF energy in DivCon at the AM1⁵² semiempirical level, which was found to work well previously:⁴¹

$$\Delta H_{\text{bind}}^{\text{gas}} = \Delta H_{\text{f}}^{\text{complex}} - \Delta H_{\text{f}}^{\text{protein}} - \Delta H_{\text{f}}^{\text{ligand}} \quad (6)$$

where

$$\Delta H_{\text{f}} = E_{\text{elec}} + E_{\text{core-core}} + \sum_{\text{atoms}} \Delta H_{\text{f}} \quad (7)$$

E_{elec} is the electronic energy, $E_{\text{core-core}}$ is the core–core repulsion, and the final term is the sum of the heat of formation for all the atoms. These terms are all calculated within DivCon.

Another important consideration for protein–ligand binding is the change in entropy of the ligand upon binding in the gas phase, $\Delta S_{\text{bind}}^{\text{gas}}$, calculated here as seen in eq 5. Protein–ligand binding is entropically unfavorable in most cases due to loss of translational, rotational, and conformational degrees of freedom.^{38,53–55} Several scoring functions estimate the conformational entropy component of the free energy change via the number of rotatable bonds in the ligand or the protein and ligand together.^{17,25,26} This measure provides a good estimate of the degrees of freedom lost by both the protein and ligand on binding.

Another way to calculate the conformational entropy component is to calculate the vibrational frequencies of the ligand in complex³⁷ with the entire QM region, as well as the protein and ligand alone. By capturing these effects, the change in translational, rotational, and vibrational degrees of freedom upon binding can be estimated, and the entropic effect on the ligand in the protein field or on the ligand and the protein side chains in the active site that interact with the ligand can be determined. This gives a more accurate estimate of the entropic penalty at the cost of computing the vibrational frequencies. In this implementation, after QM/MM conjugate gradient optimization, the frequencies of the optimized ligand and protein alone and in complex with the protein were calculated at the AM1 level using DivCon and a partial Hessian vibrational analysis (PHVA).³ This allows the frequencies of only the minimized region, either the ligand alone in the protein field or the entire QM region, to

be calculated, excluding the rest of the system, and has been found to accurately reproduce the appropriate frequencies.^{3,56}

Many of the systems being considered here are fairly large overall, and therefore, the minimization steps will not fully minimize all gradients to a near-zero value. This is due partially to the minimization scheme used and partially to our interest in keeping the binding affinity prediction quick enough for large-scale applications. Thus, a small number of imaginary frequencies may appear from the diagonalization of the Hessian matrix, especially when calculating the frequencies of the entire QM region, comprised of the ligand and protein side chains. In these cases, any imaginary vibrational frequencies found were not included in the calculation of the vibrational entropy. Some low-frequency vibrational modes may be disregarded, but this calculation can still provide an estimate of the vibrational entropy change due to binding.

From these frequencies the vibrational entropy, energy, and zero-point energy can be calculated from the normal-mode frequencies. The vibrational entropy component accounts for the change in entropy due to the gain of six vibrational degrees of freedom and loss of translational and rotational degrees of freedom when the ligand binds to the protein. The vibrational energy component represents the internal thermal energy change from molecular vibrations upon ligand binding. Here, the vibrational entropy is calculated either by finding the vibrations of the ligand in complex and free, which has been shown to be a good approximation of the degrees of freedom of the protein and ligand system,⁵³ or by calculating the frequencies of the QM/MM components including the entire QM region's side chains. In this study both of these methods are explored in the interest of measuring accuracy vs cost. The zero-point energy (ZPE) corrects the energy of the system up from the bottom of the harmonic oscillator well to the lowest vibrational quantum level, accounting for vibrations occurring even at 0 K. In this study, the vibrational energy and ZPE are already implicitly included in the calculation due to the parametric way in which the semiempirical methods are developed (fit against experimental heats of formation) and therefore are excluded from the scoring function to avoid double counting these properties. Calculating these vibrational frequencies provides the estimate to the entropy change in the gas phase that we use in our overall scoring function.

The final part of the binding free energy we consider is the solvation free energy change of the system. We calculate this as the sum of enthalpy and entropy terms:

$$\Delta G_{\text{solv}} = \Delta H_{\text{solv}} - T \Delta S_{\text{SASA}} \quad (8)$$

It has been shown that an implicit model of solvation appropriately describes the solvent interactions that occur in protein–ligand binding to account for the solvation effects.^{38,57} This term can roughly be described as the free energy associated with the desolvation of the active site and ligand upon ligand binding.

In this study an implicit solvent Poisson–Boltzmann self-consistent reaction field (PB/SCRF) solvation model was used to calculate the solvation enthalpy.⁵⁸ This method essentially calculates the energy of the solute with and

without the presence of solvent, allowing only the QM region to polarize when the solvent is added to the calculation. This gives an approximate solvation enthalpy for the active site, allowing the charges in that area to fluctuate in the solvent field, while holding the remaining charges in the system fixed. A previous study by Merz et al. found an unsigned average error of 19.2 kcal/mol for four small proteins using the same QM region cutoff size. Using a Poisson–Boltzmann method also means that an internal and external dielectric must be set to properly capture the polarization of the QM region.⁵⁹ In this study CMI⁶⁰ charges, an external dielectric constant of 80, and an internal dielectric constant of 1.0 were used. The dielectric constant is used to estimate the polarizability of the region it is assigned to. In fixed charge methods, values of 3.0 and 4.0 are commonly used for an estimate of polarizability. However, here a value of 1.0 was used because the QM region is permitted to polarize in the solvent field due to the presence of electronic degrees of freedom. Using this QM/MM PB method allowed the solvation enthalpy of the protein, ligand, and complex to be determined, yielding the overall enthalpic cost of desolvating the ligand and protein active site.

There is also an entropic solvation term to be considered on binding due to the displacement of water molecules from the active site, which plays an important role in binding.^{38,61} Changes in solvent-exposed surface area upon ligand binding have a correlation to solvent entropy,⁶² and some scoring functions successfully use the term as a measure of solvent entropy.^{11,37,38,63} In this study the solvent-accessible surface area of the heavy atoms, regardless of the polarity of the atoms, was used to estimate the solvation entropy of the binding process. This was done by running a 1.4 Å probe over the surface of the protein, ligand, and complex, which yielded the SASA⁶⁴ of each respective piece of the protein–ligand binding calculation. This surface area difference gives an estimate of the solvent entropy gained upon complexation based on the parameters from Legrand and Merz.⁶⁴ Combining the enthalpy term from the PB equations and the entropy term of the SASA approximation, the free energy change on solvation can be calculated using eq 8. The enthalpy term is simply the difference in enthalpy between the solvated and unsolvated protein, ligand, or complex, the solvation energy, and the entropy term is approximated using the SASA term.

After calculation of all of these individual components, a scoring function can be constructed to calculate the binding affinity:

$$\Delta G_{\text{bind}} = \Delta G_{\text{gas}}^{\text{bind}} + \Delta E_{\text{PB}} \quad (9)$$

In this equation, $\Delta G_{\text{gas}}^{\text{bind}}$ is the energy change from the ESCF of the unbound protein and ligand going to the complex containing the ΔS_{vib} and ΔS_{SASA} entropies and ΔE_{PB} is the Poisson–Boltzmann energy change. Combining all of these individual terms provided an overall equation that allowed the binding free energy to be calculated from individual energy components.

Regression Analysis. The method described here was assessed by comparing its predictions against experimental

data using both the square of the sample correlation coefficient, R^2 , and the standard deviation, SD:

$$R^2 = 1 - \frac{\sum_i (Y_i - f_i)^2}{\sum_i (Y_i - \bar{Y})^2} \quad (10)$$

$$\text{SD} = \sqrt{\frac{1}{N} \sum_i (Y_i - f_i)^2} \quad (11)$$

where f_i are the predicted free energies of binding, Y_i are the experimental free energies of binding, and \bar{Y} is their average.

We also attempted to improve the method by using multiple linear regression (MLR) of the terms in the energy function to predict the binding free energies. MLR defines a relationship between a dependent variable and independent variables using a least-squares method. MLR produces a linear equation where X_1, X_2, \dots are independent variables, here components of the scoring function, and Y is the dependent variable, the binding affinity:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \varepsilon_i \quad (12)$$

for data points $i = 1, 2, \dots, N$. By using these methods, not only can the predictive ability of a scoring function be estimated, but weights for individual terms (β_0, β_1, \dots) in the scoring function can be determined to give them more influence in the overall score and improve the scoring function.^{9,25} R^2 and the standard deviation can then be calculated from the values predicted by the linear regression

$$f_i = \beta_0 + \beta_1 \Delta E_{\text{SCF}} - \beta_2 T \Delta S_{\text{vib}} + \beta_3 \Delta E_{\text{PB}} + \beta_4 \Delta S_{\text{SASA}} \quad (13)$$

in comparison with the experimental data, using eqs 10 and 11.

Results and Discussion

The proteins, ligands, and complexes used in this study were taken from a full QM study undertaken by Raha and Merz,²⁵ consisting of 18 carbonic anhydrase (CA) and 5 carboxypeptidase (CPA) complexes with known experimental binding free energies and resolutions better than 2.5 Å. Table 1 summarizes the proteins and their ligands, resolution, and experimental binding affinities as well as the number of QM atoms for each. One of DivCon's features is its linear-scaling capabilities, which allows larger than usual systems to be used in semiempirical calculations.^{22,23} The DivCon method splits a protein into separate groups of atoms, in this case amino acids. These groups include a buffer region around them to account for polarization effects on the central group. The SCFs of these individual groups and their buffers are then calculated separately. This is faster than performing an SCF on the entire protein at once, since it is quicker to do many small diagonalizations than one very large one for each SCF cycle. This allows the method to be useful for large systems. Here, the crossover point for standard versus divide and conquer calculations was determined to be approximately

Table 1. PDB ID, Resolution (Å), Inhibitor, Type, and Experimental ΔG of the 18 Carbonic Anhydrase (CA) Complexes and 5 Carboxypeptidase (CPA) Complexes Used in This Study^a

PDB ID	resolution	inhibitor	type	$\Delta G(\text{exp})$	protein	ligand	complex
1a42	2.25	brinzolamide	CA	-13.66	210	44	250
1am6	2.1	methyl hydroxamate	CA	-5.98	190	10	196
1bcd	1.9	methyl sulfonamide	CA	-5.39	210	10	216
1bn1	2.1	AL5917	CA	-12.90	210	35	241
1bn3	2.2	AL6528	CA	-13.66	210	35	241
1bn4	2.1	AL5927	CA	-12.86	210	36	242
1bnn	2.3	AL7183	CA	-13.82	210	37	243
1bnq	2.4	AL4623	CA	-13.11	210	41	247
1bnt	2.15	AL5424	CA	-13.54	210	38	244
1bnu	2.15	AL5300	CA	-13.40	210	34	240
1bnv	2.4	AL7099	CA	-12.12	210	42	248
1bnw	2.25	AL5415	CA	-12.54	210	29	235
1cil	1.6	ETS	CA	-12.90	160	25	182
1cim	2.1	PTS	CA	-12.19	210	35	241
1cin	2.1	MTS	CA	-12.06	210	29	235
1cnw	2.0	EG1	CA	-10.67	210	33	239
1cnx	1.9	EG2	CA	-10.12	264	51	311
1cny	2.3	EG3	CA	-10.85	210	44	250
1cbx	1.54	L-benzylsuccinate	CPA	-8.77	210	64	270
3cpa	1.54	GY	CPA	-5.37	123	31	150
6cpa	1.54	ZAAP(O)F	CPA	-15.93	160	57	214
7cpa	2.0	BZ-FVP(O)F	CPA	-19.30	141	73	211
8cpa	1.54	BZ-AGP(O)f	CPA	-12.66	123	54	174

^a The experimental binding free energy for each complex was calculated from the K_i as $-RT \ln(K_i)$, giving the binding free energy in kilocalories per mole. The last three columns show the number of QM atoms in each system for each metalloenzyme.

600 atoms in the QM region. We therefore determined QM regions with this limitation in mind.

In the following, several different possible variables and scoring function components will be explored and compared to experimental binding free energies for these metalloenzymes to assess this function's viability. This procedure will produce the best components for use in the scoring function while also providing details on important considerations for the scoring function. We will also examine the use of MLR weights to determine the influence this statistical method will have on the predictions. Finally, an analysis of different methods to include the vibrational entropy will be undertaken to explore ways to enhance the computational performance.

ESCF Is More Predictive Than Total Energy. In the case of a QM/MM calculation an important consideration is whether to use the total energy of the system or only the QM region's energy. The QM region's energy, ESCF, encompasses only the energy of the QM region, including all the changes in the electronic terms associated with binding, while the total energy encompasses the ESCF energy as well as the MM energy terms, such as bond and angle terms. Either of these terms may be appropriate to describe protein–ligand binding, and their effect on binding affinity prediction must be examined to determine the best one to use.

In this case, we found that the ESCF energy is a marked improvement over the total energy. As calculated with eqs 10 and 11, the square of the correlation coefficient, R^2 , for the total energy is 0.56 with a standard deviation of 2.09 kcal/mol, while R^2 for the ESCF energy is 0.64 with a standard deviation of 1.89 kcal/mol. These binding affinities were calculated after two minimization runs with a nonbond cutoff of 100 Å. It can be argued that only the ESCF energy is essential in a QM/MM scoring function because the most detail is focused on the area of greatest interest. Since the

active site and ligand, when present, are defined as the QM region, it is clear that the greatest energy changes are located within the QM region and that to a first approximation the MM region is largely unaffected. Indeed, including the total energy of the protein may have no effect, or even lower the predictive ability of a scoring function as random movements unassociated with binding may occur in the MM region, introducing spurious energy terms to the total energy. These terms may especially be found at the periphery of the protein, which will have the least effect on protein–ligand binding assuming nonallosteric interactions. These distant changes may reduce the predictive ability through long-range interactions that do not reflect binding per se, but are an artifact of our computational procedure. For this reason we consider it prudent to disregard the total energy of the protein in favor of using just the ESCF energy of the smaller QM region.

A Long-Range Cutoff of 100 Å Behaves Better Than a Cutoff of 10 Å. Another important consideration is the effect of the long-range nonbond cutoff. Atoms throughout the protein may affect the electronic properties of the QM active site and ligand, which may in turn enhance or diminish protein–ligand binding. It is important to account for this when calculating the binding free energy of protein–ligand complexes. The binding affinity of the protein set was calculated using the scoring function with two steps of minimization, the ESCF energy and a cutoff of 10 and 100 Å, to explore this parameter's effects on the calculated binding affinity.

The long-range cutoff has a large impact on the predictive ability of the scoring function. Using a cutoff, for both the QM and MM regions, of 10 Å gives an R^2 of 0.36 with a standard deviation of 2.51 kcal/mol. However, using the larger cutoff of 100 Å yields an R^2 of 0.64 with a standard deviation of 1.88 kcal/mol. For these relatively small systems this cutoff includes the entire protein in the nonbond cutoff

calculation. The importance of long-range cutoffs is well-known and, not surprisingly, appears to be no less important in these binding affinity calculations than in molecular dynamics simulations.⁴⁹

A Single Minimization Cycle and the ESCF Scoring Function Perform Well. In these calculations, the QM/MM energy of the protein, the ligand, and the protein–ligand complex are minimized using the steepest decent and conjugate gradient minimizers within AMBER. The starting geometries had the hydrogens minimized while the heavy atoms were held fixed, but the rest of the protein structure is that of the crystal structure found in the PDB. Minimizing the protein allows the QM and MM regions to relax in their molecular environment, lowering the overall energy of the protein and providing a good starting point for further calculations. These QM/MM minimizations also relax the ligand in both the free and bound states, giving insight into possible structural changes of the ligand. Minimization also makes it possible to properly calculate the vibrational frequencies.

The predictive ability of the scoring function on unminimized structures was also examined. In these studies the heavy atoms were obtained directly from the PDB, while the protons were added with LEAP and allowed to minimize. From these structures the vacuum interaction energy, solvation free enthalpy, and solvation entropy were calculated. For these calculations the vibrational entropy component was estimated using the number of rotational bonds in the ligand, assigning a 1 kcal/mol penalty to each bond as used by Raha et al.^{25,26} The number of rotatable bonds is used here because using the vibrational frequencies without minimization results in large numbers of imaginary frequencies and will not provide an accurate representation of the actual entropy change.

Using the best parameters found above, namely, the ESCF energy and a long-range cutoff of 100 Å, the predictive ability of the scoring function was tested as a function of how many minimizations were done, each minimization comprising a 500-step steepest descent followed by a maximum 1500-step conjugate gradient minimization. The number of minimization steps used has an interesting effect on the predictive ability of the scoring function. If the total energy of the protein is used, one minimization results in an R^2 of 0.48 with a standard deviation of 2.27 kcal/mol. Performing two minimizations and using the total energy of the system increases the squared correlation coefficient to 0.56 with a standard deviation of 2.09 kcal/mol. Using the ESCF energy, the correlation increases to 0.59 for one minimization and to 0.64 for two minimizations, while the standard deviations are 2.01 and 1.88 kcal/mol, respectively. The results of these predictions using structures without minimization and the vibrational entropy estimated by the number of rotatable bonds gives a correlation of 0.57 with a standard deviation of 2.07 kcal/mol. Figure 3 summarizes the results of each of these minimization runs, demonstrating the differences between ESCF and total energy for each minimization scheme.

It is interesting to note that, for one minimization step, the ESCF correlation coefficient exceeds the total energy

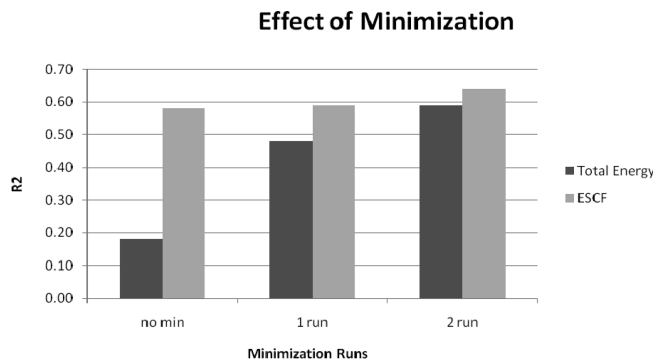


Figure 3. Effect of various minimization protocols on binding affinity prediction.

coefficient, but the two binding affinity correlations are quite close for two minimization steps. This suggests that using the ESCF energy is an adequate measure of the heat of interaction and that a single minimization may be sufficient to predict binding affinities to save time. Using one minimization cycle would obviously save time over performing two cycles, so the ability to accurately calculate binding affinity at one minimization cycle is an important consideration. Figure 3 demonstrates that using the total energy of the system at one cycle is not as predictive as the ESCF energy while there is minimal difference between one and two cycles for the ESCF energy. It is also interesting to note that scoring using ESCF without minimization also produces a result relatively close to the minimized correlation. This observation is important because if a large screening effort were undertaken, this could be used to further reduce the cost while only slightly lowering the predictive quality over partial or full minimization.

Binding Affinity Predictions Can Be Improved by MLR. The binding affinities of the 23 ligands were recalculated using multiple linear regression (eqs 12 and 13), calculating weights for the four energy terms and the constant seen in eq 13. For these calculations, the components of the scoring function were calculated using the best parameters for each term as described above. The ESCF energy was used for the heat of interaction, a long-range cutoff of 100 Å was used, and, for thoroughness, two minimization cycles were performed even though one was shown to be adequate for the ESCF energy. For these binding affinity predictions, CM1 charges were used in the SCRF solvation calculation. Figure 4 demonstrates the results of both the simple sum and the fitted function.

These results show promise for using the QM/MM method in binding affinity prediction. Calculating the binding affinities of these zinc metalloproteins with MLR yields an R^2 of 0.71 with a standard deviation of 1.69 kcal/mol, compared to the results without any fitting (R^2 of 0.64, SD = 1.88).

Using the Number of Rotatable Bonds To Estimate the Conformational Entropy Change. As mentioned above, it is fairly common to use the number of rotatable bonds to estimate the vibrational entropy penalty. The use of this estimate was also examined in this study by counting the number of rotatable bonds on the ligand using the Autotors tool from AutoDock.⁶⁵ After the number of rotatable bonds

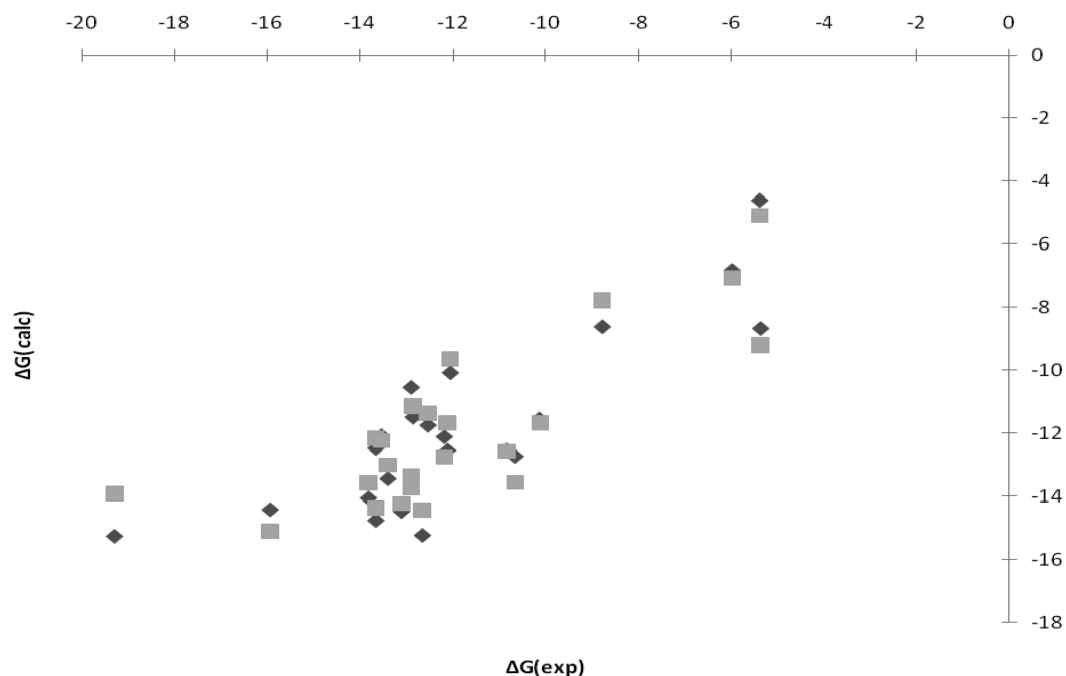


Figure 4. Calculated vs experimental ΔG of binding for the 23 zinc protein–ligand complexes. Squares represent the binding affinity calculated without any fitting to experimental data ($R^2 = 0.64$). Tilted squares represent the calculated binding affinity after fitting components of the scoring function using MLR ($R^2 = 0.71$). All units are kilocalories per mole.

in the ligand was counted, it was assumed that each of these bonds would be held fixed upon binding and each bond was responsible for 1 kcal/mol, representing the conformational degrees of freedom lost on complexation. Using energies and the surface area from the minimized structures, these calculations resulted in a correlation coefficient of 0.63 and a standard deviation of 1.92 kcal/mol without fitting. Without the minimization, these values are 0.57 and 2.07 kcal/mol (Figure 3). The correlation increases to 0.70 and the standard deviation decreases to 1.72 kcal/mol with MLR fitting. Again, the results without fitting are close to the results obtained by calculating the frequencies of the entire QM region, while the results with fitting are slightly worse than those calculated using only the ligand frequency analysis. This is an interesting result, confirming that using the number of rotatable bonds in the ligand as an estimate of the conformational entropy change may be sufficient to capture its effect on the predictive ability of this scoring function. The rotatable bond method may be used instead of a frequency calculation to save time if, for example, the ligand or data set is particularly large. It is also interesting to note that, at least for this set of 23 ligands, the number of rotatable bonds correlates with the vibrational entropy of the ligand calculated by DivCon with a correlation coefficient of 0.81, as seen in Figure 5. This presents more evidence that using the number of rotatable bonds is an acceptable estimate for the vibrational entropy of the ligand, and this can be capitalized upon to reduce the time from several hours for a full QM region vibrational calculation to a few seconds for the number of rotatable bonds if desired. Moreover, via more advanced analyses of free and bound ligands, it may be possible to create a rotatable bond model that tracks high-level computed results with an even greater accuracy.

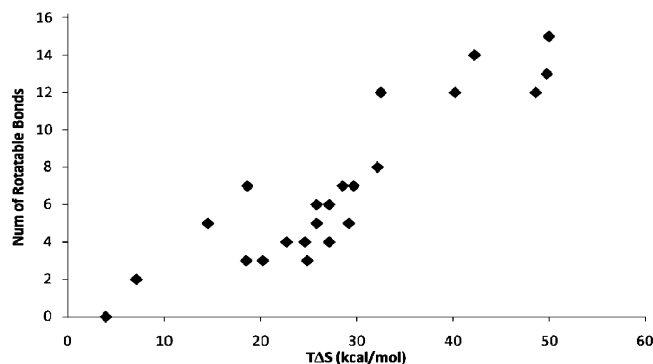


Figure 5. Correlation of $T\Delta S_{\text{vib}}$ with the number of rotatable bonds for each ligand in the set from ligand-only frequency calculations ($R^2 = 0.81$).

Using the Vibrational Frequencies of the Ligand Alone To Estimate the Conformational Entropy Change. As a way to accelerate scoring calculations, some binding affinity calculations ignore or simply estimate the vibrational entropy component. It has been found that a good estimate can be achieved using the number of rotatable bonds in the ligand^{4,8} or the change in the number of freely rotatable bonds in the ligand and protein based on solvent exposure.^{25,26,66} These two methods used to estimate vibrational entropy are extremely quick, as the number of rotatable bonds is easily calculated, but are not always as accurate as desired and do not necessarily reflect the properties of the ligand. In this situation a compromise must again be made between accuracy and computation time to decide whether this estimate is accurate enough.

To account for the entropic term in this scoring function, the vibrational frequencies of some or all of the QM atoms can be calculated. This provides a more physical representation of what is happening when the ligand binds than a simple

count of the rotatable bonds. It also provides the additional advantage of calculating the frequencies in the field of the entire protein, allowing short- and long-range interactions to influence the results. As mentioned above, due to limitations in the minimization scheme used here, imaginary frequencies were often found from the normal-mode analysis. With the minimizers used in AMBER we were unable to eliminate all imaginary frequencies. Advanced minimizers or very long minimization runs might be able to mitigate this issue and may be investigated in future efforts. These imaginary frequencies were more prevalent for larger QM regions, but still comprised only a very small portion of the overall frequencies calculated, and all were subsequently left out of the total vibrational entropy calculations. Excluding these frequencies may leave out some vibrational contributions, but overall the vibrational entropy calculated here can still provide a reasonable estimate of the entropy change associated with binding.

In the previous sections we used the vibrational entropy calculated from the entire QM region, except when no minimization was done. These frequencies capture effects in both the ligand and the protein's active site where all of the major vibrational changes will occur from binding. While in principle the most accurate, this method is also the most expensive because the active site and ligand must be fully minimized and the Hessian computation is time-consuming. This method may work for small proteins or small QM regions, but may prove to be much more difficult and time-consuming for larger proteins or QM regions; these frequency calculations are, however, fairly trivial to parallelize. It is for these reasons that in this section we will examine various ways of calculating this vibrational entropy component and the impact they have on the scoring function's predictive ability.

An alternative afforded by the QM/MM method is to calculate the vibrational frequencies of just the ligand in the protein field, as successfully applied by Grater et al.³⁷ In this calculation the entire protein is frozen and the vibrational frequencies of the ligand, in the electronic field of the protein, are calculated. These frequencies are then compared to those for the ligand alone to find the change in vibrational entropy. Since the ligand will be a more manageable size than the entire QM region, this makes the minimization requirement easier to meet, as well as reducing the number of atoms in the normal-mode calculation. These predictions were made with the same criteria as above, using the two-step minimization and the ESCF energy of the QM region while varying the manner in which the frequencies were calculated. In the case of these 23 zinc metalloenzyme complexes, using the frequency of just the ligand results in an R^2 of 0.63 with a standard deviation of 1.89 kcal/mol without fitting. This can be compared to the results of the PHVA of the entire QM region, which yielded results of $R^2 = 0.64$ with a standard deviation of 1.88 kcal/mol. With MLR fitting, the correlation coefficient of the ligand-only vibrations increases to 0.72 with a standard deviation of 1.66 kcal/mol, while the frequencies of the full QM region yield a correlation of 0.71 and a standard deviation of 1.69 kcal/mol.

Table 2. Time (min) Required To Calculate the QM Region and the Ligand-Only Vibrational Frequencies with the QM/MM Method for the Protein–Ligand Complex

PDB ID	QM region	ligand-only
1a42	868.54	8.94
1am6	416.45	0.37
1bcd	580.17	0.44
1bn1	804.06	29.05
1bn3	802.76	6.50
1bn4	794.45	6.74
1bnn	818.37	8.38
1bnq	825.55	8.09
1bnt	846.12	7.30
1bnu	785.09	10.03
1bnv	857.38	10.12
1bnw	735.04	5.76
1cbx	378.02	3.43
1cil	797.64	5.57
1cim	721.65	6.68
1cin	759.12	5.96
1cnw	1614.17	18.85
1cnx	866.89	13.80
1cny	1074.23	34.21
3cpa	199.51	4.97
6cpa	576.84	18.70
7cpa	555.22	33.83
8cpa	302.63	16.84
average	738.26	11.50

The unfitted and fitted results demonstrate that the extra information obtained from a full QM PHVA does not favorably impact our ability to predict protein–ligand binding. This result is somewhat unexpected because the full QM PHVA quantifies entropy changes in the protein side chains, but according to these results, these entropy changes do not have a large impact on the predictive ability of the scoring function for this set. Again, this impact may increase as the size of the QM region increases and more side chains are accounted for. However, these results must be viewed in terms of a cost–benefit analysis because calculating the vibrational frequencies can be quite expensive, so it is encouraging that the ligand-only results provide good predictions. Table 2 shows the time needed to calculate only the vibrational frequencies of the protein–ligand complex using the entire QM region and the ligand-only approach.

The range of vibrational entropies calculated here closely matches that from Schwarzl et al.³⁸ with a minimum of -9.85 kcal/mol and a maximum of 1.44 kcal/mol. The ligands in this study contain more rotatable bonds than those used in the study by Schwarzl, accounting for the larger range of entropies found here. Including the protein's vibrational entropy change moves the range of values to a minimum of -13.91 kcal/mol to a maximum of 4.91 kcal/mol. The appearance of this penalty in these vibrational calculations indicates a penalty to binding incurred due to loss of vibrational degrees of freedom in protein side chains. This is similar to the degrees of freedom lost in the ligand, but cannot be fully recovered with the ligand's new vibrational degrees of freedom, often leading to an overall binding penalty depending on the size of the ligand and composition of the side chains. It is also worth noting that the calculation time listed is that necessary for just one component of the vibrational entropy computation. When the protein's QM

Table 3. Average Relative Timings for the Three Main Methods Used in the Study^a

method	relative time	R^2
two minimization runs + QM vibrations	8.83	0.64
two minimization runs + ligand vibrations	4.85	0.63
no minimizations + rotatable bonds	1.00	0.57

^a Times are relative to the no minimizations and rotatable bond entropy estimate.

region is included in the vibrational analysis, an additional calculation must be undertaken to calculate the frequencies of the protein without the ligand and the ligand alone, which are additional steps that are time-consuming. These proteins are also relatively small, so the time would only increase even more with larger QM regions. This is an important consideration in the use of this scoring function, which could easily be modified to include the entire QM region frequencies or those of the ligand alone depending on the computational resources available and the amount of detail required. Overall, this analysis indicates that using the vibrations of the ligand alone may be enough to produce a good correlation, but more accuracy can potentially be gained using the frequencies of the entire QM region.

Relative Timings. To investigate the timing differences between the methods studied above, relative timings were compiled on the basis of average times for the various components. These will illustrate the timing differences between the methods in a manner that does not depend on the current level of technology, but gives an idea of the utility of these methods compared to each other. The relative timings for the QM/MM methods can be seen in Table 3 along with the correlation found using each method.

From Table 3 we can see that the full QM vibration calculation method is very expensive compared to the other two methods. Even the ligand-only vibration is expensive due to the slow minimization steps. This gap could perhaps be shortened by a quicker minimizer or perhaps using only one minimization run instead of two, which would be more acceptable for the ligand-only vibration calculations than the full QM vibrations due to negative frequencies. As the QM portion grows even larger, it could be expected that this scaling would become even worse for the minimization runs, necessitating a more efficient minimizer or skipping the minimization runs altogether and using the rotatable bond entropy estimate. The difference in the correlations is not great between all of these methods, but the timing is, so an appropriate method should be chosen depending on the number and size of the systems to be studied.

Comparison of QM/MM with QM. It is useful to compare the predictive abilities of this QM/MM study with those of the full QM study performed by Raha and Merz.²⁵ Both of these were done using the AM1 semiempirical Hamiltonian on the same set of metalloenzymes. This comparison will allow us to assess the QM/MM method's predictive ability in comparison to the full QM method and further determine the viability of the QM/MM approach for scoring. Raha and Merz reported the results without fitting and by fitting the solvation entropy term in the scoring function and calculated a correlation of 0.69 without fitting,

Table 4. Comparison of Squared Correlation Coefficients (R^2) and Standard Deviations (SDs) between Full QM Results from Raha²⁵ and the QM/MM Method Described Here

method	R^2	SD (kcal/mol)
full QM, no fitting	0.69	1.50
full QM, MLR fitting	0.80	1.18
QM/MM, no fitting	0.64	1.88
QM/MM, MLR fitting	0.71	1.69

while they report a higher correlation of 0.80 when using MLR on only the surface area terms of the organic heavy atoms in the protein. The QM/MM method here obtains a correlation of 0.71 using MLR fitting for all the terms, which is fairly close to the full QM prediction, while the predictions without fitting yield a correlation of 0.64. Not surprisingly, the QM/MM method does not do quite as well as the full QM method, but it is encouraging that it is qualitatively competitive. The correlations and standard deviations of these two studies are compiled in Table 4.

Charge Analysis. An advantage of a full quantum or QM/MM method for these binding affinity calculations is that a QM method will be able to capture polarization and charge transfer (CT) effects. Polarization is generally considered an intramolecular term, representing the internal rearrangement of electrons, whereas charge transfer is an intermolecular term, which is an external source's electronic effects on the protein. CT has been shown to play an important role in binding,^{44,45} and it is interesting to determine the degree of CT that occurs upon binding. When using a classical method, this term would generally be missed or estimated with a parametrized term, but a QM method allows the CT and polarization to have a more physical effect on the binding of the ligand. These CT effects can be especially relevant in a metalloenzyme complex due to the nature of metal ions and their bonds. Table 5 presents the charge transfer that occurs to the ligand in solution, when binding to the protein.

The table shows the charge transfer to the ligand involved in the binding process. CT was shown to contribute significantly to the interaction energy when solvating a protein, and it can safely be assumed that this is also an important factor in protein–ligand binding. Most of the carbonic anhydrase inhibitors gain approximately 0.9 electron upon binding when considering the CM1 charges, with the exception being 1am6, which only gains 0.71 electron. For the carboxypeptidase proteins the ligands give up some of their charge to the protein in all cases but 3cpa. This is due to the nature of the ligands: the ligand for 3cpa has a neutral charge, while the rest of the compounds have a negative charge. A standard molecular mechanics model would not be able to properly represent these, while a higher level QM method would better capture the effects, but might become too costly. This semiempirical QM/MM method provides a compromise for capturing CT effects, allowing them to be included while keeping the overall cost reasonable.

Conclusions

We have presented a QM/MM method to calculate the binding free energy of protein–ligand complexes. This

Table 5. Charge Transfer from the Protein to the Ligand (electrons) for Mulliken and CM1 Charges in Solution

protein type	PDB ID	Mulliken	CM1
CA	1a42	-0.57	-0.87
CA	1am6	-0.60	-0.72
CA	1bcd	-0.64	-0.91
CA	1bn1	-0.68	-0.84
CA	1bn3	-0.58	-0.88
CA	1bn4	-0.58	-0.88
CA	1bnn	-0.58	-0.88
CA	1bnq	-0.57	-0.87
CA	1bnt	-0.58	-0.88
CA	1bnu	-0.58	-0.88
CA	1bnv	-0.57	-0.87
CA	1bnw	-0.58	-0.87
CA	1cil	-0.57	-0.87
CA	1cim	-0.57	-0.87
CA	1cin	-0.61	-0.90
CA	1cnw	-0.58	-0.88
CA	1cnx	-0.58	-0.88
CA	1cny	-0.58	-0.88
CPA	1cbx	0.37	0.35
CPA	3cpa	-0.56	-0.61
CPA	6cpa	0.43	0.41
CPA	7cpa	0.49	0.46
CPA	8cpa	0.27	0.26

method takes advantage of the linear scaling capabilities of DivCon to include a large number of atoms near the ligand in the semiempirical QM region. This allows electronic effects, which would be missed in a classical calculation, to be properly represented while keeping the computational cost low enough to be performed on a large library of protein–ligand complexes. This approach was successful at predicting the binding free energies of a set of 23 zinc metalloenzyme complexes. The scoring function performs well without any fitting, but through multiple linear regression, the function can be fit to experimental data and the predictive performance increases from a squared correlation coefficient of 0.64 to one of 0.71. This may only be the case within a single protein family, and an overall set of weights for general use may be necessary as in empirical scoring functions, but it shows that the method has potential.

This method may also take into account the vibrational entropy change of the ligand upon binding. The QM/MM method allows a unique perspective on this in that a normal-mode analysis can be conducted in the field of the protein's charges while charges further from the ligand remain fixed. This gives a more accurate representation of the vibrational modes available to the ligand, and therefore a more accurate representation of the vibrational entropy contribution to the overall binding affinity of the complex, but is computationally intensive. A simple estimate of counting rotatable bonds was also examined to estimate entropy change as a way to save computation time.

The contributions of various parameters for the predicted binding affinity were also investigated including long-range cutoffs and the use of the total energy of the system versus the QM energy for the heat of interaction. These studies indicated that the long-range cutoff used makes a significant difference in the predicted binding affinity. It was also found that the use of the ESCF energy to calculate the heat of interaction was preferable to the use of the total energy of

the system. Using the ESCF energy, the number of minimization steps does not make as much of an impact, whereas increased minimization cycles greatly improve the predictive ability of the function when the total energy is used.

Although the scoring function presented is relatively good at predicting the binding affinities of zinc metalloenzyme complexes, there is room for improvement. Depending on the acceptable costs, a larger QM region can easily be chosen. This will provide a second shell of residues to interact with the ligand in the QM region, giving a better idea of the electronic effects involved in binding. Not only will this present a better representation of the charge transfer to and from the ligand, it will allow a better picture of the polarization due to the protein environment. Sampling of the system through MD snapshots similar to the MM-PBSA method could also be used to potentially improve the predictions of this scoring function. These snapshots will provide a sampling of the protein, potentially increasing the predictive ability while increasing the cost of the calculations. Using a purely classical simulation, this might present problems because each ligand would need to be parametrized properly, which is quite costly. However, a QM/MM method would allow the parametrization step to be skipped, and a QM region could be formulated to make a sampling of the different configurations tractable. All of the tools for the QM/MM scoring method are easily applied to the snapshots generated by a simulation to emulate the MM-PBSA method. A QM/MM simulation could also be constructed to include only the ligand in the QM region, allowing the protein to be sampled while removing the parametrization needs of the ligand.

Overall, the QM/MM scoring method presents good predictive trends at a reasonable cost, but does present some areas for future improvement. In its current form this method might be more useful verifying ligand poses or as a drug refinement step, but could be made more affordable through parallelization techniques and modification of the parameters used.

Acknowledgment. K.M.M. thanks the NIH (Grant GM044974) and R.D. thanks the NIH (Grants R01 GM84453 (R.D.) and P30 CA006927 (Fox Chase Cancer Center)) for financial support of this work.

References

- (1) Lyne, P. D. *Drug Discovery Today* **2002**, *7*, 1047–1055.
- (2) Jorgensen, W. L. *Science* **2004**, *303*, 1813–1818.
- (3) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. *J. Mol. Graphics Modell.* **2002**, *20*, 281–295.
- (4) Bohm, H. J. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309–323.
- (5) Wang, R.; Lu, Y.; Wang, S. *J. Med. Chem.* **2003**, *46*, 2287–2303.
- (6) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 609–623.
- (7) Trott, O.; Olson, A. J. *J. Comput. Chem.* **2009**, *31*, 455–461.

- (8) Bohm, H. J. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243–256.
- (9) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.
- (10) Gohlke, H.; Hendlich, M.; Klebe, G. *J. Mol. Biol.* **2000**, *295*, 337–356.
- (11) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. *J. Med. Chem.* **2004**, *47*, 3032–3047.
- (12) Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. *J. Med. Chem.* **2005**, *48*, 4040–4048.
- (13) Steinbrecher, T.; Case, D. A.; Labahn, A. *J. Med. Chem.* **2006**, *49*, 1837–1844.
- (14) 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., III. *Acc. Chem. Res.* **2000**, *33*, 889–897.
- (15) Kuhn, B.; Kollman, P. A. *J. Med. Chem.* **2000**, *43*, 3786–3791.
- (16) Woo, H. J.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825–6830.
- (17) Ishchenko, A. V.; Shakhnovich, E. I. *J. Med. Chem.* **2002**, *45*, 2770–2780.
- (18) Brenk, R.; Vetter, S. W.; Boyce, S. E.; Goodin, D. B.; Shoichet, B. K. *J. Mol. Biol.* **2006**, *357*, 1449–1470.
- (19) Irwin, J. J.; Raushel, F. M.; Shoichet, B. K. *Biochemistry* **2005**, *44*, 12316–12328.
- (20) Nemoto, T.; Fedorov, D. G.; Uebayasi, M.; Kanazawa, K.; Kitaura, K.; Komeiji, Y. *Comput. Biol. Chem.* **2005**, *29*, 434–439.
- (21) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T. *Chem. Phys. Lett.* **1999**, *313*, 701–706.
- (22) Dixon, S. L.; Merz, K. M., Jr. *J. Chem. Phys.* **1996**, *104*, 6643–6649.
- (23) Dixon, S. L.; Merz, K. M., Jr. *J. Chem. Phys.* **1997**, *107*, 879–893.
- (24) Zhang, D. W.; Xiang, Y.; Gao, A. M.; Zhang, J. Z. *J. Chem. Phys.* **2004**, *120*, 1145–1148.
- (25) Raha, K.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **2004**, *126*, 1020–1021.
- (26) Raha, K.; Merz, K. M., Jr. *J. Med. Chem.* **2005**, *48*, 4558–4575.
- (27) Raha, K.; Peters, M. B.; Wang, B.; Yu, N.; WollaCott, A. M.; Westerhoff, L. M.; Merz, K. M., Jr. *Drug Discovery Today* **2007**, *12*, 725–731.
- (28) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.
- (29) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700–733.
- (30) Monard, G.; Merz, K. M., Jr. *Acc. Chem. Res.* **1999**, *32*, 904–911.
- (31) Friesner, R. A. *Adv. Protein Chem.* **2005**, *72*, 79–104.
- (32) Senn, H. M.; Thiel, W. *Top. Curr. Chem.* **2007**, *268*, 173–290.
- (33) Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229.
- (34) Peters, M. B.; Raha, K.; Merz, K. M., Jr. *Curr. Opin. Drug Discovery Dev.* **2006**, *9*, 370–379.
- (35) Zhou, T.; Huang, D.; Cafilisch, A. *J. Med. Chem.* **2008**, *51*, 4280–4288.
- (36) Cho, A. E.; Rinaldo, D. *J. Comput. Chem.* **2009**, 2609–2616.
- (37) Grater, F.; Schwarzl, S. M.; Dejaegere, A.; Fischer, S.; Smith, J. C. *J. Phys. Chem. B* **2005**, *109*, 10474–10483.
- (38) Schwarzl, S. M.; Tschopp, T. B.; Smith, J. C.; Fischer, S. *J. Comput. Chem.* **2002**, *23*, 1143–1149.
- (39) Gleeson, M. P.; Gleeson, D. *J. Chem. Inf. Model.* **2009**, *49*, 670–677.
- (40) Fong, P.; McNamara, J. P.; Hillier, I. H.; Bryce, R. A. *J. Chem. Inf. Model.* **2009**, *49*, 913–924.
- (41) Anisimov, V. M.; Bugaenko, V. L. *J. Comput. Chem.* **2008**, 784–798.
- (42) Illingworth, C. J. R.; Morris, G. M.; Parkes, K. E. B.; Snell, C. R.; Reynolds, C. A. *J. Phys. Chem. A* **2008**, *112*, 12157–12163.
- (43) Cho, A. E.; Guallar, V.; Berne, B. J.; Friesner, R. *J. Comput. Chem.* **2005**, *26*, 915–931.
- (44) Garcia-Viloca, M.; Truhlar, D. G.; Gao, J. *J. Mol. Biol.* **2003**, *327*, 549–560.
- (45) Ji, C. G.; Zhang, J. Z. *J. Am. Chem. Soc.* **2008**, *130*, 17129–17133.
- (46) Illingworth, C. J. R.; Parkes, K. E.; Snell, C. R.; Marti, S.; Moliner, V.; Reynolds, C. A. *Mol. Phys.* **2008**, *106*, 1511–1515.
- (47) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712–725.
- (48) Field, M. J.; Albe, M.; Bret, C.; Proust-De Martin, F.; Thomas, A. *J. Comput. Chem.* **2000**, *21*, 1088–1100.
- (49) Walker, R. C.; Crowley, M. F.; Case, D. A. *J. Comput. Chem.* **2008**, *29*, 1019–1031.
- (50) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535–542.
- (51) 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.
- (52) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- (53) Fischer, S.; Smith, J. C.; Verma, C. S. *J. Phys. Chem. B* **2001**, *105*, 8050–8055.
- (54) Chang, C. E. A.; Chen, W.; Gilson, M. K. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1534–1539.
- (55) Murray, C. W.; Verdonk, M. L. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 741–753.
- (56) Calvin, M. D.; Head, J. D.; Jin, S. Q. *Surf. Sci.* **1996**, *345*, 161–172.
- (57) Zou, X. Q.; Sun, Y. X.; Kuntz, I. D. *J. Am. Chem. Soc.* **1999**, *121*, 8033–8043.
- (58) Hayik, S. A.; Liao, N.; Merz, K. M., Jr. *J. Chem. Theory Comput.* **2008**, *4*, 1200–1207.
- (59) Archontis, G.; Simonson, T.; Karplus, M. *J. Mol. Biol.* **2001**, *306*, 307–327.
- (60) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87–110.
- (61) Vajda, S.; Weng, Z. P.; Rosenfeld, R.; Delisi, C. *Biochemistry* **1994**, *33*, 13977–13988.

- (62) Bardi, J. S.; Luque, I.; Freire, E. *Biochemistry* **1997**, *36*, 6588–6596.
- (63) Velec, H. F.; Gohlke, H.; Klebe, G. *J. Med. Chem.* **2005**, *48*, 6296–6303.
- (64) Legrand, S. M.; Merz, K. M., Jr. *J. Comput. Chem.* **1993**, *14*, 349–352.
- (65) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639–1662.
- (66) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. *J. Comput. Chem.* **1992**, *13*, 505–524.

CT100315G

JCTC

Journal of Chemical Theory and Computation

Antisymmetric Magnetic Interactions in Oxo-Bridged Copper(II) Bimetallic Systems

R. Maurice,^{†,‡} A. M. Pradipto,[§] N. Guihéry,^{*,†} R. Broer,[§] and C. de Graaf^{*,||,‡}

Laboratoire de Chimie et Physique Quantiques, Université de Toulouse 3, 118, route de Narbonne, 31062 Toulouse France, Departament de Química Física i Inorgànica, Universitat Rovira i Virgili, Marcel·lí Domingo s/n, 43007 Tarragona, Spain, Zernike Institute for Advanced Materials, University of Groningen, Groningen 9747AG, The Netherlands, and Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010, Barcelona, Spain

Received June 15, 2010

Abstract: The antisymmetric magnetic interaction is studied using correlated wave-function-based calculations in oxo-bridged copper bimetallic complexes. All of the anisotropic multispin Hamiltonian parameters are extracted using spin–orbit state interaction and effective Hamiltonian theory. It is shown that the methodology is accurate enough to calculate the antisymmetric terms, while the small symmetric anisotropic interactions require more sophisticated calculations. The origin of the antisymmetric anisotropy is analyzed, and the effect of geometrical deformations is addressed.

1. Introduction

The combined effect of spin–orbit coupling (SOC) and spin–spin coupling (SSC) can lead to magnetic anisotropy without the necessity of applying an external magnetic field. In addition to the presence of unpaired electrons, the system should be not too symmetric to present measurable magnetic anisotropy effects and to avoid the presence of unquenched orbital momentum.^{1,2} This effect has been encountered in organic molecules,³ monometallic transition metal complexes,^{4,5} single-molecule magnets (SMMs),⁶ and extended materials related to the cuprate high- T_c superconductors.^{7,8} Antisymmetric interactions were introduced phenomenologically in 1958 by Dzyaloshinskii⁹ to describe the magnetic properties of α -Fe₂O₃. The theory was generalized by Moriya two years later,¹⁰ leading to the well-known standard multispin Hamiltonian for binuclear systems with $S = 1/2$ magnetic centers:^{2,4}

$$\hat{H} = J\hat{S}_a \cdot \hat{S}_b + \hat{S}_a \bar{D} \hat{S}_b + \vec{d} \hat{S}_a \times \hat{S}_b \quad (1)$$

This model involves an isotropic exchange term J , a symmetric zero field splitting (ZFS) tensor \bar{D} , and the antisymmetric part that is described by the Dzyaloshinskii–Moriya (DM) pseudovector \vec{d} . Interpretation of experimental data led to detailed information about spin canting in copper oxides and explained the origin of the weak ferromagnetism in some of the crystallographic phases despite the strong antiferromagnetic isotropic exchange. Recently, the norm of the DM vector was determined in SrCu₂(BO₃)₂ through electron paramagnetic resonance (EPR) spectroscopy.^{11,12} On the basis of the perturbational approach outlined by Moriya, there have been many attempts to rationalize the anisotropic interaction between two Cu²⁺ ions.^{13–16} A systematic overview has recently been published by Moskvin.¹⁷

Until now, the *ab initio* study of magnetic anisotropy has been mainly limited to the monometallic complexes or the symmetric terms in polynuclear systems. One of the first anisotropy calculations was performed on a titanium bimetallic complex, combining the complete active space self-consistent field (CASSCF) approach, multireference perturbation theory (MRPT), and effective nuclear charge SOC calculations.¹⁸ The implementation of SOC in the NRLMOL code^{19,20} triggered a major breakthrough in the application of density functional theory (DFT) to the magnetic anisotropy

* Corresponding author e-mail: nathalie.guihery@irsamc.ups-tlse.fr (N.G.); coen.degraaf@urv.cat (C.d.G.).

[†] Université de Toulouse 3.

[‡] Universitat Rovira i Virgili.

[§] University of Groningen.

^{||} ICREA.

in polymetallic SMMs.^{21–27} An *ab initio* treatment of SSC was presented by Vahtras et al.²⁸ Another important contribution was made by Neese with the implementation of SOC and SSC in the ORCA code.²⁹ This implementation not only allows the study of anisotropy with DFT but also paved the way for the use of wave-function-based methodologies.^{30–32} The latter methods have been applied successfully to several monometallic transition metal complexes and organic systems.^{33–38} A similar treatment of SOC was implemented in the MOLCAS code³⁹ based on the restricted active space state interaction spin–orbit (RASSI-SO) scheme,^{40,41} which was used to study ZFS phenomena in transition metal complexes.^{42–49} Finally, we mention the work of Gilka et al. on SSC⁵⁰ and the ZFS calculations on organic molecules by Sugisaki and co-workers.⁵¹

The binuclear copper(II) acetate complex described by Bleaney and Bowers⁵² is one of the first examples of a polynuclear system with anisotropic interactions and has been the subject of several studies.^{53–55} Since this complex presents a center of inversion, only symmetric interactions are allowed, and we will treat the anisotropy of this system in a separate study. Antisymmetric interactions in binuclear complexes are less common and difficult to probe by EPR spectroscopy.⁵⁶ Some synthetic complexes were proposed by Kahn to present antisymmetric interactions,⁵⁷ but the only clear evidence of DM interaction in a bimetallic complex was found in a diferric complex and has required the use of Mössbauer spectroscopy.⁵⁸ Important antisymmetric interactions have also been evidenced in trimetallic copper(II) complexes by both magnetic circular dichroism (MCD) and EPR spectroscopies.⁵⁹

Recently, a new extraction method of anisotropic parameters has been proposed⁴⁷ that is based on effective Hamiltonian theory.^{60,61} The method establishes a simple procedure to determine the ZFS parameters and the magnetic anisotropic axes. In addition, the method can be used to validate existing model Hamiltonians to describe the magnetic anisotropy. The standard multispin Hamiltonian for centrosymmetric bimetallic systems was found to be incomplete, lacking a non-negligible biquadratic anisotropic interaction term.⁴⁸

To add a new aspect to the understanding of the anisotropic interactions between two Cu(II) ions bridged by a diamagnetic bridge, we apply the new extraction method to a Cu(II) model complex that mimics the Cu–O–Cu units present in copper oxides and that is also relevant to molecular polynuclear Cu(II) complexes. The application of *ab initio* calculations and subsequent mapping on a model Hamiltonian through effective Hamiltonian theory allows us to extract all parameters of the general spin Hamiltonian written in eq 1 and to investigate the mechanism of anisotropy without any assumption. For example, we do not assume that the anisotropy solely arises from the interaction of the fundamental singlet and triplet with excited states. In addition, we determine the relative importance of all of the terms that were described by Moskvin by means of a decomposition of the *ab initio* wave function and study the effect on the anisotropy of the bending of the central Cu–O–Cu bond ϑ_1 and the twisting of the two CuO₄ planes defined as the dihedral angle ϑ_2 shown in Figure 1.

2. Theory and Methodology

2.1. Spin Hamiltonian in Copper(II) Bimetallic Systems.

We start our analysis with the derivation of the 4×4 interaction matrix spanned by the singlet and triplet $|S, M_S\rangle$ determinants. For convenience, we rewrite eq 1 by grouping the symmetric and antisymmetric anisotropic interaction in a single second-order tensor T :

$$\hat{H} = J\hat{S}_a \cdot \hat{S}_b + \hat{S}_a \bar{T} \hat{S}_b \quad (2)$$

The easiest way to proceed is to build the model interaction matrix in the uncoupled $|M_{S_a}, M_{S_b}\rangle$ basis taking into account all possible interactions in an arbitrary axis frame. M_{S_a} and M_{S_b} are the M_S components of the local doublets on centers a and b , respectively.

$$\hat{H}_{\text{mod}} \begin{array}{c} \left| -\frac{1}{2}, -\frac{1}{2} \right\rangle \\ \left| -\frac{1}{2}, \frac{1}{2} \right\rangle \\ \left| \frac{1}{2}, -\frac{1}{2} \right\rangle \\ \left| \frac{1}{2}, \frac{1}{2} \right\rangle \end{array} \begin{array}{c} \left| -\frac{1}{2}, -\frac{1}{2} \right\rangle \\ \left| -\frac{1}{2}, \frac{1}{2} \right\rangle \\ \left| \frac{1}{2}, -\frac{1}{2} \right\rangle \\ \left| \frac{1}{2}, \frac{1}{2} \right\rangle \end{array} \begin{array}{c} \left| -\frac{1}{2}, \frac{1}{2} \right\rangle \\ \left| \frac{1}{2}, -\frac{1}{2} \right\rangle \\ \left| \frac{1}{2}, \frac{1}{2} \right\rangle \\ \left| -\frac{1}{2}, -\frac{1}{2} \right\rangle \end{array} \begin{array}{c} \left| \frac{1}{2}, -\frac{1}{2} \right\rangle \\ \left| \frac{1}{2}, \frac{1}{2} \right\rangle \\ \left| -\frac{1}{2}, -\frac{1}{2} \right\rangle \\ \left| -\frac{1}{2}, \frac{1}{2} \right\rangle \end{array}$$

$$\begin{array}{cccc} \frac{1}{4}(J + T_{33}) & -\frac{1}{4}(T_{31} + iT_{32}) & -\frac{1}{4}(T_{13} + iT_{23}) & \frac{1}{4}[T_{11} - T_{22} + i(T_{12} + T_{21})] \\ -\frac{1}{4}(T_{31} - iT_{32}) & -\frac{1}{4}(J + T_{33}) & \frac{1}{2}J + \frac{1}{4}[T_{11} + T_{22} + i(T_{21} - T_{12})] & \frac{1}{4}(T_{13} + iT_{23}) \\ -\frac{1}{4}(T_{13} - iT_{23}) & \frac{1}{2}J + \frac{1}{4}[T_{11} + T_{22} - i(T_{21} - T_{12})] & -\frac{1}{4}(J + T_{33}) & \frac{1}{4}(T_{31} + iT_{32}) \\ \frac{1}{4}[T_{11} - T_{22} - i(T_{12} + T_{21})] & \frac{1}{4}(T_{13} - iT_{23}) & \frac{1}{4}(T_{31} - iT_{32}) & \frac{1}{4}(J + T_{33}) \end{array}$$

In a second step, the model matrix is transformed to the coupled $|S, M_S\rangle$ basis for a more straightforward understanding of the interactions:

$$\begin{array}{ccccc}
 \hat{H}_{\text{mod}} & |1, -1\rangle & |1, 0\rangle & |1, 1\rangle & |0, 0\rangle \\
 \langle 1, -1| & \frac{1}{4}(J + T_{33}) & -\frac{\sqrt{2}}{8}[T_{13} + T_{31} + i(T_{23} + T_{32})] & \frac{1}{4}[T_{11} - T_{22} + i(T_{12} + T_{21})] & -\frac{\sqrt{2}}{8}[T_{13} - T_{31} + i(T_{23} - T_{32})] \\
 \langle 1, 0| & -\frac{\sqrt{2}}{8}[T_{13} + T_{31} - i(T_{23} + T_{32})] & \frac{1}{4}(J + T_{11} + T_{22} - T_{33}) & \frac{\sqrt{2}}{8}[T_{13} + T_{31} + i(T_{23} + T_{32})] & -\frac{i}{4}(T_{12} - T_{21}) \\
 \langle 1, 1| & \frac{1}{4}[T_{11} - T_{22} - i(T_{12} + T_{21})] & \frac{\sqrt{2}}{8}[T_{13} + T_{31} - i(T_{23} + T_{32})] & \frac{1}{4}(J + T_{33}) & -\frac{\sqrt{2}}{8}[T_{13} - T_{31} - i(T_{23} - T_{32})] \\
 \langle 0, 0| & -\frac{\sqrt{2}}{8}[T_{13} - T_{31} - i(T_{23} - T_{32})] & \frac{i}{4}(T_{12} - T_{21}) & -\frac{\sqrt{2}}{8}[T_{13} - T_{31} + i(T_{23} - T_{32})] & \frac{1}{4}(-3J - T_{11} - T_{22} - T_{33})
 \end{array}$$

The symmetric and antisymmetric contributions (D_{ij} and d_{ij} , respectively) can be separated as follows:

$$\begin{aligned}
 D_{ii} &= T_{ii} \\
 D_{ij} &= D_{ji} = \frac{1}{2}(T_{ij} + T_{ji}) \\
 d_{ij} &= -d_{ji} = \frac{1}{2}(T_{ij} - T_{ji})
 \end{aligned} \quad (3)$$

From this, it is clear that the antisymmetric interactions arise from the $\langle S, M_S | H_{\text{mod}} | S', M_S' \rangle$ matrix elements and cause a direct coupling between the singlet and triplet states, which is absent in the case of symmetric interactions only. Finally, we mention that the antisymmetric second order tensor $\underline{\underline{d}}$ can be reduced to a pseudovector with the following components:

$$d_x = d_{23} \quad d_y = -d_{13} \quad d_z = d_{12} \quad (4)$$

2.2. Description of the Models and Computational Information. The model complex used in the calculations consists of a Cu–O–Cu central part using H₂O ligands to complete the coordination sphere of the copper ions. The Cu–O distances have been fixed to 2.00 Å and the O–H distances fixed to 0.96 Å, while the ϑ_1 and ϑ_2 angles are susceptible to changes. The symmetry rules for the appearance of both symmetric and antisymmetric interactions are well-known and reported in the literature.⁶² In the case where $\vartheta_1 = \vartheta_2 = 0^\circ$, the complex has an inversion center, and hence, only symmetric anisotropic interactions are allowed. When ϑ_1 is changed, the symmetry is lowered from D_{2h} to C_{2v} and a DM vector appears along the z axis. The twist of the CuO₄ planes ($\vartheta_2 \neq 0^\circ$) lowers the symmetry to D_2 and induces a DM vector aligned along the x axis. However, the interaction is strictly zero when the planes are orthogonal ($\vartheta_1 = 0^\circ$, $\vartheta_2 = 90^\circ$) and the molecule has D_{2d} symmetry. When both distortions are present, the point group symmetry is C_2 with just a 2-fold rotation axis along the y axis. In this case, the DM vector lies in the xz plane.

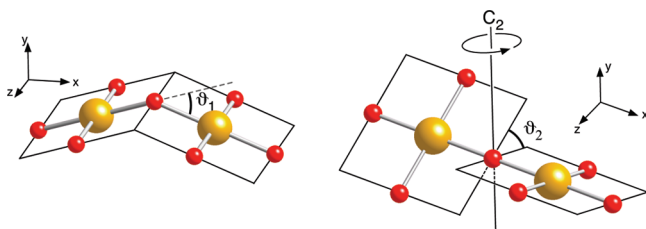


Figure 1. Schematic representation of the distortions applied to the model complex. Large spheres represent copper, and smaller spheres are oxygens.

The SSC is an important mechanism to describe anisotropy when the ZFS is on the order of a few cm^{-1} . However, it does not lead to antisymmetric interactions, as it cannot directly couple triplet with singlet states.⁶³ As our main objective of this study concerns the description of the DM interaction, that is, the effective coupling between singlet and triplet states, only the SOC has been considered. We follow a two-step procedure implemented in Molcas 7 to obtain accurate estimates of the exact N -electron wave function that account for dynamic electron correlation and spin–orbit interactions.

First, a number of spin–orbit free states is computed via the CASSCF method using the Douglas–Kroll–Hess Hamiltonian.^{64,65} The active space contains all Cu-3d orbitals and the corresponding 18 electrons. CASSCF wave functions can be defined for all 25 singlet and triplet states of the d^9 – d^9 manifold excluding the metal-to-metal charge transfer states. Second, the spin–orbit coupling is introduced *a posteriori* via the RASSI-SO method.^{40,41} This method uses the mean-field approximation and the one-center approximation,⁶⁶ through the so-called atomic-mean field integrals (AMFI).^{67,68} Dynamic correlation effects can be introduced by replacing the diagonal elements of the spin–orbit matrix by CASPT2 energies^{69,70} using the CAS(18,10)SCF wave function as a reference. Following the conclusions of a previous work on the magnetic coupling,⁷¹ the IPEA shift of the CASTP2 zeroth-order Hamiltonian is set to zero.^{72,73} An imaginary level shift of 0.2 hartree was applied to avoid the appearance of intruder states in the perturbational treatment of dynamical electron correlation.⁷⁴ The following ANO-RCC basis set⁷⁵ was used: Cu (6s 5p 4d 2f), O (4s 3p 1d), and H (2s).

In addition to the large CAS(18,10), we also performed calculations with a minimal CAS(2,2) containing the magnetic orbitals only, in which we can just define the ground state singlet and triplet states.

2.3. Extraction of Spin Hamiltonian Parameters. The interaction matrix presented in section 2.1 contains 10 parameters. This number is reduced to seven when the molecule is oriented in such a way that the magnetic axes frame coincides with the Cartesian axes frame. Since the model space is spanned by the four $|M_S\rangle$ components of the singlet and triplet, it is not possible to determine all the parameters from the energy differences only. For the same reason, it is also complicated to extract both symmetric and antisymmetric interactions in the general case from an experiment, for instance from EPR spectra.⁵⁶ The required extra information is contained in the wave function of the spin–orbit states and is used to construct an effective Hamiltonian⁶⁰ that allows us to extract all 10 parameters.

The effective Hamiltonian is determined by projecting the states that span the SO–SI space onto the model space. The projections with the largest norm are orthonormalized by the procedure of des Cloizeaux,⁶¹ and the matrix elements of the effective Hamiltonian are calculated by applying the following formula:

$$\langle \Phi_i | \hat{H}^{\text{eff}} | \Phi_j \rangle = \langle \Phi_i | \sum_{k=1}^4 |\tilde{\Psi}_k\rangle E_k \langle \tilde{\Psi}_k | \Phi_j \rangle \quad (5)$$

where $\Phi_{i,j}$ are the four M_S components arising from the singlet and triplet spanning the model space, $\tilde{\Psi}_k$ represents the orthonormalized projections of the *ab initio* wave functions, and E_k represents the corresponding energies. A more comprehensive description of the application of the effective Hamiltonian theory to extract anisotropy parameters can be found in refs 47 and 48. The comparison of these numerical matrix elements with those of the model Hamiltonian of section 2.1 leads to nine independent equations. The axial and rhombic anisotropy parameters D and E are usually defined in the magnetic axes frame as

$$\begin{aligned} D &= D_{zz} - \frac{1}{2}(D_{xx} + D_{yy}) = \frac{3}{2}D_{zz} \\ E &= \frac{1}{2}(D_{xx} - D_{yy}) \end{aligned} \quad (6)$$

The magnetic axes frame is obtained by diagonalizing the symmetric ZFS tensor and by applying the standard conventions of molecular magnetism that $|D| > 3E > 0$. While the first convention ($|D| > 3E$) fixes the attribution of the z magnetic axis as the hard or easy axis of magnetization, the second one fixes the attribution of the magnetic axes x and y by imposing E to be positive. Then, the magnetic axes frame is univocally defined.

The one-by-one comparison of the model and effective Hamiltonian establishes a way to determine the ability of the model Hamiltonian to describe the electronic interactions of the exact Hamiltonian used to obtain the *ab initio* results. This strategy revealed the existence of higher-order anisotropic interactions in bimetallic complexes with $S \geq 1$ magnetic centers.⁴⁸ Since these interactions cannot occur for the Cu(II) dimer under study, the only possible source of discrepancy between the effective and model Hamiltonian is the presence of orbital degeneracy. The model Hamiltonian should imply both spin and orbital degrees of freedom in the latter case.⁷⁶ However, the coordination of the Cu(II) ions leads to a nondegenerate ground state, and it is expected that the standard multispin Hamiltonian of eq 1 accurately accounts for the anisotropy.

3. Results and Discussion

3.1. Validation of the Spin Hamiltonian. To address the validity of the standard spin Hamiltonian and to illustrate the extraction procedure, one example will be presented in certain detail. The numbers listed in this section are calculated for the structure with $\vartheta_1 = \vartheta_2 = 45^\circ$, but the conclusions are also valid for the other structures discussed afterward. All numbers presented in the text and equations of this paragraph are given in cm^{-1} unless specified

otherwise. The RASSI-SO calculation is performed with a SI space of 25 triplet and 25 singlet spin-free states. The wave function of these states is obtained through a CAS(18/10)SCF calculation. The diagonal matrix elements of the SI matrix correspond to the CASSCF energies of the respective states. The norm of the projection onto the model space of the four low-lying spin–orbit states is approximately 98%. Hence, the model space of the spin Hamiltonian is perfectly adequate in this case. Then, the effective Hamiltonian matrix is calculated applying eq 5:

$$\begin{array}{ccccc} \hat{H}_{\text{eff}} & |1, -1\rangle & |1, 0\rangle & |1, 1\rangle & |0, 0\rangle \\ \langle 1, -1| & 50.557 & 0.024 & 0.168 & 0.657i \\ \langle 1, 0| & 0.024 & 49.781 & -0.024 & 7.015i \\ \langle 1, 1| & 0.168 & 0.024 & 50.557 & -0.657i \\ \langle 0, 0| & -0.657i & 7.015i & 0.657i & 1.006 \end{array} \quad (7)$$

The term-by-term comparison of this matrix with the model matrix presented in section 2.1 fully validates the model Hamiltonian. The effective Hamiltonian matrix does not show any deviation with respect to the model matrix. Hence, we can proceed to the extraction of the parameters of the model Hamiltonian. The trace of the effective Hamiltonian is arbitrary. As the aim of the model Hamiltonian is to reproduce the relative energies of the low-lying magnetic spectrum, we set the energy of the lowest lying eigenstate of the effective Hamiltonian to zero for convenience. The symmetric ZFS tensor is considered traceless, allowing the extraction of the 10 parameters of the model Hamiltonian. The extracted J value is 49.3 cm^{-1} , and the ZFS tensor is

$$\bar{T} = \begin{pmatrix} -0.181 & 14.030 & -0.068 \\ -14.030 & -0.853 & -1.858 \\ -0.068 & 1.858 & 1.034 \end{pmatrix} \quad (8)$$

The symmetric and antisymmetric parts are then separated:

$$\bar{D} = \begin{pmatrix} -0.181 & 0 & -0.068 \\ 0 & -0.853 & 0 \\ -0.068 & 0 & 1.034 \end{pmatrix} \quad (9)$$

$$\bar{d} = \begin{pmatrix} 0 & 14.030 & 0 \\ -14.030 & 0 & -1.858 \\ 0 & 1.858 & 0 \end{pmatrix} \quad (10)$$

Since the C_2 rotation axis coincides with the Cartesian y axis, D_{12} , D_{23} , and d_{13} are zero.⁶² The antisymmetric second-order ZFS tensor can be reduced to a pseudovector $\vec{d} = (-1.858, 0.0, 14.030)$ with a norm of 14.15 cm^{-1} . Since only its orientation and norm can be determined, the DM vector is a so-called pseudovector.

Next, we diagonalize the symmetric ZFS tensor to obtain the magnetic anisotropy axes. Taking care of the usual conventions for the definition of the x , y , and z magnetic anisotropy axes, we obtain $D_{xx} = -0.184$, $D_{yy} = -0.853$, and $D_{zz} = 1.038$.

The magnetic y axis corresponds to the Cartesian y axis (i.e., the C_2 symmetry axis), whereas the magnetic x and z axes nearly coincide with the Cartesian axes. The DM vector can be re-expressed in the magnetic axes frame in order to define its orientation with respect to the anisotropy axes: \vec{d}

Table 1. Spin-Free and RASSI-SO J Parameter (cm^{-1}) for Several Model Geometries^a

$\vartheta_1 = \vartheta_2$	J (spin-free)		J (RASSI-SO)	
	CASSCF	CASPT2	CASSCF	CASPT2
0°	153	562	151	557
15°	138	515	136	509
45°	50	236	49	232
75°	-18	-19	-17	-20
90°	-20	-97	-21	-96

^a RASSI-SO calculations were performed with 25 triplet and 25 singlet spin-free states. The energies of the spin-free states are calculated with CASSCF and CASPT2 using a CAS(18,10).

$= (-1.070, 0.0, 14.120)$. As expected from symmetry arguments, the DM vector is perpendicular to the C_2 axis, it lies in the xz plane. It makes an angle of -4.3° with the magnetic z axis.

In short, we have shown that the standard multispin Hamiltonian is valid for the Cu(II) dimer and that all ZFS parameters and magnetic axis can be extracted from the *ab initio* calculations in a straightforward manner. It remains to be determined how robust these extracted parameters are against the details of the computational scheme. Ideally, the extracted parameters should not be too sensitive to these degrees of freedom as is the case for the ZFS parameters in the mono- and bimetallic complexes studied before.^{47,48}

3.2. Dynamic Correlation Effect on Spin Hamiltonian Parameters. The isotropic magnetic coupling parameter J can be extracted either at the spin-free level or after a RASSI-SO calculation. As the magnetic triplet and singlet states interact differently with the excited states, the RASSI-SO extracted J value can be different from the spin-free extraction. However, as can be seen in Table 1, this effect is nearly negligible. As expected, J is large and antiferromagnetic for the undistorted complex ($\vartheta_1 = \vartheta_2 = 0$) but quickly decreases with the deformations. In fact, the main effect comes from the ϑ_1 deformation angle that induces the change from a large antiferromagnetic coupling (J positive) to a moderate ferromagnetic one (J negative). This result is in agreement with the Goodenough–Kanamori–Anderson rules.^{77–79} Dynamic correlation strongly affects the isotropic coupling, as observed in many other studies present in the literature.

The symmetric part of the ZFS tensor is determined by the $\langle 1, M_S | \hat{H} | 1, M_S \rangle$ and $\langle 1, M_S | \hat{H} | 1, M_S' \rangle$ terms of the effective Hamiltonian, while the antisymmetric part is determined by the $\langle S, M_S | \hat{H} | S', M_S \rangle$ and $\langle S, M_S | \hat{H} | S', M_S' \rangle$ terms. Moreover, the isotropic coupling can be extracted from the difference between the barycenter of the $\langle 1, M_S | \hat{H} | 1, M_S \rangle$ type terms and the $\langle 0, 0 | \hat{H} | 0, 0 \rangle$ term of the effective Hamiltonian. Hence, all different terms (symmetric, antisymmetric, and isotropic) are rigorously separated in the extraction. The extracted symmetric part can then only be affected computationally by changing the magnitude of mechanisms that affects directly the symmetric terms, i.e., by changing the relative energies of the excited spin–orbit free states with respect to the magnetic triplet ground state. Hence, the effect of dynamic correlation on the ZFS parameters D and E passes through the correction of these relative energies and not through the correction of the J value. Table 2 shows how the dynamic

Table 2. Symmetric Anisotropy Parameters D and E (in cm^{-1}) for Several Model Geometries Extracted from the RASSI-SO Calculations with 25 Triplet and 25 Singlet Spin-Free States^a

$\vartheta_1 = \vartheta_2$	CASSCF		CASPT2	
	D	E	D	E
0°	-0.81	0.04	-0.45	0.01
15°	-1.52	0.02	-0.71	0.14
45°	1.56	0.33	-1.66	0.20
75°	3.38	0.10	-3.27	0.57
90°	4.60	0.27	-4.10	1.22

^a The use of CAS(18,10)SCF energies for the spin-free states is compared to the use of CASPT2 energies.

correlation strongly affects D and E . The most obvious manifestation of this effect is the fact that the sign of D is changed using CASPT2 energies in three cases. This change of sign causes a reorientation of the magnetic axis frame in which the roles of the magnetic x and z axes are interchanged.

It should be noted that the extracted values are small and that the sign of D and the orientation of the magnetic axes frame obtained with the computational approach outlined in the previous section should be benchmarked against calculations at a higher level of theory. Probably, the RASSI-SO matrix elements have to be calculated with wave functions that account for dynamic correlation (i.e., beyond the presently used CASSCF wave functions). Eventually, the CASPT2 spin-free energies should also be replaced by energies obtained with variational techniques as the difference dedicated CI method. Such a study is currently being performed for the copper acetate molecule and will be the subject of another publication.

We will now show that, contrary to the symmetric interactions, the antisymmetric part of the anisotropy tensor is robust against the inclusion of dynamic correlation. The DM interaction can be split into two parts. The first one is the direct coupling between the magnetic triplet and singlet by spin–orbit interaction, the first-order contribution to the DM interaction. Being an off-diagonal element of the model Hamiltonian (see section 2.1), this interaction is hardly affected by the inclusion of dynamic correlation. The second part includes all mechanisms involving excited states, which appear in second-order perturbation theory. The effect of these mechanisms on the DM interaction depends directly on the changes in the excitation energies due to dynamic correlation.

Comparison of the results obtained with CAS(18,10)SCF and CASPT2 listed in Table 3 shows that the dynamic correlation effect influences the norm and the orientation of the DM vector in a modest way. The largest change in the angle is observed for $\vartheta_1 = \vartheta_2 = 75^\circ$, for which φ changes by $\sim 25^\circ$. Although this may indicate a rather drastic change at first sight, the effect is not so large if we compare the d_x and d_z components of the DM vector with and without taking into account the dynamic correlation. Using CASSCF energies, d_z and d_x are -4.8 and 1.4 cm^{-1} , respectively. These values change to -6.7 and -0.8 cm^{-1} when electron correlation is taken into account by CASPT2. Hence, by no means do we observe the drastic changes that occur for the symmetric anisotropy, and we conclude that the effect of

Table 3. Norm of the DM Vector $|\vec{d}|$ (in cm^{-1}) and Angle φ (in deg) of the DM Vector with the Cartesian z Axis for Several Model Geometries^a

$\vartheta_1 = \vartheta_2$	CAS(2,2)/small RASSI-SO		CAS(18,10)/large RASSI-SO		CASPT2	
	$ \vec{d} $	φ	$ \vec{d} $	φ	$ \vec{d} $	φ
0°	0.00	0.0	0.00	0.0	0.00	0.0
15°	8.35	-0.7	6.98	0.7	9.77	17.2
45°	17.58	-8.5	14.15	-7.5	17.75	7.4
75°	7.78	-17.1	4.97	-16.5	6.76	6.6
90°	7.58	-16.5	7.32	-15.3	6.75	-28.3

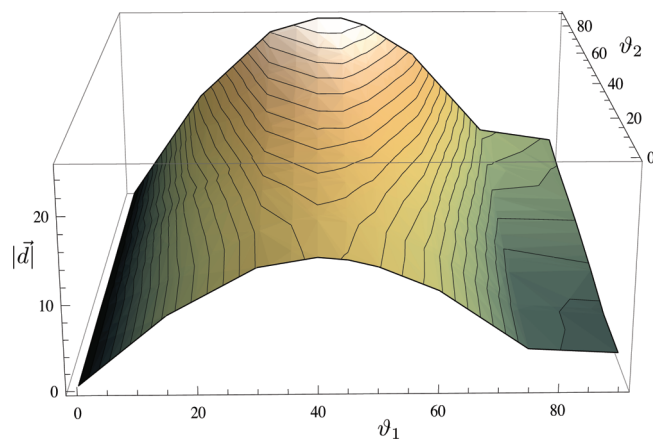
^a The small RASSI-SO space is spanned by the fundamental singlet and triplet states. The large RASSI-SO space contains 25 triplet and 25 singlet spin-free states. The energies of the spin-free states are calculated with CAS(2,2)SCF, CAS(18,10)SCF, and CASPT2.

dynamic correlation is not essential for a semiquantitative description of the DM vector in the Cu–O–Cu system.

To separate the first-order mechanisms involving the singlet and triplet ground states from the second-order mechanisms involving excited states, we performed additional CASSCF/RASSI-SO calculations in which the effect of excited states is completely eliminated. This can be achieved by reducing the active space to two orbitals with two electrons and building the RASSI-SO matrix in the space spanned by the four $|M_S\rangle$ components of the singlet and triplet ground states. Table 3 compares the results obtained with the small RASSI-SO space using the CAS(2,2)SCF wave functions and energies to those obtained with the large CAS and large RASSI-SO space used so far. Again, we observe that the essentials of the DM interaction are maintained. This means that the leading mechanism for antisymmetric anisotropy in Cu–O–Cu-based systems is the direct coupling between the singlet and triplet ground states and that the second-order processes involving excited N -electron states have a smaller effect. In the following, we will apply CAS(2,2)SCF calculations followed by RASSI-SO calculations involving only the singlet and triplet ground states to study in more detail the effect of geometrical deformations and analyze the mechanism of the DM interaction.

3.3. Magneto-Structural Correlations. To complete the study of the geometrical distortion, we calculated the norm of the DM vector as a function of the Cu–O–Cu bending (ϑ_1) and the twist angle of the two CuO_4 planes (ϑ_2). The results are shown in Figure 2. The ϑ_1 deformation leads to the C_{2v} point group symmetry with the DM vector oriented along the Cartesian z axis. As can be seen in Figure 2, this deformation creates a large DM interaction with a maximum of 14.8 cm^{-1} for $\vartheta_1 = 40^\circ$. The DM vector is obviously zero for 0° and 3.6 cm^{-1} for the other extreme when $\vartheta_1 = 90^\circ$.

The start and end points of the twist deformation ($\vartheta_2 = 0^\circ$ and 90°) do not show any DM interaction due to symmetry reasons.⁶² In between, there is a small DM vector along the C_2 axis that connects the two magnetic centers. However, according to Figure 2, this deformation on its own does not create any significant antisymmetric anisotropy. The combination of the two deformations leads to an important synergistic effect. The highest norm of the DM vector is 25.5 cm^{-1} and occurs for $\vartheta_1 = 45^\circ$ and $\vartheta_2 = 90^\circ$.

**Figure 2.** Norm of the DM vector (in cm^{-1}) as function of the ϑ_1 and ϑ_2 deformation angles (in deg) obtained at the CAS(2/2)/RASSI-SO level.

The observed behavior is apparently not due to one single, simple mechanism but to the sum of several, complementary or opposing, mechanisms. In the next section, we will describe the origin of the dominant mechanisms that lead to DM interaction and relate the findings to the shape of the surface shown in Figure 2.

3.4. Description of the Dominant Mechanisms. The results described in the previous section strongly suggest that the dominant mechanisms leading to DM interactions occur at the first order of perturbation, that is, a direct coupling between the magnetic singlet and triplet spin-free states via spin–orbit coupling. Some of these mechanisms have been described in the literature,¹⁷ but here we complete the analysis and classify the different mechanisms by increasing importance.

The CASSCF wave functions of the triplet and singlet state are

$$\begin{aligned}
 |1, 1\rangle &= |\phi_s \phi_a\rangle \\
 |1, 0\rangle &= [|\phi_s \bar{\phi}_a\rangle - |\phi_a \bar{\phi}_s\rangle] / \sqrt{2} \\
 |1, -1\rangle &= |\bar{\phi}_s \bar{\phi}_a\rangle \\
 |0, 0\rangle &= \lambda |\phi_s \bar{\phi}_s\rangle - \mu |\phi_a \bar{\phi}_a\rangle
 \end{aligned} \tag{11}$$

in which all the doubly occupied orbitals are omitted for clarity. The symmetric and antisymmetric molecular orbitals ϕ_s and ϕ_a are mainly localized on the Cu atoms but have important tails on the bridging oxygen, as shown in Figure 3. The composition of the CASSCF magnetic orbitals is as follows:

$$\begin{aligned}
 \phi_s &= \sum_i c_i [3d_i(1) \pm 3d_i(r)] + c_y 2p_y + \dots \\
 \phi_a &= \sum_i c'_i [3d_i(1) \mp 3d_i(r)] + c_x 2p_x + \dots
 \end{aligned} \tag{12}$$

where $3d_i(1,r)$ is one of the five atomic 3d orbitals centered on the left or right Cu ion, and $2p_{x,y}$ stand for the atomic $2p_{x,y}$ orbitals on the bridging oxygen. In none of the distortions considered here does the O- $2p_z$ orbital contribute to the magnetic orbitals. In addition, all of the other contributions to the magnetic orbitals are either very small or irrelevant for the anisotropy and have been removed for

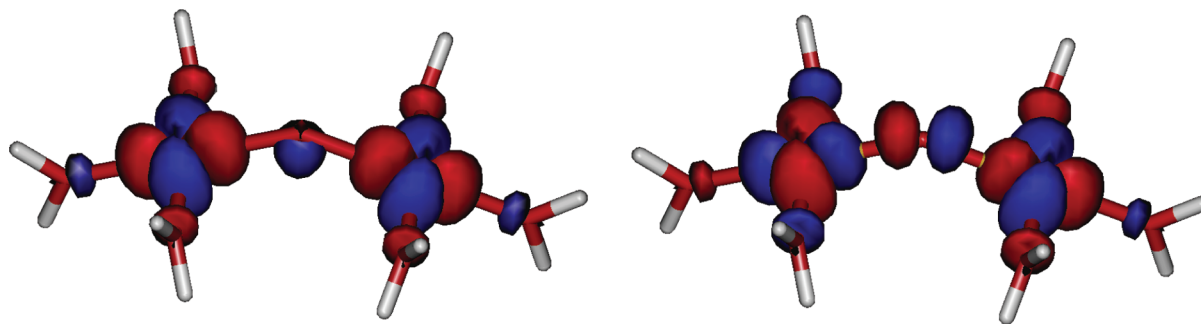


Figure 3. Symmetric (ϕ_s) and antisymmetric (ϕ_a) magnetic orbitals for the $\vartheta_1 = 40^\circ$, $\vartheta_2 = 0^\circ$ structure.

Table 4. List of Different Classes of Matrix Elements in the Direct Coupling after the Substitution of eqs 11 and 12 in $\langle 1, 0 | \hat{L}_z \cdot \hat{S}_z | 0, 0 \rangle^a$

class	electr. conf.	example matrix element	number
d–d (neutral)	$3d^9-2p^6-3d^9$	$\zeta_1 \langle 3d_{x^2-y^2}(l) 3\bar{d}_x(l) \hat{L}_z(l) \cdot \hat{S}_z(l) 3d_{xy}(l) 3\bar{d}_x(l) \rangle$	18/18
d–d (ionic)	$3d^8-2p^6-3d^{10}$	$\zeta_2 \langle 3d_{x^2-y^2}(l) 3\bar{d}_x(l) \hat{L}_z(l) \cdot \hat{S}_z(l) 3d_{xy}(l) 3\bar{d}_x(l) \rangle$	12/18
p–d (copper)	$3d^9-2p^5-3d^{10}$	$\zeta_1 \langle 3d_{x^2-y^2}(l) 2\bar{p}_x(l) \hat{L}_z(l) \cdot \hat{S}_z(l) 3d_{xy}(l) 2\bar{p}_y(l) \rangle$	24/24
p–d (oxygen)	$3d^9-2p^5-3d^{10}$	$\zeta_3 \langle 3d_x(l) 2\bar{p}_x(l) \hat{L}_z \cdot \hat{S}_z 3d_x(l) 2\bar{p}_y(l) \rangle$	24/24
p–p	$3d^{10}-2p^4-3d^{10}$	$\zeta_4 \langle 2p_x 2\bar{p}_y \hat{L}_z \cdot \hat{S}_z 2p_x 2\bar{p}_x \rangle$	2/2

^a $\bar{d}_i(l)$ indicates a 3d orbital on the left Cu center occupied with a β electron. An example matrix element is given for each class together with the total number of triplet/singlet terms. ζ_1 and ζ_2 are the atomic spin–orbit parameters of Cu^{2+} and Cu^{3+} , respectively. ζ_3 and ζ_4 are the spin–orbit parameters of O^- and O , respectively.

simplicity. The resulting orbitals are renormalized before further processing.

The second and most laborious step is the substitution of eqs 11 and 12 in the expression of the coupling between singlet and triplet through the spin–orbit operator. The derivation of the complete analytical expressions for all of the points considered in the magneto-structural correlations derived in the previous section would require consideration of the four spin–orbit states and all 12 atomic orbital contributions ($2p_x$, $2p_y$, and the 10 3d orbitals) to ϕ_s and ϕ_a . This would obviously lead to an overwhelming number of terms. Therefore, it is necessary to restrict the analysis to some special points for which the dominant mechanisms can be derived, which are then extrapolated to the other cases.

The first case to be analyzed is the structure with $\vartheta_1 \neq 0$ and $\vartheta_2 = 0$. Since, the DM vector is oriented along the z axis, only the $\langle 1, 0 | \hat{L}_z \cdot \hat{S}_z | 0, 0 \rangle$ coupling has to be considered. For symmetry reasons, the $3d_{xz}$ and $3d_{yz}$ atomic orbitals do not contribute to the magnetic orbitals. Moreover, the $\hat{L}_z \cdot \hat{S}_z$ operator does not couple the $3d_{z^2}$ to the $3d_{xy}$ or $3d_{x^2-y^2}$ orbital. These simplifications lead to a reasonable number of terms that can be classified in five different types. Table 4 gives an example of each class and enumerates the total number of terms in each class.

If we now focus on the structure with $\vartheta_1 = 40^\circ$ and $\vartheta_2 = 0$, the λ and μ CI coefficients in eq 11 are 0.7275 and 0.6861, respectively.

Using the numerical expression of the optimized orbitals of the triplet (ϕ_s , ϕ_a) and singlet (ϕ'_s , ϕ'_a) orbitals given in the Supporting Information, the DM interaction can be decomposed in the five classes of interactions mentioned before. The results are collected in Table 5. In the first place, it should be noticed that the decomposition leads to a similar norm of the DM vector as the complete RASSI-SO calculation, validating the analysis. The small

Table 5. Contributions to the d_z Component of the DM Vector (in cm^{-1}) of the Different Types of Mechanisms at the CASSCF Level for the ($\vartheta_1 = 40^\circ, 90^\circ; \vartheta_2 = 0^\circ$) Structures

class	$\vartheta_1 = 40^\circ$	$\vartheta_1 = 90^\circ$
d–d (neutral)	13.1	0.0
d–d (ionic)	–0.2	0.0
p–d (copper)	0.1	0.0
p–d (oxygen)	0.4	1.2
p–p	–0.1	0.0
total	13.3	1.2
RASSI-SO	14.8	3.6

differences arise from the use of atomic spin–orbit parameters and the simplification of the magnetic orbitals to the essential atomic orbitals contribution.

The largest contribution to the DM vector arises from the d–d (neutral) interactions. This term reaches a maximum when the contribution of the $3d_{x^2-y^2}$ and $3d_{xy}$ orbitals to ϕ_s and ϕ_a is largest. This happens for the structure with $\vartheta_1 = 45^\circ$. Nevertheless, the existence of other mechanisms and the difference between λ and μ displaces the maximum to slightly smaller angles. The d–d (ionic) contribution to the DM is small in this geometry and expected to be small in all cases. The contribution of these types of interactions may increase for smaller bending angles for which the isotropic coupling is stronger, and hence, the weight of the ionic configurations is larger. However, in these cases, the contribution of the $3d_{xy}$ orbital to the magnetic orbital is reduced, leading to a counterbalancing effect.

The p–d (copper) and p–d (oxygen) contributions are negligible in this geometry due to a numerical cancellation of various contributions. This is, however, not always the case, as will become clear for the structure with $\vartheta_1 = 90^\circ$. Finally, the contribution of the $\text{O-}2p^4$ configuration

to the wave function has such a small weight that the importance of this mechanism is negligible.

At $\vartheta_1 = 90^\circ$, only the $3d_{xy}$ and $3d_{z^2}$ orbitals have nonzero copper contributions to the magnetic orbitals. Since these orbitals are not coupled by the $\hat{l}_z \cdot \hat{s}_z$ operator, the contributions of the d–d and p–d (copper) mechanisms are strictly zero for this geometry. The dominating effect is a p–d (oxygen) mechanism, the spin–orbit coupling on the oxygen atom in the presence of a hole on one of the copper atoms. Again, the weight of the O-2p⁴ configuration is small, and as a consequence the contribution to the DM vector of the p–p mechanism is nearly zero.

Once we change ϑ_2 to values different from zero, the $\hat{l}^\pm \cdot \hat{s}^\pm$ operators come into play, and the coupling between all 3d atomic orbitals should be considered. This leads to a significant increase of the norm of the DM vector upon the increase of the twist angle of the two CuO₄ planes in the model complex. A numerical analysis of the mechanisms is simply too elaborate and would not really offer new insights. The main contribution to the DM interaction arises from the d–d (neutral) mechanism with smaller contributions from the other mechanisms. Close to $\vartheta_1 = 90^\circ$, the p–d (oxygen) term dominates.

4. Conclusions

The multispin Hamiltonian of the d⁹–d⁹ configuration contains 10 well-defined parameters. These parameters have been extracted using the effective Hamiltonian in combination with the CASSCF/CASPT2/RASSI-SO methodology. The comparison of the numerical effective Hamiltonian with the model Hamiltonian shows that the latter one accurately describes all of the magnetic interactions contained in the exact electronic Hamiltonian.

The symmetric anisotropy terms of the multispin Hamiltonian are small. The sign of the axial anisotropy and the orientation of the magnetic axis frame cannot be determined with the applied computational strategy, it being too dependent on the details of the calculation. A more sophisticated computational scheme is compulsory to benchmark the computations for these types of interactions. However, the antisymmetric anisotropic (or DM) interactions appear more robust and can be studied with the outlined strategy. Our conclusions can be summarized in three main points.

In the first place, the DM interaction is dominated at the first order of perturbation by the direct coupling between the magnetic singlet and triplet via spin–orbit coupling. This leads to an important simplification of the computational treatment, namely, the reduction of the SI space to the magnetic states and the use of the minimal active space.

Second, the main deformations of cuprate-like materials have been studied, that is, the Cu–O–Cu bending angle and the twist angle between the copper planes. The symmetry rules for the appearance of the DM interaction and the direction of the DM vector have been verified. It is shown that the twist angle alone does not produce any significant anisotropy. However, when it is combined with

the Cu–O–Cu bending, the synergic effect between both deformations can lead to a DM vector with a rather large norm.

Finally, the analysis of the leading mechanisms that contribute to the DM interactions show that the neutral d–d terms dominate and have a maximum for a bending angle close to 40°. Smaller contributions due to the spin–orbit coupling involving electrons on the oxygen bridge account for the nonzero DM interaction at 90° bending.

The approach presented here is currently being applied to *n*-center cluster of real cuprate materials in order to estimate the DM interaction and validate the model Hamiltonian for *n* > 2.

Acknowledgment. Financial support has been provided by the HPC-EUROPA2 project (project number: 228398), Spanish Ministry of Science and Innovation (Project CTQ2008-06644-C02-01), the Generalitat de Catalunya (Project 2009SGR462 and *Xarxa d'R+D+I en Química Teórica i Computacional*, XRQTC) and the Agence Nationale de la Recherche (ANR) (Project TEMAMA ANR-09-BLAN-0195-01).

Supporting Information Available: Expression of the magnetic orbitals in the $\vartheta_1 = \vartheta_2 = 40^\circ$ model structure. Values of the atomic spin–orbit constant used in the estimation of the spin–orbit coupling listed in Table 5. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Abragam, A.; Bleaney, B. *Electron Paramagnetic Resonance of Transition Ions*; Dover Publications: Dover, New York, 1986.
- (2) Boča, R. *Theoretical Foundations of Molecular Magnetism*; Elsevier: Amsterdam, 1999.
- (3) Hutchison, C. A.; Mangum, B. W. *J. Chem. Phys.* **1961**, *34*, 908–922.
- (4) Kahn, O. *Molecular Magnetism*; VCH Publishers: Weinheim, Germany, 1993.
- (5) Boča, R. *Coord. Chem. Rev.* **2004**, *248*, 757–815.
- (6) Gatteschi, D.; Sessoli, R. *Angew. Chem., Int. Ed.* **2003**, *42*, 269–297.
- (7) Thio, T.; Thurston, T. R.; Preyer, N. W.; Picone, P. J.; Kastner, M. A.; P. J. H.; Gabbe, D. R.; Chen, C. Y.; Birgeneau, R. J.; Aharony, A. *Phys. Rev. B* **1988**, *38*, 905–908.
- (8) Coffey, D.; Rice, T. M.; Zhang, F. C. *Phys. Rev. B* **1991**, *44*, 10112–10116.
- (9) Dzyaloshinskii, I. *J. Phys. Chem. Solids* **1958**, *4*, 241–255.
- (10) Moriya, T. *Phys. Rev.* **1960**, *120*, 91–98.
- (11) Zorko, A.; Arčon, D.; van Tol, H.; Brunel, L.-C. *Phys. Rev. B* **2004**, *69*, 174420.
- (12) Zorko, A.; Nellutla, S.; van Tol, J.; Brunel, L.-C.; Bert, F.; Duc, F.; Trombe, J.-C.; de Vries, M. A.; Harrison, A.; Mendels, P. *Phys. Rev. Lett.* **2008**, *101*, 026405.

- (13) Yildirim, T.; Harris, A. B.; Entin-Wohlman, O.; Aharony, A. *Phys. Rev. Lett.* **1994**, *73*, 2919–2922.
- (14) Koshibae, W.; Ohta, Y.; Maekawa, S. *Phys. Rev. B* **1994**, *50*, 3767–3778.
- (15) Entin-Wohlman, O.; Harris, A. B.; Aharony, A. *Phys. Rev. B* **1996**, *53*, 11661–11670.
- (16) Yushankai, V. Y.; Hayn, R. *Europhys. Lett.* **1999**, *47*, 116–121.
- (17) Moskvina, A. S. *J. Exp. Theor. Phys.* **2007**, *104*, 913–927.
- (18) Webb, S. P.; Gordon, M. S. *J. Chem. Phys.* **1998**, *109*, 919–927.
- (19) Pederson, M. R.; Jackson, K. A. *Phys. Rev. B* **1990**, *41*, 7453–7461.
- (20) Jackson, K. A.; Pederson, M. R. *Phys. Rev. B* **1990**, *42*, 3276–3281.
- (21) Pederson, M. R.; Khanna, S. N. *Phys. Rev. B* **1999**, *60*, 9566–9572.
- (22) Baruah, T.; Pederson, M. R. *Int. J. Quantum Chem.* **2003**, *93*, 324–331.
- (23) Kortus, J.; Pederson, M. R.; Baruah, T.; Bernstein, N.; Hellberg, C. S. *Polyhedron* **2003**, *22*, 1871–1876.
- (24) Park, K.; Pederson, M. R.; Richardson, S. L.; Aliaga-Alcalde, N.; Christou, G. *Phys. Rev. B* **2003**, *68*, 020405.
- (25) Ribas-Ariño, J.; Baruah, T.; Pederson, M. R. *J. Chem. Phys.* **2005**, *123*, 044303.
- (26) Ribas-Ariño, J.; Baruah, T.; Pederson, M. R. *J. Am. Chem. Soc.* **2006**, *128*, 9497–9505.
- (27) Ruiz, E.; Cirera, J.; Cano, J.; Alvarez, S.; Loose, C.; Kortus, J. *Chem. Commun.* **2008**, 52–54.
- (28) Vahtras, O.; Loboda, O.; Minaev, B.; Ågren, H.; Ruud, K. *Chem. Phys.* **2002**, *279*, 133–142.
- (29) Neese, F. *ORCA*, version 2.6; University of Bonn: Bonn, Germany, 2008.
- (30) Ganyushin, D.; Neese, F. *J. Chem. Phys.* **2006**, *125*, 024103.
- (31) Neese, F. *J. Am. Chem. Soc.* **2006**, *128*, 10213–10222.
- (32) Neese, F. *J. Chem. Phys.* **2007**, *127*, 164112.
- (33) Sinnecker, S.; Neese, F. *J. Phys. Chem. A* **2006**, *110*, 12267–12275.
- (34) Duboc, C.; Phoeung, T.; Zein, S.; Pécaut, J.; Collomb, M.-N.; Neese, F. *Inorg. Chem.* **2007**, *46*, 4905–4916.
- (35) Ganyushin, D.; Neese, F. *J. Chem. Phys.* **2008**, *128*, 114117.
- (36) Zein, S.; Duboc, C.; Lubitz, W.; Neese, F. *Inorg. Chem.* **2008**, *47*, 134–142.
- (37) Zein, S.; Neese, F. *J. Phys. Chem. A* **2008**, *112*, 7976–7983.
- (38) Liakos, D. G.; Ganyushin, D.; Neese, F. *Inorg. Chem.* **2009**, *48*, 10572–10580.
- (39) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222–239.
- (40) Malmqvist, P.-Å.; Roos, B. O.; Schimmelpfennig, B. *Chem. Phys. Lett.* **2002**, *357*, 230–240.
- (41) Roos, B. O.; Malmqvist, P.-Å. *Phys. Chem. Chem. Phys.* **2004**, *6*, 2919–2927.
- (42) de Graaf, C.; Sousa, C. *Int. J. Quantum Chem.* **2006**, *106*, 2470–2478.
- (43) Petit, S.; Pilet, G.; Luneau, D.; Chibotaru, L.; Ungur, L. *Dalton Trans.* **2007**, 4582–4588.
- (44) Chibotaru, L.; Ungur, L.; Aronica, C.; Elmoll, H.; Pilet, G.; Luneau, D. *J. Am. Chem. Soc.* **2008**, *130*, 12445–12455.
- (45) Chibotaru, L.; Ungur, L.; Soncini, A. *Angew. Chem., Int. Ed.* **2008**, *47*, 4126–4129.
- (46) Soncini, A.; Chibotaru, L. *Phys. Rev. B* **2008**, *77*, 220406.
- (47) Maurice, R.; Bastardis, R.; de Graaf, C.; Suaud, N.; Mallah, T.; Guihery, N. *J. Chem. Theory Comput.* **2009**, *5*, 2977–2984.
- (48) Maurice, R.; Guihéry, N.; Bastardis, R.; de Graaf, C. *J. Chem. Theory Comput.* **2010**, *6*, 55–65.
- (49) Maurice, R.; de Graaf, C.; Guihéry, N. *Phys. Rev. B* **2010**, *81*, 214427.
- (50) Gilka, N.; Taylor, P. R.; Marian, C. M. *J. Chem. Phys.* **2008**, *129*, 044102.
- (51) Sugisaki, K.; Toyota, K.; Sato, K.; Shiomi, D.; Kitagawa, M.; Takui, T. *Chem. Phys. Lett.* **2009**, *477*, 369–373.
- (52) Bleaney, B.; Bowers, K. D. *Proc. R. Soc. London, Ser. A* **1952**, *214*, 451–465.
- (53) Gregson, A. K.; Martin, R. L.; Mitra, S. *Proc. R. Soc. London, Ser. A* **1971**, *320*, 473–486.
- (54) Ross, P. K.; Allendorf, M. D.; Solomon, E. I. *J. Am. Chem. Soc.* **1989**, *111*, 4009–4021.
- (55) Ozarowski, A. *Inorg. Chem.* **2008**, *47*, 9760–9762.
- (56) Bencini, A.; Gatteschi, D. *Mol. Phys.* **1982**, *47*, 161–169.
- (57) Kahn, O. *Angew. Chem., Int. Ed.* **1985**, *24*, 834–850.
- (58) Kauffmann, K. E.; Popescu, C. V.; Dong, Y.; Lipscomb, J. D.; Que, L., Jr.; Münck, E. *J. Am. Chem. Soc.* **1998**, *120*, 8739–8746.
- (59) Yoon, J.; Mirica, L. M.; Stack, D. P.; Solomon, E. I. *J. Am. Chem. Soc.* **2004**, *126*, 12586–12595.
- (60) Bloch, C. *Nucl. Phys.* **1958**, *6*, 329–347.
- (61) des Cloizeaux, J. *Nucl. Phys.* **1960**, *20*, 321–346.
- (62) Buckingham, A. D.; Pyykko, P.; Robert, J. B.; Wiesenfeld, L. *Mol. Phys.* **1982**, *46*, 177–182.
- (63) Harriman, J. E. *Theoretical Foundations of Electron Spin Resonance*; Academic Press: New York, 1978.
- (64) Douglas, N.; Kroll, N. M. *Ann. Phys. (Leipzig)* **1974**, *82*, 89.
- (65) Hess, B. A. *Phys. Rev. A* **1986**, *33*, 3742–3748.
- (66) Hess, B. A.; Marian, C. M.; Wahlgren, U.; Gropen, O. *Chem. Phys. Lett.* **1996**, *251*, 365–371.
- (67) Schimmelpfennig, B. *AMFI*; Stockholms Universitet: Stockholm, Sweden, 1996.
- (68) Christiansen, O.; Gauss, J.; Schimmelpfennig, B. *Phys. Chem. Chem. Phys.* **2000**, *2*, 965–971.
- (69) Llusar, R.; Casarrubios, M.; Barandiarán, Z.; Seijo, L. *J. Chem. Phys.* **1996**, *105*, 5321–5330.
- (70) Barandiarán, Z.; Seijo, L. *J. Chem. Phys.* **2003**, *118*, 7439–7456.
- (71) Queralt, N.; Taratiel, D.; de Graaf, C.; Caballol, R.; Cimraglia, R.; Angeli, C. *J. Comput. Chem.* **2008**, *29*, 994–1003.
- (72) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218–1226.

- (73) Ghigo, G.; Roos, B. O.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **2004**, *396*, 142–149.
- (74) Forsberg, N.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **1997**, *274*, 196–204.
- (75) Roos, B. O.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, V.; Widmark, P.-O. *J. Phys. Chem. A* **2005**, *109*, 6575–6579.
- (76) Van den Heuvel, W.; Chibotaru, L. *Inorg. Chem.* **2009**, *48*, 7557–7563.
- (77) Anderson, P. W. *Phys. Rev.* **1950**, *79*, 350–356.
- (78) Goodenough, J. B. *Phys. Rev.* **1955**, *100*, 564–573.
- (79) Kanamori, J. *J. Phys. Chem. Solids* **1959**, *10*, 87–98.

CT100329N

Asymmetry and Electronegativity in the Electron Capture Activation of the Se–Se Bond: $\sigma^*(\text{Se–Se})$ vs $\sigma^*(\text{Se–X})$

José A. Gámez and Manuel Yáñez*

Departamento de Química, Módulo 13, Universidad Autónoma de Madrid, Campus de Excelencia UAM-CSIC, Cantoblanco, E-28049 Madrid, Spain

Received June 18, 2010

Abstract: The effects of electron capture on the structure of XSeSeX' diselenide derivatives in which the substituents attached to the selenium atoms have different electronegativities have been investigated at different levels of theory, namely, DFT, MP2, CCSD, G2, and CASSCF/CASPT2. An analysis of the bonding changes upon electron attachment shows that when the diselenides bear low-electronegativity substituents, the Se–Se bond becomes activated upon electron capture, as previous studies have shown. However, this is no longer the case for very electronegative substituents, where this bond remains practically unaltered and is the Se–X bond the one which becomes strongly activated through a preferential population of the $\sigma^*(\text{Se–X})$ antibonding orbital rather than the $\sigma^*(\text{Se–Se})$ one. When this is the case, several anionic species are also encountered, namely, *stretched*, *bent*, and *book* structures. The present findings are similar to those obtained for a series of analogous disulfide compounds, which points out that these results are not unique and could be extrapolated to a wider range of compounds than the ones covered here. The Se–Se (Se–X) linkage in CH₃SeSeOH, CH₃SeSeF, FSeSeOH, and FSeSeF bears some of the characteristics of the so-called charge-shift bonds, with a clear charge fluctuation between both selenium atoms. This is more evident in their anions where the bonding reflects the important contribution of the ionic resonant forms $\text{Se–Se}^- \leftrightarrow ^-\text{Se–Se}$ vs the covalent component $\text{Se}:\text{Se}$. This resonance changes with the nature of the substituents but also depends on the asymmetry of the substitution.

Introduction

The abundance of selenium in the Earth's crust is about 4 times lower than that of sulfur, which is reflected in the amount in which these elements are present in biological systems. However, except for tellurium, chalcogens are fundamental constituents of functional groups of amino acids and are important contributors to the chemistry and structure of peptides and proteins. Selenium is a trace element present in milligram amounts in the human body. In spite of its low abundance, selenium has been identified as an essential trace element for bacteria, birds, and mammals.¹ It is present in proteins in the form of selenocysteine and selenomethionine and has been observed in various oxidation states such as the reactive selenol, selenic acid, selenoxid, and selenylsulfide and the recently discovered diselenide bond.² Many seleno-

proteins have been identified in many living beings^{3,4} and 25 in humans.⁵ These proteins act as antioxidant agents, eliminating peroxides from the organism, and are also involved in cancer prevention and inflammation protection.^{6–8}

Sulfur and selenium have many common characteristics. Actually, in living organisms, selenium usually accompanies or substitutes sulfur thanks to its comparable physicochemical properties. Indeed, the mutation of cysteine (Cys) to selenocysteine (Sec) has been studied in a large variety of proteins.^{9–19} In all cases, the selenium analogues folded correctly, and the NMR structural analysis and CD spectroscopy confirm that the protein structure suffers from little distortion after substitution by selenium. Importantly, full biological activity was observed in all cases, confirming that the substitution of Cys by Sec is quite conservative. Nevertheless, this substitution has significant advantages over a substitution with other chemical moieties, which can import

* Corresponding author e-mail: manuel.yanez@uam.es.

structural distortions which may compromise bioactivity and selectivity.^{20–23} For example, selenoproteins with known functions are oxyreductases containing catalytic redox-active Sec,²⁴ whereas their Cys mutants are typically 100–1000 times less active.²⁵

Among the huge family of selenium-containing compounds, diselenides are of special interest. They are used in organic synthesis as precursors of organic selenium derivatives^{26–28} and as therapeutic drugs,^{29,30} and as mentioned above, they can form diselenide bridges in proteins² analogous to the disulfide linkages in Cys-containing peptides. Diselenides are well-known for their high antioxidant activity,^{31–33} which is actually higher than that of disulfides. In this respect, the work of Pearson and Boyd is noteworthy,³⁴ where they found, by means of DFT calculations, that the reduction of hydrogen peroxide by ebselene diselenide is favored with respect to its disulfur analogue due to lower energy barriers for the reactions of the former. However, the little literature available on this topic prevents a general picture of the main reasons behind this behavior. Also, in contrast with disulfides, where this process has been studied much more extensively,^{35–39} little attention has been paid to the change of the electronic structure of diselenides in the very first stage of the reduction: the electron attachment process. Previous studies^{40,41} indicate that electron attachment to dimethyldiselenide yields mainly the fragmentation of the Se–Se bond since the extra electron is accommodated in the $\sigma^*(\text{Se}–\text{Se})$ antibonding orbital, as it occurred for disulfides. However, to the best of our knowledge, there is a complete lack of data concerning asymmetric diselenides or diselenides bearing substituents with different electronegativities. We have recently shown for disulfides bearing highly electronegative substituents⁴² that the electron capture process leads to dissociations different from that of the S–S bond, as had been previously assumed. The aim of this paper is to investigate the changes in the electronic structure of asymmetric diselenides, and their consequences, upon electron capture. The CH_3SeSeX ($X = \text{NH}_2, \text{OH}, \text{and F}$) set of molecules has been chosen as a suitable model ensemble, which include substituents of increasing electronegativity that can be compared with HSeSeH and $\text{CH}_3\text{SeSeCH}_3$ to determine the influence of the asymmetry in the electron-attachment process.

Computational Methods

Density functional theory (DFT) is quite popular in the quantum chemistry community due to its high accuracy at a low computational price. However, approximate functionals suffer from the self-interaction error (an unbalanced description of the Coulomb and exchange terms), which becomes especially important for odd-electron systems,^{43–45} as is the case in this study. More recently, this problem has been renamed delocalization error,⁴⁶ present when, due to delocalization, atomic centers bear fractional charges, causing approximate functionals to underestimate the energy of such systems. As a result, DFT methods predict too large electron affinities or bond lengths, when an electron is added to a closed-shell molecule to form an open-shell anion,^{47–49} as in the present study. Additionally, DFT overestimates the

bond length and the binding energy of two-center–three-electron linkages ($2c–3e$),⁵⁰ which casts serious doubts on the reliability of DFT for this kind of system. However, since HF overestimates the energy of such charge-delocalized systems, the use of the BH&H functional proposed by Becke⁵¹ including 50% exact exchange seems a good compromise. Actually, this exchange functional combined with the LYP correlation functional⁵² has shown good performance for the kinds of systems considered here.⁵⁰ We have also used the MP2 method because it usually yields good results for the types of systems here investigated. However, it should be taken into account that sometimes the $2c–3e$ bonds are not properly described at the HF level, in which case the MP2 results would be questionable.⁵³ Hence, CCSD(T), which includes a large amount of correlation energy and will give results close to the experimental value, will be used to assess the other methods. In addition to that, G2 estimates, which are known to provide an accurate description of these systems,^{54,55} have also been included for the computation of the electron affinities.

The 6-31++G(d,p) (BS1) basis set will be used for geometry optimizations, since it provides enough flexibility to describe the bonding situations we deal with. Diffuse functions are necessary to describe the extra electron placed far from the nuclei in the anions. To ensure that the optimized geometries were true minima, the Hessian matrix has been evaluated. Although the basis set BS1 is flexible enough for geometry optimizations, final energies will be obtained with single-point calculations with a larger aug-cc-pVTZ basis set (BS2). The geometry optimizations and Hessian evaluations with BS1 and single-point DFT calculations with BS2 have been performed with the Gaussian 03 suite of programs,⁵⁶ while the single point calculations with BS2 at the MP2 and CCSD(T) levels were carried out with the MOLPRO 2009.01 package.⁵⁷

Since this study involves radicals, in some cases the nature of the wave function and the reliability of the results had to be checked to determine the single- or multiconfigurational character of the states. Multireference methods have been included in this study, in particular, the CASSCF/CASPT2 approach. Geometries were optimized at the CASSCF level with the atomic natural orbital (ANO) basis set described by Pierloot et al.⁵⁸ contracted to $\text{Se}[5s4p3d]/\text{C,N,O,F}[3s2p1d]/\text{H}[2s1p]$ (BS3). With these geometries, high-level energies were obtained with the second-order perturbation multireference CASPT2 method using the same BS2 as for the other methods. The active space was formed by distributing 10 electrons (11 in the case of anions) in eight orbitals (nine for the anions), which showed good results in a similar study involving disulfides.⁴² The multireference calculations were performed with the MOLCAS 7.2 package of programs.⁵⁹

To fully understand the nature of the bonds of the species under study, an atoms in molecules (AIM) analysis⁶⁰ has been performed. For this purpose, the electron and energy densities at the different bond critical points (BCPs) have been evaluated to gain some insight into the different changes produced upon electron attachment. This analysis has been carried out with the DGrid 4.5 program.⁶¹ Further information about the electronic rearrangement of the diselenide

Table 1. Main Internal Coordinates of the Neutral (Radical–Anionic) HSeSeH, CH₃SeSeCH₃, and CH₃SeSeNH₂ Species Calculated with Several Methods and the 6-31++G(d,p) (BS3 for CASSCF) Basis Set^a

Compound	HSeSeH				
	BH&HLYP	MP2	CCSD	CASSCF	Literature ⁵⁵
d(Se–Se)	2.324 (2.998)	2.340 (2.948)	2.354 (2.985)	2.404 (3.080)	2.355, 2.370
d(Se–H)	1.472 (1.472)	1.484 (1.483)	1.492 (1.493)	1.486 (1.489)	1.463, 1.476
∠SeSeH	97.0 (88.9)	96.6 (89.1)	96.5 (89.1)	95.5 (86.1)	
HSeSeH	91.6 (90.1)	90.3 (92.8)	90.3 (92.5)	90.0 (93.0)	90.67, 90.5
Compound	CH ₃ SeSeCH ₃				
	BH&HLYP	MP2	CCSD	CASSCF	Exp. ⁶⁶
d(Se–Se)	2.310 (2.980)	2.322 (2.929)	2.335 (2.965)	2.390 (3.069)	2.306, 2.326
d(Se–X)	1.949 (1.954)	1.966 (1.958)	1.965 (1.973)	1.999 (2.013)	1.954, 1.945
∠SeSeX	100.2 (83.3)	99.0 (85.9)	99.1 (86.3)	99.8 (90.2)	99.8, 98.9
CSeSeX	87.7 (86.9)	86.1 (85.5)	86.5 (85.7)	87.7 (88.1)	85.2, 87.5
Compound	CH ₃ SeSeNH ₂				
	BH&HLYP	MP2	CCSD	SS-CASSCF	SA-CASSCF
d(Se–Se)	2.319 (3.011)	2.331 (2.957)	2.342 (2.996)	2.344 (3.202)	2.344 (3.055)
d(Se–X)	1.826 (1.881)	1.847 (1.905)	1.889 (1.910)	1.859 (1.913)	1.859 (1.954)
∠SeSeX	106.6 (88.3)	106.9 (90.2)	106.2 (90.0)	106.8 (90.4)	106.8 (94.6)
CSeSeX	89.4 (89.9)	87.3 (82.9)	87.6 (85.1)	88.4 (89.9)	88.4 (90.6)

^a Bond lengths are given in Å and angles in degrees.

derivatives in these processes was achieved by means of the Becke and Edgecombe electron localization function⁶² (ELF) approach.⁶³ ELF is a function which measures the probability of finding an electron pair at a given region of the space, so it becomes large in regions where electron pairs are localized, either as bonding or lone pairs. By means of an appropriate Lorentzian transform, ELF can be confined in the [0,1] interval. In this way, the molecular space can be divided in polysynaptic (generally disynaptic) basins, with the participation of two (or more) atomic valence shells and monosynaptic ones, which correspond to core electrons or lone pairs.⁶⁴ ELF grids and basin integrations have been evaluated with the TopMod package.⁶⁵

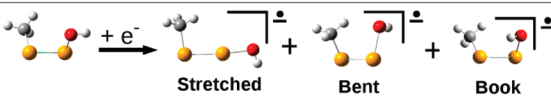
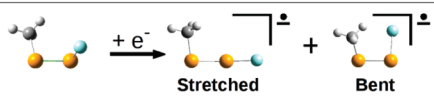
Geometrical Changes upon Electron Attachment

Tables 1 and 2 summarize the geometric changes of HSeSeH and CH₃SeSeX (X = CH₃, NH₂, OH, F) triggered upon electron capture. To the best of our knowledge, experimental

information is only available for the neutral dimethyldiselenide,⁶⁶ whereas for neutral HSeSeH just theoretical estimates have been previously reported.⁵⁵ The values calculated for these compounds agree well with those found in the literature, which confirms the reliability of the approach used.

For the compounds in Table 1 (HSeSeH, CH₃SeSeCH₃, and CH₃SeSeNH₂), the main geometrical deformation observed after electron attachment is the lengthening of the Se–Se bond by ca. 0.6 Å, while the other geometrical parameters remain practically unperturbed. All the methods employed provide rather similar geometries for the radical anions, although MP2 underestimates the lengthening of the Se–Se bond relative to the CCSD and BH&HLYP values. This was explained by Bräida and Hiberty⁵³ on the basis of a dissimilar charge distribution along the 2c–3e bond (Se–Se) at the UHF and UMP2 levels, which has not been found in this case (see Table S2 of the Supporting Information). This discrepancy could

Table 2. Main Internal Coordinates of the Neutral and Anionic Species of CH₃SeSeOH and CH₃SeSeF Calculated with Several Methods and the 6-31++G(d,p) (BS3 for CASSCF) Basis Set^a

	Isomer	CH ₃ SeSeOH					CH ₃ SeSeF				
											
		BH&HLYP	MP2	CCSD	SS-CASSCF	SA-CASSCF	BH&HLYP	MP2	CCSD	SS-CASSCF	SA-CASSCF
d(Se–Se)	Neutral	2.286	2.295	2.310	2.316		2.256	2.262	2.280	2.326	
	Stretched	2.547	2.529	2.557	2.806	2.703	2.428	2.428	2.443	2.449	2.381
	Bent	2.325	2.333	2.350	2.347		2.304	2.314	2.329	2.364	
	Book	2.938	2.894	2.929	3.023						
d(Se–X)	Neutral	1.799	1.833	1.830	1.852		1.771	1.803	1.793	1.798	
	Stretched	1.987	2.032	2.016	1.941	2.025	2.029	2.050	2.065	2.165	2.540
	Bent	2.404	2.235	2.363	2.431		2.278	2.270	2.286	2.568	
	Book	1.847	1.884	1.883	1.893						
∠SeSeX	Neutral	103.1	102.7	102.0	103.3		101.6	101.7	101.0	100.8	
	Stretched	151.6	160.2	153.0	151.1	145.8	159.9	160.1	161.3	162.7	86.8
	Bent	92.0	91.0	90.6	98.7		88.4	88.2	86.5	85.6	
	Book	82.1	81.3	81.8	85.7						
CSeSeX	Neutral	82.4	81.1	81.0	82.1		85.7	85.8	85.7	84.8	
	Stretched	80.7	79.3	79.8	82.9	82.6	85.5	83.3	83.6	89.0	49.4
	Bent	50.8	48.3	51.0	53.3		54.8	56.3	55.3	48.5	
	Book	85.2	82.0	81.6	82.5						

^a Bond lengths are given in Å and angles in degrees.

be better attributed to the large electronic correlation needed to describe the outer electrons of Se not completely recovered by MP2.

The good performance of the BH&HLYP functional is remarkable, giving results quite close to the highly correlated CCSD estimates. CASSCF presents larger deviations than MP2 due to the missing dynamic correlation. Regarding these estimates for CH₃SeSeNH₂, two values are presented. The first ones, state specific CASSCF (SS-CASSCF), were obtained by optimizing the ground state wave function, without considering any other state. However, this resulted in a somewhat large overestimation of the Se–Se lengthening upon electron capture. An optimization of the lowest root of a state average CASSCF (SA-CASSCF) comprising the six lowest states shows better results, probably due to the fact that in the SS-CASSCF wave function the $\sigma^*(\text{Se–Se})$ had an unphysical high contribution in the total wave function, corrected with the averaging procedure.

Regarding CH₃SeSeOH and CH₃SeSeF (see Table 2), the first remarkable feature is that the electron capture process leads to more than one stable radical anion: two for the fluorine derivative and three for CH₃SeSeOH, as it has been previously found for the disulfide analogues.⁴² The *stretched* isomers are characterized by a moderate elongation of about 0.2 Å for the Se–Se and Se–X bonds and by a substantial opening of the SeSeX angle, whereas the *bent* anions show a significant elongation (ca. 0.5 Å) of the Se–X bond and a decrease of the CSeSeX dihedral angle. The *book* isomer, which is unique for CH₃SeSeOH, has a book-like conformation similar to that of the neutral species, but with a longer (ca. 0.6 Å) Se–Se bond. Again, the three methods agree well in their predictions, especially BH&HLYP and CCSD. The deviations of SS-CASSCF in the *stretched* anion of

Table 3. Relative Stability in Terms of ΔH (ΔG in parentheses) at 298 K of the Different Anions Present for the CH₃SeSeOH and CH₃SeSeF Structures Calculated with the aug-cc-pVTZ Basis Set^a

	BH&HLYP	MP2	CCSD(T)	CASPT2	G2
<i>stretched</i> CH ₃ SeSeOH	0	0	0	0	0
	(0)	(0)	(0)	(0)	(0)
<i>bent</i> CH ₃ SeSeOH	32	43	36	17	36
	(36)	(46)	(40)	(20)	(44)
<i>book</i> CH ₃ SeSeOH	9	16	11	14	18
	(10)	(17)	(12)	(15)	(20)
<i>stretched</i> CH ₃ SeSeF	0	0	0	0	0
	(0)	(0)	(0)	(0)	(0)
<i>bent</i> CH ₃ SeSeF	27	29	29	34	24
	(29)	(31)	(31)	(37)	(30)

^a Values in kJ mol⁻¹.

CH₃SeSeOH are corrected with the averaging procedure but not for the *bent* isomer of CH₃SeSeF. Among this collection of anionic derivatives, the most stable species among this series of isomers correspond to the *stretched* one (Table 3), the *book* one being intermediate between the *stretched* and the *bent* structures. However, the energy gap between them suggests that upon electron attachment the only product would be the *stretched* anion. This is in contrast to what was found for disulfides, where the small energy gaps between the different anionic structures predicted that a mixture of them should be obtained.

All the anionic structures presented here correspond to minima of the potential energy surface (PES). However, to ensure that they are really stable anions, the eigenvalue of the HF SOMO has been checked, being negative in all cases (see Table S3 of the Supporting Information). Another assessment of the stability of these anions, experimentally

Table 4. Adiabatic Electron Affinities (EA_{adiab}) in terms of ΔH (ΔG in parentheses) at 298K calculated with the aug-cc-pVTZ Basis Set^a

	BH&HLYP	MP2	CCSD(T)	CASPT2	G2
HSeSeH	107 (115)	95 (103)	102 (110)	89 (96)	94 (101)
CH ₃ SeSeCH ₃	53 (59)	45 (52)	53 (60)	44 (50)	44 (55)
CH ₃ SeSeNH ₂	48 (58)	40 (50)	49 (59)	46 (53)	41 (55)
CH ₃ SeSeOH	72 (79)	75 (79)	73 (79)	78 (85)	72 (83)
CH ₃ SeSeF	124 (132)	120 (128)	124 (132)	130 (138)	121 (130)

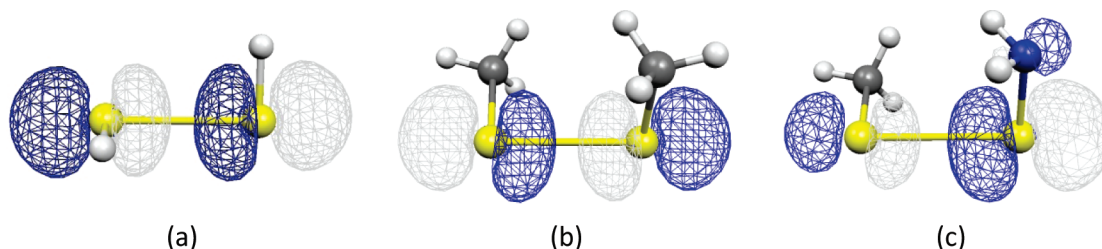
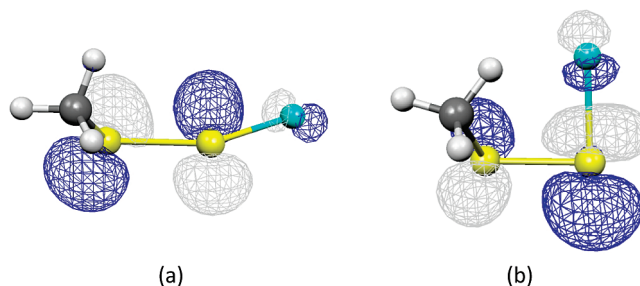
^a Values in kJ mol⁻¹.

measurable, is the adiabatic electron affinity (EA_{adiab}), which has been found positive for all of the species (see Table 4).

Although all values are similar, the very good agreement between the BH&HLYP and the CCSD(T) results is worth noting, while the MP2 and CASPT2 values bear better resemblance to the G2 ones. To the best of our knowledge, only the vertical electron affinity (EA_v) of CH₃SeSeCH₃ has been experimentally measured by Modelli et al.,⁴¹ which obtained a value of 0.27 eV, in clear contrast with our theoretical estimates: 0.55, 0.55, 0.49, and 0.57 eV for BH&HLYP, MP2, CCSD(T), and CASPT2 respectively. However, the value reported by the same authors for CH₃SSCH₃ (1.04 eV) is much smaller than the value of 1.75 eV given by Carles et al.,⁶⁷ which is nicely reproduced by our theoretical approaches (1.81 and 1.69 eV for MP2 and CASPT2, respectively), suggesting that the values of Modelli and co-workers could be underestimated. Interestingly, when these EA values are compared with those of the disulfide analogues,⁴² one finds that the EA is higher for diselenides than for disulfides, probably due to the fact that for the former the extra electron is accommodated in a more diffuse orbital, leading to a lower interelectronic repulsion. This higher EA could also be behind the stronger antioxidant activity of diselenides compared to disulfides.

The Nature of the Anions

To better understand the geometrical changes triggered by the electron capture process, it is necessary to gain some insight into the modifications that this extra electron induces in the electronic structure of the system. As indicated above, those anions presenting only one stable isomer show a significant lengthening of the Se–Se bond, whereas other geometrical parameters remain practically unperturbed.

**Figure 1.** Single occupied molecular orbital (SOMO) for the radical-anionic derivatives of (a) HSeSeH, (b) CH₃SeSeCH₃, and (c) CH₃SeSeNH₂. Selenium atoms are in yellow, carbon ones in gray, nitrogen in blue, and hydrogen in white.**Figure 2.** SOMOs of the (a) *stretched* and (b) *bent* anions of CH₃SeSeF. Selenium atoms are in yellow, carbon ones in gray, fluorine in cyan, and hydrogen in white.

Previous studies on diselenides show the liability of the Se–Se linkage upon electron attachment, which is coherent with this bond elongation. Actually, the peak of least energy of the electron-transmission spectrum of CH₃SeSeCH₃ has been attributed to the occupancy by the extra electron of the $\sigma^*(\text{Se–Se})$ antibonding orbital,⁴¹ which is confirmed by our theoretical calculations as shown in Figure 1.

However, when the electron attachment process leads to more than one stable anion, the bonding situation is very different. As the substituent becomes more and more electronegative, the contribution of its orbitals to the single-occupied molecular orbital (SOMO) increases (see Figure 2 for the particular case of the fluorine derivative), in an attempt to displace the electronic density associated with the extra electron toward the more electronegative substituent.

Hence, in both the *stretched* and the *bent* isomers of CH₃SeSeF, the SOMO arises from a linear combination of the $\sigma^*(\text{Se–F})$ and $\pi^*(\text{Se–Se})$ antibonding orbitals. This explains the moderate elongation of the Se–Se linkage, since the σ component of the bond is not affected in any case, and the increase in the Se–F distance, since the SOMO always has a significant $\sigma^*(\text{Se–F})$ antibonding character. The larger contribution of the $\sigma^*(\text{Se–F})$ MO in the SOMO would explain the longer Se–F bond of the *bent* isomer relative to the *stretched* isomer. This is not easily seen in Figure 2, but an examination of the spin density (see Table S4 of the Supporting Information) shows that for the *bent* isomer the density is accumulated preferentially at the F atom and the Se atom to which it is attached (0.21 and 0.76, respectively), whereas for the *stretched* isomer, the spin density at the F atom is very small (0.07).

The CH₃SeSeOH can be viewed as an intermediate situation between the compounds in Table 1 and CH₃SeSeF. Indeed, an anionic structure (*book anion*) can be found whose SOMO is the $\sigma^*(\text{Se–Se})$ antibonding MO (Figure 3), like

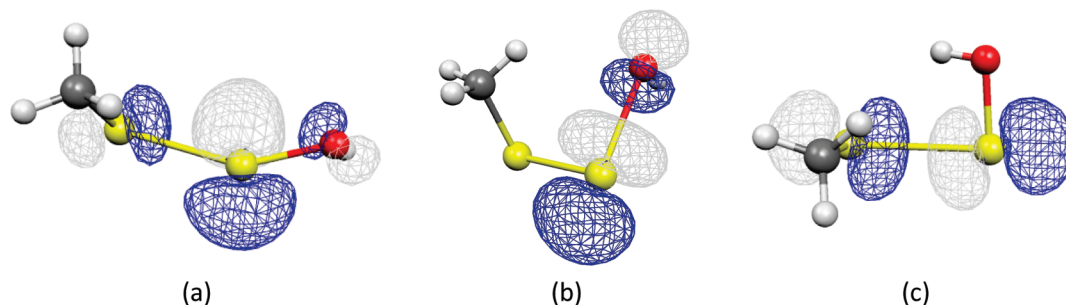


Figure 3. SOMOs of the (a) *stretched*, (b) *bent*, and (c) *book* anions of CH_3SeSeOH . Selenium atoms are in yellow, carbon ones in grey, oxygen in red, and hydrogen in white.

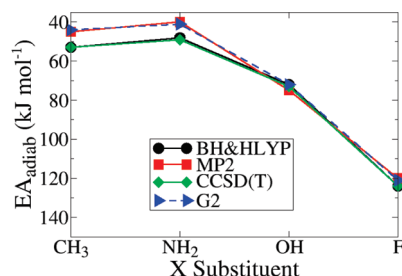


Figure 4. Adiabatic EA as a function of X for the CH_3SeSeX series of compounds. It should be noted that the energy scale is inverted from the normal order.

in $\text{CH}_3\text{SeSeNH}_2$, and another one (*bent* anion) which has the $\sigma^*(\text{Se}-\text{O})$ antibonding orbital as SOMO, similar to what has been found for CH_3SeSeF . Actually, there exists a third anion (*stretched* anion) whose SOMO is intermediate between the previous two orbitals, presenting two nodes: one at the Se–Se bond and the other at the Se–O one. The nature of the SOMOs just discussed offers useful clues to rationalize why the *book* isomer presents quite a long Se–Se bond and why for the *bent* isomer just the Se–O linkage becomes activated, whereas for the *stretched* derivative both distances partially increase.

Similarly to what has been previously reported for disulfide derivatives,^{35,42} when increasing the electronegativity of the substituent, the SOMO loses its $\sigma^*(\text{Se}-\text{Se})$ nature and gains

some $\sigma^*(\text{Se}-\text{X})$ character. This permits rationalization of the unexpected variation of the EA_{adiab} in the CH_3SeSeX series. As shown in Figure 4, while the electronegativity of X increases as $\text{CH}_3 < \text{NH}_2 < \text{OH} < \text{F}$, the same trend is not observed for the EA_{adiab} values, since surprisingly, the electron affinity of $\text{CH}_3\text{SeSeNH}_2$ is lower than that of $\text{CH}_3\text{SeSeCH}_3$. When the electronegativity of the substituent X increases, the $\sigma^*(\text{Se}-\text{X})$ MO becomes stabilized at the price of destabilizing the $\sigma^*(\text{Se}-\text{Se})$, explaining the aforementioned variation of the electron affinity between $\text{CH}_3\text{SeSeCH}_3$ and $\text{CH}_3\text{SeSeNH}_2$. Once the $\sigma^*(\text{Se}-\text{X})$ MO is lower in energy than the $\sigma^*(\text{Se}-\text{Se})$ MO, the increase of the electronegativity of X just stabilizes it further, which explains why the EA_{adiab} of CH_3SeSeF is larger than that of CH_3SeSeOH .

The Nature of the Bonding

The Se–Se Bond. As expected, the aforementioned structural changes upon electron capture just reflect concomitant changes in the electronic structure. For $\text{X} = \text{H}$, CH_3 , or NH_2 , a significant decrease of the electron density at the Se–Se BCP is observed (Table 5). This fact, together with the positive value of $\nabla^2\rho$ and the near-zero energy density at this point, seems to indicate that the Se–Se interaction comes mainly from dispersion, which is also supported by the fact that the ELF does not present a disynaptic basin $V(\text{Se},\text{Se})$ for these compounds (Figure 5).

Table 5. Electronic Density, ρ (in $e a_0^{-3}$), $\nabla^2\rho$ (in $e a_0^{-5}$), and Energy Density, H (in $E_h a_0^{-3}$), Evaluated at the BCP of the Se–Se and Se–X Bonds for the CH_3SeSeX Set of Molecules Calculated at the BH&HLYP/6-31++G(d,p) Level

		Se–Se			Se–X		
		ρ	$\nabla^2\rho$	H	ρ	$\nabla^2\rho$	H
HSeSeH	neutral	0.1070	−0.0764	−0.0529	0.1717	−0.2015	−0.1363
	anion	0.0307	0.0576	−0.0004	0.1649	−0.1497	−0.1273
$\text{CH}_3\text{SeSeCH}_3$	neutral	0.1105	−0.0820	−0.0564	0.1489	−0.1567	−0.0982
	anion	0.0321	0.0578	−0.0008	0.1416	−0.1072	−0.0907
$\text{CH}_3\text{SeSeNH}_2$	neutral	0.1098	−0.0753	−0.0556	0.1606	0.1435	−0.1120
	anion	0.0312	0.0582	−0.0006	0.1439	0.1199	−0.0912
CH_3SeSeOH	neutral	0.1161	−0.0894	−0.0621	0.1497	−0.1649	−0.0987
	stretched	0.0649	0.0482	−0.0181	0.0942	0.2089	−0.0300
	bent	0.1014	−0.0549	−0.0474	0.0509	0.1585	−0.0020
	book	0.0354	0.0613	−0.0018	0.1279	0.2122	−0.0661
CH_3SeSeF	neutral	0.1215	−0.0980	−0.0680	0.1394	0.4467	−0.0708
	stretched	0.0843	0.0263	−0.0335	0.0820	0.2362	−0.0204
	bent	0.1100	−0.0714	−0.0559	0.0552	0.1814	−0.0048
OHSeSeF	neutral	0.1235	−0.0978	−0.0701	0.1364	0.4021	−0.0694
	anion	0.0966	0.0229	−0.0449	0.0755	0.2140	−0.0167
FSeSeF	neutral	0.1341	−0.1165	−0.0826	0.1432	0.4585	−0.0742
	anion	0.1060	−0.0002	−0.0535	0.0953	0.2469	−0.0313

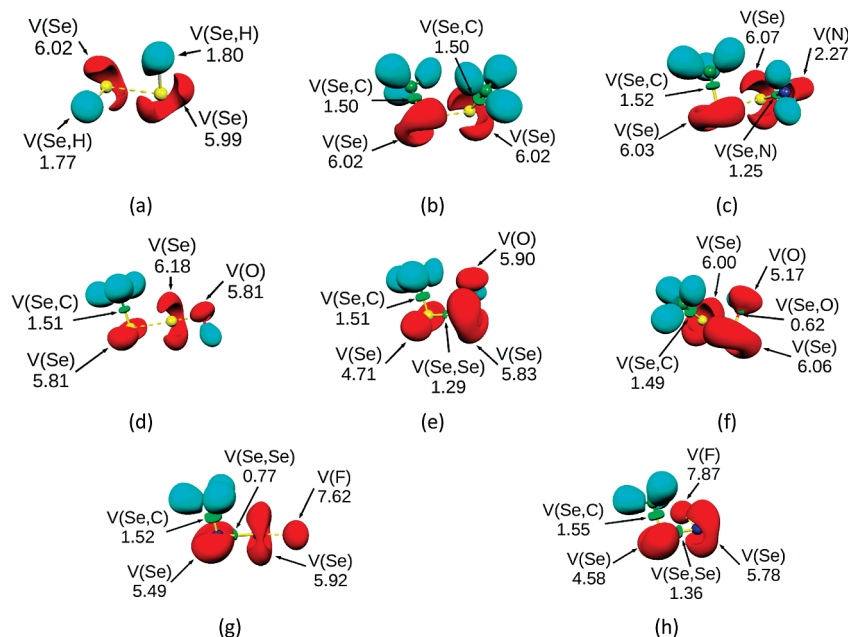


Figure 5. ELF localization domains within the isosurface $ELF = 0.8$ for the anionic derivatives of (a) HSeSeH, (b) $CH_3SeSeCH_3$, (c) $CH_3SeSeNH_2$, (d) *stretched*, (e) *bent*, and (f) *book* isomers of $CH_3SeSeOH$ and (g) *stretched* and (h) *bent* anions of CH_3SeSeF molecules. Blue lobes correspond to basins involving H atoms (protonated basins), green lobes to disynaptic basins between two bonding atoms, and red ones refer to monosynaptic lone-pair basins. The population of the different basins is given in e.

These topological features are typically encountered in $2c-3e$ bonds. However, these linkages usually present dissociation energies much larger than those expected from dispersion because a charge fluctuation between the atoms,^{68–71} due to the similar distribution of the unpaired electron between them, takes place. In fact, an exponential decrease of the dissociation energy of cationic (anionic) $2c-3e$ bonds with the relative IP (EA) of the constituent fragments was found both by experimental⁷² and theoretical⁷⁰ means, which indicates that $2c-3e$ bonds are stronger the more evenly distributed between the fragments the unpaired electron is.

Obviously for symmetric systems like HSeSeH and $CH_3SeSeCH_3$, both fragments have the same EA value and hence present strong bonds. Conversely, for $CH_3SeSeNH_2$, where the value of ΔEA increases, a bond weakening is observed (see Table 6). Since an important component of $2c-3e$ bonds is a charge fluctuation between the bonding atoms, it would be useful to estimate the extent of such delocalization. This is possible through the delocalization indexes⁷³ $\delta(A,B)$ between the lone-pair monosynaptic basins of the bound atoms A and B, which can be related to the number of electrons delocalized between both basins. Actually, high values (ca. 0.5) of these indexes have been found for $2c-3e$ bonds,⁷⁴ and an almost linear dependence between $\delta(A,B)$ and the dissociation energy of the bonds has been established.⁷⁵ For HSeSeH, $CH_3SeSeCH_3$, and $CH_3SeSeNH_2$, the increase of the $\delta(Se,Se)$ index upon electron attachment (see Table 8) indicates that this charge fluctuation is more important for the anions. Hence, a higher contribution of the ionic resonant forms $Se-Se^- \leftrightarrow ^-Se-Se$ of the bond versus the covalent component $Se:Se$ should be expected, which is coherent with the disappearance of the disynaptic basin $V(Se,Se)$ in the anions.

Table 6. Adiabatic EA (in eV) of the Fragments, Their Difference (in absolute value), and the Dissociation Energy (in kJ mol^{-1}) of the Anion of the Molecule Formed by These Fragments Calculated at the CCSD(T)/aug-cc-pVTZ Level of Theory

fragment 1	EA	fragment 2	EA	ΔEA	D_e
HSe	1.43	HSe	1.43	0.00	104
CH_3Se	1.84	CH_3Se	1.84	0.00	122
CH_3Se	1.84	NH_2Se	1.43	0.41	92
CH_3Se	1.84	OHSe	1.78	0.07	126 ^a
CH_3Se	1.84	FSe	2.34	0.49	128 ^b
CH_3SeSe	1.91	OH	1.74	0.17	137 ^c
CH_3SeSe	1.91	F	3.31	1.40	77 ^c
OHSeSe	1.94	F	3.31	1.37	140
FseSe	2.25	F	3.31	1.06	171
OHSe	1.78	FSe	2.34	0.56	156

^a Value of the *book* isomer. ^b Value of the *stretched* isomer. ^c Value of the *bent* isomer.

For the $CH_3SeSeOH$ *book* anion, the dissociation energy is greater than for $CH_3SeSeNH_2$ coherently with the lower value of ΔEA . For the $CH_3SeSeOH$ *stretched* isomer, the $V(Se,Se)$ disynaptic basin disappears, because, as indicated above, in this structure the extra electron roams between the $\sigma^*(Se-Se)$ and $\sigma^*(Se-O)$ orbitals due to the comparable electronegativity of the fragments CH_3Se , $SeOH$, CH_3SeSe , and OH . ΔEA is particularly small between the CH_3Se and $SeOH$ fragments, which explains why the spin density (Table S4 of the Supporting Information) at Se1 is higher than at O. This is also consistent with the large value of $\delta(Se,Se)$ and the consequent disappearance of the disynaptic $V(Se,Se)$ basin. The picture of the bonding is quite different, however, for the *stretched* and *bent* isomers when $X = F$. Due to the high EA value of F, the extra electron is close to it, and therefore the $Se-Se$ linkage has a low $2c-3e$ bond character and retains a large fraction of the covalent nature which it had in the neutral. This explains the very small decrease of

Table 7. NBO Charge Population Analysis (in e) of the Different Radical-Anionic Structures Calculated with the BH&HLYP/6-31+G(d,p) Density^a

		CH ₃	Se1	Se2	XH _n
HSeSeH	neutral		-0.06		0.06
	anion		-0.52		0.02
CH ₃ SeSeCH ₃	neutral		0.13		-0.13
	anion		-0.27		-0.23
CH ₃ SeNH ₂	neutral	-0.13	0.09	0.35	-0.31
	anion	-0.23	-0.33	-0.05	-0.36
CH ₃ SeSeOH	neutral	-0.12	0.12	0.44	-0.45
	stretched	-0.21	-0.18	-0.02	-0.58
	bent	-0.16	0.02	-0.29	-0.55
	book	-0.21	-0.30	0.02	-0.50
CH ₃ SeSeF	neutral	-0.11	0.16	0.52	-0.57
	stretched	-0.20	-0.11	0.06	-0.74
	bent	-0.15	0.04	-0.11	-0.77
OHSeSeF	neutral	-0.43	0.52	0.47	-0.56
	anion	-0.49	0.02	0.23	-0.76
FSeSeF	neutral		0.56		-0.56
	anion		0.19		-0.69

^a Se1 is attached to the methyl group, whereas Se2 is linked to XH_n.

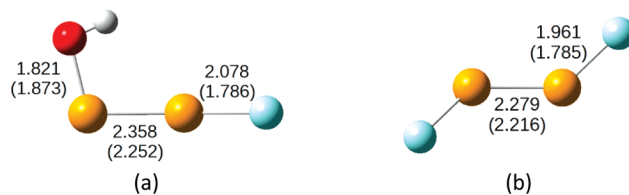
Table 8. Delocalization Indexes between the Lone-Pair Monosynaptic Basins of Both Selenium Atoms, $\delta(\text{Se},\text{Se})$, and Selenium and the Substituent X, $\delta(\text{Se},\text{X})$

		$\delta(\text{Se},\text{Se})$	$\delta(\text{Se},\text{X})$
HSeSeH	neutral	0.40	
	anion	0.52	
CH ₃ SeSeCH ₃	neutral	0.40	
	anion	0.57	
CH ₃ SeSeNH ₂	neutral	0.21	
	anion	0.55	
CH ₃ SeSeOH	neutral	0.45	0.53
	stretched	0.96	0.52
	bent	0.60	0.40
	book	0.59	0.57
CH ₃ SeSeF	neutral	0.48	0.67
	stretched	0.48	0.58
	bent	0.50	0.32
OHSeSeF	neutral	0.54	0.67
	anion	0.78	0.54
FSeSeF	neutral	0.58	0.66
	anion	1.54	0.64

the electronic density at the Se–Se BCP, the negative energy density value, and the presence of a disynaptic V(Se,Se) basin.

The Se–X Bond. As far as the Se–O bond is concerned, the similar electron affinities of the CH₃SeSe and OH fragments (see Table 6) suggest that they are prone to form 2c–3e bonds, which agrees with the high value of D_e for the *bent* anion. Actually, for both *bent* and *stretched* anions of the CH₃SeSeOH system, the negative charge is evenly distributed between the CH₃SeSe and OH fragments, which suggests that the charge fluctuation stabilization is high, in accordance with the high $\delta(\text{Se},\text{O})$ index for these species.

For both *bent* and *stretched* CH₃SeSeF anions, as well as for the neutral compound, there is no disynaptic V(Se,F) basin (see Figure S1 of the Supporting Information). In this respect, it is noteworthy that for FOOF an unusually long F–O distance was reported and explained in terms of an anomeric effect⁸¹—delocalization of the lone pairs of fluorine into the $\sigma^*(\text{O–F})$ orbital. This charge delocalization would stabilize the

**Figure 6.** Main bond lengths (in Å) for the anionic derivatives of (a) OHSeSeF and (b) FSeSeF calculated at the CCSD/6-31++G(d,p) level. Values in parentheses correspond to the neutral systems.

$\sigma^*(\text{O–F})$ MO to some extent, weakening the O–F linkages and provoking this long bond distance. In our case, a similar effect could explain the absence of the disynaptic V(Se,F) basin in the neutral CH₃SeSeF, which is in accordance with the high delocalization index $\delta(\text{Se},\text{F})$, larger than the $\delta(\text{Se},\text{Se})$ one, and pointing out a significant charge fluctuation stabilization. The large ΔEA between CH₃SeSe and F indicates, however, that the formation of 2c–3e bonds is not so favorable. This is clearly seen in the *bent* isomer, whose dissociation energy (77 kJ mol⁻¹) is close to the lower bound for the dissociation limit of 2c–3e linkages (65–85 kJ mol⁻¹). Actually, the negative charge is mainly located at the fluorine atom, and the stabilizing charge fluctuation between the V(Se) and V(F) basins decreases significantly. This limit situation is somewhat alleviated in the *stretched* isomer, where some of the extra charge goes to the Se–Se linkage (see Table 7), because the electron affinities of the fragments CH₃Se and SeF are not so different. This enhances the charge fluctuation among both Se and F atoms and stabilizes the system. However, since a large fraction of the anionic charge is still at the fluorine, the charge fluctuation between both selenium atoms is not so large, and the V(Se,Se) basin is preserved.

The Effect of Symmetry

At this point, it is worth wondering if the geometrical changes described so far are simply due to the increasing electronegativity of the substituents or whether asymmetry is also a factor to take into account. To answer this question, the molecules FSeSeOH and FSeSeF have been included in our survey. Like for CH₃SeSeOH and CH₃SeSeF, a *stretched* and a *bent* anion were found. However, due to the large F–Se bond length and the high acidity of the OH group, the *bent* FSeSeOH anion directly dissociates into OSeSe^{•-} + HF through a hydrogen transfer from the OH to the F atom. The *bent* anion of FSeSeF corresponds to a transition state, which lowers its energy by opening the FSeSeF dihedral angle, leading to the *stretched* isomer. Regarding the *stretched* derivatives, Figure 6 shows that the electron attachment process in both systems produces only a significant elongation of the Se–F linkage, whereas the Se–Se and the Se–O bonds remain practically unperturbed. In addition, the topology of the ELF shows no disynaptic V(Se,F) basin either in the anion or in the neutral as it occurred for CH₃SeSeF (Figure 7). Interestingly, upon electron capture, the delocalization index between the selenium lone pairs increases by a factor of 3 in FSeSeF,

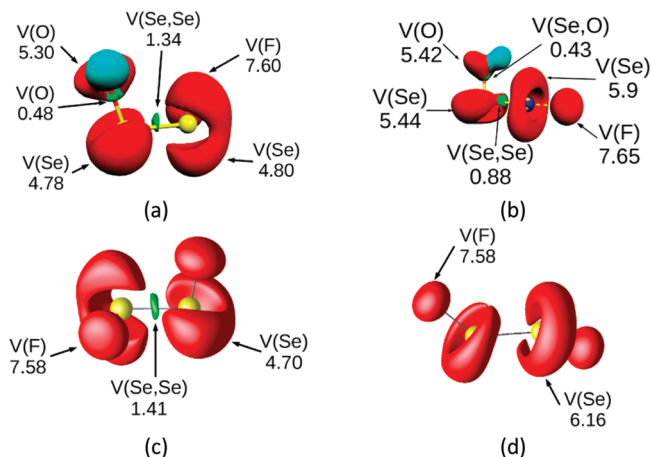


Figure 7. ELF localization domains within the isosurface $ELF = 0.8$ for the (a) neutral and (b) anionic derivatives of $FSeSeOH$ and the (c) neutral and (d) anionic structures of $FSeSeF$. The same color convention as in Figure 5 is used. The population of the different basins is given in e.

the disynaptic $V(Se,Se)$ basin disappears, and the population of the $V(F)$ basin does not change. The natural charge of the F atoms is already -0.5 in the neutral $FSeSeF$ derivative; hence an accumulation of the extra charge on the F atoms of the anion ($Se^{+0.5}F^{-1.0}$) would be highly unlikely, and only the population of the $V(Se)$ basins increases. Obviously, both F atoms bear the same negative charge ($^{-0.5}FSeSeF^{-0.5}$), a picture which is coherent with a significant participation of the $FSeSeF^- \leftrightarrow ^-FSeSeF$ resonance. As a matter of fact, the natural resonance theory shows that the resonant forms $FSeSeF^-$ and ^-FSeSeF contribute 53% to the total wave function. This is also consistent with a significant charge fluctuation along the $Se-Se$ bond, which is behind the large $\delta(Se,Se)$ for the anion and the disappearance of the disynaptic $V(Se,Se)$ basin. This high charge fluctuation along the diselenide bond is less apparent in $OHSeSeF$ since its lack of symmetry makes it possible to localize the negative charge preferentially at the fluorine atom. However, since the OH and $SeSeF$ fragments have rather similar electron affinities, the disynaptic $V(Se,Se)$ basin survives upon electron capture, and the $Se-F$ linkage keeps a $2c-3e$ nature due to the counterbalancing effect of the OH fragment, in agreement with the high value of $\delta(Se,F)$. In conclusion, the $OHSeSeF$ and $FSeSeF$ systems show that, although the bond activation triggered by electron capture depends mainly on the electronegativity of the substituent X, asymmetry turns out to be a fundamental requirement to keep the covalent nature of the $Se-Se$ linkage.

In general, the activated $Se-Se$ and $Se-X$ present the same topological features: the $\nabla^2\rho$ at the BCP is near zero or large and positive, and the ELF shows no disynaptic basin in the bonding region but a large charge fluctuation between the lone-pair basins of the bonding atoms. The positive value of $\nabla^2\rho$ as well as the absence of a disynaptic basin are typical of closed-shell interactions, but these bonds are strengthened through a charge fluctuation mechanism resulting in high D_e values. These topological features are also the signatures of the so-called *charge-shift (CS) bonds*, a new type of chemical bond proposed by Shaik and co-workers,⁷⁶⁻⁷⁸ the

only difference being that CS bonds do present a disynaptic $V(A,B)$ basin, although with a low population. In these linkages, two centers bind together not by sharing an electronic pair but by a charge fluctuation between them, exactly the same mechanism which stabilizes $2c-3e$ bonds. Therefore, these two types of bonds seem to be closely related, as previously suggested.^{74,82}

Conclusions

Through two different and complementary techniques, AIM and ELF, we have analyzed the changes of the bonding situation of a series of diselenide compounds upon electron attachment. These results have been complemented with a population analysis to locate the excess of negative charge, and the different electron affinity of the bonding fragments to rationalize their charge distribution. We have shown that, when the diselenides bear low electronegative substituents, the $Se-Se$ bond becomes activated upon electron capture, as previous studies have shown. However, this is no longer the case for very electronegative substituents, where this bond is practically unaltered and the $Se-X$ one is the one which elongates since the extra electron occupies the $\sigma^*(Se-X)$ antibonding orbital rather than the $\sigma^*(Se-Se)$. When this is the case, several anionic species are also encountered, although based on the relative energies of these isomers, only one of them, namely *stretched*, should be experimentally expected. The present findings are similar to those obtained for a series of analogous disulfide derivatives, which points out that these results are not unique and could be extrapolated to a wider range of compounds than the ones covered here.

The $Se-Se$ (and $Se-X$ in many occasions) linkage bears some of the characteristics of the so-called charge-shift bonds, with a clear charge fluctuation between both selenium atoms. This is more evident in their anions where the bonding reflects the important contribution of the ionic resonant forms $Se-Se^- \leftrightarrow ^-Se-Se$ vs the covalent component $Se::Se$. This resonance changes with the nature of the substituents but also depends on the asymmetry of the substitution.

Acknowledgment. This work has been partially supported by DGI Project No CTQ2006-08558/BQU, Project MADRISOLAR2, ref.:P2009/PPQ-1533 of the Comunidad Autonoma de Madrid and by the COST Action COST CM0702. A generous allocation of computing time at the CCC of the UAM is also acknowledged. J.A.G. acknowledges a contract from the Comunidad Autonoma de Madrid.

Supporting Information Available: Geometries (in Cartesian coordinates) of all the compounds present in this work calculated at the CCSD/6-31++G(d,p) level of theory, charge analysis and the corresponding discussion to assess the reliability of the MP2 approach, HF eigenvalues of the SOMOs of the most stable anionic species, NBO spin densities of all the anions, and the ELF analysis of all the neutral molecules. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Schwarz, K.; Foltz, C. M. *J. Am. Chem. Soc.* **1957**, *79*, 632.
- (2) Shchedrina, V. A.; Novoselov, S. V. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 13919.
- (3) Kryukov, G. V.; Castellano, S.; Novoselov, S. V.; Lobanov, A. V.; Zehtab, O.; Guigo, R.; Gladyshev, V. N. *Science* **2003**, *300*, 1439.
- (4) Behne, D.; Pfeifer, H.; Rothlein, D.; Kyriakopoulos, A. *Proceedings of the 10th International Symposium on Trace Elements in Man and Animals*; Plenum Press: New York, 2000.
- (5) Castellano, S.; Lobanov, A. V.; Chapple, C.; Novoselov, S. V.; Albercht, M.; Hua, D.; Lesc7ure, A.; Lengauer, T.; Krol, A.; Gladyshev, V. N.; Guigo, R. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 16188.
- (6) Rayman, M. P. *Lancet* **2000**, *356*, 233.
- (7) Chen, J.; Berry, M. J. *J. Neurochem.* **2003**, *86*, 1.
- (8) Schomburg, L.; Schweizer, U.; Koehrlle, J. *Cell. Mol. Life Sci.* **2004**, *61*, 1988.
- (9) Frank, W. Z. *Physiol. Chem.* **1964**, *339*, 222.
- (10) Walter, R.; du Vigneaud, V. *J. Am. Chem. Soc.* **1965**, *87*, 4192.
- (11) Walter, R.; du Vigneaud, V. *J. Am. Chem. Soc.* **1966**, *88*, 1331.
- (12) Hartrodt, B.; Neubert, K.; Bierwolf, B.; Blech, W.; Jakubke, H. D. *Tetrahedron Lett.* **1980**, *21*, 2393.
- (13) Besse, D.; Budisa, N.; Karnbrock, W.; Minks, C.; Musiol, H. J.; Pregararo, S.; Siedler, F.; Weyher, E.; Moroder, L. *J. Biol. Chem.* **1997**, *378*, 211.
- (14) Koider, T.; Itoh, H.; Otaka, A.; Furuya, M.; Kitajima, Y.; Fujii, N. *Chem. Pharm. Bull.* **1993**, *46*, 5382.
- (15) Rajarathnam, K.; Sykes, B. D.; Dewald, B.; Baggolini, M.; Clark-Lewis, I. *Biochemistry* **1999**, *38*, 7653.
- (16) Metanis, N.; Keinan, E.; Dawson, P. E. *J. Am. Chem. Soc.* **2006**, *128*, 16684.
- (17) Armishaw, C. J.; Daly, N. T.; Adams, D. J.; Craik, D. J.; Alewook, P. F. *J. Biol. Chem.* **2006**, *281*, 14136.
- (18) Fori, S.; Pregararo, S.; Rudolph-Böhner, S.; Cramer, J.; Moroder, L. *Biopolymers* **2000**, *53*, 550.
- (19) Hondal, R. J.; Nilsson, B. L.; Raines, R. T. *J. Am. Chem. Soc.* **2001**, *123*, 5140.
- (20) Nutt, R. F.; Veber, D. F.; Saperstein, R. *J. Am. Chem. Soc.* **1980**, *102*, 6539.
- (21) Stymiest, J. L.; Mitchell, B. F.; Wong, S.; Vederas, J. C. *Org. Lett.* **2003**, *5*, 47.
- (22) Hargittai, B.; Sole, N. A.; Groebe, D. R.; Abramson, S. N.; Barany, G. *Med. Chem.* **2000**, *43*, 4787.
- (23) Bondebjerg, J.; Grunnet, M.; Jespersen, T.; Meldal, M. *Chem. Biol. Chem.* **2003**, *4*, 186.
- (24) Jacob, C.; Giles, G. I.; Giles, N. M.; Sies, H. *Angew. Chem., Int. Ed.* **2003**, *42*, 4742.
- (25) Johansson, I.; Gafvelin, G.; Amér, E. S. *Biochim. Biophys. Acta* **2005**, *1726*, 1.
- (26) Sheu, C.; Sobkowiak, A.; Zhang, L.; Ozbalik, N.; Barton, D. H. R.; Sawyer, D. T. *J. Am. Chem. Soc.* **1989**, *111*, 8030.
- (27) Tian, F.; Yu, Z.; Lu, S. *J. Org. Chem.* **2004**, *69*, 4520.
- (28) Degrand, C.; Nour, M. *J. Electroanal. Chem.* **1984**, *190*, 213.
- (29) Barbosa, N. B. V.; Rocha, J. B. T.; Wondracek, D. C.; Pettroni, J.; Zeni, G.; Nogueira, C. W. *Chem. Biol. Interact.* **2006**, *3*, 230.
- (30) de Freitas, A. S.; Funck, V. R.; Rotta, M. d. S.; Bohrer, D.; Mörschbächer, V.; Puntel, R. L.; Nogueira, C. W.; Farina, M.; Aschner, M.; Rocha, J. B. T. *Brain Res. Bull.* **2009**, *79*, 77–84.
- (31) Arteel, G. E.; Sies, H. *Environ. Toxicol. Pharmacol.* **2001**, *10*, 153.
- (32) Meotti, F. C.; Stangherlin, E. C.; Zeni, G.; Nogueira, C. W.; Rocha, J. B. T. *Environ. Res.* **2004**, *94*, 276.
- (33) Kumar, B. S.; Kunwar, A.; Ahmad, A.; Kumbhare, L. B.; Jain, V. K.; Priyadarsini, K. I. *Radiat. Environ. Biophys.* **2009**, *48*, 379.
- (34) Pearson, J. K.; Boyd, R. J. *J. Phys. Chem. A* **2007**, *111*, 3152–3160.
- (35) Antonello, S.; Daasbjerg, K.; Jensen, H.; Taddei, F.; Maran, F. *J. Am. Chem. Soc.* **2003**, *125*, 14905–14916.
- (36) Sobczyk, M.; Simons, J. *Int. J. Mass Spectrom.* **2006**, *253*, 274–280.
- (37) Anusiewicz, I.; Berdys-Kochanska, J.; Simons, J. *J. Phys. Chem. A* **2005**, *109*, 5801–5813.
- (38) Modelli, A.; Jones, D. *J. Phys. Chem. A* **2006**, *110*, 13195–13201.
- (39) Yamaji, M.; Tojo, S.; Takehira, K.; Tobita, S.; Fujitsuka, M.; Majima, T. *J. Phys. Chem. A* **2006**, *110*, 13487–13491.
- (40) Meija, J.; Beck, T. L.; Caruso, J. A. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 1325–1332.
- (41) Modelli, A.; Jones, D.; Distefano, G.; Tronc, M. *Chem. Phys. Lett.* **1991**, *181*, 361.
- (42) Gámez, J. A.; Serrano-Andrés, L.; Yáñez, M. *Phys. Chem. Chem. Phys.* **2010**, *5*, 1042.
- (43) Gräfenstein, J.; Kraka, E.; Cremer, D. *J. Chem. Phys.* **2004**, *120*, 528–538.
- (44) Polo, V.; Gräfenstein, J.; Kraka, E.; Cremer, D. *Chem. Phys. Lett.* **2002**, *352*, 469–478.
- (45) Gräfenstein, J.; Kraka, E.; Cremer, D. *Phys. Chem. Chem. Phys.* **2004**, *6*, 1096–1112.
- (46) Cohen, A.; Mori-Sánchez, P.; Yang, W. *Science* **2008**, *321*, 792–794.
- (47) Curtiss, L. A.; Redfren, P. C.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 42.
- (48) Rienstra-Kiracofe, J. C.; Tschumper, G. S.; Schaefer, H. F., III; Nandi, S.; Ellison, G. B. *Chem. Rev.* **2002**, *102*, 231–282.
- (49) van Doren, J. M.; Miller, T. M.; Viggiano, V. V. *J. Chem. Phys.* **2008**, *128*, 094310–094317.
- (50) Brařda, B.; Hiberty, P. C.; Savin, A. *J. Phys. Chem. A* **1998**, *102*, 7872–7877.
- (51) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.
- (52) Lee, C.; Yang, W.; Parr, R. G. *J. Phys. Rev. B* **1988**, *37*, 785.
- (53) Brařda, B.; Hiberty, P. C. *J. Phys. Chem. A* **2000**, *104*, 4618.
- (54) Brařda, B.; Hiberty, P. C. *J. Phys. Chem. A* **2003**, *107*, 4741.

- (55) Kaur, D.; Sharma, P.; Bharatam, P. V. *THEOCHEM* **2007**, *810*, 31–37.
- (56) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*; Gaussian Inc.: Wallingford, CT, 2004.
- (57) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Mitrushenkov, A.; Rauhut, G.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Heter, G.; Hrenar, T.; Knizia, 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.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *MOLPRO 2009.01*; Cardiff University: Cardiff, U. K., 2009.
- (58) Pierloot, K.; Dumez, B.; Widmark, P.-O.; Roos, B. O. *Theor. Chem. Acta* **1995**, *90*, 87.
- (59) Veryazov, V.; Widmark, P.-O.; Serrano-Andrés, L.; Lindh, R.; Roos, B. O. *Int. J. Quantum Chem.* **2004**, *100*, 626–635.
- (60) Matta, C. F.; Boyd, R. J. In *The Quantum Theory of Atoms in Molecules*; Matta, C. F., Boyd, R. J., Eds.; Wiley: Weinheim, Germany, 2007; pp 1–30.
- (61) Kohout, M. *DGrid*, version 4.5; Federal Ministry of Education and Research: Radebeul, Germany, 2009.
- (62) Becke, A. D.; Edgecombe, K. E. *J. Chem. Phys.* **1990**, *92*, 5397.
- (63) Silvi, B.; Savin, A. *Nature* **1994**, *371*, 683.
- (64) Silvi, B. *Phys. Chem. Chem. Phys.* **2004**, *6*, 256.
- (65) Noury, S.; Krokidis, X.; Fuster, F.; Silvi, B. *Comput. Chem.* **1999**, *23*, 597.
- (66) Groner, P.; Gillies, C. W.; Gillies, J. Z.; Zhang, Y.; Block, E. *J. Mol. Spectrosc.* **2004**, *226*, 169–181.
- (67) Carles, S.; Lecomte, F.; Schermann, J. P.; Desfrancois, C.; Xu, S.; Nilles, J. M.; Bowen, K. H.; Berges, J.; Houee-Levin, C. *J. Phys. Chem. A* **2001**, *105*, 5622–5626.
- (68) Pauling, L. *J. Am. Chem. Soc.* **1931**, *53*, 3225–3237.
- (69) Pauling, L. *J. Chem. Phys.* **1933**, *1*, 56–59.
- (70) Hiberty, P. C.; Humbel, S.; Archirel, P. *J. Phys. Chem.* **1994**, *98*, 11697–11704.
- (71) Hiberty, P. C.; Shaik, S. In *Valence Bond Theory*; 1st ed.; Cooper, D. L., Ed.; Elsevier: Amsterdam, 2002; pp 207–214.
- (72) Clark, T. *J. Am. Chem. Soc.* **1988**, *110*, 1672–1678.
- (73) Fradera, X.; Auster, M. A.; Bader, R. F. W. *J. Phys. Chem. A* **1999**, *103*, 304.
- (74) Fourré, I.; Silvi, B. *Heteroatom Chem.* **2007**, *18*, 135.
- (75) Fourré, I.; Silvi, B.; Sevin, A.; Chevreau, H. *J. Phys. Chem. A* **2002**, *106*, 2651–2571.
- (76) Shaik, S.; Danovich, D.; Silvi, B.; Lauvergnat, D.; Hiberty, P. *Chem.—Eur. J.* **2005**, *11*, 6358.
- (77) Zhang, L.; Ying, F.; Wu, W.; Hiberty, P.; Shaik, S. *Chem.—Eur. J.* **2009**, *15*, 2979.
- (78) Wu, W.; Song, J.; Shaik, S.; Hiberty, P. *Angew. Chem., Int. Ed.* **2009**, *48*, 1407.
- (79) Hiberty, P. C.; Humbel, S.; Archirel, P. *J. Phys. Chem. A* **1994**, *98*, 11697.
- (80) Fourré, I.; Bergès, J. *J. Phys. Chem. A* **2004**, *108*, 898–906.
- (81) Kraka, E.; He, Y.; Cremer, D. *J. Phys. Chem. A* **2001**, *105*, 3269–3276.
- (82) Bil, A.; Berski, S.; Latajka, Z. *J. Chem. Inf. Model.* **2007**, *47*, 1021–1030.

CT100336Q

Relativistic Effects on Metal–Metal Bonding: Comparison of the Performance of ECP and Scalar DKH Description on the Picture of Metal–Metal Bonding in $\text{Re}_2\text{Cl}_8^{2-}$

Robert Ponec,^{*,†} Lukáš Bučinský,[‡] and Carlo Gatti[§]

Institute of Chemical Process Fundamentals, Academy of Sciences of the Czech Republic v.v.i., Prague 6, Suchbát 2, 165 02 Czech Republic, Institute of Physical Chemistry and Chemical Physics, Slovak University of Technology in Bratislava, Bratislava, Slovakia, and Istituto di Scienze e Tecnologie Molecolari del CNR (CNR-ISTM) e Dipartimento di Chimica Fisica ed Elettrochimica, Università di Milano, Via Golgi 19, I-20133, Milano, Italy

Received June 21, 2010

Abstract: This paper reports a systematic comparison of the performance of alternative methods of including relativistic effects on the nature of metal–metal bonding in the $\text{Re}_2\text{Cl}_8^{2-}$ anion. The comparison involved the description using a scalar relativistic Douglas–Kroll–Hess (DKH2) Hamiltonian with all-electron basis sets and the relativistic effective core potential (ECP) basis sets. The impact of the above methods on the picture of the bonding was analyzed using the so-called domain averaged Fermi holes (DAFH). Besides comparing the impact on the picture of the bonding of the two above methods, the focus was also put on the systematic comparison of the “exact” AIM generalized form of DAFH analysis with the approximate Mulliken-like approach used in an earlier DAFH study of ReRe bonding. It has been shown that in contrast to descriptions using ECP basis sets where the differences in the picture of the bonding emerging from the approximate and “exact” DAFH analysis are only marginal, the approximate DAFH approach has been found to dramatically fail in the case of all-electron basis sets required for the description in terms of the Douglas–Kroll–Hess (DKH2) Hamiltonian.

Introduction

Since its discovery in the early 1960s, the chemistry of molecules involving direct metal–metal bonds has been the subject of growing interest. The motivation for these studies not only originated from the new perspectives that the existence of molecules with metal–metal bonds opened for the synthesis of new molecules and materials but, of the same or comparable importance, the challenge was also for the chemical theory to explain often unusual multiplicities of metal–metal bonds. One of the important factors undoubtedly contributing to the peculiarity of metal–metal bonds, especially among heavy transition metals such as Re, is the

relativistic effects.^{1–4} An example in this respect can be, e.g., the contraction of atomic radii of heavier metals often referred to as lanthanoid or actinoid contraction^{5,6} and the great enhancement of the so-called inert-pair effect,⁷ which is the main cause of the specificities in the chemistry of gold and mercury,^{7–10} and an inherently relativistic effect is also spin–orbit coupling. Because of the important impact of these effects on various observable molecular properties like bond lengths, bond energies, valency changes, magnetic properties, etc., the correct description of heavy atom chemistry requires taking the relativistic phenomena into account. The most general formulation of the relativistic quantum theory of free electron is provided by the four-component Dirac equation.¹¹ Despite considerable progress in the formulation of the procedures providing the solution at this fundamental level,¹² the application of such approaches is still considerably time-consuming and, if pursued,

* Corresponding author e-mail: rponec@icpf.cas.cz.

† Academy of Sciences of the Czech Republic v.v.i.

‡ Slovak University of Technology in Bratislava.

§ CNR-ISTM-Università di Milano.

would lead to lengthy and complex calculations. For that reason, and because of the importance of a reliable description of relativistic effects for bigger molecules of real chemical interest, various approximate quasi-relativistic approaches were proposed in past years with the aim to avoid the complexity of the four-component Dirac theory without a sacrifice of the accuracy. An example in this respect can be, e.g., explicit inclusion of relativistic effects via the Douglas–Kroll–Hess Hamiltonian,^{13,15} but widespread use has found also another conceptually different approach based on the use of effective core potentials.^{16–20}

The increasing interest in the detailed scrutiny of relativistic effects on the electron structure of molecules with heavy transition metals and/or actinides finds its manifestations, e.g., in the studies given in refs 21–30.

Our aim in this study is to report the detailed comparison of the performance of the above two types of approaches for the description of the picture of the bonding of multiple metal–metal bonds in $\text{Re}_2\text{Cl}_8^{2-}$. Because of the prominent position of this ion as a paradigm for quadruple metal–metal bonding, the nature of the Re–Re bond has been the subject of numerous studies.^{31–43} Into the framework of these efforts can also be included our recent study in which the nature of Re–Re bonding was discussed using the methodology known as the analysis of domain averaged Fermi holes.⁴⁰ Its conclusions were completely consistent with the results of other theoretical studies, according to which one of the four shared electron pairs involved in the bonding is considerably weaker than the remaining three so that the bond can be best classified as an *effective* triple bond. The analysis was performed in the study⁴⁰ using the approximate Mulliken-like approach, and the relativistic effects were included via the effective core potentials at the B3LYP/LANL2DZ level of the theory. In this report, we are going to extend the original study through a detailed comparison that involves (i) the eventual impact on the picture of the bonding of the alternative inclusion of the relativistic effects via the Douglas–Kroll–Hess Hamiltonian and all-electron basis sets and (ii) a detailed inspection of the picture of the bonding resulting from the approximate Mulliken-like and “exact” AIM generalized form of DAFH analysis so as to reveal any eventual bias resulting from the original use of the approximate Mulliken-like approach.

Theoretical

The DAFH analysis was proposed some time ago^{44–46} as a new tool for the description and visualization of the bonding interactions, and its recent applications in the realm of metal–metal bonding^{47–50} have provided new insights revealing the peculiarities of this widely studied type of bonds. Because most of such bonds, especially among heavier metals, can undoubtedly be affected by the operation of relativistic effects, we decided in this study to focus on just to what extent the picture of the metal–metal bonding can be sensitive to various methods of including relativistic effects in a quantum chemical description. Although the formalism of DAFH analysis was repeatedly described in earlier studies, we consider it useful to review the basic ideas of the approach and, also, to incorporate the original

intuitively formulated formalism into the language of a strict, more elegant, mathematical description.

The original introduction of domain averaged Fermi holes, $g_\Omega(r)$, referred to the concept of the correlation hole,⁵¹ from which the corresponding holes are derived by the selective integration over the coordinates of one electron of the pair, as suggested in eq 1.

$$C(r, r') = 2\rho(r, r') - \rho(r)\rho(r')$$

$$g_\Omega(r) = - \int_\Omega C(r, r') dr' = \rho(r) \int_\Omega \rho(r') dr' - 2 \int_\Omega \rho(r, r') dr' \quad (1)$$

where $\rho(r)$ and $\rho(r, r')$ denote ordinary one-electron and pair density, respectively, and the integration is over the finite domain Ω . For the sake of mathematical elegance, it is possible to reformulate the whole approach in terms of density matrices rather than densities, although the specific choice of either formulation has no impact on the practical applicability of the DAFH analysis. For this purpose, let us introduce first the exchange-correlation density matrix (eq 2)

$$\rho^{\text{exc}}(r_1, r_1'; r_2, r_2') = \rho(r_1, r_1')\rho(r_2, r_2') - 2\rho^{(2)}(r_1, r_1'; r_2, r_2') \quad (2)$$

which measures the deviation of two electrons from independent electron behavior to truly correlated behavior via one-electron density matrices $\rho(r_1, r_1')$ and $\rho(r_2, r_2')$ and the two-electron density matrix $\rho^{(2)}(r_1, r_1'; r_2, r_2')$. On the basis of eq 2, it is possible to introduce the domain averaged Fermi-hole matrix $g_\Omega(r_1, r_1')$ by the selective integration of the matrix (eq 2) over the finite domain Ω .

$$g_\Omega(r_1, r_1') = \int_{r_2=r_2'; \Omega} \rho^{\text{exc}}(r_1, r_1'; r_2, r_2') dr_2 \quad (3)$$

Although the choice of the domain is in principle arbitrary and the corresponding Fermi hole densities and/or holes can be averaged over domains of arbitrary shape and size, in previous studies, we have demonstrated that especially interesting and chemically relevant information can be extracted from these holes if the integration domains coincide with AIM atomic domains resulting from the virial partitioning of electron density.⁵² Constructed and analyzed can be, however, also more complex domains formed by union of several atomic domains corresponding to, e.g., various functional groups or interesting molecular fragments. In such a case, the holes provide the information about the electron pairs retained in the domain as well as about broken or dangling valences created by the formal splitting of the bonds required for the isolation of the domain from the rest of the molecule. The DAFH analysis involves, as a first step, the diagonalization of the matrix representing the Fermi hole density in an appropriate basis. The eigenvectors and eigenvalues resulting from such a diagonalization are then, in the second step, subjected to the so-called isopycnic transformation,⁵³ whose aim is to transform the original eigenvectors to more localized functions that provide an appealing and highly visual description of the molecular structure in terms close to classical chemical thinking. The

structural information is primarily extracted from the resulting eigenvalues which allow detection of electron pairs (chemical bonds, core and lone pairs) retained in the domain as well as broken or dangling valences formed by the formal splitting of the bonds required for the isolation of the domain from the rest of the molecule. The above interpretation is then significantly facilitated by the graphical display of the eigenvectors associated with the corresponding eigenvalues.

Although the definition of DAFH is completely general and can be applied at any level of the theory, most of the previous applications in the realm of metal–metal bonding^{47–50} have been performed using several simplifying approximations.

The first of them concerns the pair density. The extraction of this density from post-Hartree–Fock calculations is not an easy task, and the use of DAFH analysis at a correlated level of the theory has been reported only recently and for small systems.^{54,55} For the complexes of transition metals, such an approach is still beyond the reach of present possibilities. Much more feasible is the approach based on the Kohn–Sham DFT level of the theory,^{56,57} which, especially in inorganic chemistry, represents the contemporary computational standard. In such a case, the general definition of the DAFH (eq 1) reduces to

$$g_{\Omega}(r_1, r_1') = 2 \sum_i^{\text{occ}} \sum_j^{\text{occ}} \langle ij \rangle_{\Omega} \varphi_i(r_1) \varphi_j(r_1') \quad (4)$$

where

$$\langle ij \rangle_{\Omega} = \int_{\Omega} \varphi_i(r_2) \varphi_j(r_2) dr_2 \quad (5)$$

denotes the overlap integral of molecular orbitals i and j over the domain Ω . Formula 4 results from the general definition (eq 1) if the pair density is formally constructed from the first-order density matrix using the formula exactly valid at one-determinantal Hartree–Fock level. Its use at the DFT level of theory, which provides only the first-order density and not the density matrix, thus certainly represents a kind of heuristic extension, but the results of earlier applications of this approach to other systems with metal–metal bonds provide a clear justification for the reliability of the picture of the bonding in such systems.^{47–50} This is also true for an earlier study of Re–Re bonding,⁴⁰ where, consistent with the results of recent sophisticated CASSCF calculations,^{37,39} the analysis detected the reduction of Re–Re bond multiplicity resulting from the partial population of the antibonding δ^* orbital. This result thus in a sense confirms the conclusions of the recent study by Truhlar et al.,⁵⁸ which claims that the use of density-based exchange functionals may provide a more theoretically justified way to treat transition metals than post-Hartree–Fock wave function methods and not just a cost-effective alternative to wave-function-based methods.

The second approximation which was often used in our earlier DAFH studies of transition metal systems concerned the simplification of replacing the exact integration over AIM domains by the Mulliken-like approximation, according to which the electron is in the domain of atom A if it is in the orbital attached to that atom. The simplest implementation

of such an approximate DAFH analysis is based on a simplification of formula 5 for the elements of the AOM matrix due to a transformation to a symmetrically orthogonalized AO basis

$$\mathbf{o} = \mathbf{S}^{1/2} \mathbf{c} \quad (6)$$

In such a case, the elements of the AOM matrix associated with atom A can be written as

$$\langle ij \rangle_A = \sum_{\mu} \sum_{\nu} c_{\mu} c_{\nu} \rho_{\mu\nu}^A \rightarrow \sum_{\mu} o_{\mu i} o_{\mu j} \quad (7)$$

so that the attachment of (orthogonalized) basis functions to atoms is straightforward. Using the orthogonalized AO basis, the elements of the \mathbf{g} matrix (eq 4) can be written as

$$\mathbf{G}_{\lambda\sigma}^A = \frac{1}{2} \sum_{\mu} \mathbf{D}_{\mu\lambda} \mathbf{D}_{\mu\sigma} \quad (8)$$

where \mathbf{D} , given by eq 9,

$$\mathbf{D} = \mathbf{S}^{1/2} \mathbf{P} \mathbf{S}^{1/2} \quad (9)$$

is the first-order density matrix in the orthogonalized AO basis. For the purpose of graphical display, the resulting localized eigenvectors have to be, of course, back-transformed into the original nonorthogonal basis.

The reason for the use of this approximation was that in early studies on metal–metal bonding,^{47–50} the transition metals were described using relativistic ECP basis sets, which are known to produce densities whose integration is not always straightforward.^{59–61} As we are now able to overcome the above numerical problems, the possibility to use the DAFH analysis at an exact AIM-generalized level of theory opens the way to a detailed study of the eventual bias resulting for the picture of Re–Re bonding from the use of an approximate Mulliken-like form of the approach.

In addition to this, the availability of the exact AIM generalized form of the analysis also opens the possibility for the direct comparison of the performance of various methods of inclusion of relativistic effects on the picture of the Re–Re bonding. Recently, such a comparison between the exact AIM-generalized level and the approximate Mulliken form of the DAFH analysis has been performed for two prototypical systems (homoleptic binuclear metal carbonyls) for which the 18-electron rule predicted the existence of a direct metal–metal bond. As it has been shown, the picture of the bonding provided by both forms of DAFH analysis is remarkably consistent and clearly indicates that residual bonding interactions between metals originate from multicenter bonding involving bridging ligands.⁶²

For the sake of straightforward comparison with an earlier study,⁴⁰ we focus here on the description of relativistic effects via ECP basis set LANL2DZ¹⁸ and the scalar relativistic DKH2 approach using the all-electron VTZ basis.⁶³ Such a comparison is performed for the fixed molecular geometry used in the study,⁴⁰ which resulted from the optimization at the B3LYP/LANL2DZ level of theory and which closely reproduced the experimental geometry. In addition to this, we also present a comparison of geometry optimization using

Table 1. Comparison of Experimental and Calculated (B3LYP/LANL2DZ) Geometrical Parameters of the $\text{Re}_2\text{Cl}_8^{2-}$ Ion

geometrical parameters	calculated (bond lengths in pm, angles in deg)	experimental (bond lengths in pm, angles in deg)
R_{ReRe}	221	224 ⁷⁴ 222 ⁷⁵
R_{ReCl}	243	229 ⁷⁴ 243 ⁷⁵
$\angle\text{ReReCl}$	105	104 ⁷⁴ 104 ⁷⁵

the relativistic one-component (scalar) DKH2, two-component DKH2, and nonrelativistic description in all-electron uncontracted DZ basis^{63,64} to see to what extent the inclusion and/or omission of relativistic effects and spin-orbit interactions can affect the geometrical structure of the $\text{Re}_2\text{Cl}_8^{2-}$ ion.

Computations

The application of DAFH analysis requires to perform several types of calculations. First, it is necessary to generate the wave function using ordinary quantum chemical methods. In this study, these calculations were performed for the fixed geometry of the ion generated for the purpose of an earlier DAFH study on the bonding in the $\text{Re}_2\text{Cl}_8^{2-}$ anion⁴⁰ by the geometry optimization at the B3LYP/LANL2DZ level of the theory. The comparison of calculated geometrical parameters with the experimental data is summarized in Table 1. For the sake of a straightforward comparison of the eventual impact on the picture of the bonding of various methods of including the relativistic effects, the wave functions were generated using (i) a relativistic ECP basis set at the B3LYP/LANL2DZ level of the theory and (ii) a scalar relativistic DKH2 approach using the all-electron VTZ basis sets spanned for the Re atom by the Foldy-Wouthuysen recontracted TZ basis set of Dyall⁶³ and for the Cl atom by the Douglas-Kroll recontracted cc-VTZ basis set.^{64,65} Moreover, the Re TZ basis set of Dyall⁶³ has been extended by valence correlating functions⁵⁰ with the final contraction scheme (29s24p15d11f)→[8s7p6d4f]. On the basis of this primary wave function generated by Gaussian 03,⁶⁶ the second step involves the generation of AOM matrices required, according to eq 2, for the generation of the DAFH. These calculations were performed using the PROAIM and AIMAll programs.^{67,68} After having generated the AOM matrices, the last step consists of the generation and subsequent analysis of the DAFH (eq 1), using our program WinBader, which is available upon request. In the following part, the results of these analyses are reported. In addition to the detailed study of the eventual impact on the picture of the bonding of various alternative methods of inclusion of relativistic effects for the fixed geometry of the $\text{Re}_2\text{Cl}_8^{2-}$ ion, we have also focused on the impact of the same methods on the optimized molecular geometry of the $\text{Re}_2\text{Cl}_8^{2-}$ ion. For this purpose, we performed a reoptimization of the geometry at the one-component (scalar) DKH2, two-component DKH2, and nonrelativistic B3LYP levels of the theory using the uncontracted DZ^{63,64} basis set. These calculations were performed

with the Dirac08 package.¹² At the DKH2 level, first-order methods have been used, and the resulting reoptimized geometries are summarized in Table 4.

Results and Discussion

Before pursuing the primary goal of this study, which is a comparison of the impact on the picture of the ReRe bonding of various procedures of inclusion of relativistic effects, we first briefly recall the results of the previous DAFH study in which the analysis of the bonding interactions was performed using an approximate Mulliken-like approach at the B3LYP/LANL2DZ level of the theory. Of decisive importance in this respect are the analyses of the DAFHs averaged over the fragments Re-Re and ReCl_4 . According to general interpretation, the DAFH analysis of the Fermi hole averaged over a certain fragment reveals the existence of electron pairs (chemical bonds, core or lone pairs) retained in the fragment as well as broken or dangling valences resulting from the formal splitting of the bonds required to isolate the fragment from the rest of the molecule. Thus, e.g., in the case of the hole averaged over the ReCl_4 fragment, one should see, besides electron pairs of inner shells on Re and lone pairs on Cl atoms, the existence of electron pairs of ReCl bonds as well as the broken or dangling valences whose number should indicate the number of broken ReRe bonding electron pairs, i.e., the multiplicity of the metal-metal bond. This information can then be independently checked by the analysis of the complementary hole averaged over the fragment ReRe, which directly detects the electron pairs involved in metal bonding. This implies that if the ReRe bond was indeed a quadruple bond, the information from both complementary analyses (for both the fragments ReRe and ReCl_4) should be consistent and indicate the presence of four bonding electron pairs as well as four broken valences. In the original study⁴⁰ based on the approximate Mulliken-like approach, such an ideal complementarity was not observed. Besides four bonding electron pairs (one σ , two π , and one δ) involved in ReRe bonding, the analysis detected in the same fragment the partial population of the antibonding δ^* orbital, whose contribution partly cancels the bonding contribution of the bonding δ orbital and thus reduces the effective multiplicity of the ReRe bond to 3. As, however, the above picture of the bonding resulted from the approximate Mulliken-like form of the analysis, it was of interest to see whether or to what extent the picture of the bonding could be affected if the approximate Mulliken-like DAFH analysis was replaced by the exact AIM-generalized description. In the following part, we are going to show that the differences between the approximate Mulliken-like and exact AIM-generalized description are both qualitatively and quantitatively only marginal for the analysis based on B3LYP/LANL2DZ calculations. The first indication of the close resemblance of the picture of the ReRe bonding comes already from the comparison of calculated ReRe bond orders characterized by Wiberg-Mayer bond indices^{69,70} and/or their AIM generalized counterparts.^{71,72} The actual values, equal to 2.836 and 2.690, respectively, are not only close numerically but the fact that both of them deviate significantly from the ideal value of 4 also indicates considerable

Table 2. Summary of Numerical Results of DAFH Analysis of the $\text{Re}_2\text{Cl}_8^{2-}$ Ion in Case of Relativistic Effects Included Using the ECP Basis Set

method	B3LYP/LANL2DZ				interpretation
	AIM		Mulliken		
fragment	eigenvalue	degeneracy	eigenvalue	degeneracy	
ReCl ₄	≈ 2	4	≈ 2	4	inner shells on Re
	≈ 2	4	≈ 2	4	σ_{ReCl}
	≈ 2	12	≈ 2	12	lone pairs on Cl
	1.070	1	1.030	1	broken valence σ_{ReRe}
	1.035	2	1.060	2	broken valence π_{ReRe}
	1.091	1	1.13	1	broken valence δ_{ReRe}
ReRe	≈ 2	8	≈ 2	8	inner shells on Re
	1.96	1	1.99	1	σ_{ReRe}
	1.88	2	1.99	2	π_{ReRe}
	1.83	1	1.99	1	δ_{ReRe}
	0.435	1	0.63	1	δ^*_{ReRe}
	0.407	8	0.65	8	broken valence σ_{ReCl}

reduction of the multiplicity of the ReRe bond. In order to reveal the origin of this reduction as well as demonstrate the close resemblance of both descriptions, the results of both the exact AIM-generalized and Mulliken-like DAFH analyses for the fragments ReRe and ReCl_4 are summarized in Table 2 and in graphical form in Figures 1 and 2. Thus, e.g., the analysis of the hole averaged over the fragment ReCl_4 yields in both cases 24 nonzero eigenvalues of which 20 are very close to 2. The inspection of the corresponding eigenvalues shows that four of them correspond to electron pairs of inner *ns* and *np* shells on Re. Besides this, there are another 12 electron pairs corresponding to lone pairs on chlorine ligands

as well as four electron pairs involved in four ReCl σ bonds retained in the fragment (Figure 1d,h). What remains are the four eigenvalues close to 1, and the inspection of the associated eigenvectors shows that they indeed correspond to four broken valences (one σ , two π , and one δ component) of the ReRe bond (Figure 1a,e; b,f; c,g). These results are completely consistent with the quadruple multiplicity of the ReRe bond, and the close resemblance of the picture of the bonding is also clearly evident from Figure 1, which displays the comparison of selected eigenvectors resulting from both exact AIM-generalized and approximate Mulliken-like analyses for the fragment ReCl_4 .

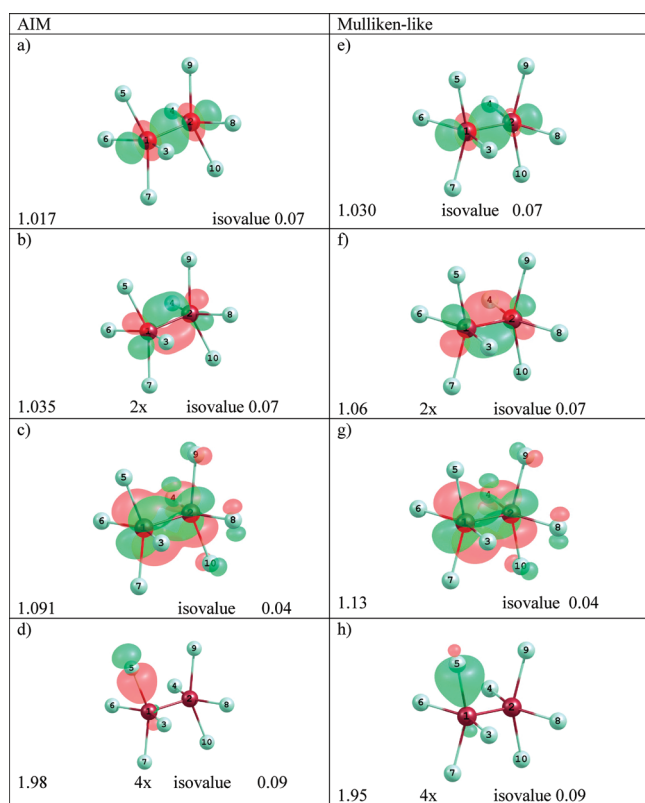
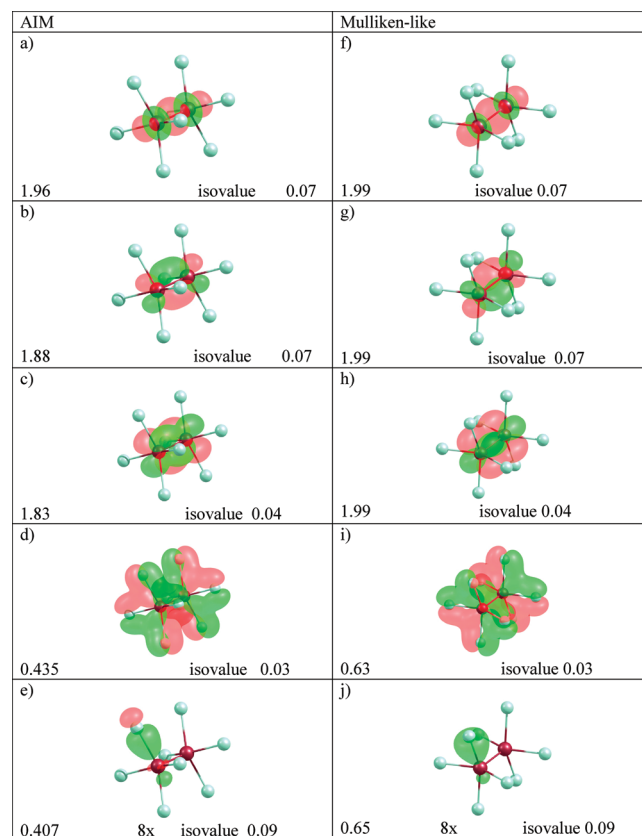
**Figure 1.** Comparison of selected eigenvectors resulting from the “exact” AIM generalized and approximate Mulliken-like DAFH analysis for the hole averaged at the B3LYP/LANL2DZ level of the theory over the fragment ReCl_4 . [Pictures were generated using ChemCraft, v. 1.6.]**Figure 2.** Comparison of selected eigenvectors resulting from the “exact” AIM generalized and approximate Mulliken-like DAFH analysis for the hole averaged at the B3LYP/LANL2DZ level of the theory over the fragment ReRe.

Table 3. Summary of Numerical Results of the DAFH Analysis of the $\text{Re}_2\text{Cl}_8^{2-}$ Ion in the Case of Relativistic Effects Included via the DKH2 Level of Theory Using the Relativistic All Electron VTZ Basis Set

base fragment	DKH2/VTZ					interpretation
	AIM		Mulliken			
	eigenvalue	degeneracy	eigenvalue	degeneracy		
ReCl ₄	≈ 2	20	≈ 2	20	inner shells on Cl atoms	
	≈ 2	12	≈ 2	12	lone pairs on Cl atoms	
	≈ 2	34	≈ 2	34	inner shells on Re	
	1.98	4	1.91	4	σ_{ReCl}	
	1.013	1	1.04	1	broken valence σ_{ReRe}	
	1.035	2	1.05	2	broken valence π_{ReRe}	
	1.08	1	1.09	1	broken valence δ_{ReRe}	
ReRe	≈ 2	68	≈ 2	68	inner shells on Re	
	1.93	1	1.98	1	σ_{ReRe}	
	1.87	2	1.97	2	π_{ReRe}	
	1.78	1	1.94	1	δ_{ReRe}	
	0.39	1			δ^*_{ReRe}	
	0.38	8		8	broken valence σ_{ReCl}	
				0.23–0.40	20 weird eigenvectors	

Similar close resemblance between the exact and approximate description is also observed for the analysis of the hole averaged over the complementary fragment ReRe. However, in contrast to the previous hole averaged over the fragment ReCl_4 whose analysis was straightforwardly consistent with quadruple multiplicity of the ReRe bond, an inspection of the results in this case indicates that the situation with metal–metal bonding is a bit more complex. The DAFH analysis of the hole averaged over the ReRe fragment yields, namely, in both cases, 21 nonzero eigenvalues, of which 12 are close to 2 and the remaining 9 vary between 0.41 and 0.65. The inspection of the eigenvectors corresponding to electron pairs shows that eight of them correspond to electron pairs of completely filled inner shells on Re atoms and as such are not involved in metal–metal bonding. What remains are four electron pairs associated with the remaining four eigenvalues close to 2, and the graphical inspection of these eigenvectors shows that they correspond to one σ , two π , and one δ component of the ReRe bond. As can be seen from the Figure 2 (Figure 2a vs f; b vs g; c vs h), the differences between the exact and approximated description are again very marginal. Although the existence of the above four bonding electron pairs is indeed consistent with the assumed quadruple multiplicity of the Re–Re bond, the problem with the above classification is that besides the above-mentioned four bonding electron pairs, the DAFH analysis again detected the existence of a partially populated δ^* orbital (Figure 2d vs i). This partially populated δ^* orbital corresponds to one of nine remaining eigenvectors, and it is associated with the eigenvalues 0.435 and/or 0.63 for the exact AIM generalized and approximate Mulliken-like description, respectively. Because the population of this antibonding δ^* orbital partially cancels the bonding contribution of the δ pair, the nature of Re–Re bond is again most consistent with a classification of an *effective* triple bond. The remaining eight partially populated eigenvectors correspond, as expected, to the broken valences of ReCl σ bonds (Figure 2e vs j). The differences in the corresponding eigenvalues (0.407 vs 0.65) only reflect the differences of AIM generalized and Mulliken-like descriptions in characterizing the uneven sharing of a bonding electron pair in the polar ReCl bond. On the basis of the above values, and consistent with known general trends, the polarity of the ReCl

bond estimated by the AIM generalized description exceeds the one from the approximate Mulliken-like approach.

After having demonstrated the close resemblance of the picture of the bonding based on the approximate Mulliken-like and “exact” AIM generalized form of the DAFH analysis in the case of relativistic effects included via the ECP basis set LANL2DZ, we are now going to scrutinize similarly the picture of the bonding resulting from the scalar relativistic DKH2 description using an all-electron VTZ basis.⁶³ The results of these analyses for the Fermi holes averaged over the fragments ReCl_4 and ReRe are summarized in Table 3 and Figures 3 and 4.

The simplest situation is again for the hole averaged over the fragment ReCl_4 where the analysis detects, besides others, the existence of four dangling valences corresponding to four formally broken electron pairs involved in ReRe bonding (Figure 3a vs e; b vs f; c vs g) as well as four electron pairs of localized ReCl σ bonds. (Figure 3d vs h). In this respect, the all-electron results are very similar to what was observed in the previous case using ECPs (Figure 1); the only difference is that because of using the all-electron basis set, the analysis now detects a much higher number (70) of eigenvalues close to 2, which correspond to completely filled inner shells on Re and Cl atoms.

The situation is, however, a bit more complex in the case of the hole averaged over the fragment ReRe. In this case, namely, the close resemblance to the results of previous analyses, using the LANL2DZ basis is observed only for the exact AIM-generalized approach, but for the approximate Mulliken-like approach, one observes clear differences that considerably question the reliability of Mulliken-like approach in the case of flexible all-electron relativistic basis sets. The first indication of possible inadequacy of the Mulliken-like approach is reflected already in the values of calculated ReRe bond orders [0.64 (Mulliken-like) vs 2.802 (AIM)]. While the value for the exact AIM generalized approach is not too much different from what was observed in the LANL2DZ basis, the Mulliken-like value is completely out of range of reasonable. The situation thus in a sense resembles the known failure of Mulliken population analysis for the extensive basis sets involving diffuse functions.⁷³

Besides the above differences, the possible inadequacy of the approximate Mulliken-like DAFH analysis using a

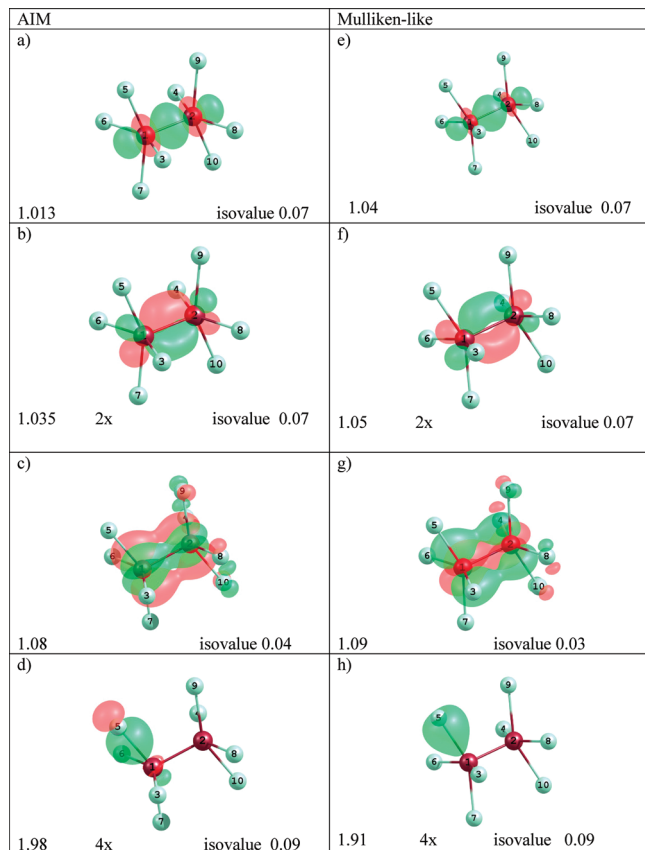


Figure 3. Comparison of selected eigenvectors resulting from the “exact” AIM generalized and approximate Mulliken-like DAFH analysis for the hole averaged at the DKH2/VTZ level of theory over the fragment ReCl_4 .

flexible relativistic all-electron basis set is also reflected in the important qualitative implications for the picture of the bonding emerging from the exact AIM generalized vs the approximate Mulliken-like form of DAFH analysis of the hole averaged over the fragment ReRe .

The results of such analyses, summarized in Table 3 and Figure 4, indicate that, while the AIM generalized approach detects, similarly as in the case of the ECP-based description, the existence of a partially populated antibonding δ^* orbital (Figure 4d), such an orbital is completely absent in the case of approximate Mulliken-like approach (Figure 4i). In addition to this, the Mulliken-like analysis also detects, however, the existence of a large number (20) of other weird eigenvectors with populations ranging between 0.23 and 0.40, whose contributions dramatically complicate a simple and transparent picture of the ReRe bonding.

This result is very important as it implies that the manifestations of relativistic effects on the picture of the ReRe bonding do not dramatically depend on the particular choice of the approach (scalar DKH2 description with all electron basis vs ECP basis) provided that the DAFH analysis is performed using the exact AIM-generalized approach. On the other hand, in the case of the approximate Mulliken-like form of the DAFH analysis, the reliable and internally consistent results are observed only in the case of relativistic effects included via small ECP basis sets, while in the case of the DKH2 description with relativistic all-electron basis sets, the use of the approximate Mulliken-like approach is

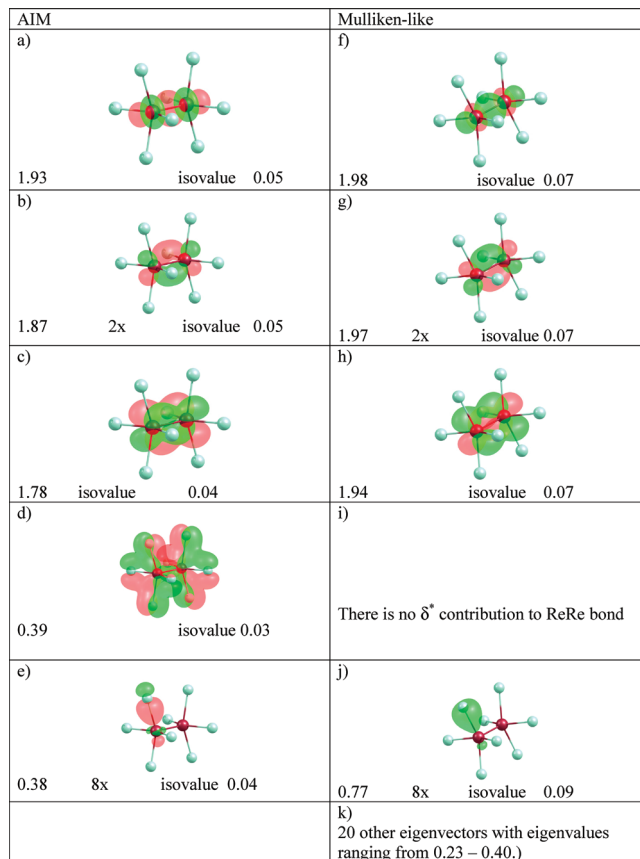


Figure 4. Comparison of selected eigenvectors resulting from the “exact” AIM generalized and approximate Mulliken-like DAFH analysis for the hole averaged at the DKH2/VTZ level of theory over the fragment ReRe .

questionable. In order to understand the reasons for the observed failure of Mulliken-like analysis for the flexible all-electron basis set, we have performed additional calculations using relativistic ECP LANL2TZ(f)^{18,76} for Re and the cc-pVTZ⁶⁴ basis for Cl, which explicitly contains one f function on Re. The results of both population and DAFH analysis in this basis are very reminiscent of the results in the LANL2DZ basis, which quite consistently keeps with the fact that the one f function in the ECP LANL2TZ(f) basis set is less diffuse, hence less “Mulliken-problematic”, than f functions in the all-electron VTZ basis.⁶³

After having presented the impact on the picture of the bonding of two alternative methods of including relativistic effects (AE-DKH2 vs ECP), we report, in the following part, the results of the systematic comparison of the effect of the same methods on the reoptimization of the molecular geometry of the $\text{Re}_2\text{Cl}_8^{2-}$ ion. The crucial geometrical parameters of the reoptimized structures are summarized in Table 4. As is possible to see from the comparison with the Table 1, the best agreement with the experimental geometry is obtained using the LANL2DZ ECP basis sets (for both Re and Cl). On the other hand, the geometry resulting from the optimization at the DKH2 level of theory using the all-electron basis set exhibits quite remarkable differences, and it is interesting that very similar structural parameters were obtained also using the LANL2TZ(f) ECP on Re and an AE cc-pVTZ basis set on the Cl atom (see Table 4)—a result

Table 4. Geometrical Parameters Resulting from the Optimization of the Molecular Geometry of the $\text{Re}_2\text{Cl}_8^{2-}$ Ion Using Various Methods of Inclusion of Relativistic Effects

method/basis	geometrical parameters	calculated values (bond lengths in pm, angles in deg)
nonrelativistic B3LYP/uncontractedDZ	R_{ReRe}	219.5
	R_{ReCl}	241.6
	$\angle\text{ReReCl}$	105.1
1c-DKH2 B3LYP/uncontractedDZ	R_{ReRe}	219.2
	R_{ReCl}	237.4
	$\angle\text{ReReCl}$	104.9
2c-DKH2 B3LYP/uncontractedDZ	R_{ReRe}	219.2
	R_{ReCl}	237.5
	$\angle\text{ReReCl}$	104.9
ECP-LANL2TZ ^a B3LYP/TZ	R_{ReRe}	219.6
	R_{ReCl}	236.0
	$\angle\text{ReReCl}$	104.9

^a Re, LANL2TZ(f); Cl, cc-pVTZ.

which raises some doubts on the physical significance of almost perfect matching between the experimental geometry and the theoretical prediction using the smaller basis set ECP model. Indeed, the geometry optimization with the LANL2DZ ECP might be influenced by error cancellations due to the fortunate combined effect of (a) the larger basis set superposition error (BSSE) inherent to a smaller basis set; (b) the specific contraction scheme of the original Hay and Wadt uncontracted Gaussian basis adopted in the LANL2DZ basis in contrast to the fully uncontracted scheme used for the LANL2TZ basis; and, finally, (c) the lack of the “f” polarization function in the LANL2DZ basis. A detailed and separate analysis of such effects, aimed at dissecting their specific role and relevance in determining the $\text{Re}_2\text{Cl}_8^{2-}$ geometry is, however, clearly outside the scope of the present paper.

Conclusions

This study reports a systematic comparison of the effect of various methods of inclusion of relativistic effects on the nature of metal–metal bonding in the $\text{Re}_2\text{Cl}_8^{2-}$ ion as reflected by the analysis of domain averaged Fermi holes. The comparison involved two widely used methods, namely, the approach based on the use of relativistic ECP basis sets and the direct Douglas–Kroll–Hess Hamiltonian using all-electron basis sets. It has been shown that the picture of the bonding emerging from both methods is practically insensitive to the actual method used provided the analysis is performed using the exact AIM-generalized form of the DAFH analysis and closely agrees with the results of earlier theoretical studies,^{32,34} according to which the partial cancellation of the bonding contribution of the δ bond due to fractional population of the antibonding δ^* orbital effectively reduces the multiplicity of the Re–Re bond, with the value close to 3. On the other hand, in the case of the approximate Mulliken-like approach of the DAFH analysis, the reliable and internally consistent results are observed only when relativistic effects are included via small ECP basis sets, while in the case of the DKH2 description with relativistic all-electron VTZ basis sets, the use of the approximate Mulliken-like approach seems questionable.

Acknowledgment. R.P. is thankful for the support of this study from the Grant Agency of the Czech Republic,

grant no. 203/118/09. The support from grants APVV (contract no. APVV-0093-07) and VEGA (contract nos. 1/0817/08 and 1/0127/09) is also gratefully acknowledged by L.B. Partial support from the Danish National Research Foundation through the Center for Materials Crystallography (CMC) is gratefully acknowledged by C.G.

References

- (1) Pitzer, K. S. *Acc. Chem. Res.* **1979**, *12*, 271–276.
- (2) Pyykkö, P. *Acc. Chem. res.* **1979**, *12*, 276–281.
- (3) Christiansen, P. A.; Ermler, W. C.; Pitzer, K. S. *Annu. Rev. Phys. Chem.* **1985**, *36*, 407–432.
- (4) Pyykkö, P. *Chem. Rev.* **1988**, *88*, 563–594.
- (5) Greenwood, N. N.; Earnshaw, A. *Chemistry of Elements*; Pergamon Press: Oxford, U. K., 1984; Chapter 30.
- (6) Seth, M.; Dolg, M.; Fulde, P.; Schwerdtfeger, P. *J. Am. Chem. Soc.* **1995**, *117*, 6597–6598.
- (7) Sidgwick, N. V. *The Covalent Link in Chemistry*; Cornell University Press: Ithaca, NY, 1933.
- (8) Wesendrup, R.; Laerdahl, J. K. *J. Chem. Phys.* **1999**, *110*, 9457–9462.
- (9) Raptis, R. G.; Fackler, J. P., Jr.; Murray, H. H.; Porter, L. C. *Inorg. Chem.* **1989**, *28*, 4059–4061.
- (10) Schwerdtfeger, P.; Dolg, M.; Schwarz, W. H. E.; Bowmaker, G. A.; Boyd, P. D. W. *J. Chem. Phys.* **1989**, *91*, 1762–1774.
- (11) Dirac, P. A. M.: *The Principles of Quantum Mechanics*, 4th ed.; Clarendon Press: Oxford, U. K., 1958; Chapter XI.
- (12) Visscher, L.; Jensen, H. J. A.; Saue, T.; Bast, R.; Dubillard, S.; Dyal, K. G.; Ekström, U.; Eliav, E.; Fleig, T.; Gomes, A. S. P.; Helgaker, T. U.; Henriksson, J.; Iliáš, M.; Jacob, C. R.; Knecht, S.; Norman, P.; Olsen, J.; Pernpointner, M.; Ruud, K.; Salek, P.; Sikkema, J. *DIRAC08*; Syddansk Universitet: Odense, Denmark, 2008. <http://dirac.chem.sdu.dk> (accessed Aug 20, 2010).
- (13) Douglas, M.; Kroll, N. M. *Ann. Phys.* **1974**, *82*, 89–155.
- (14) Hess, B. A. *Phys. Rev A* **1985**, *32*, 756–763.
- (15) Wolf, A.; Reiher, M.; Hess, B. A. *J. Chem. Phys.* **2002**, *117*, 9215–9226.
- (16) Reiher, M.; Wolf, A. *Relativistic Quantum Chemistry*; Wiley-VCH, Weinheim, Germany, 2009.
- (17) Dunning, T. H., Jr.; Hay, P. J. In *Modern Theoretical Chemistry*; Schaeffer, H. F., III, Ed.; Plenum Press: New York, 1976; Vol. 3, p 1.
- (18) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299–310.
- (19) Stevens, W.; Basch, H.; Kraus, J. *J. Chem. Phys.* **1984**, *81*, 6026–6033.
- (20) Cundari, T. R.; Stevens, W. *J. Chem. Phys.* **1993**, *98*, 5555–5565.
- (21) Onoe, J.; Nakamatsu, H.; Mukayama, T.; Sekine, R.; Adachi, H.; Takeuchi, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 2459–2462.
- (22) *J. Comput. Chem.* **2002**, vol. 23 (special issue).
- (23) Antschbach, J.; Siekierski, S.; Seth, M.; Schwerdtfeger, P. *J. Comput. Chem.* **2002**, *23*, 804–812.
- (24) Schwartz, H. *Angew. Chem., Int. Ed.* **2003**, *42*, 4442–4454.
- (25) Eickerling, G.; Mastalerz, R.; Herz, V.; Scherer, W.; Himmel, H.-J.; Reiher, M. *J. Chem. Theory Comput.* **2007**, *3*, 2182–2197.

- (26) Garin, J.; Toste, F. D. *Nature* **2007**, *446*, 395–403.
- (27) Iliáš, M.; Kellö, V.; Urban, M. *Acta Phys. Slov.* **2010**, *66*, 259–391.
- (28) Moncho, S.; Autsbach, J. *J. Chem. Theory Comput.* **2010**, *6*, 223–234.
- (29) Lein, M.; Rudolph, M.; Hashimi, S. K.; Schwedrtfeger, P. *Organometallics* **2010**, *29*, 2206–2210.
- (30) Odoh, S. O.; Schreckenbach, G. *J. Phys. Chem. A* **2010**, *114*, 1957–1963.
- (31) Cotton, F. A. *J. Mol. Struct.* **1980**, *59*, 97–108.
- (32) Mortola, A.; Moskowitz, J. W.; Rosch, N.; Cowman, C. D.; Gray, H. B. *Chem. Phys. Lett.* **1975**, *32*, 283–286.
- (33) Hay, P. J. *J. Am. Chem. Soc.* **1982**, *104*, 7007–7017.
- (34) Smith, D. C.; Goddard, W. A., III *J. Am. Chem. Soc.* **1987**, *109*, 5580–5583.
- (35) Blaudeau, J. P.; Roos, R. B.; Pitzer, R. M.; Mougenot, P.; Benard, M. *J. Phys. Chem.* **1994**, *98*, 7123–7127.
- (36) Wang, X.-B.; Wang, L.-S. *J. Am. Chem. Soc.* **2000**, *122*, 2096–2100.
- (37) Gagliardi, L.; Roos, B. O. *Inorg. Chem.* **2003**, *42*, 1599–1603.
- (38) Henandez-Avecedo, L.; Arratia-Perez, R. *J. Chil. Chem. Soc.* **2004**, *49*, 361–365.
- (39) Saito, K.; Nakao, Y.; Sato, H.; Sakaki, S. *J. Phys. Chem. A* **2006**, *110*, 9710–9717.
- (40) Ponec, R.; Yuzhakov, G. *Theor. Chem. Acc.* **2007**, *118*, 791–797.
- (41) Cavigliasso, G.; Kaltsoyannis, N. *Inorg. Chem.* **2007**, *46*, 3557–3565.
- (42) Krapp, A.; Lein, M.; Frenking, G. *Theor. Chem. Acc.* **2008**, *120*, 313–320.
- (43) Poineau, F.; Gagliardi, L.; Forster, P. M.; Sattelberger, A. P.; Czerwinski, K. R. *Dalton Trans.* **2009**, 5954–5959.
- (44) Ponec, R. *J. Math. Chem.* **1997**, *21*, 323–333.
- (45) Ponec, R. *J. Math. Chem.* **1998**, *23*, 85–103.
- (46) Ponec, R.; Duben, A. *J. Comput. Chem.* **1999**, *8*, 760–771.
- (47) Ponec, R.; Yuzhakov, G.; Carbo-Dorca, R. *J. Comput. Chem.* **2003**, *24*, 1829–1838.
- (48) Ponec, R.; Yuzhakov, G.; Girones, X.; Frenking, G. *Organometallics* **2004**, *23*, 1790–1796.
- (49) Ponec, R.; Yuzhakov, G.; Sundberg, M. *J. Comput. Chem.* **2005**, *26*, 447–454.
- (50) Ponec, R.; Feixas, F. *J. Phys. Chem. A* **2009**, *113*, 8394–8340.
- (51) McWeeny, R. *Rev. Mod. Phys.* **1960**, *32*, 335–369.
- (52) Bader, R. F. W. *Atoms in Molecules. A Quantum Theory*; Clarendon Press: Oxford, U. K., 1994.
- (53) Cioslowski, J. *Int. J. Quantum Chem.* **1990**, *S24*, 15–28.
- (54) Ponec, R.; Cooper, D. L. *Faraday Discuss.* **2007**, *135*, 31–42.
- (55) Ponec, R.; Cooper, D. L. *J. Phys. Chem. A* **2007**, *111*, 11294–11301.
- (56) Hohenberg, P.; Kohn, W. *Phys. Rev. B* **1964**, *136*, 864–871.
- (57) Kohn, W.; Sham, J. *Phys. Rev. A* **1985**, *140*, 1133–1138.
- (58) Schultz, N. E.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem A* **2005**, *109*, 4388–4403.
- (59) Vyboishchikov, S. F.; Sierralta, A.; Frenking, G. *J. Comput. Chem.* **1996**, *18*, 416–429.
- (60) Bo, C.; Costas, M.; Poblet, J. M. *J. Phys. Chem.* **1995**, *99*, 5914–5921.
- (61) Lin, Z.; Bytheway, I. *Inorg. Chem.* **1996**, *35*, 594–603.
- (62) Ponec, R.; Gatti, C. *Inorg. Chem.* **2009**, *48*, 11204–11031.
- (63) Dylla, K. G. *Theor. Chem. Acc.* **2004**, *112*, 403–409; available at <http://dirac.chem.sdu.dk> (accessed Aug 20, 2010).
- (64) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358–1371.
- (65) de Jong, W. A.; Harrison, R. J.; Dixon, D. A. *J. Chem. Phys.* **2001**, *114*, 48–53.
- (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.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.
- (67) PROAIM, version 94, rev. B: Biegler, K. F. W.; Bader, R. F. W.; Tang, T. *J. Comput. Chem.* **1982**, *13*, 317–328.
- (68) AIMAll; Keith, T. A.; TK Gristmill Software: 1997. aim.tkgristmill.com (accessed Sep 2010).
- (69) Mayer, I. *Chem. Phys. Lett.* **1983**, *97*, 270–274.
- (70) Wiberg, K. *Tetrahedron* **1968**, *24*, 1083–1096.
- (71) Bader, R. F. W.; Stephens, M. E. *J. Am. Chem. Soc.* **1975**, *97*, 7391–7397.
- (72) Fradera, X.; Austin, M.; Bader, R. F. W. *J. Phys. Chem. A* **1999**, *103*, 304–314.
- (73) Jensen, F. *Introduction to Computational Chemistry*; John Wiley: Chichester, U. K., 2001; p 218.
- (74) Cotton, F. A.; Walton, R. A. *Multiple bonds between metal atoms*, 2nd ed.; Clarendon Press: Oxford, U. K., 1993.
- (75) Kuznetsov, V. G.; Kuzmin, P. A. *Zh. Struct. Khim.* **1963**, *4*, 55–62.
- (76) Ehlers, A. W.; Bohme, M.; Dapprich, S.; Gobbi, A.; Hollwarth, A.; Jonas, V.; Kohler, K. F.; Stegmann, R.; Veldkamp, A.; Frenking, G. *Chem. Phys. Lett.* **1993**, *208*, 111–114.

Anharmonic Vibrational Analysis for the Propadienyliene Molecule ($\text{H}_2\text{C}=\text{C}=\text{C}:$)

Qunyan Wu,^{†,‡} Qiang Hao,^{‡,§} Jeremiah J. Wilke,[‡] Andrew C. Simmonett,[‡]
Yukio Yamaguchi,[‡] Qianshu Li,^{†,||} De-Cai Fang,[§] and Henry F. Schaefer III^{*,‡}

*Institute of Chemical Physics, Beijing Institute of Technology, Beijing, P. R. China 100081,
Center for Computational Quantum Chemistry, University of Georgia, Athens, Georgia 30602,
College of Chemistry, Beijing Normal University, Beijing, P. R. China 100875, and
Center for Computational Quantum Chemistry, South China Normal University,
Guangzhou, P. R. China 510631*

Received June 22, 2010

Abstract: Maier et al. found that photolysis of singlet cyclopropenylidene (**1S**) in a matrix yields triplet propargylene (**2T**), which upon further irradiation is converted to singlet propadienyliene (vinylidenecarbene, **3S**). Their discovery was followed by interstellar identification of **3S** by Cernicharo et al. An accurate quartic force field for propadienyliene (**3S**) has been determined employing the *ab initio* coupled-cluster (CC) with single and double excitations and perturbative triple excitations [CCSD(T)] method and the correlation-consistent core–valence quadruple- ζ (cc-pCVQZ) basis set. Utilizing vibrational second-order perturbation theory (VPT2), vibration–rotation coupling constants, rotational constants, centrifugal distortion constants, vibrational anharmonic constants, and fundamental vibrational frequencies are determined. The predicted fundamental frequencies for **3S** as well as its ¹³C and deuterium isotopologues are in good agreement with experimental values. The theoretical zero-point vibration corrected rotational constants **B**₀ are consistent with experimental values within 0.3% of errors. The isotopic shifts of **B**₀ are in close to exact agreement with experimental observations. The mean absolute deviation between theoretical anharmonic and experimental fundamental vibrational frequencies for 24 modes (excluding CH₂ s-str.) is only 2.6 cm^{−1}. The isotopic shifts of the vibrational frequencies are also in excellent agreement with the available experimental values. However, a large discrepancy is observed for the CH₂ symmetric stretch, casting doubt on the experimental assignment for this mode.

1. Introduction

The singlet state of cyclopropenylidene (**1S**) is known to be the global minimum on the C₃H₂ potential energy surfaces (PESs). The first laboratory detection of **1S** was achieved by Reisenauer et al. in 1984.¹ This was enabled by the *ab initio* prediction of its vibrational frequencies and infrared (IR) intensities provided by Lee et al.² Shortly thereafter,

Reisenauer et al.³ were able to show that, upon irradiation, **1S** is photoisomerized into triplet propynylidene (propargylene, **2T**) and, in a second photostep, into singlet propadienyliene (vinylidenecarbene, **3S**). The second photolysis can be reversed by using shorter wavelength light (254 nm) to regenerate **2T**. By a repeated photolysis of the newly formed **2T** with 313 nm light, complete reversibility of the isomerization to **1S** was demonstrated, as shown in Scheme 1. The same photochemical cycle was observed when the dideuteriocyclopropenylidene (**1S-D**₂) was treated in the same way.^{3,4} **1S** has also been detected in interstellar space, and it appears to be the most abundant of all hydrocarbons and plays an important role in the chemistry of the interstellar

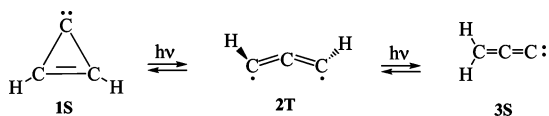
* Corresponding author e-mail: sch@uga.edu.

† Beijing Institute of Technology.

‡ University of Georgia.

§ Beijing Normal University.

|| South China Normal University.

Scheme 1. Interconversion of the Three Lowest-Lying C₃H₂ Species

medium.⁵ Dateo and Lee⁶ reported an *ab initio* quartic force field for cyclopropenylidene (**1S**) at the cc-pVTZ CCSD(T) level of theory. They used second-order vibrational perturbation theory (VPT2) to predict fundamental vibrational frequencies of **1S** and its ¹³C and deuterium isotopologues. Their theoretical anharmonic quantities were in good agreement with the available experimental observations. Very recently, Lee et al.⁷ also published newer fundamental vibrational frequencies and rovibrational spectroscopic constants for isotopologues of cyclopropenylidene (**1S**) with the cc-pVQZ CCSD(T) method.

Propadienylidene (**3S**) is the first member of the cumulene carbene series to exhibit great stability and has a singlet ground state. It is characterized by carbon–carbon double bonds, terminal nonbonded electrons, and a large dipole moment. Propadienylidene (**3S**) was first detected in the photolysis products of cyclopropenylidene (**1S**) as mentioned above.³ In 1990, **3S** was produced in a laboratory discharge, and its rotational spectrum was determined to high precision by Vrtilek et al.⁸ In the following year (1991), Cernicharo et al. reported the astronomical detection of **3S** in Tmc-1 (one of the best astronomical sources of carbon chains) and tentative detection in the molecular envelope of IRC+10216.⁹ Subsequently, Gottlieb et al.¹⁰ reported the millimeter-wave rotational spectra of four isotopic species of the propadienylidene (H₂¹³CCC, H₂C¹³CC, H₂CC¹³C, and D₂CCC) and the same set of rotational and centrifugal distortion constants. From the observed rotational constants, the *r_s* structure was determined: *r_s*(C₁C₂) = 1.326 ± 0.003 Å, *r_s*(C₂C₃) = 1.287 ± 0.003 Å, *r_s*(CH) = 1.084 ± 0.004 Å, and θ_s(HCH) = 117.7 ± 0.02°. Furthermore, vibration–rotation coupling constants calculated in the CEPA-1 approximation were combined with the experimental rotational constants for the five isotopologues to yield an equilibrium geometry:¹⁰ *r_e*(C₁C₂) = 1.3283 ± 0.0005 Å, *r_e*(C₂C₃) = 1.291 ± 0.001 Å, *r_e*(CH) = 1.083 ± 0.001 Å, and θ_e(HCH) = 117.6 ± 0.2°.

Seburg et al.¹¹ investigated photochemical automerizations and isomerizations of C₃H₂ isomers. Photolysis of [¹³C] diazopropynes under matrix isolation conditions produced C₃H₂ isomers containing a single ¹³C label. Monitoring the distribution of the ¹³C label during photolysis at either λ = 313 nm or λ > 444 nm revealed the involvement of two photochemical automerization processes. At λ = 313 ± 10 nm, triplet propynylidene (**2T**) and singlet cyclopropenylidene (**1S**) photoequilibrate. The interconversion does not occur by a simple ring-closure/ring-opening mechanism, as hydrogen migration accompanies the interconversion. At λ > 444 nm, H₂C=C=¹³C and H₂C=¹³C=C rapidly equilibrate. Various lines of evidence suggested that the equilibration occurs through a cyclopropyne transition state. Seburg et al.'s *ab initio* theoretical study¹¹ confirmed that the planar isomer of singlet cyclopropyne is the transition state for the

interconversion of two ¹³C isotopologues. Stanton et al.¹² reported the first electronic absorption spectrum of **3S** in an argon matrix by the striking photochemical automerization of ¹³C-labeled **3S**. The electronic spectrum of **3S** exhibits rich vibronic structure with absorption maxima that span virtually the visible spectrum. Later, Hodges et al.¹³ reinvestigated the electronic spectrum and recorded the vibrationally resolved spectrum of **3S** in a neon matrix at 6 K. Three electronic transitions were observed: a strong $\tilde{\text{C}}-\tilde{\text{X}}$ band system in the 39 051–47 156 cm⁻¹ range, a weaker $\tilde{\text{B}}-\tilde{\text{X}}$ transition in the 16 161–24 802 cm⁻¹ region, and the hardly detectable (forbidden) $\tilde{\text{A}}-\tilde{\text{X}}$ transition at 13 885–16 389 cm⁻¹. On the basis of these observations, one can search for these absorptions in the gas phase. Recently, the rotationally resolved vibronic bands in the forbidden $\tilde{\text{A}}^1\text{A}_2-\tilde{\text{X}}^1\text{A}_1$ electronic transition of **3S** have been observed in the gas phase by cavity ring down (CRD) absorption spectroscopy, through a supersonic planar plasma with allene as a precursor, by Achkasova et al.¹⁴

There have been several theoretical studies^{8–14} on the singlet state of propadienylidene (**3S**). Quite recently, Klopner's group¹⁵ determined the atomization energies of 19 C₃H_x (x = 0–4) molecules and radicals using explicitly correlated coupled-cluster theory. For the singlet propadienylidene (**3S**), they determined harmonic and anharmonic zero-point vibrational energy (ZPVE) at the cc-pCVTZ(C)/cc-pCVDZ(H) CCSD(T) level of theory. On the other hand, Vázquez et al.¹⁶ reported high-accuracy extrapolated *ab initio* thermochemistry (HEAT) of the propargyl radical and the singlet C₃H₂ carbenes. Their HEAT scheme included harmonic and anharmonic contributions to ZPVE correction at the cc-pVQZ CCSD(T) level of theory. However, neither of these two excellent papers presented detailed analyses of rovibrational anharmonic quantities.

In the present study, anharmonic vibrational analyses of the electronic ground state of propadienylidene (**3S**) are performed employing vibrational second-order perturbation (VPT2) theory.^{17–23} The molecular parameters are determined using the *ab initio* coupled cluster with single, double, and perturbative triple excitations [CCSD(T)] method^{24–26} with the correlation-consistent polarized core–valence quadruple-ζ (cc-pCVQZ) basis set.^{27,28} The theoretically determined harmonic and anharmonic rovibrational quantities will be compared with the available experimental measurements. The present research should stimulate further characterization of the propadienylidene molecule (**3S**), which is involved in hydrocarbon chemistry, combustion chemistry, chemical dynamics, interstellar chemistry, and high-resolution spectroscopy.

2. Electronic Structure Considerations

The ¹A₁ state of the **3S** isomer arises from the following electronic configuration, when oriented in the *yz* plane, with the C₂ axis aligned with the *z* axis:

$$[\text{core}]4a_1^25a_1^26a_1^21b_2^21b_1^27a_1^22b_2^2 \tilde{\text{X}}^1\text{A}_1 \quad (1)$$

where [core] denotes the three lowest-lying core (C: 1s-like) orbitals. In eq 1, the 2b₂ molecular orbital (MO) describes the C₂–C₃ in-plane π bond, while the 1b₁ MO is related to

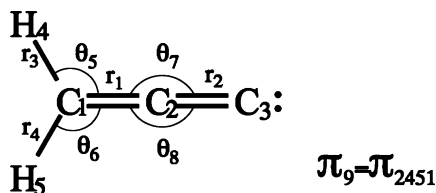


Figure 1. The internal coordinates of the propadienylidene molecule ($\text{H}_2\text{C}=\text{C}=\text{C}:$).

the out-of-plane $\text{C}_1\text{C}_2\text{C}_3$ π bonding. The $7a_1$ MO is associated with the lone-pair orbital localized on the carbene (C_3) atom.

3. Symmetry Internal Coordinates

The vibrational potential energy (\bar{V}) for propadienylidene (**3S**) may be expanded in terms of displacement symmetry internal coordinates (ΔS_i) in the vicinity of the equilibrium point (E_0) as

$$\bar{V} = E_0 + \frac{1}{2} \sum_{ij} F_{ij} \Delta S_i \Delta S_j + \frac{1}{6} \sum_{ijk} F_{ijk} \Delta S_i \Delta S_j \Delta S_k + \frac{1}{24} \sum_{ijkl} F_{ijkl} \Delta S_i \Delta S_j \Delta S_k \Delta S_l \quad (2)$$

In eq 2, F_{ij} , F_{ijk} , and F_{ijkl} denote quadratic, cubic, and quartic force constants. The nine symmetry internal coordinates for C_{2v} species (e.g., C_3H_2 , C_3D_2) are defined by

$$\begin{aligned} S_1(a_1) &= r_1 && \text{C}_1\text{C}_2 \text{ stretch} \\ S_2(a_1) &= r_2 && \text{C}_2\text{C}_3 \text{ stretch} \\ S_3(a_1) &= \frac{1}{\sqrt{2}}(r_3 + r_4) && \text{CH}_2 \text{ s-stretch} \\ S_4(a_1) &= \frac{1}{\sqrt{2}}(\theta_5 + \theta_6) && \text{CH}_2 \text{ scissor} \\ S_5(b_1) &= \theta_8 && \text{CCC bend (oop)} \\ S_6(b_1) &= \pi_9 && \text{CH}_2 \text{ wag} \\ S_7(b_1) &= \frac{1}{\sqrt{2}}(r_3 - r_4) && \text{CH}_2 \text{ a-stretch} \\ S_8(b_2) &= \frac{1}{\sqrt{2}}(\theta_5 - \theta_6) && \text{CH}_2 \text{ rock} \\ S_9(b_2) &= \theta_7 && \text{CCC bend (ip)} \end{aligned} \quad (3)$$

where the internal coordinates (r_1 – r_4 , θ_5 – θ_8 , and π_9) are depicted in Figure 1. The CCC bending coordinates θ_7 and θ_8 are described by linear bending coordinates of the form $\sin^{-1}[\mathbf{e}_y(\mathbf{e}_{23} \times \mathbf{e}_{21})]$, where \mathbf{e}_y is a fixed unit vector aligned with either the x or y axis and \mathbf{e}_{23} is a unit vector directed from atom C_3 to atom C_2 . The out-of-plane coordinate S_6 can be more explicitly written as $\sin^{-1}[\mathbf{e}_{12}(\mathbf{e}_{14} \times \mathbf{e}_{15})/\sin \theta_{415}]$, where the \mathbf{e} vectors have the previously stated meaning, and θ_{415} is the valence bond angle between atoms H_4 , C_1 , and H_5 . The displacement sizes used for the finite difference procedure (in Å and rad units) are 0.01, 0.01, 0.01, 0.02, 0.03, 0.03, 0.01, 0.02, and 0.03 for coordinates S_1 – S_9 , which yielded a fit with just a 1.9×10^{-9} E_h error.

4. Theoretical Procedures

In the present research, the correlation-consistent polarized core–valence quadruple- ζ (cc-pCVQZ) basis set developed by Dunning and Woon^{27,28} was employed to optimize the

geometry and to determine analytical potentials. The zeroth-order description of the ground state of propadienylidene (**3S**) was obtained using single configuration self-consistent-field (SCF) [restricted Hartree–Fock (RHF)] wave functions. The coupled cluster with single, double, and perturbative triple excitations [CCSD(T)] wave functions^{24–26} were constructed without freezing any core orbitals. In our previous study, it was found that the ground state of **3S**, in the vicinity of the equilibrium point, may be described adequately by the CCSD(T) method, which is based on a single determinant RHF wave function.²⁹

The structure of propadienylidene (**3S**) was optimized using analytic derivative methods.^{30–32} Its dipole moment, harmonic vibrational frequencies, and corresponding IR intensities were determined analytically. Electronic structure computations were carried out using the ACESII (Mainz–Austin–Budapest version),^{33,34} MOLPRO,³⁵ and PSI2³⁶ suites of quantum chemistry software.

The C++ program GRENDL³⁷ was used to generate perturbed geometries and to compute force constants in symmetry internal coordinates. The INTDER 2005^{38,39} code of Allen was employed to perform nonlinear coordinate transformations of quadratic, cubic, and quartic force constants between symmetry internal and Cartesian coordinates. The VPT2 analyses were then performed upon the Cartesian force constants by the ANHARM program.⁴⁰

5. Results and Discussion

5.1. Equilibrium Geometry and Dipole Moment. At the cc-pCVQZ CCSD(T) level of theory, the propadienylidene (**3S**) molecule is predicted to lie 14.5 (13.6) kcal mol^{−1} (the zero-point vibrational energy corrected value in parentheses) above the global minimum on the C_3H_2 PES, cyclopropenylidene (**1S**). At the same level of theory, the structure of propadienylidene (**3S**) has been determined to be $r_e(\text{C}_1\text{C}_2) = 1.3281$ Å, $r_e(\text{C}_2\text{C}_3) = 1.2879$ Å, $r_e(\text{CH}) = 1.0837$ Å, and $\theta_e(\text{HCH}) = 117.45^\circ$. This theoretical r_e structure is in good agreement with the recommended r_e structure (theory + experiment) of Gottlieb et al.:¹⁰ $r_e(\text{C}_1\text{C}_2) = 1.3283 \pm 0.0005$ Å, $r_e(\text{C}_2\text{C}_3) = 1.291 \pm 0.001$ Å, $r_e(\text{CH}) = 1.083 \pm 0.001$ Å, and $\theta_e(\text{HCH}) = 117.6 \pm 0.2^\circ$. Our zero-point corrected structure, which includes the effects of anharmonicity through the cubic terms in the force field is $r_z(\text{C}_1\text{C}_2) = 1.3289$ Å, $r_z(\text{C}_2\text{C}_3) = 1.2870$ Å, $r_z(\text{CH}) = 1.0907$ Å, and $\theta_z(\text{HCH}) = 117.55^\circ$, which compares very favorably with the r_s structure of Gottlieb and co-workers. The dipole moment of isomer **3S** is predicted to be 4.174 debye along the C_2 axis with sign $^+\text{H}_2\text{CCC}^-$. Our value is consistent with the dipole moment of 4.135 debye at the CCSD(T) level of theory with the 131 cGTO basis set of Gottlieb et al.¹⁰ The magnitude of the dipole moment for propadienylidene (**3S**) is larger than that of cyclopropenylidene (**1S**) [3.420 debye at the cc-pCVQZ CCSD(T) level of theory, 3.43(2) debye exptl.⁴¹], probably due to a longer distance (charge separation) between the positive (CH_2) and negative ($\text{C}:$) ends of the molecule.

5.2. Vibration–Rotation Coupling Constants. The vibrational dependence of a rotational constant B_v , v being a vibrational quantum number, has the general form

$$B_v = B_e - \sum_r \alpha_r^B \left(\nu + \frac{1}{2} \right) + \text{higher terms} \quad (4)$$

where B_e is the equilibrium rotational constant and the sums run over all normal modes. Similar expressions hold for the vibrational dependence of A_v and C_v .

The vibration–rotation coupling constants α_r^B for an asymmetric top from perturbation theory are given by

$$-\alpha_r^B = \frac{2B_e^2}{\omega_r} \left[\sum_{\xi} \frac{3(a_r^{(b\xi)})^2}{4I_{\xi}} + \sum_s (\zeta_{rs}^{(b)})^2 \frac{(3\omega_r^2 + \omega_s^2)}{\omega_r^2 - \omega_s^2} + \pi \left(\frac{c}{h} \right)^{1/2} \sum_s \phi_{rrs} a_s^{(bb)} \left(\frac{\omega_r}{\omega_s^{3/2}} \right) \right] \quad (5)$$

where ω_r is the r th harmonic vibrational frequency, I_{ξ} is the ξ th principal moment of inertia, $\zeta_{rs}^{(b)}$ is the Coriolis coupling constant about the b axis, and ϕ_{rrs} is the cubic force constant. In this equation, the $a_r^{(\alpha\beta)}$ constants are derivatives of the equilibrium moments and products of inertia with respect to the r th normal coordinate Q_r

$$a_r^{(\alpha\beta)} = \left(\frac{\partial I_{\alpha\beta}}{\partial Q_r} \right)_e \quad (6)$$

The vibration–rotation coupling constants (α_r^A , α_r^B , and α_r^C) for six isotopologues of propadienyldiene are presented in Table 1. The second term in the square brackets of eq 5 sometimes suffers from Coriolis resonances when two vibrational frequencies are very close, i.e., $\omega_r \approx \omega_s$. About the a axis of propadienyldiene, strong Coriolis interactions are expected between the CH₂ wagging (ν_5) and CH₂ rocking (ν_8) and between the CCC out-of-plane (ν_6) and in-plane (ν_9) bending vibrations.¹⁰ The corresponding Coriolis coupling constants are determined to be $\zeta_{5,8}^{(a)} = 0.527$ and $\zeta_{6,9}^{(a)} = 0.937$. Anomalous α_r^A constants are indeed observed for these four vibrational modes. In Table 1, therefore, deperturbed values^{20,21} are reported. However, the B_0 constant ($\nu = 0$) may be determined without Coriolis resonances by taking the sums of α_r^B constants over all normal modes rather than their individual values, as pointed out by East et al.⁴²

5.3. Rotational Constants and Centrifugal Distortion Constants. Since propadienyldiene (**3S**) is a near symmetric top molecule (a highly prolate asymmetric top, $\kappa = -0.9972$),¹⁰ Gottlieb's experiment¹⁰ analyzed the transition frequencies using Watson's S-reduced Hamiltonian.¹⁹ Consequently, the centrifugal distortion constants and the zero-point vibration corrected rotational constants are determined using Watson's S-reduction in the I^f representation in our theoretical study. In Table 2, rotational constants and centrifugal distortion constants for the standard H₂CCC species and its four isotopologues are provided. In the following discussion, the equilibrium (A_e , B_e , and C_e) and zero-point vibration corrected (A_0 , B_0 , and C_0) rotational constants are abbreviated as \mathbf{B}_e and \mathbf{B}_0 , respectively. For the standard H₂CCC species, the differences between theoretical equilibrium and zero-point vibration corrected rotational constants are $\Delta[\mathbf{B}_e(\text{theor.}) - \mathbf{B}_0(\text{exptl.})] = (+3484, -15, +1)$ MHz and $\Delta[\mathbf{B}_0(\text{theor.}) - \mathbf{B}_0(\text{exptl.})] = (+999, -8, -7)$ MHz, respectively. Improvement in agreement between theoretical

Table 1. Theoretical Predictions of Vibration–Rotation Coupling Constants (in MHz) for C₃H₂ and Its D and ¹³C Labeled Isotopologues at the cc-pCVQZ CCSD(T) Level of Theory

	C ₃ H ₂	C ₃ D ₂	C ₃ HD	H ₂ ¹³ CCC	H ₂ CC ¹³ C	H ₂ C ¹³ CC
α_1^A	4785.8	2028.0	1486.4	4711.2	4785.8	4784.7
α_2^A	177.9	76.4	2998.9	194.9	178.8	154.4
α_3^A	-3002.5	-707.0	13.5	-3008.5	-3008.3	-2988.2
α_4^A	134.1	-470.5	-1704.1	162.5	137.4	135.4
α_5^A	3736.3	1210.6	-1378.8	3754.8	3736.8	3747.4
α_6^A	4091.9	1710.6	-3708.1	4058.1	4145.5	3987.9
α_7^A	2867.3	1018.6	-2126.0	2867.7	2867.3	2867.3
α_8^A	-5492.9	-2022.3	3802.1	-5509.4	-5490.7	-5452.2
α_9^A	-2328.8	-1171.0	2655.9	-2281.1	-2367.6	-2223.3
$1/2\sum\alpha_r^A$	2484.5	836.7	1019.9	2475.1	2492.5	2506.8
α_1^B	7.1	14.1	11.1	6.2	6.7	7.1
α_2^B	78.5	61.9	13.6	76.1	75.6	75.6
α_3^B	-18.8	11.1	70.4	-18.6	-17.7	-18.7
α_4^B	27.8	-15.0	-13.1	27.8	26.5	27.8
α_5^B	25.3	24.1	17.8	24.6	23.5	25.3
α_6^B	-64.3	-55.1	-8.1	-62.9	-62.5	-60.6
α_7^B	10.0	16.4	-66.2	8.6	9.4	10.0
α_8^B	-8.5	-9.9	20.3	-8.0	-7.7	-8.6
α_9^B	-70.5	-62.7	-58.3	-67.7	-67.2	-68.8
$1/2\sum\alpha_r^B$	-6.7	-7.6	-6.3	-7.0	-6.7	-5.5
α_1^C	11.8	19.4	12.3	10.6	11.1	11.8
α_2^C	73.3	55.2	17.6	71.1	70.9	70.7
α_3^C	11.9	24.4	64.3	11.0	10.8	11.9
α_4^C	27.7	-0.9	13.0	27.5	26.3	27.7
α_5^C	-2.9	-8.0	18.5	-2.2	-2.5	-2.8
α_6^C	-94.5	-75.7	11.2	-91.5	-90.7	-92.0
α_7^C	11.2	16.6	-28.9	9.8	10.5	11.2
α_8^C	10.3	14.4	-5.1	9.8	9.8	10.3
α_9^C	-32.6	-25.8	-83.5	-32.0	-31.8	-30.6
$1/2\sum\alpha_r^C$	8.2	9.9	9.7	7.0	7.2	9.1

and experimental rotational constants due to zero-point vibration correction is significant; 1.2% deviation for the former $\Delta[\mathbf{B}_e(\text{theor.}) - \mathbf{B}_0(\text{exptl.})]$ and 0.3% for the latter $\Delta[\mathbf{B}_0(\text{theor.}) - \mathbf{B}_0(\text{exptl.})]$. For the dideutero isotopologue (D₂CCC), the corresponding differences are $\Delta[\mathbf{B}_e(\text{theor.}) - \mathbf{B}_0(\text{exptl.})] = (+1091, -15, +3)$ MHz and $\Delta[\mathbf{B}_0(\text{theor.}) - \mathbf{B}_0(\text{exptl.})] = (+253, -8, -6)$ MHz, respectively. Again, the improvement in agreement between theoretical and experimental rotational constants is evident: 0.8% for the former $\Delta[\mathbf{B}_e(\text{theor.}) - \mathbf{B}_0(\text{exptl.})]$ and 0.2% for the latter $\Delta[\mathbf{B}_0(\text{theor.}) - \mathbf{B}_0(\text{exptl.})]$.

The quartic centrifugal distortion constants of the standard H₂CCC species at the cc-pCVQZ CCSD(T) level of theory are predicted to be $D_J = 3.748$ (4.248) kHz, $D_{JK} = 0.5822$ (0.5164) MHz, $D_K = 22.026$ (23.535) MHz, $d_1 = -0.146$ (-0.153) kHz, and $d_2 = -0.054$ (-0.070) kHz, where the experimental values are shown in parentheses. Our predictions are in reasonable agreement with the theoretical values at the CEPA-1/131 cGTO level of theory and experimental observations.¹⁰ For the other four isotopologues, the theoretically determined centrifugal distortion constants are also consistent with the corresponding experimental values. In this light, it should be noted that during their least-squares fit procedures Gottlieb et al. constrained D_k to the values in H₂CCO and D₂CCO, whereas in each of the ¹³C species, they constrained it to the value in normal H₂CCO.¹⁰

In Table 3, isotopic shifts for rotational constants and centrifugal distortion constants for the four isotopologues with respect to the standard H₂CCC species are presented. Both the theoretical equilibrium rotational constants \mathbf{B}_e (A_e ,

Table 2. Theoretical and Experimental Rotational Constants, and Quartic Centrifugal Distortion Constants (in MHz) for H₂CCC and Its Four Isotopologues at the cc-pCVQZ CCSD(T) Level of Theory

	H ₂ CCC		D ₂ CCC		H ₂ ¹³ CCC		H ₂ C ¹³ CC		H ₂ CC ¹³ C	
	theory	exptl. ^a	theory	exptl. ^a	theory	exptl. ^a	theory	exptl. ^a	theory	exptl. ^a
rotational constants										
A _e	292267		146246		292267		292267		292267	
B _e	10574		9388		10264		10571		10166	
C _e	10205		8821		9916		10202		9824	
A ₀	289782	288783	145408	145155	289792	288880	289760	288610	289774	288860
B ₀	10581	10589	9395	9403	10271	10279	10577	10585	10173	10180
C ₀	10197	10204	8812	8818	9909	9916	10193	10200	9817	9824
quartic centrifugal distortion constants										
10 ³ D _J	3.748	4.248	2.747	3.128	3.552	4.027	3.748	4.240	3.473	3.955
D _{JK}	0.5822	0.5164	0.3848	0.3699	0.5477	0.485	0.5790	0.518	0.5485	0.485
D _K	22.026	23.535	5.272	5.391	22.060	23.535	22.028	23.535	22.060	23.535
10 ³ d ₁	-0.146	-0.153	-0.214	-0.251	-0.133	-0.135	-0.146	-0.147	-0.130	-0.127
10 ³ d ₂	-0.054	-0.070	-0.116	-0.135	-0.048	-0.058	-0.054	-0.078	-0.047	-0.063

^a Ref 10.**Table 3.** Theoretical and Experimental Isotopic Shifts for Rotational Constants and Quartic Centrifugal Distortion Constants (in MHz) of the Four Isotopologues with Respect to Those of the Standard H₂CCC Species at the cc-pCVQZ CCSD(T) Level of Theory

	D ₂ CCC		H ₂ ¹³ CCC		H ₂ C ¹³ CC		H ₂ CC ¹³ C	
	theory	exptl.	theory	exptl.	theory	exptl.	theory	exptl.
rotational constants								
ΔA _e	-146021		0		0		0	
ΔB _e	-1186		-310		-3		-408	
ΔC _e	-1384		-289		-3		-381	
ΔA ₀	-144374	-143628	10	97	-22	-173	-8	77
ΔB ₀	-1186	-1186	-310	-310	-4	-4	-408	-409
ΔC ₀	-1385	-1386	-288	-288	-4	-4	-380	-380
quartic centrifugal distortion constants								
10 ³ ΔD _J	-1.001	-1.12	-0.196	-0.221	0.0	-0.008	-0.275	-0.293
ΔD _{JK}	-0.1974	-0.1465	-0.0345	-0.0314	-0.0032	0.0016	-0.0337	-0.0314
ΔD _K	-16.754	-18.144	0.034	0.0	0.002	0.0	0.034	0.0
10 ³ Δd ₁	-0.068	-0.098	0.013	0.018	0.0	0.006	0.016	0.026
10 ³ Δd ₂	-0.062	-0.065	0.006	0.012	0.0	-0.008	0.007	0.007

B_e, and C_e) and zero-point vibration corrected rotational constants B₀ (A₀, B₀, and C₀) reproduce the experimental isotopic shifts extremely well. Specifically, the isotopic shift predictions for the B₀ and C₀ constants are almost exact for all four isotopologues. The largest deviation between theory and experiment is 151 MHz for the A₀ constant of the H₂C¹³CC isotopologue. However, since experimental uncertainty in determining the A₀ constant for the H₂C¹³CC isotopologue is ±130 MHz, our theoretical isotopic shift is still considered to be in excellent agreement.

Isotopic shifts for the five centrifugal distortion constants relative to those for normal H₂CCC species are largest for D₂CCC, and the magnitudes of these shifts are reasonably consistent with the experimental observations. Among the ¹³C species, the D_J constant is sensitive to ¹³C labeling in the methylene carbon (H₂¹³CCC) and the terminus carbon (H₂CC¹³C). The ¹³C labeling in the central carbon (H₂C¹³CC) is least sensitive for all five constants.

5.4. Harmonic Vibrational Frequencies and Infrared (IR) Intensities. In Tables 4–6, harmonic vibrational frequencies and the associated infrared (IR) intensities for C₃H₂, C₃D₂, and C₃HD are presented. The corresponding quantities for the H₂¹³C=C=C, H₂C=C=¹³C, and H₂C=¹³C=C species are deposited in Tables S1–S3 as

Supporting Information. In these tables, the IR intensities, computed within the double harmonic approximation, are given as units of km mol⁻¹ in parentheses and as ratios relative to the largest intensity in square brackets for direct comparison with Maier's experimental observations.³ For the standard isotopologue of propadienylidene (**3S**), five fundamental frequencies have been observed via matrix isolation IR spectroscopy.³ Our harmonic vibrational frequencies for these five modes are predicted to be (differences from the experimental fundamental values in parentheses): ω₁(a₁) = 3126 (+77), ω₂(a₁) = 2005 (+53), ω₃(a₁) = 1492 (+45), ω₅(b₁) = 1023 (+23), and ω₈(b₂) = 1052 (+27) cm⁻¹. An averaged percent deviation relative to the harmonic vibrational frequencies for the standard H₂CCC (**3S**) isotopologue is 2.6%. For the dideutero isotopologue (D₂CCC), six fundamental vibrational frequencies have been experimentally observed. Theoretical harmonic vibrational frequencies for these six modes are determined to be ω₁(a₁) = 2288 (+87), ω₂(a₁) = 1981 (+48), ω₃(a₁) = 1229 (+20), ω₄(a₁) = 968 (+17), ω₅(b₁) = 816 (+16), and ω₈(b₂) = 846 (+17) cm⁻¹. Average percent deviation relative to the harmonic vibrational frequencies for the dideutero D₂CCC isotopologue is 2.3%.

Table 4. Theoretical Predictions of the Harmonic Vibrational Frequencies (in cm^{-1}), Anharmonic (Fundamental) Vibrational Frequencies (in cm^{-1}), and Infrared Intensities (in km mol^{-1}) [Ratio Relative to the Largest Intensity] for the \tilde{X}^1A_1 State of the Propadienylidene (Vinylidene carbene, **3S**) C_3H_2 Molecule at the cc-pCVQZ CCSD(T) Level of Theory

mode number (assignment)	harmonic	$\Delta(\text{anh.} - \text{harm.})$	anharmonic	exptl. ^a	$\Delta(\text{harm.} - \text{exptl.})$	$\Delta(\text{anh.} - \text{exptl.})$
1 (a_1) CH_2 s-str.	3126.1(4.4)[0.02]	-135.6	2990.5	3049.5[0.02]	76.6	-59.0
2 (a_1) CC a-str.	2004.8(255.5)[1.00]	-47.0	1957.8	1952.2[1.00]	52.6	5.6
3 (a_1) CH_2 sciss.	1492.0(10.8)[0.04]	-44.4	1447.6	1446.9[0.14]	45.1	0.7
4 (a_1) CC s-str.	1122.3(2.0)[0.01]	-9.6	1112.7			
5 (b_1) CH_2 wag.	1022.5(18.6)[0.07]	-20.3	1002.2	999.5[0.10]	23.0	2.7
6 (b_1) CCC oop-bend	210.7(3.5)[0.01]	-3.5	207.2			
7 (b_2) CH_2 a-str.	3215.6(0.0)[0.00]	-160.0	3055.6			
8 (b_2) CH_2 rock	1051.6(3.4)[0.01]	-28.1	1023.6	1025.0[<0.01]	26.6	-1.4
9 (b_2) CCC ip-bend	273.9(1.4)[0.01]	-3.6	270.3			

^a Ref 3.**Table 5.** Theoretical Predictions of the Harmonic Vibrational Frequencies (in cm^{-1}), Anharmonic (Fundamental) Vibrational Frequencies (in cm^{-1}), and Infrared Intensities (in km mol^{-1}) [Ratio Relative to the Largest Intensity] for the \tilde{X}^1A_1 State of the Dideuterio-Propadienylidene (Vinylidene carbene, **3S**) C_3D_2 Molecule at the cc-pCVQZ CCSD(T) Level of Theory

mode number (assignment)	harmonic	$\Delta(\text{anh.} - \text{harm.})$	anharmonic	exptl. ^a	$\Delta(\text{harm.} - \text{exptl.})$	$\Delta(\text{anh.} - \text{exptl.})$
1 (a_1) CD_2 s-str.	2287.9(29.2)[0.13]	-82.0	2205.9	2200.5[0.07]	87.4	5.4
2 (a_1) CC a-str.	1981.4(232.4)[1.00]	-43.7	1937.6	1933.4[1.00]	48.0	4.2
3 (a_1) CD_2 sciss.	1228.5(8.5)[0.04]	-18.3	1210.2	1208.7[0.08]	19.8	1.5
4 (a_1) CC s-str.	967.6(0.0)[0.00]	-13.8	953.8	950.8[0.01]	16.8	3.0
5 (b_1) CD_2 wag.	816.3(8.9)[0.04]	-14.2	802.1	800.3[0.06]	16.0	1.8
6 (b_1) CCC oop-bend	203.3(4.5)[0.02]	-2.8	200.5			
7 (b_2) CD_2 a-str.	2395.6(0.7)[0.00]	-92.3	2303.3			
8 (b_2) CD_2 rock	845.7(5.2)[0.02]	-17.5	828.1	829.2[0.02]	16.5	-1.1
9 (b_2) CCC ip-bend	252.5(0.3)[0.00]	-3.4	249.1			

^a Ref 3.**Table 6.** Theoretical Predictions of the Harmonic Vibrational Frequencies (in cm^{-1}), Anharmonic (Fundamental) Vibrational Frequencies (in cm^{-1}), and Infrared Intensities (in km mol^{-1}) [Ratio Relative to the Largest Intensity] for the \tilde{X}^1A_1 State of the Monodeuterio-Propadienylidene (Vinylidene carbene, **3S**) C_3HD Molecule at the cc-pCVQZ CCSD(T) Level of Theory

mode number (assignment)	harmonic	$\Delta(\text{anh.} - \text{harm.})$	anharmonic	exptl. ^a	$\Delta(\text{harm.} - \text{exptl.})$	$\Delta(\text{anh.} - \text{exptl.})$
1 (a') CH str.	3173.9(1.9)[0.01]	-154.5	3019.4			
2 (a') CD str.	2337.2(13.2)[0.05]	-80.7	2256.5	2254.5[0.02]	82.7	2.0
3 (a') CC a-str.	1994.4(246.3)[1.00]	-44.8	1949.6	1940.6[1.00]	53.8	9.0
4 (a') CH bend	1367.1(9.1)[0.04]	-34.7	1332.4	1331.6[0.07]	35.5	0.8
5 (a') CC s-str.	1109.3(1.8)[0.01]	-17.5	1091.8			
6 (a') CD bend	882.5(4.1)[0.02]	-18.8	863.7	865.4[0.03]	17.1	-1.7
7 (a') CCC ip-bend	262.0(0.7)[0.00]	-3.5	258.5			
8 (a'') CHD wag.	925.0(13.7)[0.06]	-17.6	907.4	904.0[0.03]	21.0	3.4
9 (a'') CCC oop-bend	207.9(3.9)[0.02]	-3.1	204.8			

^a Ref 3.

Five fundamental vibrational frequencies have been identified for the monodeuterio isotopologue (HDCCC). For this isotopologue, the experimental assignments of the CD in-plane bending (ν_6, a') at 904.0 cm^{-1} and the CD out-of-plane bending (ν_8, a'') at 865.4 cm^{-1} should be reversed as the CD in-plane bending (ν_6, a') at 865.4 cm^{-1} and the CD out-of-plane bending (ν_8, a'') at 904.0 cm^{-1} . Predicted harmonic vibrational frequencies for these five modes are $\omega_2(a') = 2337 (+83)$, $\omega_3(a') = 1994 (+54)$, $\omega_4(a') = 1367 (+36)$, $\omega_6(a') = 883 (+17)$, and $\omega_8(a'') = 925 (+21) \text{ cm}^{-1}$. An averaged percent deviation relative to the harmonic vibrational frequencies for the monodeuterio HDCCC isotopologue is 2.6%. The differences between the theoretical (harmonic) and experimental (fundamental) vibrational frequencies may be mainly attributed to the anharmonicity of molecular vibrations. The anharmonicity effect for the vibrational frequencies will be addressed in the following three sections.

The IR intensity of the $\omega_2(a_1)$ mode is extraordinarily strong compared to those of the other eight modes, since it corresponds to a CC antisymmetric stretching motion with a large change in dipole moment.³ In Maier et al.'s paper,³ this I_2 intensity is regarded as a reference to determine relative intensities of other vibrational modes. In Seburg's experimental work, they were able to observe the three fundamental frequencies (ν_2, ν_3 , and ν_5) with the strongest intensities in an argon matrix at 8 K.¹¹ Theoretically predicted relative intensities provided in Tables 4–6 and Tables S1–S3 are in reasonable agreement with experimentally observed values.

5.5. Vibrational Anharmonic Constants. The r th anharmonic (fundamental) vibrational frequency (ν_r) is determined using the following equation:

$$\nu_r = \omega_r + 2\chi_{rr} + \frac{1}{2} \sum_{s \neq r} \chi_{rs} \quad (7)$$

Table 7. Theoretical Predictions of Anharmonic Vibrational Constants (in cm^{-1}) for the C_3H_2 , C_3D_2 , and C_3HD Molecules at the cc-pCVQZ CCSD(T) Level of Theory

	C_3H_2	C_3D_2	C_3HD		C_3H_2	C_3D_2	C_3HD
χ_{11}	-30.197	-14.725	-63.988	χ_{34}	-2.827	-4.767	-5.689
χ_{22}	-10.543	-9.262	-33.096	χ_{35}	-4.202	-1.022	-10.707
χ_{33}	-10.527	-0.146	-10.015	χ_{36}	-0.021	5.367	-5.815
χ_{44}	-2.054	-1.071	-6.335	χ_{37}	-21.493	-2.881	-7.510
χ_{55}	-0.965	-2.004	-2.431	χ_{38}	-14.767	-9.523	-3.694
χ_{66}	-0.559	-0.572	-1.736	χ_{39}	-0.131	2.599	-8.377
χ_{77}	-35.583	-21.117	-21.117	χ_{45}	-3.182	-1.758	-3.857
χ_{88}	-2.593	-1.058	-1.898	χ_{46}	10.091	5.294	-11.609
χ_{99}	-0.516	-0.344	-0.561	χ_{47}	-1.403	-8.943	0.158
χ_{12}	-2.722	-7.578	-1.152	χ_{48}	-3.064	-2.987	-0.433
χ_{13}	1.863	-15.354	-2.383	χ_{49}	4.511	1.676	0.594
χ_{14}	-1.843	-6.489	-17.477	χ_{56}	-1.239	-1.620	-6.961
χ_{15}	-8.242	-3.988	-7.394	χ_{57}	-19.980	-12.215	-1.069
χ_{16}	-2.240	-2.288	-5.725	χ_{58}	4.545	1.733	-2.491
χ_{17}	-123.602	-61.080	-1.308	χ_{59}	1.397	0.902	7.835
χ_{18}	-12.206	-6.874	-15.224	χ_{67}	-2.292	-1.761	4.601
χ_{19}	-1.443	-1.393	-2.352	χ_{68}	-0.309	-0.585	0.368
χ_{23}	-5.022	-10.435	-5.318	χ_{69}^a	0.0	0.0	0.631
χ_{24}	-13.223	-5.311	-5.800	χ_{78}	-9.674	-6.682	1.134
χ_{25}	-5.835	-2.465	-0.665	χ_{79}^b	-1.403	-1.220	0.0
χ_{26}	-8.845	-7.673	-6.138	χ_{89}	0.068	-1.133	-0.443
χ_{27}	2.200	-5.342	-1.298				
χ_{28}	-10.374	-4.791	-6.794				
χ_{29}	-8.094	-6.813	-1.894				

^a $\chi_{69} = \chi_{96} = 0.0$ (adjusted) for C_3H_2 and C_3D_2 , whose original values were 17.7 and 8.1 cm^{-1} , respectively. ^b $\chi_{79} = \chi_{97} = 0.0$ (adjusted) for C_3HD , whose original value was 11.3 cm^{-1} .

where ω_r is an r th harmonic vibrational frequency and χ_{rs} are anharmonic vibrational constants. During the anharmonicity analyses, it was found that the anharmonic vibrational constant χ_{69} , the coupling constant between two CCC bending modes, is unphysically too large, probably due to the very low frequencies. Therefore, we adjusted this χ_{69} constant to be zero, i.e., $\chi_{69} = \chi_{96} = 0$ for the isotopologues with C_{2v} point group symmetry and $\chi_{79} = \chi_{97} = 0$ for the HDCCC isotopologue. The anharmonic vibrational constants (χ_{rs}) for C_3H_2 , C_3D_2 , and C_3HD are presented in Table 7, whereas those for the three ^{13}C labeled isotopologues are reported in Table S4 (Supporting Information).

5.6. Fundamental Vibrational Frequencies. In Tables 4–6, anharmonic vibrational frequencies for C_3H_2 , C_3D_2 , and C_3HD are included, while those for $\text{H}_2^{13}\text{C}=\text{C}=\text{C}$, $\text{H}_2\text{C}=\text{C}=\text{C}=\text{C}$, and $\text{H}_2\text{C}=\text{C}=\text{C}=\text{C}$ are shown in Tables S1–S3. In the last column of Table 4, the anharmonic vibrational frequencies for the standard H_2CCC species are compared with the experimentally observed fundamental vibrational frequencies. The largest discrepancy (59.0 cm^{-1}) is seen for the CH_2 symmetric stretching mode, possibly due to higher anharmonicity, or an experimental misassignment, given that our computed value for the CC antisymmetric stretch lies only 5.6 cm^{-1} away from the experimentally assigned CC antisymmetric stretching frequency. Given the close agreement between experiment and theory for the C–D stretching frequencies in the deuterated isotopologs, further spectroscopic studies of the C_3H_2 (**3S**) species are highly desirable, to confirm the laboratory assignment.

For the dideutero isotopologue (D_2CCC), the agreement between theoretical anharmonic and experimental fundamental frequencies is again excellent, as shown in Table 5. The largest deviation is 5.4 cm^{-1} for the CD_2 symmetric

stretching mode. In the last column of Table 6, a similar comparison has been made for the monodeutero isotopologue (HDCCC). The largest difference of 9.0 cm^{-1} is seen for the CC antisymmetric stretching mode. The theoretical anharmonic vibrational frequencies well reproduce the experimental fundamental frequencies. Specifically, our new assignments for the CD in-plane bending (ν_6, a') and CD out-of-plane bending (ν_8, a'') modes, as mentioned above (in section 5.4), are in excellent agreement with the experimental observations.

In the last columns of Tables S1–S3 (Supporting Information), the differences between the theoretical anharmonic and experimental fundamental vibrational frequencies for the three ^{13}C -labeled C_3H_2 isotopologues are presented. For each of the three isotopologues, the largest deviation is seen for the CC antisymmetric stretching mode (ν_2): 4.4 ($\text{H}_2^{13}\text{C}=\text{C}=\text{C}$), 4.7 ($\text{H}_2\text{C}=\text{C}=\text{C}=\text{C}$), and 4.6 cm^{-1} ($\text{H}_2\text{C}=\text{C}=\text{C}=\text{C}$). The mean absolute deviation between theoretical anharmonic and experimental fundamental vibrational frequencies for 24 modes of the six isotopologues (excluding CH_2 symmetric stretching) is only 2.6 cm^{-1} .

5.7. Isotopic Shifts of Vibrational Frequencies. In Table 8, theoretical isotopic shifts (in cm^{-1}) for the harmonic and anharmonic vibrational frequencies of the D_2CCC and HDCCC isotopologues with respect to the standard H_2CCC species are compared with the corresponding shifts of the experimental fundamental vibrational frequencies. Upon deuteration, the five vibrational modes involving the H atoms [mode 1 (a_1 , CH_2 s-str.), mode 3 (a_1 , CH_2 sciss.), mode 5 (b_1 , CH_2 wag.), mode 7 (b_2 , CH_2 a-str.), and mode 8 (b_2 , CH_2 rock)] are significantly red-shifted, as expected. Isotopic shifts from the theoretical harmonic vibrational frequencies are reasonably consistent with the experimental observations. On the other hand, isotopic shifts based on the theoretical anharmonic vibrational frequencies are in excellent agreement with the experimental values, within 2 cm^{-1} of deviations except for the ν_1 (a_1 , CH_2 s-str.) mode of D_2CCC and ν_2 (a_1 , CC str.) mode of HDCCC.

In Table 9, theoretical isotopic shifts for harmonic and anharmonic vibrational frequencies of the $\text{H}_2^{13}\text{C}=\text{C}=\text{C}$, $\text{H}_2\text{C}=\text{C}=\text{C}=\text{C}$, and $\text{H}_2\text{C}=\text{C}=\text{C}=\text{C}$ isotopologues with respect to the standard H_2CCC molecule are compared with those of the experimental fundamentals. The ^{13}C -labeling is most significant for the antisymmetric CC stretching mode (ν_2, a_1) of the $\text{H}_2\text{C}=\text{C}=\text{C}=\text{C}$ and $\text{H}_2\text{C}=\text{C}=\text{C}=\text{C}$ isotopologues. On the other hand, the CH_2 scissoring mode (ν_3, a_1) and CH_2 wagging mode (ν_5, b_1) are affected notably by the ^{13}C -labeling of the CH_2 group ($\text{H}_2^{13}\text{C}=\text{C}=\text{C}$). Isotopic shifts from both theoretical harmonic and anharmonic vibrational frequencies are in excellent agreement with the experimental measurements of the ^{13}C -labeled isotopologues.

Concluding Remarks

Anharmonic rotational–vibrational analysis has been carried out for the electronic singlet state of the propadienylidene (**3S**) molecule employing vibrational second-order perturbation theory (VPT2). The equilibrium geometry and a quartic

Table 8. Theoretical and Experimental Isotopic Shifts of Vibrational Frequencies (cm^{-1}) for C_3D_2 and C_3HD with Respect to C_3H_2 at the cc-pCVQZ CCSD(T) Level of Theory

mode (symmetry)	$\text{D}_2\text{C}=\text{C}=\text{C}$ theory	$\text{D}_2\text{C}=\text{C}=\text{C}$ exptl.	mode (symmetry)	$\text{HDC}=\text{C}=\text{C}$ theory	$\text{HDC}=\text{C}=\text{C}$ exptl.
$\Delta\omega_1(a_1)$	-838.2		$\Delta\omega_1(a')$		
$\Delta\omega_2(a_1)$	-23.4		$\Delta\omega_2(a')$	-10.4	
$\Delta\omega_3(a_1)$	-263.5		$\Delta\omega_3(a')$	-124.9	
$\Delta\omega_4(a_1)$	-154.7		$\Delta\omega_4(a')$	-13.0	
$\Delta\omega_5(b_1)$	-206.2		$\Delta\omega_5(a'')$	-97.5	
$\Delta\omega_6(b_1)$	-7.4		$\Delta\omega_6(a'')$	-2.8	
$\Delta\omega_7(b_2)$	-820.0		$\Delta\omega_7(a')$		
$\Delta\omega_8(b_2)$	-205.9		$\Delta\omega_8(a')$	-169.1	
$\Delta\omega_9(b_2)$	-21.4		$\Delta\omega_9(a')$	-11.9	
$\Delta\nu_1(a_1)$	-784.6	-849.0	$\Delta\nu_1(a')$		
$\Delta\nu_2(a_1)$	-20.2	-18.8	$\Delta\nu_2(a')$	-8.2	-11.6
$\Delta\nu_3(a_1)$	-237.4	-238.2	$\Delta\nu_3(a')$	-115.2	-115.3
$\Delta\nu_4(a_1)$	-158.9		$\Delta\nu_4(a')$	-20.9	
$\Delta\nu_5(b_1)$	-200.1	-199.2	$\Delta\nu_5(a'')$	-94.8	-95.5
$\Delta\nu_6(b_1)$	-6.7		$\Delta\nu_6(a'')$	-2.4	
$\Delta\nu_7(b_2)$	-752.3		$\Delta\nu_7(a')$		
$\Delta\nu_8(b_2)$	-195.5	-195.8	$\Delta\nu_8(a')$	-159.9	-159.6
$\Delta\nu_9(b_2)$	-21.2		$\Delta\nu_9(a')$	-11.8	

Table 9. Theoretical and Experimental Isotopic Shifts of Vibrational Frequencies (cm^{-1}) for $\text{H}_2^{13}\text{C}=\text{C}=\text{C}$, $\text{H}_2\text{C}=\text{C}=\text{C}^{13}$, and $\text{H}_2\text{C}=\text{C}=\text{C}^{13}$ with Respect to C_3H_2 at the cc-pCVQZ CCSD(T) Level of Theory

mode (symmetry)	$\text{H}_2^{13}\text{C}=\text{C}=\text{C}$ theory	$\text{H}_2^{13}\text{C}=\text{C}=\text{C}$ exptl.	$\text{H}_2\text{C}=\text{C}=\text{C}^{13}$ theory	$\text{H}_2\text{C}=\text{C}=\text{C}^{13}$ exptl.	$\text{H}_2\text{C}=\text{C}=\text{C}^{13}$ theory	$\text{H}_2\text{C}=\text{C}=\text{C}^{13}$ exptl.
$\Delta\omega_2(a_1)$	-9		-16		-52	
$\Delta\omega_3(a_1)$	-8		-2		-1	
$\Delta\omega_5(b_1)$	-10		0		-1	
$\Delta\nu_2(a_1)$	-9	-8	-16	-15	-50	-49
$\Delta\nu_3(a_1)$	-9	-9	-2	-2	-1	-2
$\Delta\nu_5(b_1)$	-9	-7	0	2	-1	-3

force field have been determined at the all-electron cc-pCVQZ CCSD(T) level of theory. The predicted fundamental frequencies for propadienylidene as well as its ^{13}C and deuterium isotopologues are in good agreement with available experimental values. The mean absolute deviation between theoretical anharmonic and experimental fundamental vibrational frequencies for 24 modes of seven isotopologues (excluding CH_2 s-str.) is only 2.6 cm^{-1} . The isotopic shifts of the vibrational frequencies are also in excellent agreement with the available experimental values, agreeing within 2.0 cm^{-1} . It has been unambiguously demonstrated that second-order vibrational perturbation theory (VPT2) using a highly accurate quartic force field accompanied with careful detailed analysis does provide quite reliable information for a penta-atomic molecule (C_3H_2) and its ^{13}C and deuterium isotopologues. We hope that the present research will aid further characterization of the propadienylidene molecule and encourage theoretical and experimental studies in the areas of hydrocarbon chemistry, combustion chemistry, chemical dynamics, interstellar chemistry, and high-resolution spectroscopy. We encourage experimental reinvestigation of the CH_2 symmetric stretch to confirm its assignment.

Acknowledgment. This research was supported by the Department of Energy, Office of Basic Energy Sciences, Division of Chemistry, Fundamental Interactions Branch, Grant No. DEFG02-97ER14748, and used resources of 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. Q.W. and Q.H. gratefully acknowledge the support provided by the China Scholarship Council (CSC) [2008] 3019, and the University of Georgia Center for Computational Quantum Chemistry for hospitality during their one-year visit. We thank Dr. Justin M. Turney for many helpful discussions. We are indebted to the 111 Project (B07012) in China.

Note Added in Proof. We recently became aware of a complimentary study of the vibrational modes of propadienylidene [Botschwina, P.; Oswald, R. *J. Phys. Chem. A* **2010**, *114*, 9782], performed independently of our work, but at the same time. Their results are in very good agreement with ours, and a thorough analysis of the ν_1 fundamental misassignment is provided.

Supporting Information Available: Theoretical predictions of the harmonic vibrational frequencies, anharmonic vibrational frequencies, and infrared intensities and theoretical predictions of anharmonic vibrational constants. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Reisenauer, H. P.; Maier, G.; Riemann, A.; Hoffmann, R. W. *Angew. Chem., Int. Ed. Engl.* **1984**, *23*, 641.
- (2) Lee, T. J.; Bunge, A.; Schaefer, H. F. *J. Am. Chem. Soc.* **1985**, *107*, 137.

- (3) Maier, G.; Reisenauer, H. P.; Schwab, W.; Carsky, P.; Hess, B. A.; Schaad, L. J. *J. Am. Chem. Soc.* **1987**, *109*, 5183.
- (4) Maier, G.; Reisenauer, H. P.; Schwab, W.; Carsky, P.; Spirko, V.; Hess, B. A.; Schaad, L. J. *J. Chem. Phys.* **1989**, *91*, 4763.
- (5) Thaddeus, P.; Vrtilek, J. M.; Gottlieb, C. A. *Astrophys. J.* **1985**, *299*, L63.
- (6) Dateo, C. E.; Lee, T. J. *Spectrochim. Acta, Part A.* **1997**, *53*, 1065.
- (7) Lee, T. J.; Huang, X.; Dateo, C. E. *Mol. Phys.* **2009**, *107*, 1139.
- (8) Vrtilek, J. M.; Gottlieb, C. A.; Gottlieb, E. W.; Killian, T. C.; Thaddeus, P. *Astrophys. J.* **1990**, *364*, L53.
- (9) Cernicharo, J.; Gottlieb, C. A.; Guélin, M.; Killian, T. C.; Paubert, G.; Thaddeus, P.; Vrtilek, J. M. *Astrophys. J.* **1991**, *368*, L39.
- (10) Gottlieb, C. A.; Killian, T. C.; Thaddeus, P.; Botschwina, P.; Flüge, J.; Oswald, M. *J. Chem. Phys.* **1993**, *98*, 4478.
- (11) Seburg, R. A.; Patterson, E. V.; Stanton, J. F.; McMahon, R. J. *J. Am. Chem. Soc.* **1997**, *119*, 5847.
- (12) Stanton, J. F.; DePinto, J. T.; Seburg, R. A.; Hodges, J. A.; McMahon, R. J. *J. Am. Chem. Soc.* **1997**, *119*, 429.
- (13) Hodges, J. A.; McMahon, R. J.; Sattelmeyer, K. W.; Stanton, J. F. *Astrophys. J.* **2000**, *544*, 838.
- (14) Achkasova, E.; Araki, M.; Denisov, A.; Maier, J. P. *J. Mol. Spectrosc.* **2006**, *237*, 70.
- (15) Aguilera-Iparraguirre, J.; Boese, A. D.; Klopper, W.; Ruscic, B. *Chem. Phys.* **2008**, *346*, 56.
- (16) Vázquez, J.; Harding, M. E.; Gauss, J.; Stanton, J. F. *J. Phys. Chem. A* **2009**, *113*, 12447.
- (17) Nielsen, H. H. *Rev. Mod. Phys.* **1951**, *23*, 90.
- (18) Mills, I. M. In *Molecular Spectroscopy: Modern Research*; Rao, K. N., Mathews, C. W., Eds.; Academic: New York, 1972; pp 115–140.
- (19) Watson, J. K. G. In *Vibrational Spectra and Structure*; Durig, J. R., Ed.; Elsevier: Amsterdam, 1977; Vol. 6, pp 1–89.
- (20) Papoušek, D.; Aliev, M. R. *Molecular Vibrational-Rotational Spectra*; Elsevier: Amsterdam, 1982.
- (21) Clabo, D. A.; Allen, W. D.; Remington, R. B.; Yamaguchi, Y.; Schaefer, H. F. *Chem. Phys.* **1988**, *123*, 187.
- (22) Allen, W. D.; Yamaguchi, Y.; Császár, A. G.; Clabo, D. A.; Remington, R. B.; Schaefer, H. F. *Chem. Phys.* **1990**, *145*, 427.
- (23) Aarset, K.; Császár, A. G.; Sibert, E. L.; Allen, W. D.; Schaefer, H. F.; Klopper, W.; Noga, J. *J. Chem. Phys.* **2000**, *112*, 4053.
- (24) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.
- (25) Scuseria, G. E. *Chem. Phys. Lett.* **1991**, *176*, 27.
- (26) Kaiser, R. I.; Ochsenfeld, C.; Head-Gordon, M.; Lee, Y. T.; Suits, A. G. *Science* **1996**, *274*, 1508.
- (27) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.
- (28) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1995**, *103*, 4572.
- (29) Wu, Q.; Cheng, Q.; Yamaguchi, Y.; Li, Q.; Schaefer, H. F. *J. Chem. Phys.* **2010**, *132*, 044308.
- (30) Pulay, P. *Mol. Phys.* **1969**, *17*, 197.
- (31) Pulay, P. In *Modern Theoretical Chemistry*; Schaefer, H. F., Ed.; Plenum: New York, 1977; Vol. 4, pp 153–185.
- (32) Yamaguchi, Y.; Osamura, Y.; Goddard, J. D.; Schaefer, H. F. *A New Dimension to Quantum Chemistry: Analytic Derivative Methods in Ab Initio Molecular Electronic Structure Theory*; Oxford University Press: New York, 1994.
- (33) Stanton, J. F.; Gauss, J.; Watts, J. D.; Lauderdale, W. J.; Bartlett, R. J. *Int. J. Quantum Chem., Symp.* **1992**, *S26*, 879.
- (34) Stanton, J. F.; Gauss, J.; Watts, J. D.; Szalay, P. G.; Bartlett, R. J.; Auer, A. A.; Bernholdt, D. E.; Christiansen, O.; Harding, M. E.; Heckert, M.; Heun, O.; Huber, C.; Jonsson, D.; Jusélius, J.; Lauderdale, W. J.; Metzroth, T.; Michauk, C.; O'Neill, D. P.; Price, D. R.; Ruud, K.; Schiffmann, F.; Tajti, A.; Varner, M. E.; Vázquez, J. *ACES II* and the integral packages: *MOLECULE* (Almlöf, J.; Taylor, P. R.), *PROPS* (Taylor, P. R.), and *ABACUS* (Helgaker T.; Jensen, H. J. Aa.; Jørgensen, P.; Olsen, J.). Current version, see <http://www.aces2.de> (accessed Sep 2010).
- (35) Werner, H. J.; Knowles, P. J. *MOLPRO*, version 2006.1, a package of ab initio programs; see <http://www.molpro.net> (accessed Sep 2010).
- (36) Janssen, C. L.; Seidl, E. T.; Scuseria, G. E.; Hamilton, T. P.; Yamaguchi, Y.; Remington, R. B.; Xie, Y.; Vacek, G.; Sherrill, C. D.; Crawford, T. D.; Fermann, J. T.; Allen, W. D.; Brooks, B. R.; Fitzgerald, G. B.; Fox, D. J.; Gaw, J. F.; Handy, N. C.; Laidig, W. D.; Lee, T. J.; Pitzer, R. M.; Rice, J. E.; Saxe, P.; Scheiner, A. C.; Schaefer, H. F. *PSI 2.0.8*; PSITECH, Inc.: Watkinsville, GA, 1994.
- (37) GRENDL (GeneRAL ENergy Derivatives for Electronic structure) is a C++ program written by J. J. Wilke to perform general numerical differentiations to high orders of electronic structure results.
- (38) INTDER2005 is a general program written by W. D. Allen which performs various vibrational analyses and higher-order nonlinear transformations among force field representations.
- (39) Allen, W. D.; Császár, A. G.; Szalay, V.; Mills, I. M. *Mol. Phys.* **1996**, *89*, 1213.
- (40) ANHARM is a FORTRAN program for VPT2 analysis written by Yamaguchi, Y. Schaefer, H. F. Center for Computational Chemistry, University of Georgia: Athens, GA.
- (41) Kanata, H.; Yamamoto, S.; Saito, S. *Chem. Phys. Lett.* **1987**, *140*, 221.
- (42) East, A. L. L.; Johnson, C. S.; Allen, W. D. *J. Chem. Phys.* **1993**, *98*, 1299.

CT100347R

Charge Transfer Across ONIOM QM:QM Boundaries: The Impact of Model System Preparation

Nicholas J. Mayhall and Krishnan Raghavachari*

Department of Chemistry, Indiana University, 800 E. Kirkwood Avenue,
Bloomington, Indiana 47405

Received July 24, 2010

Abstract: The inability to describe charge redistribution from regions I to II at the high level of theory imposes limitations on the general applicability of the our own N-layered integrated molecular orbital and molecular mechanics (ONIOM) method. In this report, we exploit the most inexpensive components of an ONIOM QM:QM calculation to provide a new method which has the ability to describe such charge-transfer effects with only a nominal increase in computational effort. Central to this method is the model system preparation step, in which an one-electron potential is optimized to shift density into or out of a defined buffer region. In this initial effort, we treat the link atoms on the model subsystem as the electron buffer region and swell or diminish the link-atom nuclear charges to shift electron density into or out of the buffer region. Due to the relatively small computational cost of the model-low calculation, this procedure can be iteratively optimized to produce a charge distribution equal to the real-low calculation. Initial results for a test set of 20 reaction energies and 8 different combinations of high and low levels of theory show improvements of more than 35% over the standard ONIOM QM:QM approach, with improvements of up to 50% for some high and low combinations.

1. Introduction

The steep scaling of computational cost with molecular size restricts application of the most accurate quantum chemical methods to relatively small molecules.^{1–6} More approximate methods, while applicable to much larger chemical systems, often fail to achieve the accuracy required to obtain qualitatively correct results. One of the most popular approaches aimed at achieving a better balance of efficiency and accuracy has been through the use of hybrid energy methods.^{7–16} By restricting expensive calculations to only the chemically interesting regions of a large molecule, hybrid methods try to combine accuracy and efficiency to extend the range of systems able to be treated computationally. To perform a hybrid energy calculation, the chemical system is first partitioned into different regions, and in the simplest case, two regions, I and II. The energy is then defined as

$$E_{\text{hybrid}} = E(\text{I}) + E(\text{II}) + E_{\text{Interaction}} \quad (1)$$

where $E(\text{I})$ and $E(\text{II})$ are the regions I and II energy, and $E_{\text{Interaction}}$ is the interaction energy of the two regions. By judiciously selecting a region which contains the chemically relevant atoms, a high-level quantum mechanical (QM) method is used to obtain $E(\text{I})$, and a cheaper computational method is used to compute $E(\text{II})$ and $E_{\text{Interaction}}$. Morokuma and co-workers have developed a particularly useful hybrid energy method called our own N-layered integrated molecular orbital and molecular mechanics (ONIOM).^{17–23} This method is obtained by using the following definition of the interaction energy:

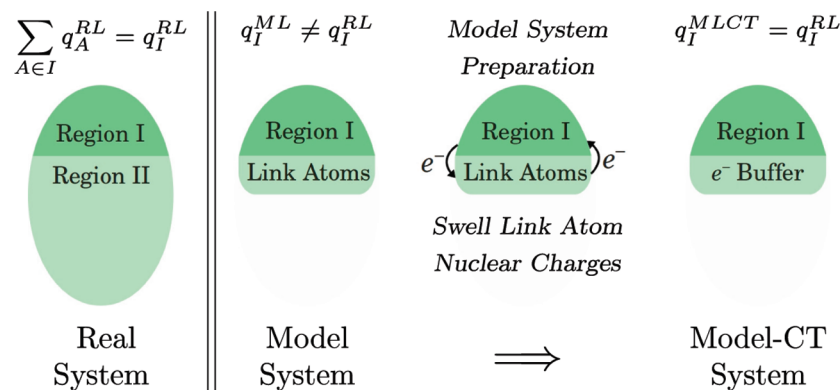
$$E_{\text{Interaction}} = E_{\text{LL}}(\text{I} + \text{II}) - E_{\text{LL}}(\text{I}) - E_{\text{LL}}(\text{II}) \quad (2)$$

where the LL subscripts indicate that the low level of theory is used. The standard ONIOM energy is therefore obtained from the following expression:

$$E_{\text{ONIOM}} = E_{\text{RL}} - E_{\text{ML}} + E_{\text{MH}} \quad (3)$$

where E_{RL} , E_{ML} , and E_{MH} denote the real system: low-level energy, the model system: low-level energy, and the model

* Corresponding author. E-mail: kraghava@indiana.edu.

Scheme 1. Schematic Illustration of the Model System Preparation Procedure^a

^a Note that all of these steps are done with the low level of theory, and q_A denotes atomic charge for atom A, while q_I denotes regional charges for region I.

system: high-level energy, respectively. The real system, being the full, unpartitioned molecule, is composed of both regions I and II. The model system, however, consists of only region I with the addition of link atoms which cap “dangling bonds” created from severed covalent bonds. This type of energy expression allows one to couple different computational chemistry methods without modification.

Despite its success, the ONIOM truncation of the model system can impose some artificial effects. While all energetic contributions (coulomb, exchange, charge-transfer, etc.) are fully included via the E_{RL} subcalculation, they are done so only at the low level of theory. Difficulties may arise when the link atoms have different electron donating/accepting properties from the region II atoms. This can generate a significantly different amount of electron density in region I of the model system. In such a case, the high-level correction ($E_{MH} - E_{ML}$) is performed on a model system, which is either partially oxidized or reduced relative to the untruncated system. To treat this problem, one would need to effectively change the number of electrons to achieve a fractional charge in region I. In this direction, Merz and co-workers²⁴ have developed an approach to match chemical potentials of different regions in a divide and conquer method^{25,26} by transferring charge from one region to another. While this was successfully done to couple density functional theory (DFT) and semiempirical methods, it is not clear whether this can be easily extended to post-self-consistent field (SCF) methods. Lin and co-workers have also addressed this problem in the context of quantum mechanical/molecular mechanical (QM/MM) methods.²⁷ In this approach, the fractional number of electrons is obtained by taking an ensemble average of integer charge states, whose weights are determined by the equalization of chemical potentials. Another hybrid method which also has the ability to describe some charge transfer from one region to the next is the generalized hybrid orbital (GHO) method of Gao and co-workers,^{28–30} which has also found use in computing the QM units in biomolecular application of the X-Pol method.³¹ Other workers have also treated the problem of cross-system charge redistribution, though in other contexts.^{32–34}

In this paper, we provide an efficient method to include charge redistribution across regional boundaries in an

ONIOM QM:QM calculation to reduce the unwanted electronic effects from truncation of the model system. This is done by means of a model system preparation step which provides a general approach to improving the standard ONIOM method.

2. Methods

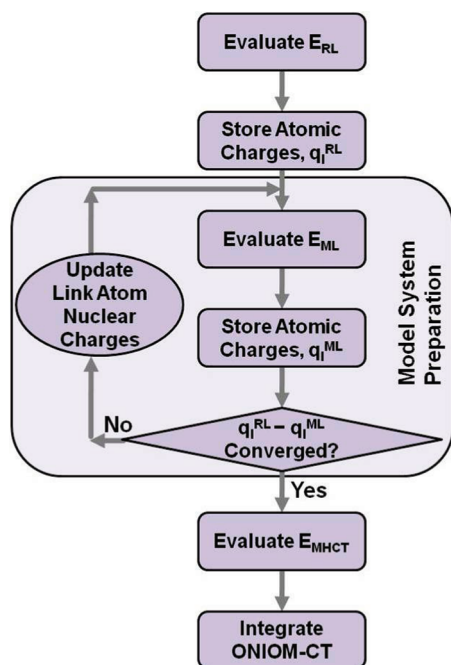
In Scheme 1, we present a schematic illustration of the currently presented method in which we optimize the model system to minimize the electronic differences between the RL and ML subcalculations. To begin, the real system is partitioned into a chemically active region I and inactive region II. By summing the atomic charges over the atoms in region I, we obtain the region I charge at the low level of theory, q_I^{RL} . Similarly for the ML subcalculation, the region I charge, q_I^{ML} , is obtained by summing the atomic charges from the ML calculation, which are in region I (i.e., all but the link atom charges). Generally

$$q_I^{RL} \neq q_I^{ML} \quad (4)$$

and thus the amount of region I charge in the model system is not the same as that in the real system. As shown in the third column of Scheme 1, we swell (or diminish) the nuclear charges on the link atoms to pull (or push) electron density out of (or into) region I. The magnitude of the resulting link atom nuclear charges is chosen such that the following becomes true:

$$q_I^{RL} = q_I^{MLCT} \quad (5)$$

where q_I^{MLCT} denotes the region I charge from the low-level model CT calculation. Eq 5 is satisfied by iteratively optimizing the link-atom nuclear charges and, in practice, requires around 4–5 ML subcalculations to obtain convergence to $10^{-5}e$. Due to the relatively low cost of the ML calculation, this can be done without significantly increasing the overall computational effort. While the region I charge can be defined with any population analysis, we have chosen to use Löwdin charges³⁵ due to their computational simplicity and proven performance for use in other applications.^{36,37} Later in the text we provide a performance comparison of a few commonly used population analyses.

Scheme 2. Flow Chart of an ONIOM-CT Calculation

Once the optimal link-atom nuclear charges are obtained for the MLCT subcalculation, the MHCT subcalculation is then performed using the same link-atom nuclear charges with no further buffer region optimization. The ONIOM-CT energy expression is then given as

$$E_{\text{ONIOM-CT}} = E_{\text{RL}} - E_{\text{MLCT}} + E_{\text{MHCT}} \quad (6)$$

where E_{MLCT} and E_{MHCT} denote the model CT energies with the low and high levels of theory, respectively. Note that the only difference between eqs 3 and 6 is that the model system has link atoms with modified nuclear charges. The individual steps of the ONIOM-CT calculation are shown in Scheme 2. We plan to make available a simple utility program, compatible with both Gaussian 03 and 09,^{38,39} to facilitate the computation of the ONIOM-CT energy.

Although we have defined the buffer region as the link atoms in this paper, the buffer region could also be chosen to include entire functional groups as well. Clearly, one would need to choose a different buffer region to describe charge-transfer effects for nonbonded model systems (i.e., no link atoms). However, the scaling procedure for the nuclear charges in such cases is not unique and has to be defined carefully. Just as the standard ONIOM approach requires model and real system definitions, the ONIOM-CT method requires model system, real system, and buffer region definitions. In future work, we will investigate the effect of different buffer region choices. While we currently only consider QM:QM models, it may be possible to perform an ONIOM-CT calculation with a MM low level (ONIOM-CT QM:MM) by using a polarizable force field to obtain the region I charges q_i^{RL} and q_i^{ML} . We plan to investigate the viability of such an approach in the future.

3. Results

To assess the performance of the ONIOM-CT method, we have carried out a set of calculations for a representative

Table 1. Test Set of Reactions^a

Deprotonation		
1)	$\text{X}_3\text{C}-\text{X}-\text{CH}_2\text{OH}_2^+$	$\rightarrow \text{X}_3\text{C}-\text{X}-\text{CH}_2\text{OH} + \text{H}^+$
2)	$\text{X}_3\text{C}-\text{X}-\text{CH}_2\text{OH}$	$\rightarrow \text{X}_3\text{C}-\text{X}-\text{CH}_2\text{O}^- + \text{H}^+$
3)	$\text{X}_3\text{C}-\text{X}-\text{CH}_2\text{NH}_3^+$	$\rightarrow \text{X}_3\text{C}-\text{X}-\text{CH}_2\text{NH}_2 + \text{H}^+$
4)	$\text{X}_3\text{C}-\text{X}-\text{CH}_2\text{NH}_2$	$\rightarrow \text{X}_3\text{C}-\text{X}-\text{CH}_2\text{NH}^- + \text{H}^+$
5)	$\text{X}_3\text{C}-\text{X}-\text{COOH}$	$\rightarrow \text{X}_3\text{C}-\text{X}-\text{COO}^- + \text{H}^+$
H-Abstraction		
6)	$\text{X}_3\text{C}-\text{X}-\text{CH}_2\text{OH}$	$\rightarrow \text{X}_3\text{C}-\text{X}-\text{CH}_2\text{O}^\bullet + \text{H}^\bullet$
7)	$\text{X}_3\text{C}-\text{X}-\text{CH}_2\text{NH}_2$	$\rightarrow \text{X}_3\text{C}-\text{X}-\text{CH}_2\text{NH}^\bullet + \text{H}^\bullet$
Electron Affinity		
8)	$\text{X}_3\text{C}-\text{X}-\text{CH}_2\text{O}^-$	$\rightarrow \text{X}_3\text{C}-\text{X}-\text{CH}_2\text{O}^\bullet + \text{e}^-$
9)	$\text{X}_3\text{C}-\text{X}-\text{CH}_2\text{NH}^-$	$\rightarrow \text{X}_3\text{C}-\text{X}-\text{CH}_2\text{NH}^\bullet + \text{e}^-$
S _N 2 Reactions		
10)	$\text{X}_3\text{C}-\text{X}-\text{CH}_2\text{F} + \text{Cl}^-$	$\rightarrow \text{X}_3\text{C}-\text{X}-\text{CH}_2\text{Cl} + \text{F}^-$

^a All 20 reactions have been carried out with X = F and CH₃. Location of ONIOM system partition is denoted by wavy lines.

Table 2. List of the Eight Combinations of High and Low Levels of Theory Used Throughout This Paper

entry	high level	:	low level
1a	MP2/6-311+G(d,p)	:	HF/3-21G
1b	MP2/6-31+G(d)	:	HF/3-21G
2a	B3LYP/6-311+G(d,p)	:	B3LYP/3-21G
2b	B3LYP/6-31+G(d)	:	B3LYP/3-21G
3a	MP2/6-311+G(d,p)	:	B3LYP/3-21G
3b	MP2/6-31+G(d)	:	B3LYP/3-21G
4a	B3LYP/6-311+G(d,p)	:	HF/3-21G
4b	B3LYP/6-31+G(d)	:	HF/3-21G

test set of 20 chemical reactions using several different combinations of high and low levels of theory. The test set is listed in Table 1 and includes not only different types of reactions but also multiple electron-donating/-accepting properties of the region II subsystem (X = F and CH₃). The different high:low combinations, which are listed in Table 2, are chosen to highlight some of the effects of choosing combinations with varying disparities both in electron correlation (MP2:HF, B3LYP:B3LYP, MP2:B3LYP, and B3LYP:HF) and in basis set size (6-311+G(d,p):3-21G and 6-31+G(d):3-21G). Note that for the low level of theory, only the 3-21G basis set has been considered in this study. Since the regional charges are obtained from a population analysis of the low-level subcalculations, we have chosen to use a small low-level basis set to avoid the known problems associated with performing Mulliken and Löwdin population analyses with large basis sets.⁴⁰

In Table 3, we provide the results for the full test set of reactions using each high:low combination listed in Table 2. Each quantity reported (e.g., mean absolute deviation (MAD), standard deviation, etc.) is given for each high:low combination with the averaged value given at the bottom. Throughout this paper, a deviation is defined as the difference between reaction energies obtained from the real high calculations and those from the ONIOM or ONIOM-CT calculations. Therefore, a zero deviation would indicate that the ONIOM or ONIOM-CT method produced a reaction

Table 3. Comparison of the Performance of the ONIOM and ONIOM-CT Methods^a

high:low	MAD		standard deviation		max deviation		MAD(X = F)		MAD(X = CH ₃)	
	ONIOM	ONIOM-CT	ONIOM	ONIOM-CT	ONIOM	ONIOM-CT	ONIOM	ONIOM-CT	ONIOM	ONIOM-CT
1a	3.26	1.58	3.99	2.14	6.65	5.45	3.90	1.83	2.62	1.34
1b	2.96	1.40	3.59	1.85	5.64	4.55	3.13	1.35	2.79	1.44
2a	3.80	2.77	5.67	3.51	11.77	6.67	1.45	2.01	6.15	3.53
2b	3.72	2.33	5.61	2.92	11.71	5.49	1.37	2.38	6.07	2.28
3a	3.70	2.36	5.29	3.09	10.86	7.06	1.87	3.00	5.53	1.73
3b	3.63	3.10	5.48	4.09	11.72	8.96	1.45	4.50	5.80	1.71
4a	3.49	2.34	4.14	2.75	6.71	4.78	3.53	2.34	3.46	2.33
4b	3.49	1.98	4.13	2.39	6.62	4.89	3.60	2.15	3.38	1.80
average	3.51	2.23	4.74	2.84	8.96	5.98	2.54	2.45	4.48	2.02

^a Mean absolute deviation is MAD. MAD(X = F) indicates the MAD for the subset of reactions for which X = F. MAD(X = CH₃) indicates the MAD for the subset of reactions for which X = CH₃. Units are in kcal/mol.

energy identical to that of the high level of theory on the real system.

By inspection of the overall mean absolute deviations (MAD's) in the first two columns in Table 3, it is clear that the ONIOM-CT method provides significant overall improvements over the standard ONIOM method. While it is notable that the mean absolute deviation (MAD) for all 160 data points is reduced from 3.51 to 2.23 kcal/mol, it is especially encouraging to see that for every high:low combination the ONIOM-CT method provides significant improvements, at times reducing the MAD by over 50%.

The inclusion of charge redistribution during the model system preparation step not only reduces the MAD but also makes the performance of the ONIOM-CT method more reliable, as suggested by the significantly lower standard deviations of the ONIOM-CT method. For each high:low combination the standard deviation is significantly decreased with the averaged standard deviations being 4.74 and 2.84 kcal/mol for the ONIOM and ONIOM-CT methods, respectively. Also providing an indication of reliability, the maximum deviation for each high:low combination is also decreased with improvements of up to 6 kcal/mol with high:low 2b.

By separating the test set into two subsets, X = F and CH₃, it is observed that the accuracy with which the standard ONIOM method treats these two groups is quite uneven. Using the ONIOM method, the X = F reactions have an overall MAD of only 2.54 kcal/mol, while the analogous reactions with X = CH₃ have a MAD of 4.48 kcal/mol. Use of the ONIOM-CT method reduces the performance discrepancy between the two subsets with MAD's of 2.45 and 2.02 kcal/mol for the X = F and CH₃ reactions, respectively. These are listed in the last 4 columns of Table 3.

3.1. Effect on Region I Density. To demonstrate the physical effect of the model system preparation, we provide in Figure 1 a density difference plot of the model system densities for both F₃C-*l*-COOH and (CH₃)₃C-*l*-COOH molecules. Here, the density of the MLCT subsystem (ρ^{MLCT}) is obtained from the low-level calculation on the model system with the optimized link-atom nuclear charges, while the density of the ML subsystem (ρ^{ML}) is obtained from the low-level calculation on the model system before optimization (i.e., link-atom nuclear charges = 1). The density difference ($\rho^{\text{MLCT}} - \rho^{\text{ML}}$) is shown with the black and gray surfaces representing negative and positive density

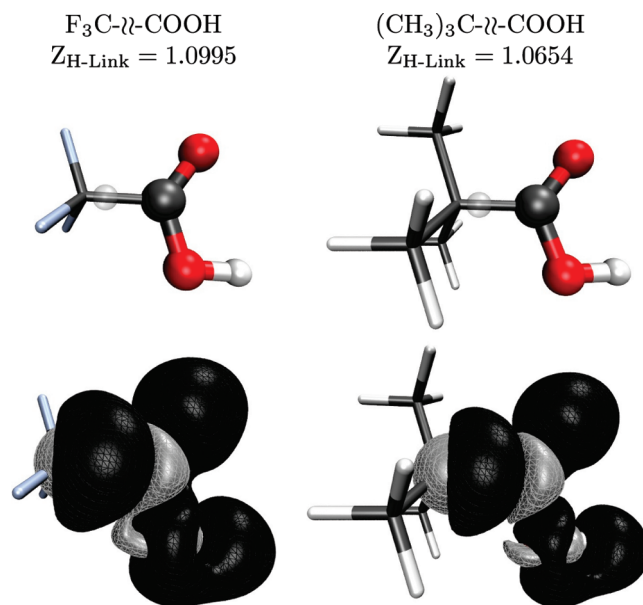


Figure 1. Change in density due to the model system preparation step, $\rho^{\text{MLCT}} - \rho^{\text{ML}}$. Black and gray surfaces represent negative and positive density changes, respectively. Isovalue = 0.0002. Translucent spheres indicate location of hydrogen link atoms. $Z_{\text{H-Link}}$ is the optimized value of the link-atom nuclear charge.

changes, respectively. To identify the locations of all the atomic centers, the bare molecules are given in the top of the figure, with the link-atom position shown as a translucent center. The ONIOM-CT method treats the link-atoms as a charge buffer region where density is either pushed out of or pulled into.

The effect on these two molecules is quite clear. As evident by the large amount of negative change in density around the oxygens, there is an overall transfer of negative charge out of region I into the buffer region. While both molecules show a qualitatively similar effect, the amount of charge redistributed is noticeably greater in F₃C-*l*-COOH. As F is a stronger electron acceptor than CH₃, this difference between the two molecules is consistent with what would be expected based on chemical intuition. The optimized link-atom nuclear charges giving rise to this transfer of density are 1.100 and 1.065 for F₃C-*l*-COOH and (CH₃)₃C-*l*-COOH, respectively.

3.2. Dependence on Population Analysis. The nonexistence of physical observables corresponding to chemical

Table 4. Comparison of the ONIOM-CT Method Using Either Löwdin, Mulliken, or Hirshfeld Population Analysis to Define the Region I Charges^a

reaction type	ONIOM	ONIOM-CT		
		Löwdin	Mulliken	Hirshfeld
deprotonation	3.83	2.24	2.62	2.86
H abstraction	0.52	1.04	1.84	1.14
electron affinity	5.71	3.12	4.46	3.81
S _N 2	3.46	2.83	2.75	2.99
X = F	2.54	2.45	3.33	2.82
X = Me	4.48	2.02	2.36	2.62
σ (std dev)	4.74	2.84	3.68	3.42
overall MAD	3.51	2.23	2.84	2.72

^a Results are given for all 160 reaction energies and are in kcal/mol.

quantities, such as the atomic charge, imposes a certain degree of arbitrariness upon any model which makes use of them. In the previous section, we report results for the ONIOM-CT method using Löwdin charges to define the region I charges, q_i^{RL} and q_i^{MLCT} , which are used during the model system preparation step. However, while Löwdin charges have been found to be favorable compared to other commonly used population analyses, benefit from their use is not obvious in this context, and thus we provide a comparison of the performance of a few different population analyses as used in the ONIOM-CT method.

In Table 4 we list the overall results (20 reactions and 8 high:low combinations) for the standard ONIOM and the ONIOM-CT methods using either Löwdin, Mulliken,⁴¹ or Hirshfeld charges⁴² during the model system preparation step.

From an inspection of the overall MAD's and average standard deviations for the different ONIOM-CT results, use of Löwdin charges provides the best performance, with a MAD of only 2.23 kcal/mol compared to 2.84 and 2.72 kcal/mol for Mulliken and Hirshfeld charges, respectively. Note however that while the performance does indeed have a dependence on the type of atomic charges used, all three population analyses provide significant improvements over the standard ONIOM-CT method in both the overall MAD and the standard deviations.

In all the examples listed in Table 2, we have only used the 3-21G basis set in the low level to avoid potential problems associated with using charges from a density matrix-based population analysis with large basis sets. We also carried out a few additional calculations to test the stability of the ONIOM-CT energy with increasing the low-level basis set size. Using the high:low combination MP2/6-311+G(d,p):HF/6-311+G(d,p) as an example, we found a decrease in the performance of the ONIOM-CT method compared to the standard ONIOM method with MADs of 2.53 and 1.62 kcal/mol for ONIOM-CT and ONIOM, respectively. Based on these results, it is suggested that large basis sets in the low level of theory be used with caution. If large low-level basis sets are required, then it may be advantageous to first project the associated wave functions onto a minimal basis set prior to performing the population analysis. We plan to investigate this carefully in future works.

4. Conclusions

In this paper, we have provided an inexpensive approach to describing inter-region charge-transfer effects in an ONIOM formalism. In this method, link-atom centers are treated as an electron buffer region, and their nuclear charges are swelled or diminished to transfer density into or out of the buffer region until the model system region I charge is identical to the real system region I charge.

To assess the performance of the ONIOM-CT method, we have performed calculations on a test set of 20 different reaction energies using 8 different combinations of high and low levels of theory. From these results it was found that for each high:low pair, the ONIOM-CT method provided significantly improved results as compared to the standard ONIOM method. For two of the high:low pairs, ONIOM-CT was found to be over 50% more accurate than ONIOM. Averaging over all high:low pairs, the overall MAD was decreased from 3.51 to 2.23 kcal/mol. In addition to improvements in accuracy, ONIOM-CT also provided precision gains as reflected by the greatly reduced standard and maximum deviations.

We have also performed a comparative assessment of using different population analyses in the model system preparation step. Among the Löwdin, Mulliken, and Hirshfeld atomic charge definitions, use of Löwdin charges provided the greatest improvements in both accuracy and precision over the standard ONIOM method.

While the focus of this work was to include the effects of charge redistribution from one ONIOM layer to another, the model system preparation step used in the ONIOM-CT method provides a general framework in which one may improve the ONIOM method.

Future plans include the development of an efficient procedure for obtaining gradients, coupling of the model system preparation step with our previous work on electronic embedding,^{36,43,44} and also generalizing the definition of the buffer region to extend beyond link atoms.

Acknowledgment. This work was supported by an National Science Foundation grant, CHE-0911454, at Indiana University. The authors would like to thank Dr. Mike Frisch for helpful suggestions related to this work and Dr. Hrant Hratchian for useful comments on the manuscript.

References

- (1) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2007**, *126*, 084108–084119.
- (2) DeYonker, N. J.; Cundari, T. R.; Wilson, A. K. *J. Chem. Phys.* **2006**, *124*, 114104.
- (3) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kallay, M.; Gauss, J. *J. Chem. Phys.* **2004**, *120*, 4129–4141.
- (4) Ochterski, J. W.; Petersson, G. A.; Montgomery, J. A., Jr. *J. Chem. Phys.* **1996**, *104*, 2598.
- (5) Karton, A.; Rabinovich, E.; Martin, J. M.; Ruscic, B. *J. Chem. Phys.* **2006**, *125*, 144108.
- (6) Tajti, A.; Szalay, P.; Csaszar, A.; Kallay, M.; Gauss, J.; Valeev, E.; Flowers, B.; Vazquez, J.; Stanton, J. *J. Chem. Phys.* **2004**, *121*, 11599.

- (7) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.
- (8) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986**, *7*, 718–730.
- (9) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700–733.
- (10) Aqvist, J.; Warshel, A. *Chem. Rev.* **1993**, *93*, 2523–2544.
- (11) Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170–1179.
- (12) Mordasini, T.; Thiel, W. *Chimia* **1998**, *52*, 288–291.
- (13) Monard, G.; Merz, K. M. *Acc. Chem. Res.* **1999**, *32*, 904–911.
- (14) Gao, J. L.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467–505.
- (15) Field, M. J. *J. Comput. Chem.* **2002**, *23*, 48–58.
- (16) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *117*, 185–199.
- (17) Humbel, S.; Sieber, S.; Morokuma, K. *J. Chem. Phys.* **1996**, *105*, 1959–1967.
- (18) Svensson, M.; Humbel, S.; Froese, R.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357–19363.
- (19) Karadakov, P. B.; Morokuma, K. *Chem. Phys. Lett.* **2000**, *317*, 589–596.
- (20) Vreven, T.; Morokuma, K. *J. Comput. Chem.* **2000**, *21*, 1419–1432.
- (21) Vreven, T.; Mennucci, B.; da Silva, C.; Morokuma, K.; Tomasi, J. *J. Chem. Phys.* **2001**, *115*, 62–72.
- (22) Vreven, T.; Morokuma, K. *Theor. Chem. Acc.* **2003**, *109*, 125–132.
- (23) Rega, N.; Iyengar, S.; Voth, G.; Schlegel, H.; Vreven, T.; Frisch, M. *J. Phys. Chem. B* **2004**, *108*, 4210–4220.
- (24) Gogonea, V.; Westerhoff, L. M.; Merz, K. M. *J. Chem. Phys.* **2000**, *113*, 5604.
- (25) Yang, W.; Lee, T. *J. Chem. Phys.* **1995**, *103*, 5674.
- (26) Dixon, S. L.; Merz, K. M. *J. Chem. Phys.* **1996**, *104*, 6643.
- (27) Zhang, Y.; Lin, H. *J. Chem. Theor. Comp.* **2008**, *4*, 414–425.
- (28) Gao, J. L.; Amara, P.; Alhambra, C.; Field, M. J. *J. Phys. Chem. A* **1998**, *102*, 4714–4721.
- (29) Pu, J. Z.; Gao, G. L.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 632–650.
- (30) Pu, J. Z.; Gao, J. L.; Truhlar, D. G. *Chem. Phys. Chem.* **2005**, *6*, 1853–1865.
- (31) Xie, W. S.; Gao, J. L. *J. Comp. Theor. Comp.* **2007**, *3*, 1890–1900.
- (32) Aviram, A.; Ratner, M. A. *Chem. Phys. Lett.* **1974**, *29*, 277–283.
- (33) Nitzan, A.; Ratner, M. A. *Science* **2003**, *300*, 1384–1389.
- (34) Pacheco, A.; Iyengar, S. S. *J. Chem. Phys.* **2010**, *133*, 044106.
- (35) Löwdin, P. O. *Adv. Quantum Chem.* **1970**, *5*, 185–199.
- (36) Mayhall, N. J.; Raghavachari, K.; Hratchian, H. P. *J. Chem. Phys.* **2010**, *132*, 114107.
- (37) Wu, Q.; Van Voorhis, T. *J. Phys. Chem. A* **2006**, *110*, 9212.
- (38) 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.
- (39) 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, A.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.1; Gaussian Inc.: Wallingford, CT, 2009.
- (40) Jensen, F. *Introduction to Computational Chemistry*, 2nd ed.; John Wiley & Sons: Chichester, England, 2007; pp 293–296.
- (41) Mulliken, R. S. *J. Chem. Phys.* **1962**, *36*, 3428.
- (42) Hirshfeld, F. L. *Theor. Chem. Acc.* **1977**, *44*, 129.
- (43) Hratchian, H. P.; Parandekar, P. V.; Raghavachari, K.; Frisch, M. J.; Vreven, T. *J. Chem. Phys.* **2008**, *128*, 034107.
- (44) Parandekar, P. V.; Hratchian, H. P.; Raghavachari, K. *J. Chem. Phys.* **2008**, *129*, 145101.

A Local Pair Natural Orbital Coupled Cluster Study of Rh Catalyzed Asymmetric Olefin Hydrogenation

Anakuthil Anoop,[†] Walter Thiel,[‡] and Frank Neese^{*†}

*Institut für Physikalische und Theoretische Chemie, Wegelerstr 12,
Bonn 53115 Germany and Max-Planck-Institut für Kohlenforschung,
Kaiser-Wilhelm-Platz 1, Mülheim an der Ruhr, 45470 Germany*

Received June 20, 2010

Abstract: The recently developed local pair natural orbital coupled cluster theory with single and double excitations (LPNO–CCSD) was used to study the rhodium-catalyzed asymmetric hydrogenation of two prochiral enamides. The method was carefully calibrated with respect to its accuracy. According to calculations on a truncated model system, the effects of perturbative triples (T) on the reaction energetics are very limited, the LPNO approximation is accurate, and complete basis set extrapolation (CBS) causes only minor changes in the relative energies computed with a standard basis set (def2-TZVP). The results for the full system are thus believed to be within 1–2 kcal/mol of the CCSD(T)/CBS limit for the present systems. Relativistic effects were treated by a scalar relativistic Hamiltonian using the zeroth order regular approximation (ZORA). The results of the study were compared to density functional calculations on the same systems and with calculations available in the literature. All calculations predict the correct stereochemical outcome of the reaction that is determined by the relative energies of the transition states in the early stages of the catalytic cycle. In general, DFT calculations using the B3LYP functional are in reasonable agreement with the LPNO–CCSD results, although deviations of 3–5 kcal/mol exist that are also not entirely systematic in the minor and major reaction branches. The present case study thus demonstrates that catalytic reactions, which are well described by single-reference electronic structure theory, can now be routinely studied with confidence in systems with 50–100 atoms applying local correlation methods that are as easy to use as DFT methods.

1. Introduction

New quantum chemical methods have to prove their utility in practical chemical applications. This is particularly true for local correlation approximations or otherwise simplified electron correlation methods where it is crucial that a consistent accuracy is maintained over the entire potential energy surface. Hence, species that differ in the number and nature of chemical bonds must be treated in a balanced way in order to obtain accurate results. At the same time, the methods should be easy to apply so that the user can focus

attention on the chemical problem at hand rather than worry about technical details. Thus, it should not be necessary to readjust truncation parameters or to run long series of preliminary calculations.

According to previous test calculations, the newly developed local pair natural orbital coupled cluster method with single and double excitations (LPNO–CCSD)¹ is believed to fulfill these criteria. The method has three conservatively chosen truncation parameters and recovers 99.8–99.9% of the canonical CCSD correlation energy. It has been claimed that this accuracy will be generally reached and is not dependent on the particular chemical system or the basis set. In order to test whether the LPNO–CCSD method lives up to these ambitious expectations, we have chosen the Rh-based asymmetric hydrogenation as a subtle example of a

* Corresponding author phone: +49-228-732351, fax: +49-228-739064, e-mail: neese@thch.uni-bonn.de.

[†] Institut für Physikalische und Theoretische Chemie.

[‡] Max-Planck-Institut für Kohlenforschung.

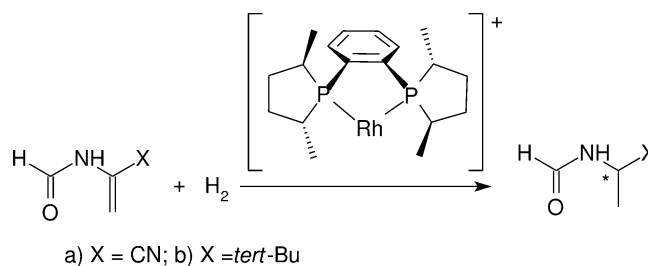
transition-metal-catalyzed reaction.² In this catalytic cycle, there is a large number of intermediates and transition states. The outcome of the reaction depends on small energy differences between these species. Hence, it represents a challenging test case for the LPNO–CCSD method.

Besides balance, accuracy, and black-box character, another important aspect of actual chemical applications is the pronounced basis set dependence of electron correlation methods. The slow convergence of the correlation energy to the basis set limit is well-known and needs to be taken into account if comparison with experimental results is the goal. Unfortunately, most local correlation methods that have been developed to date increase dramatically in their computational cost if they are used in conjunction with at least triply polarized triple- ζ basis sets that probably represent the lowest basis set level at which one obtains reliable results. This is different for the LPNO methods that behave excellently with basis set extension such that calculations with extended basis sets remain affordable. Explicitly correlated local correlation methods may change this situation in the future.³

The asymmetric hydrogenation of prochiral olefins using Rh catalysts is the prototype of an enantioselective transition-metal-catalyzed reaction. It has received considerable interest both from theory and experiment because of its intriguing mechanistic aspects and its strong industrial and academic impact.^{2b} Even though there are several successful catalysts for the efficient hydrogenation of several substrates, research efforts are continuing in this area due to the lack of a universal catalyst, and hence for each substrate an optimal catalyst has to be designed. The field has gained additional momentum from the discovery that highly enantioselective olefin hydrogenation can be achieved not only by the classic Rh catalysts with bidentate phosphorus ligands but also by Rh complexes containing BINOL-based monodentate ligands such as phosphites,⁴ phosphonites,⁵ and phosphoramidites.⁶ For the classic catalysts with bidentate ligands, the major aspects of the mechanism appear to be well established by now,⁷ and it is generally accepted that enantioselectivity results from kinetic control. That is, the minor diastereomer of the initially formed catalyst–substrate adduct reacts faster and thus yields the major product (anti-lock-and-key principle). By contrast, in the more recent catalysts with monodentate phosphorus ligands, there is evidence for the opposite behavior, with the major diastereomer leading to the favored enantiomeric product.⁸ Computational studies of such subtle mechanistic issues require methods that are efficient enough to be applied to complex transition-metal systems, and at the same time accurate enough to unravel the specific features of any particular catalyst–substrate combination.

In this article, we have selected two classic examples of asymmetric Rh-catalyzed olefin hydrogenation that have been studied previously using a variety of methods including DFT and DFT/MM hybrid methods.^{9,10} Since one of the major goals of the study is to compare local correlation with DFT methods, we focus on the established pathways in the catalytic cycle of two substrates, (a) α -formamidoacrylonitrile (hereafter called the cyano system)⁹ and (b) N(1-*tert*-

Scheme 1. Asymmetric Hydrogenation of Enamides



butylvinyl)formamide (hereafter called the butyl system).¹⁰ The catalyst used in both reactions is [(*R,R*)-MeDuPHOS]⁺ (Scheme 1).

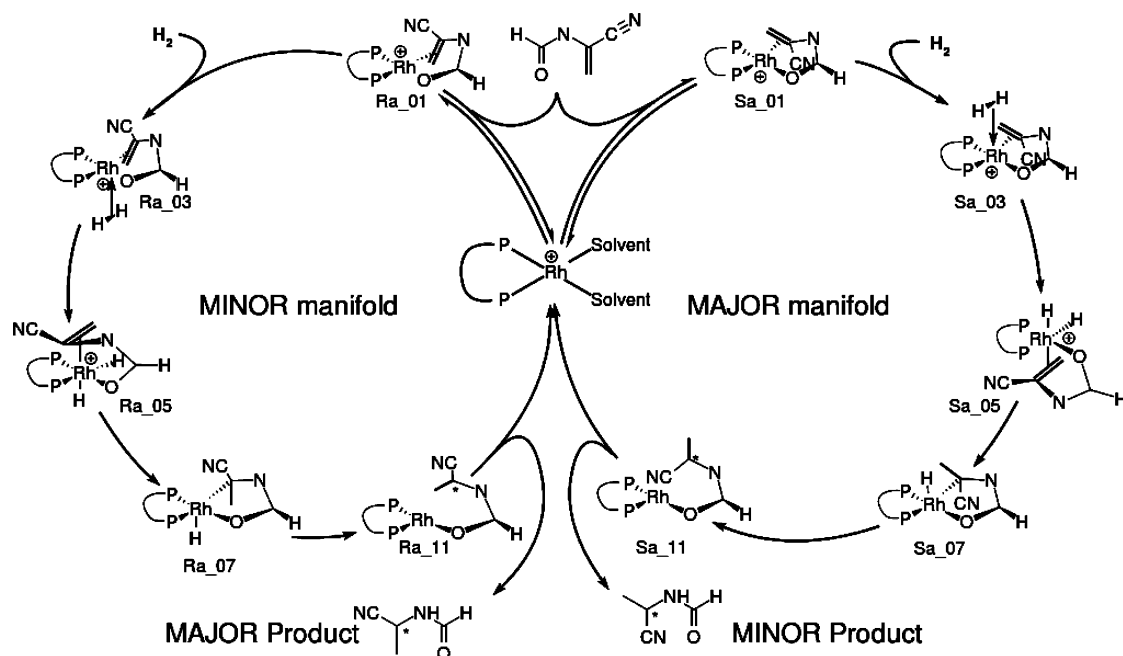
The generally accepted catalytic cycle (Scheme 2) involves the following elementary steps: (a) the formation of the diastereomeric catalyst–substrate adducts, (b) the addition of H₂ to the catalyst–substrate adduct, (c) the oxidative addition of H₂ to form the dihydride complex, (d) migratory insertion of the alkene into the Rh–H bonds in two consecutive steps, and (e) reductive elimination of the hydrogenated product.

Various reaction pathways originating from the different modes of coordination of the substrate to the catalyst and further from the different modes of addition of H₂ have been analyzed in detail by Feldgus and Landis.^{9,10} In this study, we have included pathways that are important in the catalytic cycle. As shown by Feldgus and Landis, the enantioselectivity of these reactions is dominated by the reactivity of the catalyst–substrate adduct rather than by its stability.^{9,10} The enantioselectivity and the rate-determining step have been shown to be variable and to depend on the specific combination of ligands and substrates.^{9,10} The previous theoretical studies by Landis and co-workers have led to the conclusion that for the cyano and butyl systems investigated here, the rate-determining step is indeed the oxidative addition. Thus, the barriers toward the formation of the dihydride intermediates determine the enantioselectivity and are therefore of particular interest for the present study.

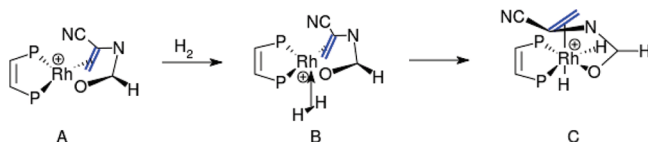
2. Computational Details

Density functional theory (DFT) was used for geometry optimizations. The optimizations for the butyl system were done by employing the BP86¹¹ functional in combination with the resolution of identity (RI) approximation¹² as implemented in Turbomole 5.71.¹³ The Rh atom was described using the ecp-28-mwb¹⁴ effective core potential and the associated basis set. All other atoms were described with the 6-31G*¹⁵ basis set. Transition state optimizations were performed with ChemShell (version 3.0ac)¹⁶ in conjunction with energy and gradient evaluations by Turbomole 5.71. For the cyano system, all optimizations were done using the ORCA program package¹⁷ together with the RI-BP86 method (within the Split-RI-J variant¹⁸) and the segmented all-electron relativistically contracted (SARC¹⁹) TZVP basis set.²⁰ Unless otherwise stated, all ORCA calculations employed the scalar relativistic, all-electron ZORA approach in conjunction with the model potential idea of van Wüllen.²¹

Single-point energies for all species were computed with ORCA at various levels of theory. They include the DFT

Scheme 2. Catalytic Cycle for the Cyano System^a

^a One pathway each for the formation of *R* and *S* products is shown.

Scheme 3. Model System Used for the Calibration Study

methods BP86 and B3LYP as well as the coupled cluster variants, CCSD, CCSD(T), LPNO-CCSD, LPNO-CEPA/1, and LPNO-QCISD.^{1a} For B3LYP calculations, the RIJCOSX approximation²² was applied. Ahlrichs basis sets of double-, triple-, and quadruple- ζ quality were used as described in the text. The truncation thresholds of the LPNO methods were left at their default values, $T_{\text{CutPNO}} = 3.33 \times 10^{-7}$, $T_{\text{CutPairs}} = 1 \times 10^{-4}$, and $T_{\text{CutMKN}} = 1 \times 10^{-3}$.¹

3. Results and Analysis

3.1. Calibration. For calibration purposes, the newly developed LPNO-CCSD approach was compared with canonical CCSD and CCSD(T) results. To this end, a truncated model system was constructed for which the bulkier substituents were replaced by hydrogen atoms (Scheme 3). The first two steps in the catalytic cycle of the minor manifold in the cyano system were modeled in this way. These two steps involve some of the more difficult bonding situations in the catalytic cycle. Since they also include some fairly major structural rearrangements (e.g., switch from a square planar to an octahedral rhodium species), these reaction steps should be representative of the entire catalytic cycle. The computed reaction and activation energies for the two steps are summarized in Table 1.

The Effect of Perturbative Triples. To study the influence of perturbative triple excitations, CCSD and CCSD(T) results were compared. In the canonical case, this can be done even

for the small model system only in conjunction with a small basis set (def2-SVP). With the triples correction included, the exothermicity of the first step increases by 1.66 kcal/mol, and the second step becomes less endothermic by 0.64 kcal/mol. The activation energies are reduced by 0.49 and 0.93 kcal/mol, respectively. The negative activation barrier for the first step (-0.20 kcal/mol) is probably an artifact of the small basis set since the use of the TZVP basis set for the same step yields a positive barrier of 0.75 kcal/mol. We conclude that the effects of connected triples are fairly limited and can be disregarded in the present system. This is fortunate because an accurate (T) approximation has not yet been developed in the LPNO framework, and canonical calculations on the target systems with saturated basis sets are presently not feasible. We shall thus proceed with CCSD for the remainder of this study.

The Effect of LPNO Approximation. The errors arising from the LPNO approximation are found to be very small in our model system. The largest deviation from the canonical CCSD results is 0.28 kcal/mol, which can be considered negligible. This attests once more to the reliability of the LPNO approximation that has so far always faithfully reproduced the canonical coupled cluster results.

Basis Set Convergence. There are various procedures for the extrapolation to the complete basis set (CBS) limit. We have chosen to extrapolate the Hartree-Fock energy using²³

$$E_{\text{SCF}}^{(X)} = E_{\text{SCF}}^{(\infty)} + A \exp(-\alpha\sqrt{X}) \quad (1)$$

where $E_{\text{SCF}}^{(X)}$ is the SCF energy for the basis set with cardinal number "X". We have employed the def2-TZVPP and def2-QZVPP basis sets for the extrapolation, and hence $X = 3$ and 4. $E_{\text{SCF}}^{(\infty)}$ is the basis set limit SCF energy obtained with $\alpha = 7.88$, the optimized value for the def2 basis sets.²⁴

Table 1. Reaction Energies and Activation Energies (kcal/mol) for the Model System (See Scheme 3)

method	basis	activation energy	reaction energy	activation energy	reaction energy
		A→B	A→B	B→C	B→C
BP86(RI)	def2-SVP	0.90	-7.20	4.02	2.30
B3LYP	def2-SVP	1.55	-2.71	6.17	3.79
CCSD	def2-SVP	0.29	-5.11	6.44	4.56
CCSD(T)	def2-SVP	-0.20	-6.77	5.51	3.92
CCSD(T)	ECP, TZVP	0.75	-6.09	5.80	4.52
LPNO-CCSD	def2-SVP	0.13	-4.83	6.48	4.44
LPNO-CCSD	def2-TZVPP	1.30	-3.17	6.53	4.36
LPNO-CCSD	def2-QZVPP	1.39	-3.34	6.56	4.33
LPNO-CCSD	CBS	1.36	-3.50	6.56	4.31
LPNO-CCSD	def2-TZVP	1.80	-3.04	6.70	4.67
LPNO-CCSD (full system)	def2-TZVP	2.07	-4.59	5.19	4.04
B3LYP(RIJCOSX)	def2-TZVP	2.75	-0.28	6.13	3.20
BP86(RI)	def2-TZVP	2.18	-4.38	3.93	1.48
BP86(RI)	CBS	1.84	-4.74	4.26	1.70
BP86(RI)	ECP,def2-TZVP	2.29	-4.14	4.28	1.86

For the extrapolation of the correlation energy, we use

$$E_{\text{corr}}^{(\infty)} = \frac{X^{\beta} E_{\text{corr}}^{(X)} - Y^{\beta} E_{\text{corr}}^{(Y)}}{X^{\beta} - Y^{\beta}} \quad (2)$$

with $\beta = 3.0$, which is the best choice for any combination of triple- and quadruple- ζ basis sets.^{24,25} The extrapolation from the def2-TZVP basis to the CBS limit according to eqs 1 and 2 leads to overall changes of 0.46 and 0.36 kcal/mol in the reaction energies, and of 0.44 and 0.14 kcal/mol in the activation energies, respectively (Scheme 3). These changes are small, which indicates that def2-TZVP is an appropriate basis set for our purposes. Therefore, we have not applied any such extrapolation procedures in the remainder of this study.

Comparison of DFT and CCSD(T) (def2-SVP Basis). The BP86 and B3LYP results differ by 0.65 and 2.15 kcal/mol for the activation energies, and by 4.49 and 1.49 kcal/mol for the reaction energies. Compared to CCSD(T), the deviations are as follows: BP86 activation energies, 1.10 and 1.49 kcal/mol; BP86 reaction energies, 0.53 and 1.62 kcal/mol; B3LYP activation energies, 1.75 and -0.66 kcal/mol; B3LYP reaction energies, -4.06 and 0.13 kcal/mol. Thus, if CCSD(T) is accepted as a reference, the DFT errors are acceptably small. However, one should keep in mind that the basis set dependence is generally less pronounced for DFT than for CCSD(T), so that it is not clear whether the deviations between DFT and CCSD(T) will decrease or increase for larger basis sets.

Comparison of the Full System and the Truncated Model System. In the final step of the calibration, we checked that the use of a truncated model system does not affect our conclusions. Thus, we compared the energetics of the two investigated steps in the model system (Scheme 3) against those in the full system at the LPNO-CCSD/def2-TZVP level. The activation energy for hydrogen addition increased by merely 0.27 kcal/mol in the full system. Contrary to the expectations derived from steric arguments, the formation of the molecular hydrogen complex is *more* favorable for the full system by 1.55 kcal/mol. Similarly, the barrier for the oxidative addition is reduced in the full system by 1.51 kcal/mol. The formation of the dihydride complex is slightly

less endothermic for the full system (0.63 kcal/mol). These changes are small enough to conclude that the validation achieved in the model system (with regard to perturbative triples, LPNO approximation, and basis set) will also hold for the full system. In an overall assessment, we thus expect that the deviations of the reported LPNO-CCSD results from the CCSD(T)/CBS limit are not larger than 1–2 kcal/mol.

3.2. Cyano System. We studied the two catalytically relevant pathways *R* and *S* that had previously been denoted as pathways **a** and **A**.⁹ The energies of all species in the reaction cycle were calculated at the LPNO-CEPA/1, LPNO-QCISD, and LPNO-CCSD levels with the def2-TZVP basis set. The computed relative energies are given in Table 2 relative to Sa_01 + H₂, where Sa_01 denotes the *proS* conformer of the catalyst-substrate adduct (Scheme 2). The energy difference between the major (*proR*) and minor (*proS*) adducts is 3.24, 3.37, and 3.38 kcal/mol at the LPNO-CEPA/1, LPNO-QCISD, and LPNO-CCSD levels, respectively, in reasonable agreement with the DFT values of 2.30 (BP86) and 2.20 (B3LYP) kcal/mol and the B3LYP/ONIOM literature value⁹ of 4.57 kcal/mol. The higher stability of the *proS* conformation (Sa_01) of the adduct compared to the *proR* conformation (Ra_01) is consistent with the experimental observation that the *proS* conformation is dominantly formed in solution.

In the following discussion, we focus on the computed energy profiles and disregard zero-point vibrational energies as well as finite-temperature and entropic effects. According to previous work,⁹ the initial addition of molecular hydrogen is associated with a large entropic penalty, but there are only minor differences between the energies and the free energies at 300 K for the other reaction steps. In particular, these differences are very similar for corresponding steps on the *R* and *S* pathways so that we can assess the relative ease of the steps in the *R* and *S* manifold on the basis of the computed energy profiles. A quantitative evaluation of reaction rates would of course require the inclusion of zero-point, finite-temperature, and entropic corrections, which is beyond the scope of this article.

Given these caveats, the present calculations confirm that the *R* pathway is kinetically favored over the *S* pathway, since the highest point on the corresponding reaction profiles

Table 2. Relative Energies (kcal/mol) for the Hydrogenation of α -Formamidoacrylonitrile (Cyano System) Using [(*R,R*)-MeDuPHOS]⁺^a

molecule	BP86	B3LYP	LPNO-CEPA/1	LPNO-QCISD	LPNO-CCSD	B3LYP/ONIOM
Ra_01 + H ₂	2.30	2.20	3.24	3.37	3.39	4.57
Ra_02	6.91	7.48	4.87	5.33	5.46	4.88
Ra_03	-0.40	3.66	-1.22	-1.01	-1.21	0.13
Ra_04	1.78	8.11	4.50	4.42	3.99	4.41
Ra_05	0.32	6.08	2.92	2.94	2.83	1.24
Ra_06	1.74	7.30	3.94	4.00	3.93	2.16
Ra_07	-6.94	-4.34	-8.78	-9.11	-9.14	-12.29
Ra_08	-5.32	-4.35	-9.22	-9.80	-9.88	-12.16
Ra_09	-19.19	-17.63	-21.06	-21.22	-21.52	-24.79
Ra_10	-5.94	-2.87	-4.54	-4.78	-4.55	-4.79
Ra_11	-24.78	-26.29	-25.57	-25.82	-25.49	-28.87
Sa_01 + H ₂	0.00	0.00	0.00	0.00	0.00	0.00
Sa_02	9.89	10.81	8.62	9.26	9.36	9.67
Sa_03	3.44	7.38	3.16	3.50	3.33	5.24
Sa_04	5.24	11.23	8.27	8.30	7.95	8.95
Sa_05	2.41	7.81	4.89	5.02	5.00	5.74
Sa_06	4.48	9.79	7.02	7.14	7.13	5.89
Sa_09	-16.86	-15.26	-18.64	-18.74	-18.99	-22.37
Sa_10	-3.75	-1.38	-1.70	-1.92	-1.85	-2.67
Sa_11	-24.98	-26.30	-25.79	-25.95	-25.61	-29.23
Mean	0.19	-2.47	0.04	-0.04		0.92
rms	2.04	3.16	0.34	0.21		1.98
MAD	1.73	2.73	0.25	0.16		1.61
MAX	4.56	5.53	0.74	0.43		3.62

^a The structures with odd numbers (e.g., Ra_01, Ra_03) represent minima (see Scheme 2), while the structures with even numbers (e.g., Ra_02, Ra_04) represent the intervening transition states. A statistical evaluation of the computed relative energies with respect to the LPNO-CCSD reference values provides the mean, root-mean-square (RMS), mean absolute (MAD), and maximum (MAX) deviations listed at the bottom. B3LYP/ONIOM data taken from Table 2 of ref 9.

is lower by 3.9 (LPNO-CCSD), 3.9 (LPNO-QCISD), 3.8 (LPNO-CEPA/1), 3.1 (B3LYP), 3.0 (BP86), and 4.8 kcal/mol (B3LYP/ONIOM⁹). In most cases (except for B3LYP), this is the transition state for the formation of the molecular hydrogen complex: the barrier for this process (relative to the catalyst-substrate adduct and H₂) is lower for the *R* pathway compared to the *S* pathway by 7.3 (LPNO-CCSD), 7.3 (LPNO-QCISD), 7.0 (LPNO-CEPA/1), 5.5 (B3LYP), 5.3 (BP86), and 5.7 kcal/mol (B3LYP/ONIOM⁹). The transition state for the subsequent oxidative addition (dihydride formation) has a similar energy that lies slightly below that of the initial transition state, typically within 1 kcal/mol (except for B3LYP, where it lies slightly higher, by less than 0.6 kcal/mol). These findings nicely explain why experimentally the *R* conformation of the final reaction product is formed in >99% enantiomeric excess. The higher activation energy found in the *S* pathway arises from the higher steric demand upon going from the square-planar catalyst-substrate adduct to the pentacoordinate molecular hydrogen complex and then to the dihydride. The proximity of the methyl group in the phosphine ligand to the substrate makes the transition state in the *S* pathway unfavorable relative to the one in the *R* pathway where the methyl group is rather distant from the substrate. This is a general feature of bidentate C₂ symmetric phosphine ligands which impose different steric demands on different enantiomeric pathways and thus facilitate chirality transfer.²

A statistical evaluation of the computed relative energies shows that LPNO-CEPA/1 and LPNO-QCISD reproduce the LPNO-CCSD reference values faithfully, with mean absolute deviations (MAD) of 0.25 and 0.16 kcal/mol, respectively, whereas BP86 and B3LYP show somewhat

larger deviations (MAD of 1.73 and 2.73 kcal/mol). The maximum deviations are below 1 kcal/mol for the LPNO-based methods and around 5 kcal/mol for the DFT methods. Compared with LPNO-CCSD, the energies of the transition states and intermediates generally tend to be too low for BP86 and too high for B3LYP, by up to several kilocalories per mole. The deviations for B3LYP are reasonably systematic (see Figure 1), but not entirely uniform; for example, B3LYP underestimates the difference between the barriers for the initial hydrogen addition on the minor and major pathways by 2 kcal/mol, relative to LPNO-CCSD (see above). The deviations for BP86 are generally even less uniform (see Table 2).

3.3. Butyl System. We studied four catalytically relevant pathways which had previously been denoted as pathways **A** and **C** in the *R* and *S* manifolds.¹⁰ The energies of all relevant species were computed at the BP86, B3LYP, LPNO-CEPA/1, and LPNO-CCSD levels. LPNO-QCISD calculations were not performed for the butyl systems since the LPNO-QCISD and LPNO-CCSD results had been very similar for the cyano system (Table 2). The calculated energies are given in Table 3 relative to Ra_01 + H₂.

Experimentally, the substitution of the cyano group in the substrate by a *tert*-butyl group (Scheme 1) changes the outcome of the reaction since it is exclusively the *S*-enantiomeric product that is observed in the butyl system (instead of the *R*-enantiomeric product in the cyano system). This has been rationalized previously¹⁰ by a careful analysis of the possible stereochemical pathways for hydrogenation. Using the terminology introduced previously,^{9,10} the favored mechanism involves pathway **C** in the butyl system and pathway **A** in the cyano system. These two pathways **A** and

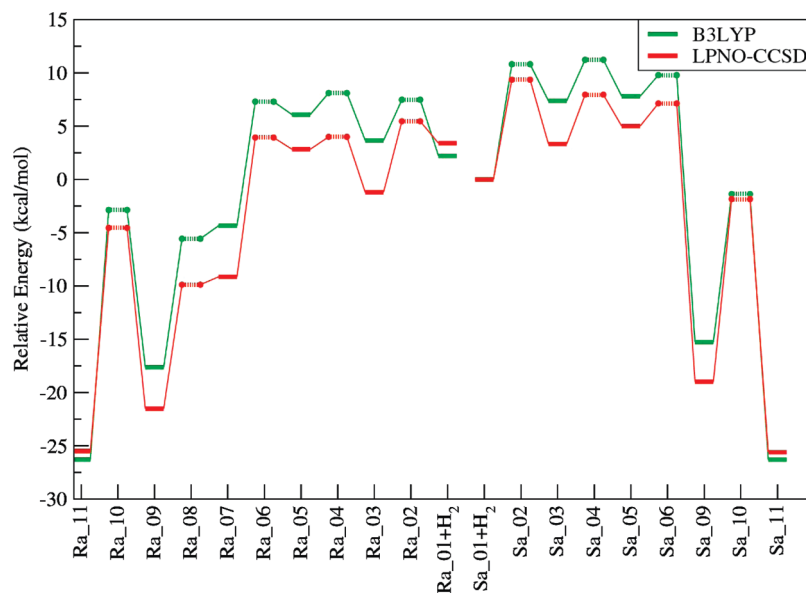


Figure 1. Reaction profile for the hydrogenation of α -formamidoacrylonitrile (cyano system) using $[(R,R)\text{-MeDuPHOS}]^+$ from B3LYP/def2-TZVP and LPNO-CCSD/def2-TZVP calculations.

C differ in the orientation of the incoming H_2 molecule with respect to the P-Rh-P plane and the C=C double bond of the substrate in the catalyst-substrate adduct. The geometry of this adduct is influenced by the substituents at the C=C double bond. The electron-deficient cyano group at the nonterminal α -carbon atom favors an orientation of the substrate with the terminal alkenyl carbon atom close to the P-Rh-P plane, whereas the *tert*-butyl group in the α position favors an orientation with the α -carbon close to this plane. The H_2 molecule prefers an approach from the side close to the terminal carbon (pathway A) in the former case and from the side of the nonterminal carbon (pathway C) in the latter case.

The *proR* conformer of the catalyst-substrate adduct has previously been found to be more stable than the *proS* conformer at the B3LYP/ONIOM level, but only by a small margin (0.46 kcal/mol).¹⁰ This is confirmed by the present LPNO-CCSD, LPNO/CEPA-1, and B3LYP calculations that yield energy differences of 1.84, 1.26, and 0.10 kcal/mol, respectively, whereas BP86 gives the opposite stability order (*proS* more stable by 1.06 kcal/mol). At all levels applied, and in agreement with previous work,¹⁰ pathway C in the *S* manifold is most favorable since the highest-energy barriers are much lower than those in the other pathways considered, because of less steric congestion.¹⁰ In the present calculations, the highest point in the reaction profile of the favored pathway is reached at the transition state for the initial migratory insertion (Sc_06), which lies slightly above that for oxidative addition (Sc_04), by 1.2 (LPNO-CCSD), 1.2 (LPNO-CEPA/1), 0.6 (B3LYP), and 2.1 (BP86) kcal/mol, contrary to the B3LYP/ONIOM results where Sc_04 is highest in energy (0.3 kcal/mol above Sc_06). According to the current calculations, the best route to the *R* product also involves pathway C. The highest point of the energy profile corresponds again to the transition state for the initial migratory insertion (Rc_06), which lies 4.5 (LPNO-CCSD), 4.5 (LPNO-CEPA/1), 2.5 (B3LYP), and 2.4 (BP86) kcal/mol above Sc_06. The preference for enantioselective

formation of the *S* product is thus more pronounced in the LPNO than in the DFT calculations. The B3LYP/ONIOM results¹⁰ agree with the LPNO-CCSD results with regard to the predicted enantioselectivity since pathway C in the *S* manifold is favored by an even larger margin, but there are some discrepancies with regard to the other three pathways (Table 3): contrary to LPNO-CCSD, B3LYP/ONIOM prefers pathway A as the best route to the *R* product and gives different highest-energy transition states (Ra_04, Rc_04, and Sa_04 instead of Ra_06, Rc_06, and Sa_06, see Table 3), although it should be noted that the corresponding energy differences are not large.

The statistical evaluation of the computed relative energies confirms again that the LPNO-CEPA/1 results are close to the LPNO-CCSD reference values (MAD of 0.74 kcal/mol) while the DFT results are less reliable, with MAD values of 3.76 (2.22) kcal/mol and maximum deviations of more than 9 (5) kcal/mol for BP86 (B3LYP). Compared with LPNO-CCSD, the relative energies of the transition states and intermediates are generally underestimated by BP86. The corresponding B3LYP values are often in the right ballpark, especially in the mechanistically crucial first three steps of the reaction cycle, although they overestimate the LPNO-CCSD values by 2–4 kcal/mol in the case of the mechanistically most important pathway (see Table 3, from Sc_03 to Sc_06).

4. Conclusions

In this study, we have used recently developed local correlation methods to study a well established catalytic reaction cycle that involves subtle stereoelectronic effects. The validation for a truncated model system shows that the LPNO-CCSD approach faithfully reproduces the canonical CCSD results and that the effects of connected triple excitations are small. The LPNO-CCSD method is thus suitable for investigating such systems. It has been applied to compute energy profiles of the complete catalytic cycle for rhodium-catalyzed asymmetric hydrogenation of two

Table 3. Relative Energies (kcal/mol) for the Hydrogenation of (1-*tert*-butylvinyl)formamide (Butyl System) Using [(*R,R*)-MeDuPHOS]⁺^a

	BP86	B3LYP	LPNO- CEPA/1	LPNO- CCSD	B3LYP/ ONIOM
Ra_01+ H ₂	0.00	0.00	0.00	0.00	0.00
Ra_02	7.65	9.73	7.82	9.34	
Ra_03	4.39	9.87	7.10	8.04	11.18
Ra_04	8.40	16.73	15.34	15.59	15.08
Ra_05	0.61	1.35	7.85	6.69	0.83
Ra_06	8.49	15.09	14.22	14.37	9.47
Ra_07	-2.27	3.64	3.64	3.84	1.29
Ra_08	4.87	8.44	5.12	4.80	
Ra_09	-8.73	-6.00	-10.75	-11.43	-8.95
Ra_10	-0.34	3.81	3.33	2.92	
Ra_11	-26.87	-30.43	-26.15	-27.59	
Rc_02	1.95	4.84	5.39	6.08	
Rc_03	-2.17	3.55	2.36	2.67	8.09
Rc_04	0.56	9.26	9.96	9.90	17.70
Rc_05	1.82	1.08	5.94	4.70	-2.87
Rc_06	6.35	12.45	11.56	11.49	13.83
Rc_09	-17.31	-16.80	-12.67	-14.44	-15.48
Rc_10	-8.12	-6.53	-2.74	-4.31	
Rc_11	-23.22	-27.48	-23.68	-25.05	
Sa_01+ H ₂	-1.06	0.10	1.26	1.84	
Sa_02	9.37	12.55	12.56	13.88	
Sa_03	6.62	12.36	11.41	12.34	19.18
Sa_04	12.15	20.45	19.12	19.35	25.16
Sa_05	0.22	0.97	7.10	5.95	0.38
Sa_06	10.69	17.70	18.98	19.40	15.41
Sa_07	5.44	10.85	8.32	8.92	7.25
Sa_08	6.42	11.04	8.00	7.83	
Sa_09	-10.99	-7.88	-10.63	-11.28	-7.95
Sa_10	1.01	4.93	3.48	3.09	
Sa_11	-27.03	-30.43	-26.96	-28.31	
Sc_02	0.80	3.49	2.33	2.98	
Sc_03	-4.41	1.27	-1.11	-0.91	2.37
Sc_04	1.82	9.53	5.87	5.75	7.53
Sc_05	-2.22	3.38	0.67	0.63	-2.66
Sc_06	3.96	10.09	7.10	6.99	7.20
Sc_09	-17.23	-16.62	-12.79	-14.52	-15.53
Sc_10	-11.87	-9.70	-2.91	-4.60	
Sc_11	-24.25	-27.22	-20.74	-22.39	
Mean	3.40	0.13	-0.27		-0.19
rms	4.41	2.70	0.94		4.18
MAD	3.76	2.22	0.74		3.50
MAX	9.34	5.43	1.77		7.80

^a The structures with odd numbers (e.g. Ra_01, Ra_03) represent minima (Scheme 2), while the structures with even numbers (e. g. Ra_02, Ra_04) represent the intervening transition states. A statistical evaluation of the computed relative energies with respect to the LPNO-CCSD reference values provides the mean, root-mean-square (RMS), mean absolute (MAD), and maximum (MAX) deviations listed at the bottom. B3LYP/ONIOM data from Table 1 of ref 10.

prochiral enamide substrates. The calculations for the cyano and butyl systems are in qualitative agreement with experimental results in predicting the correct stereochemical outcome of the reaction in both cases. The qualitative picture is consistent with earlier studies that were done with density functional theory using the B3LYP/ONIOM approach.

The present B3LYP results are similar to those previously obtained by Landis and co-workers with B3LYP/ONIOM. Comparison of the B3LYP and LPNO-CCSD relative energies demonstrates mean absolute deviations of 2–3 kcal/mol (Tables 2 and 3). While this is in the realm commonly assumed for B3LYP, the deviations for individual reaction steps can be as large as 5 kcal/mol. In the case of BP86, the mean absolute deviations are 2–4 kcal/mol, and the maximum deviations range up to 9 kcal/mol in the butyl system.

Moreover, the deviations between the DFT and LPNO-CCSD results are *not* constant over the potential energy surface (see the discussion in sections 3.2 and 3.3). These deviations should probably be considered as typical. They can be regarded as an indicator for the uncertainty in the computed DFT energies, since the results from high-level wavefunction based methods such as LPNO-CCSD are overall more reliable and more balanced.²⁶

It is reassuring that all calculations reported here provide qualitatively consistent scenarios for the reaction mechanism and the same qualitative explanation for the origin of enantioselectivity. On the other hand, the significant and nonuniform differences between the DFT and LPNO-CCSD relative energies call for extreme caution in attempts to come up with quantitative predictions of enantioselectivity. According to rate theory, rather small changes in the computed barriers will result in significantly different relative reaction rates into the minor and major channels and will hence influence the relative amounts of minor and major reaction products formed. For example, in a single-step reaction with two competing paths for *R* and *S*, a difference of 1 (3) kcal/mol in the activation free energy between two paths gives rise to 69% (99%) enantiomeric excess. While it is not certain that the LPNO-CCSD energies are accurate enough for the reliable assessment of the relative rates in competing channels, it is clearly preferable to use them for this purpose (rather than DFT energies).

It may be considered as remarkable progress in theoretical methodology that calculations that had to be performed by an ONIOM-type DFT/MM approach with small basis sets 10 years ago can now be done in an all-electron scalar relativistic fashion with coupled cluster theory and large basis sets. We believe that this sets the stage for a more widespread use of the more generally reliable wave function-based *ab initio* technology in the study of catalytic reactions of industrial and biochemical relevance. Of high importance in this respect is the development of a consistent connected triples approximation in the LPNO framework as triples effects are often important for predicting accurate energetics. Efforts along these lines are underway in our laboratories.

Acknowledgment. We gratefully acknowledge the SFB 813 (“Chemistry at Spin Centers”) for financial support of this work. A.A. thanks Martin Graf for helpful discussions.

References

- (1) (a) Neese, F.; Wennmohs, F.; Hansen, A. *J. Chem. Phys.* **2009**, *130*, 114108–18. (b) Neese, F.; Hansen, A.; Liakos, D. G. *J. Chem. Phys.* **2009**, *131*, 064103–15.
- (2) (a) de Vries, J. G.; Elsevier, C. J. *The Handbook of Homogeneous Hydrogenation*; Wiley-VCH: Weinheim, Germany, 2007. (b) Evans, P. A.; Tsuji, J. *Modern Rhodium-Catalyzed Organic Reactions*; Wiley-VCH: Weinheim, Germany, 2005.
- (3) (a) Kutzelnigg, W. *Theor. Chem. Acc. (Theor. Chim. Acta)* **1985**, *68*, 445–469. (b) Marchetti, O.; Werner, H. J. *J. Phys. Chem. A* **2009**, *113*, 11580–11585. (c) Höfener, S.; Tew, D. P.; Klopper, W.; Helgaker, T. *Chem. Phys.* **2009**, *356*, 25–30.

- (4) Reetz, M. T.; Mehler, G. *Angew. Chem., Int. Ed.* **2000**, *39*, 3889–3890.
- (5) (a) Reetz, M. T.; Sell, T. *Tetrahedron Lett.* **2000**, *41*, 6333–6337. (b) Claver, C.; Fernandez, E.; Gillon, A.; Heslop, K.; Hyett, D. J.; Martorell, A.; Orpen, A. G.; Pringle, P. G. *Chem. Commun.* **2000**, 961–962.
- (6) Van den Berg, M.; Minnaard, A. J.; Schudde, E. P.; van Esch, J.; de Vries, A. H. M.; Feringa, B. L. *J. Am. Chem. Soc.* **2000**, *122*, 11539–11540.
- (7) Brown, J. M. Mechanism of Enantioselective Hydrogenation. In *The Handbook of Homogeneous Hydrogenation*; de Vries, J. G., Elsevier, C. J., Ed.; Wiley-VCH: Weinheim, Germany, 2007; p 1073.
- (8) Reetz, M. T.; Meiswinkel, A.; Mehler, G.; Angermund, K.; Graf, M.; Thiel, W.; Mynott, R.; Blackmond, D. G. *J. Am. Chem. Soc.* **2005**, *127*, 10305–10313.
- (9) Feldgus, S.; Landis, C. R. *J. Am. Chem. Soc.* **2000**, *122*, 12714–12727.
- (10) Feldgus, S.; Landis, C. R. *Organometallics* **2001**, *20*, 2374–2386.
- (11) (a) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098. (b) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.
- (12) (a) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283–289. (b) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283–289. (c) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *242*, 652–660.
- (13) *TURBOMOLE*, V5.71; University of Karlsruhe and Forschungszentrum Karlsruhe GmbH: Karlsruhe, Germany, 2009. Available from <http://www.turbomole.com> (accessed Sep 2010).
- (14) Andrae, D.; Häußermann, U.; Dolg, M.; Stoll, H.; Preuß, H. *Theor. Chem. Acc. (Theor. Chim. Acta)* **1990**, *77*, 123–141.
- (15) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. *J. Chem. Phys.* **1982**, *77*, 3654–3665.
- (16) (a) Sherwood, P.; de Vries, A. H.; Guest, M. F.; Schreckenbach, G.; Catlow, C. R. A.; French, S. A.; Sokol, A. A.; Bromley, S. T.; Thiel, W.; Turner, A. J.; Billeter, S.; Terstegen, F.; Thiel, S.; Kendrick, J.; Rogers, S. C.; Casci, J.; Watson, M.; King, F.; Karlson, E.; Sjøvoll, M.; Fahmi, A.; Schäfer, A.; Lennartz, C. *THEOCHEM* **2003**, *632*, 1–28. (b) See <http://www.chemshell.org> (accessed Sep 2010).
- (17) Neese, F. *ORCA*, version 2.7; University of Bonn: Bonn, Germany, 2010. <http://www.thch.uni-bonn.de/tc/orca> (accessed Aug 28, 2010).
- (18) Neese, F. *J. Comput. Chem.* **2003**, *24*, 1740–1747.
- (19) Pantazis, D. A.; Chen, X.; Landis, C. R.; Neese, F. *J. Chem. Theory Comput.* **2008**, *4*, 908–919.
- (20) (a) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571–2577. (b) Weigend, F. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065.
- (21) van Wüllen, C. *J. Chem. Phys.* **1998**, *109*, 392–399.
- (22) (a) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. *Chem. Phys.* **2009**, *356*, 98–109. (b) Kossmann, S.; Neese, F. *Chem. Phys. Lett.* **2009**, *481*, 240–243.
- (23) (a) Klopper, W.; Kutzelnigg, W. *THEOCHEM* **1986**, *135*, 339–356. (b) Kutzelnigg, W. *Int. J. Quantum Chem.* **1994**, *51*, 447–463. (c) Zhong, S. J.; Barnes, E. C.; Petersson, G. A. *J. Chem. Phys.* **2008**, *129*, 184116.
- (24) Neese F. Unpublished results.
- (25) Truhlar, D. G. *Chem. Phys. Lett.* **1998**, *294*, 45–48.
- (26) Neese, F.; Hansen, A.; Wennmohs, F.; Grimme, S. *Acc. Chem. Res.* **2009**, *42*, 641–648.

CT100337M

Density Functional Calculations of E2 and S_N2 Reactions: Effects of the Choice of Method, Algorithm, and Numerical Accuracy

Marcel Swart,^{†,‡} Miquel Solà,[†] and F. Matthias Bickelhaupt^{*,§}

Institut de Química Computacional and Departament de Química, Universitat de Girona, Campus Montilivi, E-17071 Girona, Spain, Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, E-08010 Barcelona, Spain, and Department of Theoretical Chemistry and Amsterdam Center for Multiscale Modeling (ACMM), Scheikundig Laboratorium der Vrije Universiteit, De Boelelaan 1083, NL-1081 HV Amsterdam, The Netherlands

Received August 13, 2010

Abstract: Herein we provide a detailed account on how the potential energy surfaces of the E2 and S_N2 reactions of X[−] + CH₃CH₂X (X = F, Cl) depend on various methodological and technical choices in density functional calculations. We cover a choice of density functionals (OLYP, various M06-types, and the new SSB-D), basis sets (up to quintuple- and quadruple- ζ for Gaussian- and Slater-type orbitals, respectively, plus polarization and diffuse functions), and other aspects of the computations (among others: nonrelativistic versus zeroth-order regular approximation relativistic; numerical integration accuracy; all-electron versus frozen core; self-consistent field (SCF) versus post-SCF). The program codes ADF and NWChem are used for calculations with Slater- and Gaussian-type basis sets, respectively. The fluoride systems (X = F) appear to not only depend extremely sensitively on the basis set size (especially the presence of diffuse functions) but also on other technical settings, especially in the case of hybrid meta-generalized gradient approximation functionals. This work complements a recent contribution (Y. Zhao, D. G. Truhlar, *J. Chem. Theory Comput.* **2010**, *6*, 1104) and provides recommendations for density functionals, basis sets, and technical settings.

1. Introduction

Base-induced elimination (E2) and nucleophilic substitution (S_N2) constitute two fundamental types of chemical reactions that play an important role in organic synthesis.^{1,2} E2 elimination is, in principle, always in competition with S_N2 substitution, and the two pathways may occur as unwanted side reactions of each other (see Scheme 1).

Recently,³ some of us reported a benchmark study on the potential energy surfaces (PESs) for the *anti*- and *syn*-E2 and S_N2 pathways of X[−] with CH₃CH₂X (X = F, Cl) in the gas phase. A hierarchical series of ab initio methods was

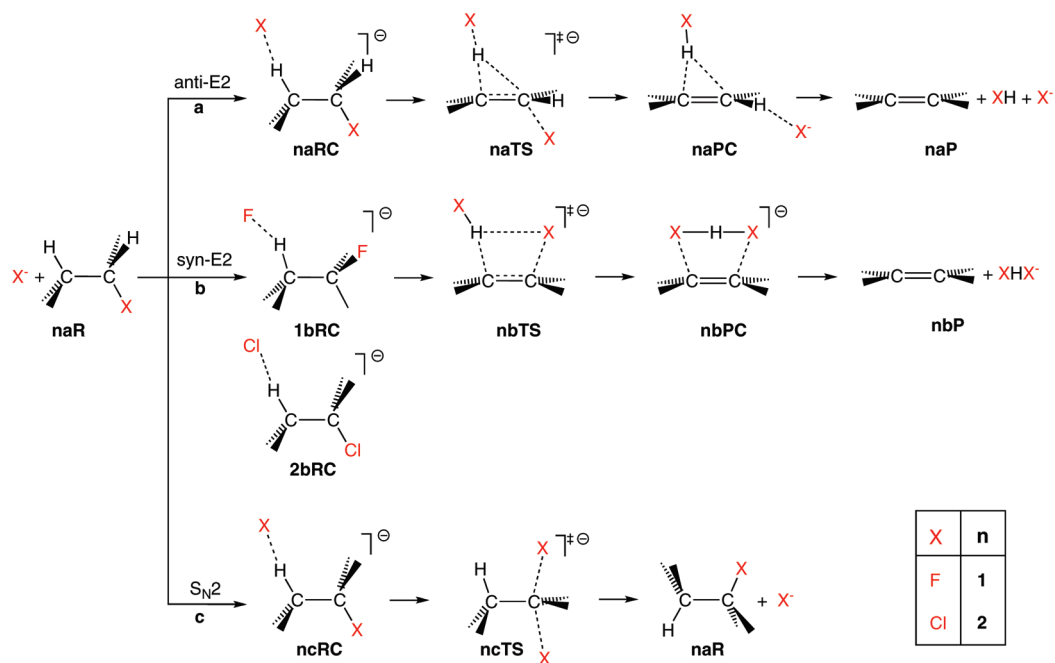
employed (Hartree–Fock (HF), second- and fourth-order Møller–Plesset (MP2 and MP4), coupled-cluster singles and doubles (CCSD), and CCSD(T)) in combination with a hierarchical series of Gaussian-type orbital (GTO) basis sets (up to quadruple- ζ + diffuse functions for reactions involving F, and up to (triple+d)- ζ + diffuse functions for reactions involving Cl). The performance of several popular density functionals were also evaluated using a TZ2P basis set of Slater-type orbitals (STOs).⁴ This TZ2P basis set is in general sufficiently large to be able to draw meaningful conclusions and normally gives results that are close to the basis set limit. In the case of the very demanding systems studied here (see below),⁵ the energies obtained may differ by a few kcal·mol^{−1} from the basis set limit results.⁶ In a recent paper, this is correctly pointed out by Zhao and Truhlar,⁷ who report mean unsigned errors (MUE) for several density functionals

* Corresponding author: fm.bickelhaupt@few.vu.nl.

[†] Universitat de Girona.

[‡] Institució Catalana de Recerca i Estudis Avançats.

[§] Scheikundig Laboratorium der Vrije Universiteit.

Scheme 1. E2 and S_N2 Reaction Pathways for X[−] + CH₃CH₂X

that were 1.5–2.0 kcal·mol^{−1} lower than those obtained with the TZ2P basis set. Nevertheless, the relative ordering in performance of the density functionals is much less affected, with the smallest MUE values observed for the Minnesota M06-2X functional,⁸ followed by M06,⁸ SSB-D,⁹ M06-L,¹⁰ OLYP,^{11,12} and TPSS.¹³

In the current contribution, we report a detailed study of the PESs for these competing pathways as obtained with a variety of density functionals and with both GTO and STO basis sets. In particular, we highlight the importance of several aspects of the computational setup that are shown to have a great impact on the results obtained, especially concerning the numerical accuracy. Our contribution not only provides full details referred to in the recent article of Zhao and Truhlar but also augments the latter with the above-mentioned numerical analyses and recommendations for density functionals, basis sets, and technical settings.

2. Computational Details

2.1. Basis Sets and Programs. All calculations using the STOs were obtained with the Amsterdam Density Functional (ADF)^{14,15} program (version 2009.01) using a variety of basis sets ranging in size from double- ζ valence plus polarization (DZP) to an even-tempered quadruple- ζ plus five polarization functions (ET-QZ+5P).¹⁶ The workhorse TZ2P (STO) basis set is of triple- ζ quality, augmented by two sets of polarization functions (2p and 3d on H and 3d and 4f on C, F, and Cl) and contains diffuse functions. In some cases, the zeroth-order regular approximation (ZORA)¹⁷ scalar relativistic Hamiltonian was used, but it was shown to have a small effect on the PES results (see below). All calculations with GTOs were obtained using the NWChem program¹⁸ (version 5.1.1 but locally modified to include SSB-D). The CCSD(T) calculations on the isolated fluoride and chloride anions used a frozen core, similar to those obtained in ref 3. A number of Dunning's^{19,20} correlation-consistent basis sets have been

used, ranging from double- ζ plus polarization (cc-pV(D+d)Z for Cl,²⁰ cc-pVDZ on others, abbreviated as dz) to quintuple- ζ plus diffuse functions (aug-cc-pV(5+d)Z on Cl,²⁰ aug-cc-pV5Z on others, abbreviated as a5z).

2.2. Numerical Integration. Unless noted otherwise, all of these calculations used a precise numerical integration scheme with a fine grid in NWChem (consisting of 70 radial and 590 angular shells for first-row elements and 123 radial and 770 angular shells for second-row elements) and ACCINT = 8.0 within ADF. These grids are used within NWChem only for the matrix elements of the exchange–correlation potential, while within ADF it is also used for other integrals¹⁵ (e.g., Coulomb). This allows for the use of the intrinsically most suitable basis sets (STOs) but is accompanied by a larger dependency on numerical integration. In most cases however, this poses no problem as the grid is generated in such a way that one can smoothly increase the accuracy by using more integration (grid) points. The size of the grid is determined by one parameter, the accuracy level (ACCINT).

2.3. Geometries and Mean Unsigned Errors. All energies of stationary points were obtained using the geometries optimized at OLYP/TZ2P from ref 3. Note, in Scheme 1, that all reactant complexes are trans-coplanar, except for **1bRC** in which [F[−]⋯H]–C^β–C^α–[F] is in a gauche conformation (i.e., **2aRC**, **2bRC**, and **2cRC** are identical species, and **1aRC** and **1cRC** are identical species). Mean unsigned errors (MUEs) in the energies relative to reactants are calculated as the mean absolute deviations with respect to the CCSD(T)/CBS energies for **1aRC**, **1aTS**, **1aPC**, **1aP**, **1bRC**, **1bTS**, **1bPC**, **1bP**, **1cRC**, and **1cTS** (MUE-F, F systems) and **2aRC**, **2aTS**, **2aPC**, **2aP**, **2bRC**, **2bTS**, **2bPC**, **2bP**, **2cRC**, and **2cTS** (MUE-Cl, Cl systems), and MUE is the average of MUE-F and MUE-Cl.

Table 1. Energies Relative to Reactants of Stationary Points Along the *Anti*-E2 Reaction of F⁻ + CH₃CH₂F and MUE for All Six Reactions Compared to CCSD(T)/CBS Data^a

method, basis	<i>E</i>					MUE		
	1aR	1aRC	1aTS	1aPC	1aP	MUE-F ^b	MUE-Cl ^c	MUE ^d
CCSD(T) ^a								
adz	0.0	-15.81	-2.03	-7.16	16.11	0.97	1.16	1.07
atz	0.0	-15.17	-1.31	-6.28	16.51	0.29	0.18	0.24
aqz	0.0	-14.99	-1.33	-6.39	15.95	0.11	n/a	n/a
CBS-limit	0.0	-14.89	-1.27	-6.35	15.77	-	-	-
OLYP								
qz ^e	0.0	-22.58	-10.19	-14.32	14.84	8.02	4.33	6.17
5z ^f	0.0	-17.44	-5.37	-9.55	13.30	3.51	4.17	3.84
aqz ^g	0.0	-12.21	-0.39	-4.25	12.94	2.16	3.64	2.90
a5z ^h	0.0	-11.84	-0.06	-3.88	12.87	2.42	3.67	3.05
TZ2P ⁱ	0.0	-19.90	-7.58	-12.00	13.22	5.94	4.69	5.32
QZ4P ^j	0.0	-17.83	-5.79	-9.83	12.97	3.88	3.72	3.80
df-ATZ2P ^k	0.0	-11.83	-0.22	-4.06	13.07	2.33	3.86	3.10
ET-QZ+5P ^l	0.0	-12.12	-0.45	-4.29	12.79	2.21	3.70	2.95
M06								
qz ^e	0.0	-21.44	-6.64	-11.11	18.79	5.76	2.15	3.96
5z ^f	0.0	-16.86	-2.57	-6.55	18.09	2.03	1.71	1.87
aqz ^g	0.0	-15.14	-0.63	-4.71	18.24	1.56	1.48	1.52
a5z ^h	0.0	-15.41	-1.25	-5.34	17.85	1.44	1.60	1.52
TZ2P ⁱ	0.0	-18.78	-3.00	-8.04	18.81	3.01	2.05	2.53
QZ4P ^j	0.0	-16.74	-2.37	-6.51	17.97	1.85	1.71	1.78
df-ATZ2P ^k	0.0	-14.83	-0.45	-4.63	17.80	1.29	1.51	1.40
ET-QZ+5P ^l	0.0	-14.74	-0.28	-4.24	18.13	1.57	1.74	1.65
M06-2X								
qz ^e	0.0	-20.55	-5.16	-10.31	19.18	5.21	1.12	3.16
5z ^f	0.0	-17.69	-2.62	-7.49	18.69	2.41	0.88	1.65
aqz ^g	0.0	-15.06	-0.26	-4.73	18.71	0.88	0.62	0.75
a5z ^h	0.0	-15.11	-0.41	-4.94	18.56	0.85	0.57	0.71
TZ2P ⁱ	0.0	-18.03	-2.00	-7.49	19.35	2.71	0.84	1.77
QZ4P ^j	0.0	-16.40	-1.79	-6.42	18.34	1.38	0.87	1.13
df-ATZ2P ^k	0.0	-14.81	-0.39	-4.80	18.31	0.85	1.07	0.96
ET-QZ+5P ^l	0.0	-15.11	-0.43	-5.01	18.57	0.86	0.86	0.86
M06-L								
qz ^e	0.0	-21.50	-5.91	-9.27	20.17	5.41	2.58	3.99
5z ^f	0.0	-17.15	-1.70	-4.81	19.62	2.95	2.10	2.52
aqz ^g	0.0	-15.71	0.05	-3.22	19.89	2.72	2.18	2.45
a5z ^h	0.0	-15.81	-0.39	-3.55	19.56	2.63	2.38	2.50
TZ2P ⁱ	0.0	-19.23	-3.11	-7.02	20.45	3.42	2.95	3.18
QZ4P ^j	0.0	-17.55	-1.98	-5.31	19.78	3.02	2.91	2.97
df-ATZ2P ^k	0.0	-15.50	-0.35	-3.72	19.95	2.42	2.66	2.54
ET-QZ+5P ^l	0.0	-15.68	0.00	-3.12	19.94	2.77	3.08	2.93
SSB-D								
qz ^e	0.0	-25.85	-10.61	-13.82	21.55	9.24	3.38	6.31
5z ^f	0.0	-20.32	-5.32	-8.46	20.48	4.56	2.80	3.68
aqz ^g	0.0	-15.92	-0.98	-3.88	20.40	2.83	2.39	2.61
a5z ^h	0.0	-15.74	-0.84	-3.74	20.33	2.78	2.33	2.56
TZ2P ⁱ	0.0	-23.29	-7.62	-11.57	19.40	6.91	3.16	5.04
QZ4P ^j	0.0	-21.55	-6.38	-9.68	20.45	5.33	2.61	3.97
df-ATZ2P ^k	0.0	-15.68	-1.04	-3.97	20.49	2.68	2.45	2.57
ET-QZ+5P ^l	0.0	-15.83	-1.03	-3.95	20.19	2.76	2.45	2.61

^a *E* and MUE are in kcal·mol⁻¹. CCSD(T)/CBS data from ref 3. See Scheme 1 for definition of species. See Computational Details for technical settings. ^b MUE for F-systems (reactions **1a–c**) with respect to CCSD(T)/CBS. ^c MUE for Cl-systems (reactions **2a–c**) with respect to CCSD(T)/CBS. ^d Total MUE for F- and Cl-systems (reactions **1a–c** and **2a–c**). ^e GTO basis: cc-pV(Q+d)Z on Cl and cc-pVQZ on others. ^f GTO basis: cc-pV(5+d)Z on Cl and cc-pV5Z on others. ^g GTO basis: aug-cc-pV(Q+d)Z on Cl and aug-cc-pVQZ on others. ^h GTO basis: aug-cc-pV(5+d)Z on Cl and aug-cc-pV5Z on others. ⁱ STO basis: TZ2P(all-electron). ^j STO basis: QZ4P(all-electron). ^k STO basis: augmented TZ2P:ATZ2P(all-electron) with AddDiffuseFit keyword. ^l STO basis: even-tempered ET-QZ+5P(all-electron).

3. Results and Discussion

3.1. Influence of the Basis Set. Recently,³ some of us reported coupled cluster benchmark results for the *anti*- and *syn*-E2 and S_N2 pathways of X⁻ with CH₃CH₂X (X = F, Cl). Based on an interpolation procedure, estimates for the complete basis set (CBS) limit were obtained³ for the coupled cluster results (see Table 1). Note that in ref 3, the CCSD(T)/CBS value for **1bPC** was inadvertently reported as -37.39 kcal·mol⁻¹ instead of the correct value -34.27 kcal·mol⁻¹.²¹ The CCSD(T) results with the various correlation-

consistent basis sets are also given in Table 1; see Scheme 1 for the definition of the various stationary points. They indicate that even with the quite large aug-cc-pVTZ basis (aug-cc-pV(T+d)Z for Cl) still a deviation of 0.74 kcal·mol⁻¹ can be observed compared to the complete basis set results (see Supporting Information and refs 5 and 22).

The same strong dependence on the size of the basis set is observed for the density functionals. This is shown for both the correlation-consistent GTO and the STO basis sets (see Table 1). However, unlike the case for the CCSD(T)

Table 2. Relative Energies for Fluoride and Chloride Anions with CCSD(T) and Different Density Functionals and Basis Sets^a

	CCSD(T)	OLYP	M06	M06-2X	M06-L	SSB-D
Fluoride Anion						
dz	143.1	90.0	71.5	74.8	66.2	85.1
tz	55.3	41.4	32.5	26.1	29.6	41.9
qz	22.3	23.1	14.7	13.1	12.4	21.2
5z	6.4	9.5	3.3	4.3	2.7	7.5
adz	74.0	21.7	17.7	23.0	19.3	20.9
atz	23.2	6.6	7.5	4.0	6.6	7.6
aqz	6.0	1.7	2.0	1.8	1.4	1.5
a5z	0.0	0.0	0.0	0.0	0.0	0.0
Chloride Anion						
dz	88.4	31.3	23.4	29.8	19.4	27.2
tz	30.0	13.9	10.6	8.8	8.1	13.0
qz	9.1	7.5	3.6	4.5	1.9	5.4
5z	2.7	4.1	1.1	1.7	0.4	2.6
adz	57.1	10.7	6.3	13.4	7.4	9.7
atz	17.7	3.4	4.2	2.4	3.4	3.5
aqz	4.0	1.4	1.2	1.7	0.7	0.8
a5z	0.0	0.0	0.0	0.0	0.0	0.0

^a E is in kcal·mol⁻¹. Relative to results obtained with a5z basis set.

data that seems to systematically lead to a lower MUE value with increasing basis set, there is no such systematical improvement for the density functional theory (DFT) data (see Table 1). For example, with OLYP, the MUE for the F systems decreases from 8.02 (qz) to 3.51 (5z) to 2.16 (aqz) but then increases to 2.42 kcal·mol⁻¹ (a5z). The same is observed for other functionals (see, e.g., MUE-Cl for M06-L with the qz, 5z, aqz, and a5z basis sets) and the STO basis sets (see, e.g., MUE-F of M06 with TZ2P, QZ4P, df-ATZ2P, and ET-QZ+5P). The M06-2X and SSB-D functionals seem to more systematically yield lower MUE values with increasing basis set size, although small jumps of less than 0.05 kcal·mol⁻¹ still occur (see Supporting Information).

The lowering of the MUE with increasing basis set size is in part due to a better description of the anionic nucleophiles (F⁻ and Cl⁻), where in particular, the description of the fluoride anion warrants a more flexible and large basis set. This is shown in Table 2 which, for the five functionals and the eight correlation-consistent GTO basis sets, shows the energies of fluoride and chloride anions, evaluated with the various basis sets, relative to the corresponding energy obtained with the largest basis set. Given that the relative energies in Table 1 are taken with respect to the reactants, which include the anionic halides, it matters greatly how well these anionic reactants are described. For instance for OLYP with the dz basis, the reactant complex energies of the F systems are too low (because the energy of F⁻ is too high), while those of the Cl systems are quite good. The MUEs with respect to the CCSD(T)/CBS data for the reactant complexes at the OLYP/dz level are therefore 21.3 and 1.6 kcal·mol⁻¹ for the F and Cl systems, respectively (see Supporting Information). The same is observed for other functionals, such as M06-2X, which in combination with the dz basis shows a MUE value for the RC complexes of the F systems of 16.9 and 3.2 kcal·mol⁻¹ for the Cl systems. As soon as the basis set allows for a better description of the charge distribution in the anionic halides, the MUE values decrease accordingly (see Table 2). A

similarly delicate dependence on the quality of the basis set was found in the case of the heterolytic dissociation energies of alkalimetal fluorides MF (M = Li, Na, K, Rb, Cs).²³ The extreme sensitivity of bond and reaction energies for systems involving fluorine can be ascribed to the large effective expansion (orbital “breathing effect”), as the very compact fluorine atom takes up excess negative charge.²⁴ This effect is still present for chlorine and heavier halogens but to a much lesser extent.

It should be noted that not just the basis set size but in particular the presence of diffuse functions is very important for these systems. This is not that surprising since the systems studied here are mainly of anionic character, for which diffuse functions are known to be important.⁶ Thus, for these systems, going from the tz to the qz or even the 5z basis yields less of an improvement than simply adding a layer of diffuse functions in the atz basis. For instance, in the case of M06-2X, going from the tz basis to the qz and 5z basis corresponds to a drop in the MUE value from 5.61 to 3.16 (qz) and 1.65 kcal·mol⁻¹ (5z), respectively (see Table 1 and Supporting Information). With the atz basis, a MUE value of just 0.72 kcal·mol⁻¹ is observed (see Supporting Information).

3.2. Evaluation of the Different Functionals. For any basis set, the lowest MUE value is found for the M06-2X functional, which is then followed by M06, M06-L, SSB-D, OLYP, and TPSS. This ordering of the different functionals is in agreement with the trends mentioned in the previous studies by some of us³ and by Zhao and Truhlar⁷ and situates our recently reported SSB-D⁹ as the best GGA functional among the ones applied to the E2 and S_N2 reactions in this study. This holds especially for the central barrier for which, with the ET-QZ+5P basis, an MUE compared to CCSD(T) of 3.1 kcal·mol⁻¹ is observed (see Supporting Information). This value can then be compared to those of the other functionals, i.e., 1.2 (M06-2X), 1.8 (M06), 3.9 (M06-L), 3.8 (OPBE), 4.5 (OLYP), and 6.2 kcal·mol⁻¹ (TPSS). Although the MUE for SSB-D is larger than the M06-2X and M06 results, it is significantly better than M06-L and OPBE which were the best meta-GGA and GGA functionals, respectively, of our previous study.³

The competition between the three pathways (*anti*-E2 vs *syn*-E2 vs S_N2) was shown previously, at CCSD(T)/CBS, to be different for the F and Cl systems.³ For the Cl systems, the S_N2 pathway is found to be the one with the lowest barrier. In contrast, for the F systems the *anti*-E2 pathway has the lowest barrier at the CCSD(T)/CBS level (see Table 3). This is correctly reproduced by most functionals³ but with some exceptions, like LDA, BP86, PBE, VS98, TPSS, M06-L, and BHandH. Again, our SSB-D functional works well for both the F and Cl systems (see Table 3), for which it predicts the lowest barrier to be *anti*-E2 and S_N2, respectively.

The effect of the basis set on the relative ordering of the reaction barriers (i.e., TS energies) is again found to be larger for the F systems than for the Cl systems (see Table 4). Although the energies relative to the reactants show dramatic changes of more than 20 kcal·mol⁻¹, this is mainly due to the description of the isolated halide ions (vide supra, Table 2). More important for the chemistry of these systems is the

Table 3. Energies Relative to Reactants for Transition States of All Six Reactions^a

	1aTS	1bTS	1cTS	2aTS	2bTS	2cTS
Ab Initio/CBS ^b						
CCSD(T)	-1.27	5.68	2.20	18.18	30.92	5.81
DFT/a5z						
OLYP	-0.06	4.90	5.02	13.85	22.98	7.32
M06	-1.25	3.61	0.71	15.57	25.27	2.31
M06-2X	-0.41	4.94	4.02	18.03	30.11	6.31
M06-L	-0.39	3.93	-2.11	13.22	22.32	0.49
SSB-D	-0.84	2.43	0.04	13.88	22.94	5.07
DFT/ET-QZ+5P						
OLYP	-0.45	4.51	4.66	13.62	22.73	7.08
M06	-0.28	4.62	1.42	15.68	27.28	2.67
M06-2X	-0.43	5.00	4.34	18.68	28.77	6.72
M06-L	0.00	4.24	-2.16	10.31	24.18	0.13
SSB-D	-1.03	2.30	-0.13	13.60	22.63	4.68
OPBE	2.90	5.77	7.25	16.38	24.86	10.03
TPSS	-0.14	2.62	-3.64	11.27	20.19	-0.99

^a E is in kcal·mol⁻¹. In bold: pathways with lowest reaction barriers. See Scheme 1 for definition of species. See computational details for technical settings. ^b Based on GTO basis sets, from ref 3.

relative height of the reaction barriers because they determine which one of the three pathways is most competitive (see Scheme 1). It is found that at least the atz (or df-ATZ2P for STO basis sets) basis set is needed for the correct ordering of the pathways for the F systems. This is observed for, e.g., SSB-D (see Table 4) and M06 (Supporting Information). The Cl systems are much less demanding. For example, already with the dz and DZP basis sets does one correctly obtain S_N2 substitution as the most favorable pathway (see Table 4 and Supporting Information).

3.3. Numerical Accuracy. Apart from the effect of the basis set size and the presence of diffuse basis functions, there are a number of other (technical) parameters that influence the results obtained. Some of these are due to the numerical accuracy and integration grid and are therefore more of technical nature, while others are physical. An example of the latter is the use of scalar relativistic corrections. These relativistic corrections are usually relevant mainly for elements starting from the fourth period onward and should have a smaller impact for the systems studied here. This is indeed reconfirmed by our computed ZORA relativistic corrections, which amount to only 0.03 kcal·mol⁻¹ or less (see Table 5).

The size of the integration grid is less stringent for GTO-based programs, like NWChem, in which these grids are only used for the comparatively small exchange-correlation potential and all other terms, especially the large Coulomb potential, are calculated analytically using integrals involving the GTO basis functions. However, the grid should be sufficiently large, which is not always the default option as was recently shown by Wheeler and Houk.²⁵ Within the STO-based ADF program, many more energy terms are calculated numerically. This has clear benefits in terms of, for example, scalability (parallelization). There are also drawbacks because the dependence of the results on the numerical accuracy is larger. Exactly how important the numerical accuracy is depends to a large extent on the combination of the density functional and the basis set. The

Table 4. TS Energies Relative to Reactants of All Six Reactions Computed with 13 Different Basis Sets^a

	1aTS	1bTS	1cTS	2aTS	2bTS	2cTS
CCSD(T) ³						
CBS	-1.27	5.68	2.20	18.18	30.92	5.81
M06-2X						
dz	-12.76	-17.28	-15.98	12.75	23.37	0.46
tz	-8.51	-7.10	-7.28	15.40	27.09	3.09
qz	-5.16	-1.84	-2.51	16.94	28.91	5.01
5z	-2.62	1.85	0.98	17.18	29.17	5.30
adz	-0.82	4.40	2.76	16.69	28.43	5.20
atz	-0.46	4.89	3.62	17.77	29.81	5.95
aqz	-0.26	5.18	4.07	18.16	30.27	6.47
a5z	-0.41	4.94	4.02	18.03	30.11	6.31
DZP	-8.03	-7.98	-6.59	14.17	22.68	0.92
TZ2P	-2.00	2.24	0.90	19.21	28.91	6.06
QZ4P	-1.79	3.35	2.08	18.53	28.72	6.58
df-ATZ2P	-0.39	5.49	4.41	19.87	29.70	7.31
ET-QZ+5P	-0.43	5.00	4.34	18.68	28.77	6.72
M06-L						
dz	-17.36	-20.59	-24.71	7.51	15.47	-5.61
tz	-10.68	-9.91	-14.98	10.86	19.79	-2.41
qz	-5.91	-3.27	-9.36	12.75	21.98	-0.05
5z	-1.70	2.38	-3.93	13.61	22.84	0.65
adz	-2.00	0.93	-5.47	11.92	20.32	-1.12
atz	-0.37	3.70	-2.85	13.42	22.48	0.41
aqz	0.05	4.45	-1.88	13.35	22.59	0.94
a5z	-0.39	3.93	-2.11	13.22	22.32	0.49
DZP	-11.75	-12.06	-15.75	4.29	17.87	-5.35
TZ2P	-3.11	-0.41	-5.56	9.58	23.58	0.09
QZ4P	-1.98	1.79	-4.35	10.58	24.82	0.67
df-ATZ2P	-0.35	3.71	-1.24	10.37	24.50	1.10
ET-QZ+5P	0.00	4.24	-2.16	10.31	24.18	0.13
SSB-D						
dz	-22.08	-26.70	-29.62	6.56	14.77	-2.43
tz	-15.52	-16.48	-17.97	9.87	18.70	0.24
qz	-10.61	-10.20	-12.08	11.56	20.46	2.35
5z	-5.32	-3.23	-5.42	12.47	21.41	3.46
adz	-1.31	1.87	-3.03	12.71	21.64	3.98
atz	-1.14	2.14	-0.42	13.59	22.65	4.49
aqz	-0.98	2.27	-0.22	13.90	22.93	4.98
a5z	-0.84	2.43	0.04	13.88	22.94	5.07
DZP	-15.71	-17.39	-18.28	6.72	15.39	-2.85
TZ2P	-7.62	-6.27	-7.89	12.16	20.75	3.39
QZ4P	-6.38	-4.65	-6.80	13.54	22.59	4.47
df-ATZ2P	-1.04	2.35	0.01	13.13	22.16	4.18
ET-QZ+5P	-1.03	2.30	-0.13	13.60	22.63	4.68

^a E is in kcal·mol⁻¹. In bold: pathways with lowest reaction barriers. See Scheme 1 for definition of species. See computational details for technical settings.

standard TZ2P basis set is robust and can be used for standard production work without any special precaution. One should be more cautious for, e.g., the ATZ2P basis set, which was designed for use in TD-DFT calculations to obtain reasonably accurate excitation energies with a relatively small basis set. Its direct use on ground-state properties, like the energy profiles studied here, should be carried out with care. Importantly, an additional set of diffuse functions (df) is needed in the auxiliary fit set which is employed for fitting the molecular density and to represent the Coulomb and exchange potentials (c.f., “density fitting” or “resolution of identity” scheme).²⁶ This particular combination of basis set (ATZ2P) and diffuse fit set (df) is designated df-ATZ2P.

The ADF program employs the numerical procedure of density fitting also for evaluating the HF exchange potential.²⁷ This implies that results obtained with hybrid func-

Table 5. Influence of Technical and Other Parameters on Energies Relative to Reactants of the *Anti*-E2 Reaction of $F^- + CH_3CH_2F$ and on the MUE of All Six Reactions. Other Parameters (grid size, frozen-core, diffuse functions in fit set, Dependency, ZORA)^a

	<i>E</i>					MUE ^b
	1aR	1aRC	1aTS	1aPC	1aP	
OLYP/aoz						
medium ^c	0.0	-12.20	-0.39	-4.24	12.95	2.90
fine ^d	0.0	-12.21	-0.39	-4.25	12.94	2.90
xfine ^e	0.0	-12.21	-0.39	-4.25	12.94	2.90
gfine ^f	0.0	-12.21	-0.39	-4.25	12.94	2.90
gultra ^g	0.0	-12.21	-0.39	-4.25	12.94	2.90
M06-2X/aoz						
medium ^c	0.0	-14.96	-0.06	-4.59	18.68	0.80
fine ^d	0.0	-15.06	-0.26	-4.73	18.71	0.75
xfine ^e	0.0	-15.06	-0.26	-4.74	18.70	0.75
gfine ^f	0.0	-15.06	-0.27	-4.74	18.70	0.75
gultra ^g	0.0	-15.06	-0.27	-4.74	18.70	0.75
OLYP/TZ2P(Bento et al.) ³						
frozen-core	0.0	-20.01	-7.95	-12.49	12.85	5.50 ^h
OLYP/TZ2P(2009.01)						
frozen-core	0.0	-20.01	-7.94	-12.47	12.87	5.48
all-electron	0.0	-19.90	-7.58	-12.00	13.22	5.32
OLYP/TZ2P						
Accint 6.0, ZORA	0.0	-19.91	-7.60	-12.03	13.19	5.35
Accint 8.0, ZORA	0.0	-19.91	-7.61	-12.03	13.19	5.35
Accint 8.0	0.0	-19.90	-7.58	-12.00	13.22	5.32
Accint 8.0, diff_fit	0.0	-19.89	-7.49	-11.92	13.27	5.26
Accint 8.0, dependency	0.0	-19.90	-7.58	-12.00	13.22	5.32
Accint 10.0	0.0	-19.90	-7.58	-12.00	13.22	5.32
OLYP/ATZ2P						
Accint 8.0	0.0	-11.83	-1.12	-5.33	11.91	3.29
Accint 8.0, diff_fit ⁱ	0.0	-11.83	-0.22	-4.06	13.07	3.10
Accint 8.0, dependency	0.0	-11.75	-0.09	-3.94	12.96	3.13
Accint 10.0, diff_fit ⁱ	0.0	-11.83	-0.23	-4.06	13.07	3.10
M06-2X/TZ2P, Accint 6.0, ZORA						
post-SCF (Bento et al.) ³	0.0	-15.67	1.49	-5.62	18.37	2.42
post-SCF (2009.01)	0.0	-15.67	1.49	-5.62	18.37	2.44
SCF (2009.01)	0.0	-18.28	-1.41	-6.89	19.79	2.73
M06-2X/TZ2P						
Accint 8.0, ZORA	0.0	-18.05	-2.01	-7.50	19.32	1.75
Accint 8.0	0.0	-18.03	-2.00	-7.49	19.35	1.77
Accint 8.0, diff_fit	0.0	-18.09	-2.06	-7.49	19.45	1.81
Accint 8.0, dependency	0.0	-18.03	-2.17	-7.59	19.16	1.71
Accint 10.0	0.0	-18.27	-2.01	-7.45	19.52	1.87
M06-2X/ATZ2P						
Accint 8.0	0.0	-14.82	-0.38	-4.81	18.22	0.95
Accint 8.0, diff_fit ⁱ	0.0	-14.81	-0.39	-4.80	18.31	0.96
Accint 8.0, dependency	0.0	-14.79	-0.15	-4.86	18.55	0.98
Accint 10.0, diff_fit ⁱ	0.0	-15.07	-0.40	-4.77	18.49	0.77

^a *E* and MUE are in kcal/mol. See Scheme 1 for definition of species. ^b Total MUE for F- and Cl-systems (reactions **1a–c** and **2a–c**). ^c Standard medium grid of NWChem (49 radial/434 angular shells for first row and 88/434 for second row elements). ^d Standard fine grid of NWChem (70/590 for first row and 123/770 for second row elements). ^e Standard xfine grid of NWChem (100/1202 for first row and 125/1454 for second row elements). ^f Obtained with NWChem using settings for fine grid in Gaussian09 (75/302 for all elements). ^g Obtained with NWChem using settings for fine grid in Gaussian09 (99/590 for all elements). ^h Total MUE recalculated from data given in ref 3. ⁱ The df-ATZ2P basis set, i.e., combination of AddDiffuseFit keyword with ATZ2P basis set.

tionals have an increased dependence on the size and quality of the auxiliary fit set and the associated numerical procedure. For the calculation of the HF exchange, this sensitivity becomes more important for larger basis sets.

In the case of very large regular or auxiliary basis sets, overcompleteness in certain parts of function space may occur, and the set can thus become (nearly) linearly dependent. In such cases, counter measures can be taken by removing linear combinations of basis functions that correspond to small eigenvalues of the virtual symmetrized fragment orbitals (SFO) overlap matrix; this procedure is

invoked with the ADF input keyword Dependency in combination with some threshold-type technical parameters. We stress at this point that this procedure should be used with great caution, and individual cases must be carefully analyzed. If the threshold parameters are chosen too large, then not only the overcompleteness is tackled but also “useful” portions of the basis-function space are removed which corresponds effectively to a reduction of the basis set quality and therefore the computed energy.

In our present computations, the Dependency procedure (with technical parameters: bas = 1e - 3, fit = 1e - 10,

and $\text{eig} = 1\text{e}8$) has a negligible effect on the results for pure density functionals such as OLYP (see Table 5). But the results obtained with hybrid (meta-GGA) functionals such as M06-2X appear to depend more delicately on this and other numerical procedures. For example, the product energy (**1aP**) of reaction 1a is found in a range between 19.16 and 19.52 kcal·mol⁻¹ for M06-2X/TZ2P, depending on several numerical parameters (see Table 5). Choices that affect the M06-2X/TZ2P energy are, among others, the removal of nearly linear-dependent combinations of basis functions (cf., Dependency keyword), the addition of a set of diffuse fit functions in the fit set, or the integration grid size. For OLYP/TZ2P, this variation is much smaller, i.e., between 13.19 and 13.27 kcal·mol⁻¹. Note that this variation in the obtained energies has a much smaller impact on the MUE values, which for, e.g., M06-2X/TZ2P vary from 1.71 to 1.87 kcal·mol⁻¹. Note also that varying the standard basis sets from a smaller to a larger one causes larger changes in the computed energies. For example, the M06-2X energy of product **1aP** relative to reactants varies some 0.8 kcal·mol⁻¹ between the TZ2P and the ET-QZ+5P basis sets (see Table 1).

3.4. Influence of Self-Consistency. The ADF program contains the option to calculate the energies of density functionals in a post-SCF manner, i.e., to evaluate the density functionals with densities and orbitals as obtained with another density functional. This is a very useful feature because it allows a fast determination of these energies for a large number of density functionals, which we have used very often in the past. Previous studies^{28–34} have shown that for GGA functionals the effect of self-consistency is of the order of 0.1–0.3 kcal·mol⁻¹, depending on the basis set, the numerical accuracy used, etc. For meta-GGA and hybrid (meta-GGA) functionals, the magnitude of this effect was unknown because, until the 2009.01 version, it was not possible to use a meta-GGA or hybrid meta-GGA during the SCF. In a recent study,³⁵ we already observed that for hybrid functionals, like B3LYP or X3LYP, the effect of self-consistency depends on the functional and on the system and may range from 0.3 to 2.2 kcal·mol⁻¹ for B3LYP* and X3LYP, respectively.³⁵ The same trend is observed here, where the effect of self-consistency is larger for hybrid meta-GGA functionals, like M06-2X, as already described by Zhao and Truhlar.⁷ As can be seen in Table 5, for these functionals, the effect can be even larger, i.e., up to 2.8 kcal·mol⁻¹ for any one of the stationary points (R, RC, TS, PC, and P). However, the overall effect of self-consistency is much smaller, as is shown by the MUE value for M06-2X/TZ2P (2.73 kcal·mol⁻¹) and M06-2X@OLYP/TZ2P (2.44 kcal·mol⁻¹), i.e., a difference of only 0.3 kcal·mol⁻¹ (see Table 5). A similar difference is observed (Supporting Information) for the effect of self-consistency for M06-L (0.25 kcal·mol⁻¹), but in that case, the effect for individual stationary points is much smaller (maximum 0.6 kcal·mol⁻¹, see Supporting Information).

4. Conclusions

We have investigated how the potential energy surfaces of the E2 and S_N2 reactions of X⁻ + CH₃CH₂X (X = F, Cl)

depend on various methodological choices in density functional calculations with the ADF and NWChem programs which employ Slater- and Gaussian-type basis sets, respectively. The present work complements the recent contribution by Zhao and Truhlar⁷ by providing full numerical details and recommendations for the usage of the different functionals, basis sets, and technical settings, a selection of which is compiled in this section.

In the case of both programs (and types of basis sets), the fluoride systems (X = F) appear to depend extremely sensitively on basis set size, especially the presence of diffuse functions. For ADF calculations on problems involving fluoride anions, we recommend the QZ4P basis or the even better (but also computationally more expensive) even-tempered ET-QZ+5P basis set. The TZ2P performs excellently for neutral species, but it leads to a significant exaggeration of the fluoride anion's binding capability and reactivity.

Other technical settings are also important, especially for energies computed with hybrid meta-GGA functionals, which depend more strongly, e.g., on the accuracy of the numerical integration than energies computed with GGA functionals. The mean unsigned error (MUE) computed with ADF over all six reaction profiles of this study for M06-2X, for example, decreases from 2.42 to 1.77 to 0.86 kcal/mol as we go from basis set/integration accuracy combinations TZ2P/accint = 6 (similar to ref 3) to TZ2P/accint = 8 to ET-QZ+5P/accint = 8.

Thus, we find M06-2X as the best performing functional in the ADF and NWChem programs. Note that SSB-D is the best GGA functional studied. It improves upon both M06-L and OPBE which were the best meta-GGA and GGA functional, respectively, in our previous study.³ For calculations with the ADF program, we recommend that standard GGA functionals (e.g., OLYP or SSB-D) are evaluated with a numerical integration accuracy parameter accint = 6.0, whereas (hybrid) meta-GGA and hybrid functionals should be evaluated with a higher numerical accuracy, accint = 8.0. Note also that the (hybrid) meta-GGA and hybrid functional implementation in ADF requires all-electron basis sets, i.e., the frozen-core approximation must not be used in these cases.

Acknowledgment. We thank the following organizations for financial support: The Netherlands organization for Scientific Research (NWO-CW and NWO-NCF), the Spanish Ministry of Science and Innovation (MICINN projects CTQ2008-03077/BQU and CTQ2008-06532/BQU), and the Catalan Ministry of Innovation, Universities, and Enterprise (DIUE projects 2009SGR637 and 2009SGR528). M. Solà is also indebted to the Catalan DIUE for financial support through the ICREA Academia Prize 2009.

Supporting Information Available: Full details of relative energies with Dunning's dz, tz, qz, 5z, adz, atz, aqz, and a5z GTO basis sets and with the DZ, DZP, TZ2P, df-ATZ2P, QZ4P, ET-QZ+5P, and ET-pVQZ STO basis sets; the corresponding mean absolute deviations; full details of the influence of technical parameters on the results obtained.

This information is available free of charge via the Internet at <http://pubs.acs.org/>.

References

- (1) Smith, M. B.; March, J. *Reactions, Mechanisms and Structure. March's Advanced Organic Chemistry*; Wiley: New York, 2007; pp 425–656 and 1477–1558.
- (2) Carey, F. A.; Sundberg, R. J. *Advanced Organic Chemistry, Part A: Structure and Mechanisms*; Springer: New York, 2007; pp 389–472 and 548–578.
- (3) Bento, A. P.; Solà, M.; Bickelhaupt, F. M. *J. Chem. Theory Comp.* **2008**, *4*, 929.
- (4) van Lenthe, E.; Baerends, E. J. *J. Comput. Chem.* **2003**, *24*, 1142.
- (5) Gonzales, J. M.; Cox, R. S., III; Brown, S. T.; Allen, W. D.; Schaefer, H. F., III. *J. Phys. Chem. A* **2001**, *105*, 11327.
- (6) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 1384.
- (7) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comp.* **2010**, *6*, 1104.
- (8) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (9) Swart, M.; Solà, M.; Bickelhaupt, F. M. *J. Chem. Phys.* **2009**, *131*, 094103.
- (10) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.
- (11) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403.
- (12) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.
- (13) Tao, J. M.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (14) Baerends, E. J.; Autschbach, J.; Bashford, D.; Berger, J. A.; Bérces, A.; Bickelhaupt, F. M.; Bo, C.; de Boeij, P. L.; Boerrigter, P. M.; Cavallo, L.; Chong, D. P.; Deng, L.; Dickson, R. M.; Ellis, D. E.; van Faassen, M.; Fan, L.; Fischer, T. H.; Fonseca Guerra, C.; Giammona, A.; Ghysels, A.; van Gisbergen, S. J. A.; Götz, A. W.; Groeneveld, J. A.; Gritsenko, O. V.; Grüning, M.; Harris, F. E.; van den Hoek, P.; Jacob, C. R.; Jacobsen, H.; Jensen, L.; Kadantsev, E. S.; van Kessel, G.; Klooster, R.; Kootstra, F.; Krykunov, M. V.; van Lenthe, E.; Louwen, J. N.; McCormack, D. A.; Michalak, A.; Mitoraj, M.; Neugebauer, J.; Nicu, V. P.; Noodleman, L.; Osinga, V. P.; Patchkovskii, S.; Philippsen, P. H. T.; Post, D.; Pye, C. C.; Ravenek, W.; Rodríguez, J. I.; Romaniello, P.; Ros, P.; Schipper, P. R. T.; Schreckenbach, G.; Seth, M.; Snijders, J. G.; Solà, M.; Swart, M.; Swerhone, D.; te Velde, G.; Vernooijs, P.; Versluis, L.; Visscher, L.; Visser, O.; Wang, F.; Wesolowski, T. A.; van Wezenbeek, E. M.; Wiesenecker, G.; Wolff, S. K.; Woo, T. K.; Yakovlev, A. L.; Ziegler, T. *ADF 2009. 01*; SCM: Amsterdam, The Netherlands, 2009.
- (15) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931.
- (16) Chong, D. P.; van Lenthe, E.; van Gisbergen, S. J. A.; Baerends, E. J. *J. Comput. Chem.* **2004**, *25*, 1030.
- (17) van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1993**, *99*, 4597.
- (18) Bylaska, E. J.; de Jong, W. A.; Kowalski, K.; Straatsma, T. P.; Valiev, M.; Wang, D.; Apra, E.; Windus, T. L.; Hirata, S.; Hackler, M. T.; Zhao, Y.; Fan, P.-D.; Harrison, R. J.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Auer, A. A.; Nooijen, M.; Brown, E.; Cisneros, G.; Fann, G. I.; Fruchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J. A.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Pollack, L.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers*; Pacific Northwest National Laboratory: Richland, Washington, 2008.
- (19) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- (20) Dunning, T. H., Jr.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244.
- (21) Bento, A. P.; Solà, M.; Bickelhaupt, F. M. *J. Chem. Theory Comp.* **2010**, *6*, 1445.
- (22) Botschwina, P.; Horn, M.; Seeger, S.; Oswald, R. *Ber. Bunsenges. Phys. Chem.* **1997**, *101*, 387.
- (23) Osuna, S.; Swart, M.; Baerends, E. J.; Bickelhaupt, F. M.; Solà, M. *ChemPhysChem* **2009**, *10*, 2955.
- (24) van Zeist, W.-J.; Yi, R.; Bickelhaupt, F. M. *Sci. China, Ser. B: Chem.* **2010**, *53*, 210.
- (25) Wheeler, S. E.; Houk, K. N. *J. Chem. Theory Comp.* **2010**, *6*, 395.
- (26) Baerends, E. J.; Ellis, D. E.; Ros, P. *Chem. Phys.* **1973**, *2*, 41.
- (27) Watson, M. A.; Handy, N. C.; Cohen, A. J. *J. Chem. Phys.* **2003**, *119*, 6475.
- (28) Swart, M.; Groenhof, A. R.; Ehlers, A. W.; Lammertsma, K. *J. Phys. Chem. A* **2004**, *108*, 5479.
- (29) Swart, M.; Solà, M.; Bickelhaupt, F. M. *J. Comput. Chem.* **2007**, *28*, 1551.
- (30) de Jong, G. T.; Geerke, D. P.; Diefenbach, A.; Bickelhaupt, F. M. *Chem. Phys.* **2005**, *313*, 261.
- (31) de Jong, G. T.; Geerke, D. P.; Diefenbach, A.; Solà, M.; Bickelhaupt, F. M. *J. Comput. Chem.* **2005**, *26*, 1006.
- (32) de Jong, G. T.; Bickelhaupt, F. M. *J. Phys. Chem. A* **2005**, *109*, 9685.
- (33) de Jong, G. T.; Bickelhaupt, F. M. *J. Chem. Theory Comp.* **2006**, *2*, 322.
- (34) Swart, M.; van der Wijst, T.; Fonseca Guerra, C.; Bickelhaupt, F. M. *J. Mol. Model.* **2007**, *13*, 1245.
- (35) Swart, M.; Güell, M.; Solà, M. Accurate description of spin states and its implications for catalysis. In *Quantum Biochemistry: Electronic structure and biological activity*; Matta, C. F., Ed.; Wiley-VCH: Weinheim, Germany, 2010; Vol. 2, pp 551–583.

Electronic Polarizability and the Effective Pair Potentials of Water

I. V. Leontyev and A. A. Stuchebrukhov*

Department of Chemistry, University of California Davis, One Shields Avenue,
Davis, California 95616

Received April 15, 2010

Abstract: Employing the continuum dielectric model for electronic polarizability, we have developed a new consistent procedure for parametrization of the effective nonpolarizable potential of liquid water. The model explains the striking difference between the value of water dipole moment $\mu \approx 3D$ reported in recent ab initio and experimental studies with the value $\mu^{\text{eff}} \approx 2.3D$ typically used in the empirical potentials, such as TIP3P or SPC/E. It is shown that the consistency of the parametrization scheme can be achieved if the magnitude of the effective dipole of water is understood as a scaled value $\mu^{\text{eff}} = \mu/\sqrt{\epsilon_{\text{el}}}$, where $\epsilon_{\text{el}} = 1.78$ is the electronic (high-frequency) dielectric constant of water, and a new electronic polarization energy term, missing in the previous theories, is included. The new term is evaluated by using Kirkwood–Onsager theory. The new scheme is fully consistent with experimental data on enthalpy of vaporization, density, diffusion coefficient, and static dielectric constant. The new theoretical framework provides important insights into the nature of the effective parameters, which is crucial when the computational models of liquid water are used for simulations in different environments, such as proteins, or for interaction with solutes.

1. Introduction

Simple nonpolarizable models of water, such as SPC¹ (simple point charge), TIP3P² (transferable intermolecular potential with 3 points), TIP4P² (transferable intermolecular potential with 4 points), and their modifications, have been in use in computer simulations for over forty years. More recently, a number of more sophisticated models, including flexible and polarizable ones, have also been developed;³ however, the majority of current simulations, in particular biological ones, still rely upon those simple empirical models.

An important conceptual step in the development of empirical potentials was made in 1987 by Berendsen and co-workers,⁴ who recognized that, for a correct comparison between a model and experimental vaporization enthalpy, one of the experimental data points typically used to calibrate the empirical parameters, an additional energy term, a constant, should be added in the effective potential (or equivalently experimental enthalpy should be corrected) to

reflect the energy of repolarization of water upon the transfer from the liquid to gas phase. Indeed, the experimental enthalpy of vaporization involves two different states of the molecule: the liquid, and the gas, and the dipole moment of a gas molecule (1.855D⁵) is much different from that in the liquid state.

This new energy term, which was missing in the previous models, is known as self-polarization energy and is given by

$$E_{\text{pol}} = (\mu_l - \mu_g)^2/2\alpha \quad (1.1)$$

where μ_l and μ_g are liquid- and gas-phase dipole moments of water, respectively, and α is electronic polarizability of water. Most of the conventional nonpolarizable models of liquid water have the dipole moment around 2.3D: TIP3P (2.35), SPC (2.27), TIP4P (2.18); the difference between 2.3D in liquid and 1.85D in gas amount to approximately to 1.25 kcal/mol of self-polarization energy, or about ten percent of the total vaporization enthalpy (10.5 kcal/mol⁶). The self-

* To whom correspondence should be addressed. E-mail: stuchebr@chem.ucdavis.edu.

polarization energy accounted for in a modified SPC/E model has been shown to produce significant improvements in the model.⁴

It should be noticed, however, that the dipole moment used in the above estimate is the effective dipole moment of the model, which can be different from the value of actual dipole of liquid water molecule.^{7,8} In fact, it is the value of actual dipole of liquid water that should be used in the analysis of experimental vaporization enthalpy.

The exact value of the actual dipole moment of liquid water, surprisingly, is still a matter of debate. Recent estimates deduced from experiments (these are not however direct measurements of the dipole moment) suggest a value around 2.7–2.9D.^{9,10} There are no direct measurements of the liquid dipole moment, unfortunately, and the estimates in ref 10 have significant uncertainty of $\pm 0.6D$. The ab initio calculations and first principles simulations do not agree on the actual value but mostly point to a much greater value than the effective 2.35D of TIP3P and SPC/E: 2.43D in ref 11, 2.65D in ref 12, 2.70D in ref 13, 2.95D in refs 14 and 15, and 3.09 in ref 16. We argued recently¹⁷ that on the basis of a simple Kirkwood-Onsager model the actual dipole moment should be around 3.0D; moreover our MDEC model^{17–19} predicts that the value of actual dipole of water is related to the value of effective dipole of the model as $\mu_l = \sqrt{\epsilon_{el}} \cdot \mu^{\text{eff}}$ where $\epsilon_{el} = 1.78$ is the electronic (high-frequency) dielectric constant of water; thus if $\mu^{\text{eff}} = 2.3D$, the value of actual dipole of liquid water should be about 3.0D; these estimates are in agreement with first principles simulations of liquid water.^{14,16}

If indeed the actual dipole moment is much greater than 2.3D, then the self-polarization term is much greater than the previously estimated 1.25 kcal/mol. Assuming for the actual dipole of liquid water a value 3.0D, the self-polarization term amounts in fact to 6–7 kcal/mol, which is more than half of the vaporization enthalpy 10.5 kcal/mol, dramatically changing the vaporization energy data relevant for model building. Therefore the high value of the dipole moment of liquid water appears to be in striking contradiction with the existing parametrization schemes of the effective potentials. [Needless to say that such problems would not appear in the fully polarizable models and only specific for nonpolarizable effective models discussed in this paper.]

In this paper, we show that the disagreement between the existing models and the actual vaporization energy data is due to an important missing term in theory, which is related to electronic polarization of the medium. The missing term is the energy of solvation of water molecule in the electronic continuum representing polarization of bulk water, which needs to be accounted for in an accurate analysis of the vaporization process. This energy also amounts to about 6–7 kcal/mol, of opposite sign to the increase in the self-polarization term, when a water molecule is transferred from the liquid state (which involves polarization of the electronic continuum) to a gas state (no electronic polarization at all).

On the basis of this observation and assuming that neither the exact value of the dipole moment of liquid water nor its polarizability are known, we developed three closely related new self-consistent models that fit accurately experimental

data for density, diffusion coefficient, radial distribution functions, as well as the static dielectric constant and vaporization energy. It is shown that the consistency of the parametrization scheme can be achieved if magnitude of the effective dipole of water is understood as a scaled value $\mu^{\text{eff}} = \mu_l/\sqrt{\epsilon_{el}}$, and the new electronic polarization energy term is included. The new term is evaluated by using Kirkwood–Onsager theory.

The effective parameters (charges and Lennard-Jones parameters) of the models derived from the new principles are slightly different from those of the known models; more importantly, however, the new theory provides important insights into their actual nature, which is of significant interest for applications of the effective water models in simulations of biological environments, or for simulation of solute–solvent interactions. We show that self-consistency of the whole theory, and reconciliation of experimental evaporation data with other data (diffusion coefficient, radial distributions, density) is only possible if the value of the actual dipole of liquid water is around 3.0 (± 0.1) D.

2. Theory

2.1. Electronic Screening. MDEC Model. There are two effects caused by the electronic polarizability of water molecules important for the following discussion: the enhancement of the dipole in the liquid state, and the screening of electrostatic interactions. As commonly described in the literature,³ the dipole enhancement is reflected in the partial atomic charges q_i^{eff} of nonpolarizable models, so that the effective dipole in the liquid state is greater than that in the gas phase 1.855D: for example, TIP3P, SPC and TIP4P models have an effective dipole moments of 2.35D, 2.27D, and 2.18D, respectively; there is less clarity in the literature on the screening effect, however.

A simple way to account for the electronic screening effect is to consider all nuclear charges q_i as moving in polarizable electronic continuum with dielectric constant ϵ_{el} . In this case, all electrostatic interactions are scaled by a factor $1/\epsilon_{el}$. The empirical parameter ϵ_{el} is known from experiment as the high-frequency dielectric permittivity ($\epsilon_{el} = n^2$, where n is a refraction index of the medium). Such a model, which combines a nonpolarizable (fixed-charge) force field for nuclear dynamics (MD) and a phenomenological electronic continuum (EC) for the electronic effects, is referred to as MDEC.¹⁹

Since electrostatic interactions are quadratic in charges, the effects of electronic dielectric screening can be taken into account implicitly by using scaled partial charges, $q_i^{\text{eff}} = q_i/\sqrt{\epsilon_{el}}$; in this case the Coulomb interactions automatically have the correct form $q_i^{\text{eff}} q_j^{\text{eff}}/r_{ij} = q_i q_j/\epsilon_{el} r_{ij}$ without explicitly introducing the factor $1/\epsilon_{el}$.

The scaling of charges has a direct consequence for the effective dipole moment of liquid water. As we argued earlier,^{17–19} the effective dipole of liquid water should be also understood as a scaled value, $\mu^{\text{eff}} = \mu_l/\sqrt{\epsilon_{el}}$. Thus, if the effective dipole value of a model is $\mu^{\text{eff}} \approx 2.3D$, as in TIP3P or SPC/E, the actual dipole is $\mu_l = \sqrt{\epsilon_{el}} \cdot \mu^{\text{eff}}$; for water $n = 1.33336^6$ results in $\epsilon_{el} = n^2 = 1.78$, $\mu_l \approx 3.0D$, which

is in agreement with recent first principles simulations of liquid water^{14,16} and within the uncertainty of experimental data.^{9,10} As we already pointed out in the Introduction, the high value of the actual dipole moment of liquid water gives rise to a significant problem for the existing parametrization schemes in interpretation of the vaporization experimental data, which is considered next.

2.2. Vaporization Energy and Self-Polarization Term.

One of the main benchmark tests characterizing the quality of the effective potential of water $U^{\text{eff}}(r)$ is a comparison of the average energy of molecular configurations $\langle U^{\text{eff}} \rangle_{\text{bulk}}$ with the heat of vaporization given by the relation:³

$$\Delta H_{\text{vap}} = -\langle U^{\text{eff}}(r) \rangle_{\text{bulk}} + E^{\text{intra}} + RT \quad (2.1)$$

Here, the total potential energy difference between phases $(\langle U \rangle_{\text{gas}} - \langle U \rangle_{\text{bulk}})$ was partitioned into intermolecular $(\langle U^{\text{inter}} \rangle_{\text{gas}} - \langle U^{\text{inter}} \rangle_{\text{bulk}}) = -\langle U^{\text{eff}}(r) \rangle_{\text{bulk}}$ and intramolecular $(\langle U^{\text{intra}} \rangle_{\text{gas}} - \langle U^{\text{intra}} \rangle_{\text{bulk}}) = E^{\text{intra}}$ components; the first term is directly described by the effective nonpolarizable potential, while the second term E^{intra} is not explicitly present in the nonpolarizable model. As pointed out by Berendsen and co-workers,⁴ for a correct comparison with the experiment⁶ ($\Delta U_{\text{vap}} \equiv \Delta H_{\text{vap}} - RT = 9.92$ kcal/mol) both terms, the simulated energy $\langle U^{\text{eff}}(r) \rangle_{\text{bulk}}$ and E^{intra} , should be taken into account.

The physical meaning of the correction E^{intra} can be understood by partitioning the vaporization process onto two steps illustrated in Figure 1a: first, a polarized water molecule (i.e., polarized as in liquid state) is moved from the bulk to gas phase (keeping its dipole frozen), then the molecule is relaxed to its equilibrium configuration in the gas phase. The work required for the first step is given by the effective potential energy $-\langle U^{\text{eff}}(r) \rangle_{\text{bulk}}$, while the energy of the second step is not explicitly present in the nonpolarizable models and therefore should be added separately. Berendsen et al.⁴ argued that the main contribution to E^{intra} is because of the difference of the dipole moment of a water molecule in gas and in liquid state and the repolarization self-energy correction, $E^{\text{intra}} = -E_{\text{pol}}$, where E_{pol} is given by eq 1.1.

In the conventional picture, following Berendsen et al., the molecular electric moment μ in the liquid state is identified with the effective dipole μ^{eff} of the model; for example for SPC/E $\mu^{\text{eff}} = 2.35\text{D}$, and the correction E_{pol} is only 1.25 kcal/mol. However, if one assumes that the actual dipole of liquid water is not 2.35D, but rather closer to 3.0D, as recent data indicate, the correction is much greater. For example, using MDEC model, the SPC/E actual dipole moment in the bulk is $\mu_l = \sqrt{\epsilon_{\text{cl}}} \cdot \mu^{\text{eff}} = 3.14\text{D}$, which produces the correction $E_{\text{pol}} \approx 8$ kcal/mol, that is, more than half of the total vaporization energy ΔU_{vap} .

It is obvious that if the parameter calibration is done in the standard way, that is, as in ref 4 or 8, the resulting model will be completely at odds with the experimental data for density, structural properties, diffusion coefficients etc. To reconcile the high value of the actual dipole moment of liquid water, and the vaporization data, one needs to consider an additional polarization term.

2.3. Missing Term. According to MDEC model, the actual charges of the effective model should be understood

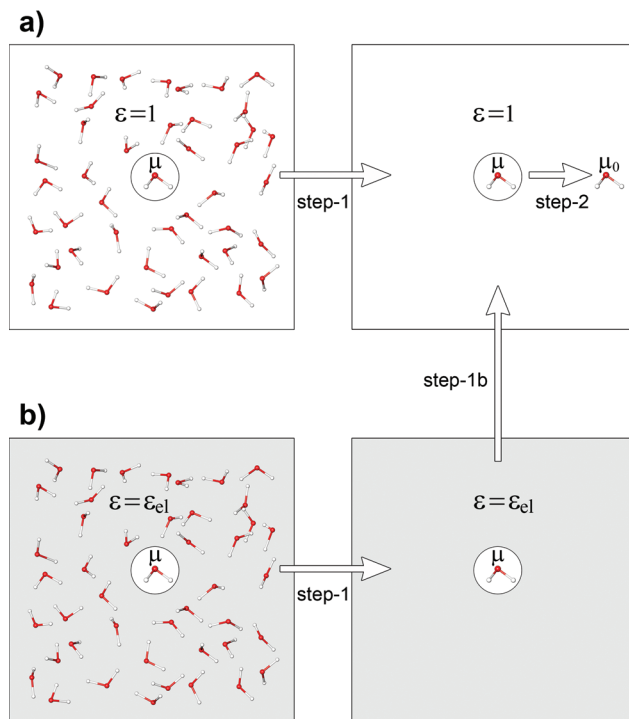


Figure 1. Partitioning the vaporization process onto steps. (a) Standard nonpolarizable model of the medium: on the step 1, the polarized water molecule is moved from the bulk to infinity (keeping its dipole μ frozen), and on the step 2, the molecule relaxes and equilibrates its polarization to the gas-phase value in the absence of polar environment μ_0 . (b) MDEC model, charges are moving in the electronic continuum: on the step 1, the polarized water molecule is moved from the bulk to infinity (keeping its dipole μ frozen and remaining in the electronic continuum); on the step 1b, the molecule is moved from the electronic continuum to vacuum (still keeping its dipole μ frozen), and on the step 2, the molecule relaxes and equilibrates its polarization to the gas-phase value in the absence of polar environment μ_0 .

as immersed in the electronic dielectric continuum; the vaporization process then is represented by the steps shown in Figure 1b: (1) a polarized water molecule is moved from the bulk to infinity (i.e., is separated from other water molecules), keeping its dipole frozen and remaining in the electronic continuum; (1b) the molecule is moved from the electronic continuum to vacuum (keeping its dipole frozen); and finally (2) the molecule is relaxed to its equilibrium gas-phase state, in the absence of polar environment. As in SPC/E model,⁴ the energy associated with step (1) is calculated by the nonpolarizable effective potential as $-\langle U^{\text{eff}}(r) \rangle_{\text{bulk}}$, and the energy of step (2) is given by the negative self-polarization energy, eq 1.1, while the energy of step 1b is a new energy term, $-E_{\text{sol}}$. This new term is the solvation energy of a water molecule, polarized as it is in the liquid state, in the electronic continuum. Therefore, a new expression for E^{intra} in the relation of the vaporization enthalpy and model parameters, eq 2.1, is

$$E^{\text{intra}} = -E_{\text{pol}} - E_{\text{sol}} \quad (2.2)$$

2.4. Kirkwood–Onsager Model for Electronic Solvation Energy. To estimate the missing term, E_{sol} , that is, interaction of a single molecule with the polarizable environ-

ment, we employ the Kirkwood–Onsager model,^{20,21} in which the medium is represented by a continuum dielectric of ϵ and the solvent molecule is modeled by a point polarizable dipole, placed in a spherical cavity of radius R ; the permanent dipole is μ_0 and the polarizability is α .

The reaction field (rf) due to polarization of the environment induces additional polarization of the dipole. Suppose the resulting dipole is μ , the polarization energy of the medium then is²⁰

$$\Delta F_{\text{rf}}(\mu) = -\frac{2(\epsilon - 1)\mu^2}{(2\epsilon + 1)2R^3} \quad (2.3)$$

Combining this term with the self-polarization energy we obtain the total electrostatic (solvation) free energy ΔF_{elec} of the dipole in the polarizable environment

$$\Delta F_{\text{elec}}(\mu) = -\frac{2(\epsilon - 1)\mu^2}{(2\epsilon + 1)2R^3} + \frac{(\mu - \mu_0)^2}{2\alpha} \quad (2.4)$$

The correction term to vaporization energy (contributions from steps 1b and 2 of the Figure 1b) E^{intra} is given by eq 2.4 with $\epsilon = \epsilon_{\text{el}}$, μ_0 being the gas dipole value, and μ being the actual dipole moment of liquid water molecule. It is seen that the two terms in the above expression have different signs.

One can also use the above expression for an estimate of the liquid molecule dipole itself. In this case $\epsilon = \epsilon_0$, where ϵ_0 is the static dielectric constant of the solvent, and μ is unknown. At equilibrium, the dipole of the solvent molecule in the bulk satisfies the condition $(\delta\Delta F_{\text{elec}}(\mu))/(\delta\mu) = 0$. This condition results in

$$\mu_l = \frac{\mu_0}{1 - \frac{2(\epsilon_0 - 1)\alpha}{(2\epsilon_0 + 1)R^3}} \quad (2.5)$$

If one assumes the above value of the equilibrium dipole of liquid water, the total solvation free energy is

$$\Delta F_{\text{elec}}(\mu_l) = -\frac{1}{\frac{(2\epsilon_0 + 1)R^3}{2(\epsilon_0 - 1)\alpha} - 1} \frac{\mu_0^2}{2\alpha} \quad (2.6)$$

Equations 2.3–2.6 represent a complete set of relations needed to describe polarization effects in a transition of a molecule from gas-phase to dielectric medium of ϵ . Equation 2.5, derived by Kirkwood²⁰ in 1939, expresses the magnitude of the molecular dipole enhancement in a polarizable environment, and eq 2.3 with $\epsilon = \epsilon_{\text{el}} = 1.78$ expresses the energy correction E_{sol} of step 1b (see Figure 1b), while eq 2.4 gives the total correction E^{intra} to the vaporization energy.

The theory would be complete if one knew the value of the radius R of the molecular cavity. This is obviously a phenomenological parameter, which needs to be fixed in comparison with experimental data, or derived from a suitable theoretical model. In addition to R , we recall that the polarizability α can in principle be different from the gas polarizability of water. [Recently, molecular polarizability of liquid water has been examined in ab initio studies; the

calculations^{12,22} show a reduction of polarizability in liquid state in the range of 10–15%; however, in ref 13 a slight increase was observed.] In this case α can also be considered as an unknown phenomenological parameter to be fixed together with R .

In the next section, we describe three related computational models that are based on the above relations, and in which the phenomenological parameters R and α are self-consistently found together with other parameters of the effective potential of water (charges and Lennard-Jones parameters) using the complete experimental data set: density, radial distributions, diffusion coefficient, static dielectric constant, and vaporization energy.

We close this section with an estimate of the polarization term using a simple model for R , and assuming a gas value for polarizability α .

One reasonable estimate of the radius of molecular sphere R is to assume that $2R$ is the average distance between the liquid water molecules, $a = 2R$. The distance between the molecules is related to their number density, N_α , as $N_\alpha = a^{-3} = 1/8R^3$. Employing now the Clausius-Massotti relation between N_α and ϵ_{el} :

$$\frac{4\pi}{3}\alpha N_\alpha = \frac{\epsilon_{\text{el}} - 1}{\epsilon_{\text{el}} + 2} \quad (2.7)$$

the radius R is expressed as

$$R^3 = \frac{\pi(\epsilon_{\text{el}} + 2)}{6(\epsilon_{\text{el}} - 1)}\alpha \quad (2.8)$$

and the solvation free energy of the dipole μ in electronic continuum, eq 2.3, that is, the new energy correction term in eq 2.2, becomes:

$$E_{\text{sol}}(\mu) = -\frac{6}{\pi(2\epsilon_{\text{el}} + 1)(\epsilon_{\text{el}} + 2)} \frac{\mu^2}{\alpha} \quad (2.9)$$

If we take now the dipole value corresponding to SPC/E model $\mu = \sqrt{\epsilon_{\text{el}}}\mu^{\text{eff}} = 3.14$ D, and $\alpha = 1.47$ Å³, the electronic solvation energy E_{sol} is about -6.5 kcal/mol, which almost completely cancels the self-polarization term E_{pol} of 8.0 kcal/mol in eq 2.2, and the total correction E^{intra} in eq 2.2 is only -1.5 kcal/mol. Thus, the electronic term E_{sol} , which was missing in the standard nonpolarizable models, resolves the problem of extremely large correction to the vaporization energy.

With the radius R given by eq 2.8, the equilibrium dipole moment in the liquid phase is expressed by eq 2.5 as

$$\mu_l = \frac{\mu_0}{1 - \frac{6(\epsilon_{\text{el}} - 1)2(\epsilon_0 - 1)}{\pi(\epsilon_{\text{el}} + 2)(2\epsilon_0 + 1)}} \quad (2.10)$$

which gives the value 3.0 D for the bulk water ($\epsilon_{\text{el}} = 1.78$, $\epsilon_0 = 78$, $\mu_0 = 1.85$ D). This value is in good agreement with the recent experimental estimations; however, it is slightly different from the empirical SPC/E MDEC value, $2.35\text{D}\sqrt{1.78} = 3.14\text{D}$. This slight inconsistency will be corrected in our computational model I, in the next section.

The approximations involved in the above model are evident: for example, the charge distribution in the actual water molecule is not the same as that of a point dipole located at the center of the van-der-Waals sphere; also, the polarization of the electronic environment might not be perfectly described by the continuum model. Moreover, the parameter R , and to some extent α , are not well-defined for the bulk water. The cavity radius R is in fact completely phenomenological. Yet, despite a number of approximations involved in the Kirkwood–Onsager model, this theory captures qualitatively all the effects of molecular polarization, while quantitative accuracy can be achieved by an appropriate adjustment of the model parameters by fitting experimental data.

In the following section, a computational procedure based on relations 2.3–2.10 is described to develop a new microscopic effective potential of liquid water. Three possible strategies to achieve complete self-consistency of a model are considered.

3. Computational Models for the Effective Potentials of Water

Following the above theory, we will now describe the corresponding new parametrization of pairwise effective potential for water, which accounts for the electronic solvation and polarization effects. Three different schemes: model I, II, and III specified below, are used for the evaluation of the new E^{intra} term; the models are summarized by eqs 3.2, 3.3, and 3.5 below. The conventional 3-site form for the water intermolecular potential is considered

$$u_{a,b} = \sum_{ij} \frac{q_i^{\text{eff}} q_j^{\text{eff}}}{r_{ij}} - 4\epsilon \left[\left(\frac{\sigma}{r_{\text{OO}}} \right)^6 - \left(\frac{\sigma}{r_{\text{OO}}} \right)^{12} \right] \quad (3.1)$$

here i and j indices run over all charge sites of the molecule a and b , respectively, while, the last term stands for the Lennard-Jones interaction between the two oxygen (O) atoms separated by the distance r_{OO} . The geometry of rigid SPC water model is adopted, namely, $r(\text{OH}) = 1.0 \text{ \AA}$ and an H–O–H angle of 109.47° .

The adjustable parameters are: partial atomic charges (which for the fixed SPC geometry of water are uniquely defined by a value of the effective dipole moment μ^{eff} , which is in fact varied), and the Lennard-Jones (LJ) parameters ϵ and σ ; the experimental data sets are: density, self-diffusion coefficient, and vaporization energy ($T = 298 \text{ K}$ and $P = 1 \text{ atm}$).

Model I. In this model, we assume the relation ($N_\alpha = a^{-3} = 1/8R^3$) resulting in eqs 2.7–2.9. The corresponding expression for the correction term E^{intra} in this model is

$$E^{\text{intra}}(\mu) = -\frac{(\mu - \mu_0)^2}{2\alpha} + \frac{6}{\pi} \frac{(\epsilon_{\text{el}} - 1)^2}{(2\epsilon_{\text{el}} + 1)(\epsilon_{\text{el}} + 2)} \frac{\mu^2}{\alpha} \quad (3.2)$$

Using the above correction term, LJ parameters of the effective potential, and the value of actual dipole of liquid water μ , related to the effective model dipole as $\mu = \sqrt{\epsilon_{\text{el}}} \cdot \mu^{\text{eff}}$, are adjusted to reproduce in MD simulations the

liquid water density, self-diffusion coefficient, and vaporization energy. The total number of independently variable parameters in this scheme is 3 (μ^{eff} , ϵ , and σ), while number of experimental properties to be reproduced is also 3 (density, self-diffusion coefficient and vaporization energy), which provides a nonredundant system of equations to uniquely define these parameters.

Model II. In the above scheme, the dipole moment of liquid water is a free adjustable parameter, and only the electronic solvation is calculated by Kirkwood–Onsager model. To follow the logic of Kirkwood–Onsager model more deeply, we can consider that the actual dipole of liquid water is defined by eq 2.5, which requires the knowledge of R , which itself is unknown. In this scheme, we use the relation 2.5 for definition of R , still considering the value of water dipole μ as a free parameter. Then the solvation energy E_{sol} is expressed by eq 2.3, with parameter R taken from eq 2.5 (instead of eq 2.8 for model I). The resulting expressions for R and E^{intra} of model II are

$$R^3 = \frac{2(\epsilon_0 - 1)}{(2\epsilon_0 + 1)} \frac{\mu}{(\mu - \mu_0)} \alpha$$

$$E^{\text{intra}}(\mu) = -\frac{(\mu - \mu_0)^2}{2\alpha} + \frac{(\epsilon_{\text{el}} - 1)(2\epsilon_0 + 1)}{(2\epsilon_{\text{el}} + 1)(\epsilon_0 - 1)} \left(1 - \frac{\mu_0}{\mu} \right) \frac{\mu^2}{2\alpha} \quad (3.3)$$

where value of the polarizability α is still identified with its gas phase value ($\alpha = 1.47^{23} \text{ \AA}^3$). The self-consistent values of R and the correction E^{intra} are obtained by adjusting $\mu = \sqrt{\epsilon_{\text{el}}} \cdot \mu^{\text{eff}}$ and LJ parameters of the effective potential to reproduce the liquid water density, self-diffusion coefficient, and vaporization energy, as in the first model.

In this scheme, there are same 3 variable parameters (μ^{eff} , ϵ , and σ); the number of experimental data to be reproduced is also 3, which provides a nonredundant system of equations to uniquely define the unknown parameters.

Model III. The use of the Kirkwood–Onsager model can be extended even further to allow for a possibility that the molecular polarizability α of bulk water is not equal to its gas-phase value ($\alpha = 1.47 \text{ \AA}^3$), as is assumed in the previous two models. Here the polarizability α is treated as an adjustable parameter. This is achieved as follows.

Here, we still keep eq 3.3 for E^{intra} , but consider α in that expression as unknown. We then use eq 2.6 for total electrostatic solvation energy and assume it to be equal the corresponding energy of the microscopic effective potential: $\Delta F_{\text{elec}}(\mu_l) = \langle U_{\text{elec}}^{\text{eff}} \rangle_{\text{bulk}} - E^{\text{intra}}$, where $U_{\text{elec}}^{\text{eff}}$ is the electrostatic part of the effective potential: $U^{\text{eff}} = U_{\text{elec}}^{\text{eff}} + U_{\text{vdW}}^{\text{eff}}$. In this case, the microscopic vaporization energy $\Delta U_{\text{vap}} \equiv \Delta H_{\text{vap}} - RT = 9.92 \text{ kcal/mol}$ is expressed (employing eqs 2.1 and 2.6 as

$$\Delta U_{\text{vap}} = -\langle U_{\text{vdW}}^{\text{eff}}(r) \rangle_{\text{bulk}} + \frac{1}{(2\epsilon_0 + 1)R^3} \frac{\mu_0^2}{2\alpha} - 1 \quad (3.4)$$

This relation is then used for a consistent definition of α (assuming R defined by eq 2.5). The self-consistent set of equations for α , R , and E^{intra} in this model are

$$\alpha = \frac{(\mu - \mu_0)\mu_0}{2} \frac{1}{(\Delta U_{\text{vap}} + \langle U_{\text{vdW}}^{\text{eff}} \rangle)}$$

$$R^3 = \frac{2(\varepsilon_0 - 1)\mu_0\mu}{(2\varepsilon_0 + 1)} \frac{1}{2(\Delta U_{\text{vap}} + \langle U_{\text{vdW}}^{\text{eff}} \rangle)}$$

$$E^{\text{intra}}(\mu) = -(\Delta U_{\text{vap}} + \langle U_{\text{vdW}}^{\text{eff}} \rangle) \frac{\mu}{\mu_0} \left(1 - \frac{\mu_0}{\mu} - \frac{(\varepsilon_{\text{cl}} - 1)(2\varepsilon_0 + 1)}{(2\varepsilon_{\text{cl}} + 1)(\varepsilon_0 - 1)} \right) \quad (3.5)$$

The consistent values of the parameters α , R , and correction E^{intra} are obtained in this parametrization procedure adjusting $\mu = \sqrt{\varepsilon_{\text{cl}}}\mu^{\text{eff}}$ and LJ parameters of the effective potential to reproduce the liquid water density, self-diffusion coefficient, and vaporization energy, as in the previous two models.

The number of independently variable parameters in this scheme is still 3 (μ^{eff} , ε , and σ) the same as for models I and II because the value of α depends on these 3 parameters as given by eq 3.5, where the value of $\langle U_{\text{vdW}}^{\text{eff}} \rangle_{\text{bulk}}$ extracted from MD is also function of the effective potential parameters. The number of experimental properties to be reproduced is also 3, this provides a nonredundant system of equations to uniquely define all 3 adjustable parameters of the effective potential.

4. Results

The parameters of the effective potentials obtained by using models I, II, III are summarized in Table 1; for reference, the parameters of TIP4P² and SPC/E⁴ models are also shown. The optimized parameters of the models, cavity radius R and polarizability α , obtained in the fitting procedures, are given in the Table 2. The quality of fitting of the experimental

data for liquid state properties of water by different effective potentials is shown in Table 2. The comparison of radial distribution functions for different models with experiment is shown in Figure 2. Details of the simulations are given in the Appendix.

It is seen that for all models the electronic solvation term E_{sol} essentially cancels the significant (7–8 kcal/mol) self-polarization term E_{pol} , so that the total energy correction E^{intra} becomes approximately 1 kcal/mol. It is remarkable that in the new scheme, the overall vaporization energy correction E^{intra} is essentially the same as in the original SPC/E model because of a cancellation of the corrected self-polarization term, E_{pol} , and the new electronic solvation energy term E_{sol} . Indeed, for example, eq 3.5 of model III yields a correction 1.21 kcal/mol for SPC/E potential, while in the original SPC/E model⁴ the correction is 1.25 kcal/mol. The almost complete cancellation of the two terms perhaps is not fortuitous, since both the electronic solvation term E_{sol} and the polarization term E_{pol} are of the same origin, namely due to electronic polarizability of water molecules.

As a result of the effective cancellations of the two correction terms E_{pol} and E_{sol} the obtained parameters of the effective potentials are not much different from those of the SPC/E or TIP4P, which means that these models are consistent with the concepts of electronic continuum model MDEC,¹⁷ on which the present theory is entirely based. In other words, SPC/E and similar effective potentials, which were developed purely empirically and have no reference to electronic polarizability, in fact are

Table 1. Parameters for Water Model Potentials

parameter	experiment ^a	TIP4P	SPC/E	model I	model II	model III
$r(\text{OH})$, Å	0.9572 ²⁴	0.9572	1.0000	1.0000	1.0000	1.0000
$\angle\text{HOH}$, deg	104.52 ²⁴	104.52	109.47	109.47	109.47	109.47
$r(\text{OM})$, Å		0.15				
$q(\text{O})$, e		0	-0.8476	-0.7894	-0.8384	-0.8456
$q(\text{H})$, e		0.52	0.4238	0.3947	0.4192	0.4228
$q(\text{M})$, e		-1.04				
μ^{eff} , D	1.855 ⁵	2.177	2.351	2.189	2.325	2.345
ε , kcal/mol		0.155	0.155	0.0818	0.146	0.155
σ , Å		3.1536	3.1656	3.2227	3.1689	3.1659

^a The experimental data^{5,24} correspond to the gas phase.

Table 2. Liquid State Properties of Different Water Models at 1 atm and 298.15 K

property	experiment	TIP4P ^e	SPC/E ^e	model I ^a	model II ^b	model III ^c
ρ , g/cm ³	0.997 ⁶	0.995	1.000	0.998	0.999	0.999
D_{self} , 10 ⁻⁵ cm ² /s	2.3 ²⁵	3.2	2.4	2.6	2.4	2.4
U_{vap} , kcal/mol	9.92 ^d	9.70	9.99	9.94	9.95	9.93
$\langle U^{\text{eff}} \rangle$, kcal/mol		-9.92	-11.20	-9.87	-10.96	-11.11
E_{pol} , kcal/mol		6.60	8.32	5.56	7.62	8.25
E_{sol} , kcal/mol		-6.38 ^c	-7.11 ^c	-5.62	-6.61	-7.07
$\langle U_{\text{vdW}}^{\text{eff}} \rangle$, kcal/mol		1.77	2.13	2.56	2.15	2.10
μ , D	2.9 ¹⁰	2.904	3.137	2.920	3.102	3.129
R , Å		1.482 ^c	1.505 ^c	1.551	1.531	1.505
α , Å ³	1.47 ²³	1.20 ^c	1.42 ^c	1.47 ²³	1.47 ²³	1.42
ε (dielectric constant)	78.4	51	71	60	71	71

^a The values of R and E_{sol} are calculated according to eqs 2.8 and 2.9, respectively. ^b The values of R and E_{sol} are calculated according to eq 3.3. ^c The values of R , α , and E_{sol} are calculated according to eq 3.5. ^d The value ΔH_{vap} from ref 6 corrected by RT term. ^e TIP4P and SPC/E liquid state properties were recalculated according to the modern simulation protocol, for example, using PME treatment of electrostatic interactions and long-range dispersion correction for pressure and energy (see Appendix). Accordingly, some data shown for TIP4P and SPC/E models are slightly different from those reported in the original papers.^{2,4}

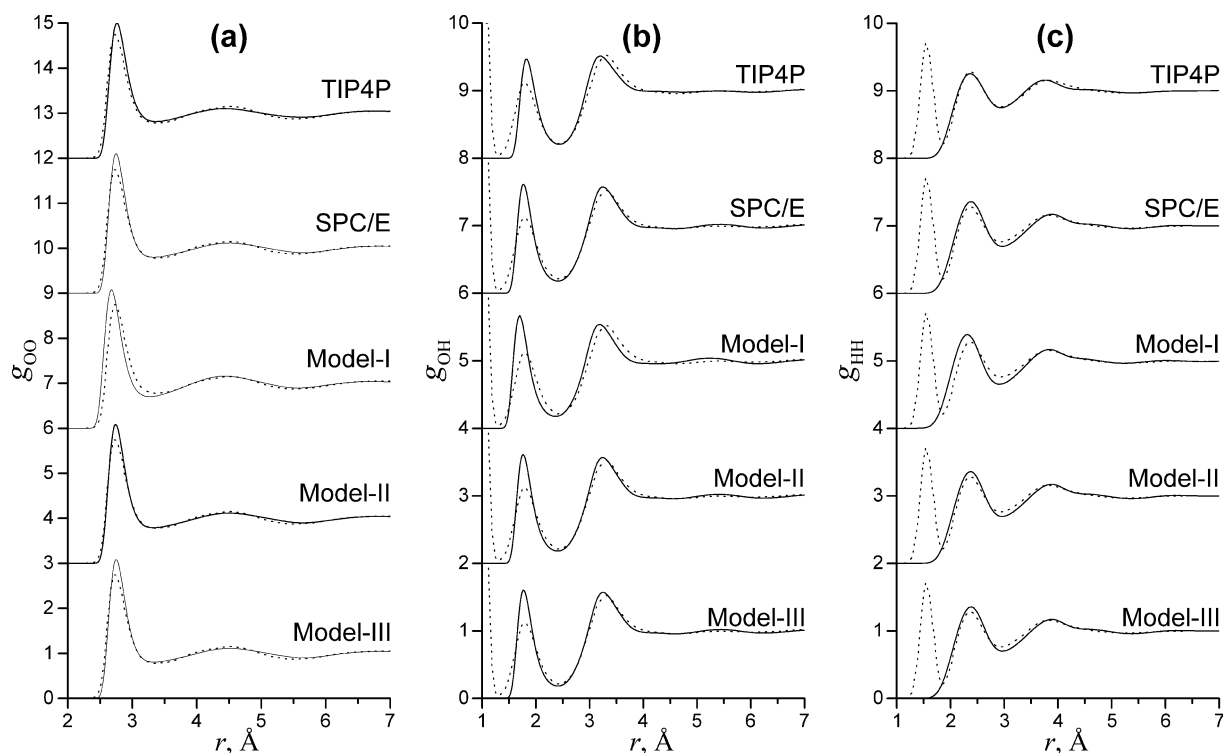


Figure 2. Radial distribution functions of water: (a) oxygen–oxygen; (b) oxygen–hydrogen and (c) hydrogen–hydrogen. The simulated curves (solid lines, this work) are compared with the re-evaluated²⁶ Soper neutron diffraction data²⁷ (dotted lines).

MDEC-type models, in the sense that they use the effective, scaled charges.

This equivalence to MDEC model sheds new light on the physical meaning of the effective charges of standard nonpolarizable models such as SPC/E, providing their interpretation as *scaled* actual charges. It also helps to reconcile those models with the data on the larger than previously thought dipole moment of liquid water. Namely, the apparent inconsistency of the value of water dipole $\sim 2.3D$ of TIP3P, SPC or SPC/E, and the value $\sim 3D$ observed in the recent ab initio simulations^{14,16} and experimental study,¹⁰ is eliminated once the dipole, that is, charges, of effective potential are considered as the real charges scaled by the factor of $1/\sqrt{\epsilon_{el}}$ (0.75 for $\epsilon_{el} = 1.78$).

5. Discussion

The majority of current simulations, in particular biological ones, still rely upon simple nonpolarizable models of water, such as TIP3P, SPC/E, TIP4P and their modifications. Since these models describe the properties of pure liquid water reasonably well, it is often tacitly assumed that they will work equally well for solutions or in different environments, such as proteins, channels, or membranes. In fact, the transferability of such models often is an open question, and their use in conditions different from those in a pure liquid state, for which the empirical parameters of the models were optimized, requires great care. In general, a clear understanding of the physical nature of the empirical parameters involved is a key issue. This paper provides important insights into the nature of the effective parameters of liquid water.

The much discussed, but still unsettled issue is that of water electronic polarizability and its role in the effective

potentials of water. The empirical models are typically developed so as to match the experimental data for density, radial distribution functions, diffusion coefficient, static dielectric constant, and the enthalpy of vaporization. It should be noticed that the first three types of data require only a correct description of the nuclear dynamics of water in simulations, thus an empirical choice of the appropriate effective internuclear potentials of increasing complexity (i.e., using more adjustable parameters) can, in principle, produce a model ideally matching the experimental data. The last two data types are more subtle, however, since the experimental observables involve the behavior of the electronic degrees of freedom, i.e. polarization, which is not explicitly present in the model.

In case of the dielectric constant, the electronic polarizability is directly involved. Typically it is expected that a good nonpolarizable model should reproduce the total dielectric response of the bulk water, despite the fact that only nuclear motions are involved in the model. We have recently argued¹⁹ that in fact a consistent model can (and should) reproduce only the nuclear part of dielectric response of the real system, $\epsilon_{MD} = \epsilon_0/\epsilon_{el}$, where ϵ_{MD} is the dielectric constant of the effective model, $\epsilon_0 = 78.4$ is the actual static dielectric constant of liquid water, and ϵ_{el} is the electronic part of the dielectric constant. In that case the effective charges of the model should be understood as real charges “scaled” by the factor $1/\sqrt{\epsilon_{el}}$. The results of this paper fully support this picture.

A related issue concerns the vaporization data matching by an empirical model. Namely, the actual experimental enthalpy of vaporization involves two different states of the molecule: the liquid, and the gas. The dipole moment of a

gas molecule is much different from that in the liquid state, because in the liquid state water is additionally polarized by the reaction field of the environment of other water molecules. The problem is that the effective model deals only with the liquid state molecule, thus the direct comparison with experiment is not possible. Moreover, the exact value of the dipole moment of liquid water molecule is still unknown.

In this and the previous papers of this series, we have shown that on the basis of a simple Kirkwood–Onsager model, the actual dipole moment should be around 3.0D; MDEC model predicts that the value of actual dipole of water is related to the value of effective dipole of the model as $\mu_l = \mu^{\text{eff}}\sqrt{\epsilon_{\text{el}}}$; thus if $\mu^{\text{eff}} = 2.3\text{D}$, the actual dipole of liquid water should be about 3.0D; these estimates are in agreement with first principles simulations of liquid water.^{14,16}

The self-polarization energy correction term of Berendsen and co-workers accounts for the energy of repolarization of water upon transfer from liquid- to gas-phase. If one identifies the effective dipole moment of the model, which is about 2.3 D or less, with the value of actual dipole of liquid model the correction is only about 1.25 kcal/mol, or about 10% of the total vaporization enthalpy. However, the use of the actual dipole moment results in much higher values (6–8 kcal/mol) and gives rise to an apparent problem for theory. We have shown that the problem can be resolved by including the effects of electronic polarization, which are only implicitly present in the effective models. The new correction term introduced in this paper, E_{sol} , is the electronic solvation energy which, in addition to the Berendsen term,⁴ E_{pol} , is also an inherent part of the vaporization energy. The new term is the energy of repolarization of the electronic continuum of liquid water upon transfer of a molecule from the liquid to the gas phase.

MDEC model is based on a simple idea of describing the effects of electronic polarization by considering a system of point charges immersed in polarizable electronic continuum. The combination of the MDEC model and Kirkwood–Onsager theory allowed us to develop a self-consistent model that naturally incorporates the effects of electronic polarization into the effective potential of liquid water. The effective parameters of the models derived from the new principles, turns out to be not much different from those of known models; more importantly, however, is that the theory sheds new light onto their actual nature, which is of significant interest for applications of the effective water models in simulations of biological environments, or for simulation of solute–solvent interactions. We showed that self-consistency of the whole theory, and reconciliation of experimental evaporation data with other data (diffusion coefficient, radial distributions, density) is only possible if the value of the actual dipole of liquid water is around 3.0 (± 0.1) D.

6. Conclusions

Concluding, the paper can be summarized as follows:

(1) A new set of parameters for the effective potential of liquid water consistent with the idea of dynamics in polarizable electronic continuum has been developed; the theory

illuminates the nature of the effective charges and the dipole moment of standard nonpolarizable models as scaled values.

(2) The average dipole moment of water molecule in the liquid μ_l is about 3.0(± 0.1) D. This high value is fully consistent with the value of effective dipole $\mu^{\text{eff}} = 2.35\text{D}$ of SPC/E and TIP3P models, or smaller values of other models, for example, TIP4P. The reason for the discrepancy is that the effective dipole moment is a scaled value: $\mu^{\text{eff}} = \mu_l/\sqrt{\epsilon_{\text{el}}}$.

(3) The apparent contradiction of the high dipole moment of a liquid water molecule and the vaporization data is resolved by including the new electronic solvation energy term, E_{sol} . The effects of electronic polarization and screening are present in the effective empirical models in the form of scaled charges.

(4) A combination of MDEC model and Kirkwood–Onsager theory can satisfactorily account for the effects of electronic polarization, including the difference between the gas phase dipole of water (1.85D) and the liquid state dipole, around 3.0D.

(5) The effective charges of SPC/E or TIP3P models are scaled values of the actual charges by a factor of $1/\sqrt{\epsilon_{\text{el}}}$. The scaled nature of these charges should be taken into account in developing the effective potentials of interactions with solutes, or using these models in different (e.g., biological) environments. For example, the charges of ionized groups (± 1) in the aqueous solutions, or in a protein, should be also scaled by the same factor $1/\sqrt{\epsilon_{\text{el}}}$.

Acknowledgment. This work has been supported in part by the NSF grant PHY 0646273 and NIH GM054052.

Appendix

Computational Details. The properties of liquid water were obtained in molecular dynamics simulations by Gromacs²⁸ package. The cubic MD box was formed by 512 water molecules. The electrostatic interactions were treated by the PME technique with a real space cutoff of 12 Å and sixth order spline for mesh interpolation. van der Waals interactions were switched at 10 Å to vanish at 12 Å. The long-range dispersion correction was accounted for pressure and energy. Nonbonded pair list was updated each 5 fs.

The new Berendsen thermostat with a stochastic term and coupling constant 0.5 ps and Berendsen barostat with coupling constant 0.4 ps were applied to keep the temperature at 298 K and pressure 1 atm. For each parameter set, the system was equilibrated first during a 1 ns run, followed by 10 ns data collection run with the MD time step 1 fs. The trajectory was saved each 0.1 ps resulting in total to 100000 configurations.

The self-diffusion coefficient was calculated from mass weighted mean square displacements of atoms from their initial positions using the Einstein relation as implemented in the `g_msd` program.²⁸ The dielectric constant was obtained from fluctuations of the total dipole moment of the MD box as implemented in `g_dipoles` program.²⁸ The vaporization energy was calculated according to the theory described in the manuscript.

The calculated liquid water properties are summarized in the Table 2.

References

- (1) 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*; Pullmann, B., Ed.; D. Reidel Publishing Company: Dordrecht, The Netherlands, 1981; pp 331–342.
- (2) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (3) Guillot, B. A Reappraisal of What We Have Learnt during Three Decades of Computer Simulations on Water. *J. Mol. Liq.* **2002**, *101*, 219–260.
- (4) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
- (5) Lovas, F. J. Microwave Spectral Tables. II. Triatomic Molecules. *J. Phys. Chem. Ref. Data* **1978**, *7*, 1445–1750, no.4.
- (6) Lide, D. R., Ed. *CRC Handbook of Chemistry and Physics (Internet Version 2010)*, 90th ed.; CRC Press/Taylor and Francis: Boca Raton, FL, 2010.
- (7) Carnie, S. L.; Patey, G. N. Fluids of Polarizable Hard-Spheres with Dipoles and Tetrahedral Quadrupoles—Integral-Equation Results with Application to Liquid Water. *Mol. Phys.* **1982**, *47*, 1129–1151.
- (8) Watanabe, K.; Klein, M. L. Effective Pair Potentials and the Properties of Water. *Chem. Phys.* **1989**, *131*, 157–167.
- (9) Gregory, J. K.; Clary, D. C.; Liu, K.; Brown, M. G.; Saykally, R. J. The Water Dipole Moment in Water Clusters. *Science* **1997**, *275*, 814–817.
- (10) Badyal, Y. S.; Saboungi, M. L.; Price, D. L.; Shastri, S. D.; Haefner, D. R.; Soper, A. K. Electron Distribution in Water. *J. Chem. Phys.* **2000**, *112*, 9206–9208.
- (11) Delle Site, L.; Alavi, A.; Lynden-Bell, R. M. The Electrostatic Properties of Water Molecules in Condensed Phases: An Ab Initio Study. *Mol. Phys.* **1999**, *96*, 1683–1693.
- (12) Tu, Y. Q.; Laaksonen, A. The Electronic Properties of Water Molecules in Water Clusters and Liquid Water. *Chem. Phys. Lett.* **2000**, *329*, 283–288.
- (13) Gubskaya, A. V.; Kusalik, P. G. The Multipole Polarizabilities and Hyperpolarizabilities of the Water Molecule in Liquid State: An Ab Initio Study. *Mol. Phys.* **2001**, *99*, 1107–1120.
- (14) Silvestrelli, P. L.; Parrinello, M. Structural, Electronic, and Bonding Properties of Liquid Water from First Principles. *J. Chem. Phys.* **1999**, *111*, 3572–3580.
- (15) Gubskaya, A. V.; Kusalik, P. G. The Total Molecular Dipole Moment for Liquid Water. *J. Chem. Phys.* **2002**, *117*, 5290–5302.
- (16) Sharma, M.; Resta, R.; Car, R. Dipolar correlations and the dielectric permittivity of water. *Phys. Rev. Lett.* **2007**, *98*, 247401.
- (17) Leontyev, I. V.; Stuchebrukhov, A. A. Electronic Continuum Model for Molecular Dynamics Simulations of Biological Molecules. *J. Chem. Theory Comput.* **2010**, *6*, 1498–1508.
- (18) Leontyev, I. V.; Vener, M. V.; Rostov, I. V.; Basilevsky, M. V.; Newton, M. D. Continuum Level Treatment of Electronic Polarization in the Framework of Molecular Simulations of Solvation Effects. *J. Chem. Phys.* **2003**, *119*, 8024–8037.
- (19) Leontyev, I. V.; Stuchebrukhov, A. A. Electronic Continuum Model for Molecular Dynamics Simulations. *J. Chem. Phys.* **2009**, *130*, 085102.
- (20) Kirkwood, J. G. The Dielectric Polarization of Polar Liquids. *J. Chem. Phys.* **1939**, *7*, 911–919.
- (21) Onsager, L. Electric Moments of Molecules in Liquids. *J. Am. Chem. Soc.* **1936**, *58*, 1486–1493.
- (22) Morita, A.; Kato, S. An Ab Initio Analysis of Medium Perturbation on Molecular Polarizabilities. *J. Chem. Phys.* **1999**, *110*, 11987–11998.
- (23) Murphy, W. F. Rayleigh Depolarization Ratio and Rotational Raman-Spectrum of Water-Vapor and Polarizability Components for Water Molecule. *J. Chem. Phys.* **1977**, *67*, 5877–5882.
- (24) Benedict, W. S.; Gailar, N.; Plyler, E. K. Rotation-Vibration Spectra of Deuterated Water Vapor. *J. Chem. Phys.* **1956**, *24*, 1139–1165.
- (25) Krynicki, K.; Green, C. D.; Sawyer, D. W. Pressure and Temperature-Dependence of Self-Diffusion in Water. *Faraday Discuss.* **1978**, *66*, 199–208.
- (26) Soper, A. K. The Radial Distribution Functions of Water and Ice from 220 to 673 K and at Pressures up to 400 MPa. *Chem. Phys.* **2000**, *258*, 121–137.
- (27) Soper, A. K.; Phillips, M. G. A New Determination of the Structure of Water at 25-Degrees-C. *Chem. Phys.* **1986**, *107*, 47–60.
- (28) van der Spoel, D.; Lindahl, E.; Hess, B.; van Buuren, A. R.; Apol, E.; Meulenhoff, P. J.; Tieleman, D. P.; Sijbers, A. L. T. M.; Feenstra, K. A.; van Drunen, R.; Berendsen, H. J. C. Gromacs User Manual version 4.0, www.gromacs.org (accessed September, 2010).

CT1002048

JCTC

Journal of Chemical Theory and Computation

Using Electronic Energy Derivative Information in Automated Potential Energy Surface Construction for Vibrational Calculations

Manuel Sparta,^{*,†} Mikkel B. Hansen,[†] Eduard Matito,^{†,‡} Daniele Toffoli,^{†,§} and Ove Christiansen[†]

The Lundbeck Foundation Center for Theoretical Chemistry, Center for Oxygen Microscopy and Imaging, Department of Chemistry, University of Aarhus, Langelandsgade 140, DK-8000 Aarhus C, Denmark, Institute of Physics, University of Szczecin, Wielkopolska 15, 70-451 Szczecin, Poland, and Department of Chemistry, Middle East Technical University, 06531 Ankara, Turkey

Received May 4, 2010

Abstract: The availability of an accurate representation of the potential energy surface (PES) is an essential prerequisite in an anharmonic vibrational calculation. At the same time, the high dimensionality of the fully coupled PES and the adverse scaling properties with respect to the molecular size make the construction of an accurate PES a computationally demanding task. In the past few years, our group tested and developed a series of tools and techniques aimed at defining computationally efficient, black-box protocols for the construction of PESs for use in vibrational calculations. This includes the definition of an adaptive density-guided approach (ADGA) for the construction of PESs from an automatically generated set of evaluation points. Another separate aspect has been the exploration of the use of derivative information through modified Shepard (MS) interpolation/extrapolation procedures. With this article, we present an assembled machinery where these methods are embedded in an efficient way to provide both a general machinery as well as concrete computational protocols. In this framework we introduce and discuss the accuracy and computational efficiency of two methods, called ADGA[2gx3M] and ADGA[2hx3M], where the ADGA recipe is used (with MS interpolation) to automatically define modest sized grids for up to two-mode couplings, while MS extrapolation based on, respectively, gradients only and gradients and Hessians from the ADGA determined points provides access to sufficiently accurate three-mode couplings. The performance of the resulting potentials is investigated in vibrational coupled cluster (VCC) calculations. Three molecular systems serve as benchmarks: a trisubstituted methane (CHFCIBr), methanimine (CH₂NH), and oxazole (C₃H₃NO). Furthermore, methanimine and oxazole are addressed in accurate calculations aiming to reproduce experimental results.

1. Introduction

During the past few years, accurate calculation of vibrational spectra and vibrational corrections to molecular properties have become increasingly feasible due to the development of ad hoc efficient computational methods. Restricting

attention to time-independent explicit wave function methods, the vibrational self-consistent field (VSCF)^{1–4} method is the first to be mentioned. In a VSCF calculation a mean-field description of the multimode dynamics is achieved. In analogy with the Hartree–Fock method of electronic structure theory, a VSCF calculation may serve as the starting point for more elaborate correlated calculations. For instance, explicit mode–mode correlation may be accounted for with vibrational Møller–Plesset perturbation theory (VMP, by some denoted correlation-corrected VSCF, cc-VSCF),^{5–10}

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

[†] University of Aarhus.

[‡] University of Szczecin.

[§] Middle East Technical University.

vibrational configuration interaction (VCI),^{2,3,11–16} and vibrational coupled-cluster (VCC),¹⁴ and for the latter two vibrational response theory can be applied to calculate excitation energies and properties.^{17,18}

A fundamental prerequisite for the accuracy of any vibrational calculation is the availability of a precise representation of the anharmonic part of the PES. This means that a sufficiently large portion of the Born–Oppenheimer (BO) hypersurface has to be sampled with electronic structure methods, which represents one of the bottlenecks in molecular vibrational dynamics. In fact, one should notice that the dimensionality of the hypersurface grows linearly with the number of atoms in the molecule, and hence, the calculation of a fully coupled PES is prohibitive except for the smallest molecules.

To overcome this problem, the full-dimensional PES can be approximated as a sum of potential energy functions (PEFs) of lower dimensionality and high-order mode couplings are included in a hierarchical way. In the vibrational context, this approximation goes under the name of *n*-mode representation of the potential. It was originally suggested and implemented by Carter et al. up to four-mode couplings,¹⁹ while around the same time, a similar approach restricted to two-mode couplings was suggested and since used extensively by Gerber and co-workers.^{6,20} The restricted mode-coupling approach is becoming a standard way for constructing accurate approximate PESs and has been generalized and extended in various ways and combined with various techniques for obtaining efficiency in the calculations.^{6,15,21–27} At least for fairly rigid molecules the restricted mode-coupling expansion of the potential converges fast with the mode-coupling level and inclusion of up to the three-mode-coupling terms provides sufficiently accurate fundamental vibrations. However, inclusion of three-mode couplings soon becomes computationally costly as the size of the molecular system is increased.

Recently, we implemented an adaptive density-guided approach (ADGA) for accurate and flexible representation of the potential energy functions, (PEFs) relevant to quantum dynamics calculations.²⁸ The method allows for a dynamical and automatic generation of the grid of evaluation points for each PEF included in the approximate representation of the fully coupled PES. The term “dynamical” relates to the adaptive strategy for determining the grid boundaries. This is opposed to a “static” determination of the grid of points, where the grid boundaries are specified by the user on input, see, e.g., ref 25. In the ADGA, the densities of the vibrational wave functions are used to guide the dynamic generation of the grids of evaluation points with respect to both the size of the grid and the mesh of evaluation points. Subsequently, the procedure was improved by adding multilevel or multiresolution capabilities, i.e., information from single-point calculations executed with different electronic structure methods are combined to increase the accuracy and/or reduce computational cost.²⁹ The attractiveness of adaptive procedures has also been emphasized in a quite different context by Iyengar and co-workers.^{30,31} In ref 31 it is shown how the use of a partial nondirect product grid in an adaptive scheme to construct PESs has a significant effect on the

computational accuracy and efficiency. As the appropriate definition of nondirect grids is not clear in this context, we use direct product grids for our PEFs. We also point out that Rauhut¹⁵ discussed iterative approaches as means of error control and, in conjunction with other clever tricks, reported huge computational savings. Finally, a completely different approach toward building up the PES is the use of neural networks as pursued by Manzhos and Carrington.^{32,33}

In a concurrent project, our lab has developed a procedure for the generation of molecular PESs which integrates the aforementioned restricted mode-coupling representation of the PES with efficient use of derivative information by means of a modified Shepard (MS) interpolation/extrapolation. In particular, the method aims at calculating higher mode couplings by extrapolation of lower mode couplings in a grid-based approach. The method does not require any a priori knowledge of the potential and has, using second-order derivatives, been shown to give an accuracy competitive with that of the explicitly calculated higher order mode-coupling PES. Using only gradients already leads to significant improvements. For a detailed explanation and an overview on the use of derivatives information in the PES construction we refer to ref 36 and references therein as well as Dawes et al.^{34,35} for interesting alternative approaches and further references.

In this article we integrate the use of the MS interpolation/extrapolation with the multilevel implementation of the ADGA. This carries both theoretical and technical challenges. First, the procedures for interpolation and extrapolation were found solid and useful when based on rather dense and homogeneous grids constructed with a “static” approach; within the ADGA framework, the iterative procedure provides relatively coarse and dishomogeneous grids; hence, both accuracy and efficiency must be thoroughly investigated. From the implementational point of view, it is not unique how to define and use the interpolating functions during the iterative procedure nor is the resulting effect on the convergence rate clear. Once these issues are investigated and addressed, the goal is (i) to achieve a faster convergence of the iterative approach and (ii) establishing higher order extrapolated PEFs from parental lower order mode couplings converged with the ADGA as accurate but low-cost approximations to explicitly calculated higher order terms.

From these general aspects we can define concrete PES construction protocols, for instance, ADGA[2gx3M] where the ADGA recipe is used (with MS interpolation) to automatically define accurate representation for up to two-mode couplings, while the MS extrapolation trick based on gradients is used to construct the three-mode couplings. Alternatively, if both gradient and Hessian are used we will refer to the PES with the code ADGA[2hx3M]. As a further example, a 2M potential where the one mode part is explicitly constructed and the two-mode couplings are extrapolated using both gradients and Hessians will be referred by ADGA[1hx2M]. Certainly, the possibility to obtain three-mode couplings with only two-mode grids is a unique opportunity. In passing we note a very interesting procedure recently developed by Rauhut:³⁷ A semiempirical method is adjusted for the particular molecule and the particular set

of one- and two-mode calculation points. Using this adjusted and fast semiempirical method provides another way of constructing three-mode couplings without the explicit ab initio electronic structure calculations of the potentially large number of three-mode grids.

While calculation of the PES is a costly procedure, calculation of the anharmonic wave function for a given potential can also be extremely costly as the size of the molecule increases. Recently, some of us derived concrete optimal formal scalings and accompanying implementations for vibrational wave function methods such as VCC and VCI.^{4,38} In this study our main vibrational wave function method is the recently developed VCC[2pt3] approach,³⁹ which in the framework of VCC response theory gives a full description of two-mode couplings and a perturbational motivated approximate inclusion of three-mode excitations. For each state calculated, VCC[2pt3] comes with a modest cubic scaling with respect to the number of degrees of freedom.³⁹ Clearly, the combined use of, e.g., ADGA[2gx3M] and VCC[2pt3] is natural as these methodologies are unique in providing a computationally realistic inclusion of three-mode couplings in the PES and wave function parts, respectively. Thus, their combined use allows for accurate calculations on larger molecules where calculations otherwise would have to be limited to approaches including only up to two-mode couplings. As a further point, ADGA has been shown to lead to reduced scaling also in the wave function part due to its ability to provide compact representation of the PES.⁴

The paper is organized as follows. In section 2 we describe the ADGA algorithm used for generation of the grids of evaluation points relating it to previous works as well as the features and implementation of the MS procedures. Furthermore, the way the two concepts are integrated is fully accounted for. A brief summary of the computational details follows in section 3. The results of the benchmark calculations are given in section 4 and organized as follows: first, we present the convergence properties of the algorithm described by means of a systematic survey on 3 smaller molecules (trisubstituted methane, methanimine, and oxazole). Second, we apply the knowledge gained to attempt an accurate calculation of the vibrational spectra of methanimine and oxazole with high-level electronic structure methods. Finally, conclusions and perspectives are given in section 5.

2. Description of the Method

2.1. Vibrational Hamiltonian and Potential Energy Operator. Vibrational energies and vibrational contributions to molecular properties are obtained from solution of the Born–Oppenheimer vibrational Schrödinger equation. The Hamiltonian in mass-weighted rectilinear normal coordinates ($q_m \in \mathbf{Q}$) can be written as

$$H = T(\mathbf{Q}) + V(\mathbf{Q}) \quad (1)$$

where $T(\mathbf{Q})$ is the kinetic energy operator (either in the Watson form⁴⁰ or simply as a sum of $-(1/2)(d^2/dq_m^2)$ terms; for details and further references we refer to our previous ref 25). $V(\mathbf{Q})$ is the BO potential energy term.

In order to reduce the large dimensionality of the problem associated with computation of the potential energy surface the restricted mode-coupling representation of the fully coupled potential is adopted.^{6,15,19,21,23} One starts by defining a set of potential energy functions (PEFs) which include the coupling among a subset n of the M vibrational degrees of freedom

$$\begin{aligned} V^{m_i} &= V(0, \dots, 0, q_{m_i}, 0, \dots, 0) \\ V^{m_i, m_j} &= V(0, \dots, 0, q_{m_i}, 0, \dots, 0, q_{m_j}, 0, \dots, 0) \end{aligned} \quad (2)$$

and so forth up to V^{m_1, m_2, \dots, m_M} , the fully coupled potential, $V(\mathbf{Q})$. In eq 2 $m_i \neq m_j$. For the sake of simplicity the set of modes (referred to as a mode combination, MC, hereafter) defining the particular PEF are collected in an n -dimensional vector \mathbf{m}_n . In this way, an n -dimensional PEF is denoted as $V^{\mathbf{m}_n}$. Furthermore, one notes that by summing over all MCs, overcounting is introduced since each n -dimensional PEF includes all the lower dimensional PEFs corresponding to the set of $\mathbf{m}_n' \subset \mathbf{m}_n$. In order to avoid overcounting, potentials $\bar{V}^{\mathbf{m}_n}$ are conveniently introduced (see, e.g., ref 23 for details) such that

$$V = \sum_{\mathbf{m}_n \in \text{MCR}\{V\}} \bar{V}^{\mathbf{m}_n} \quad (3)$$

where MCR is a mode-combination range, the set of MCs we want to include in the potential.

Within this framework we developed some efficient computational protocols. The overall setup is rather general. For a given mode-combination level n , all the n -dimensional PEFs, $V^{\mathbf{m}_n}$, are sampled in a set of grid points by means of ab initio electronic structure calculations. An analytic representation is then obtained by fitting each of the PEFs to a multivariate polynomial in frequency-scaled mass-weighted normal coordinates (for more details we refer to ref 25). To appreciate the overall setup in this paper we shall now briefly describe the adaptive density-guided approach for construction of the PES.

2.2. Adaptive Density-Guided Approach for Construction of the PES. The basic ideas of ADGA²⁸ are (i) the boundaries of the sampling grids are determined by the spatial extension of the vibrational wave functions studied and therefore depend naturally on the specific chemical problem under investigation, (ii) each of the PEFs is initially explored with a “minimal” sampling grid, the mesh of the grid is iteratively refined until the reference vibrational wave function calculation is stable toward further extension within a threshold, and (iii) vibrational wave functions are used to construct vibrational densities (ρ) which are used as key quantities in the construction of $\rho \times V$ “energy-like” contributions. These contributions help to identify the importance of each mode coupling as well as the optimized grid mesh for its description.

The PES construction and wave function density calculation now become interdependent of each other. In this context, our recently proposed fast VSCF algorithm is useful due to the (possibly) large number of VSCF optimizations required.⁴ The original papers detail the used densities, and

we note only that the user in input specifies the vibrational levels taken into account in the density calculation and that the VSCF one-mode densities integrate to one over the full one-mode configuration space. This feature is used to ensure that the grid boundaries for construction of the PES covers the space explored by the wave function.

In the ADGA for PES construction grid points are systematically added until convergence: The n th iteration starts with evaluation of the PEFs in correspondence to the list of points generated in the $(n - 1)$ th iteration. An analytical representation of the $V^{\mathbf{m}, n\text{th}}$ terms is provided via a polynomial fitting (superscript n th refers to the n th iteration). The potential is then used in a VSCF calculation for the ground vibrational state, and the VSCF modals are then used to construct the vibrational density ρ_{av}^{nth} for mode q_m . The grid is then considered as composed of subsectors (wherein 1D a subsector is defined as the interval between two adjacent sampling points). For each subsector we can define an “energy-like” contribution, $(\rho V)_i^{nth} = \int_{\Omega_i} \rho_{av}^{nth}(q_m) V^{\mathbf{m}, nth}(q_m) dq_m$, where Ω_i denotes the i th interval. A further testing point is requested at the middle point of the subsectors for which the condition

$$\frac{(\rho V)_i^{nth} - (\rho V)_i^{(n-1)\text{th}}}{(\rho V)_i^{nth}} < \epsilon_{\text{rel}} \quad (4)$$

is not fulfilled. Here, $(\rho V)_i^{(n-1)\text{th}}$ is taken equal to zero at the first iteration. An interval where the relative variation of the integral value $(\rho V)_i$ between two iterations is found to be larger than a specified threshold (usually of the order of 1%) is further subdivided with addition of a new testing point. Furthermore, intervals with a small contribution to the energy are not further subdivided, regardless of the relative error computed by means of eq 4. In particular, a subsector is not further divided if the conditions

$$(\rho V)_i^{nth} - (\rho V)_i^{(n-1)\text{th}} < \epsilon_{\text{abs}} \wedge (\rho V)_i^{nth} < \epsilon_{\text{abs}} \quad (5)$$

are fulfilled, with ϵ_{abs} being usually on the order of 10^{-6} au. The n th iteration ends with the definition of the new subsectors, calculation and storage of the integrals to be used in the subsequent iteration, and a list of new sampling points is compiled.

The ADGA converges hierarchically up to a user-specified maximum mode-combination level, i.e., monodimensional PEFs are converged before the bidimensional ones and so forth. The adaptive construction of the n -dimensional grids is a straightforward generalization of the procedure outlined above:⁴¹ the multidimensional grid domains are partitioned in subsectors defined by 2^n adjacent points (see ref 28), the convergence criteria are n -dimensional generalization of the criteria outlined above, and the n -mode densities for construction of the $(\rho V)_i$ quantities are taken to be a direct product of the converged one-mode densities $(\rho(q_m))$.

In a multilevel or multiresolution procedure for PES construction, PEFs computed with different methods and/or approximations (electronic structure methods, basis sets, analytic representations, and electronic structure programs) are combined to obtain a hybrid PES.^{15,24,42-44} We refer to

ref 24 for an exhaustive discussion of this formalism adapted to the intrinsic mode-coupling framework. In the multiresolution implementation, the ADGA is used for construction of the various PEFs entering in the final PES allowing for the following two options: (i) the PEFs for a given mode coupling term can be combined to include extrapolation procedures and/or linear corrections and (ii) a PES with a lower maximum mode-combination level can be corrected adding higher mode-combination terms from PEFs constructed with computationally less expensive electronic structure methods. In passing, we note that the use of a multiresolution procedure is well suited for a preoptimization of the grid boundaries. A detailed description of the implementation and integration with the ADGA can be found in ref 29.

2.3. Use of Derivative Information. The quartic force field representation of the PES is an obvious example of the use of derivatives in PES constructions. However, such a representation relies on a local Taylor series expansion, and thus, the description of the potential may not be accurate in regions far from the ground state equilibrium geometry. To overcome this problem, the modified Shepard interpolation has been adopted by several authors. Yagi et al.⁴⁵⁻⁴⁷ showed that with some knowledge of the vibrational structure of the molecule, one might obtain a very accurate PES by using a few Taylor expansions centered in the vicinity of higher mode-coupling regions. This method proved very effective even for PESs with moderately strong mode couplings. Other studies have also emphasized that derivative information can be very useful if one has an effective method to generate the candidate points.^{48,49}

Recently, our group developed a similar procedure for generation of the PES following a different spirit: rather than aiming at a PES described by a few high-order Taylor expansions suitably located, the focus was on a dense grid using in this way many low-order Taylor expansions. The approach does not require prior knowledge about the PES and simplifies input preparation. Such an approach was integrated with the restricted mode-coupling representation, thus producing a natural extrapolation scheme for higher mode couplings based solely on calculations on grids for lower mode couplings. By merging both techniques one introduces additional steps in the hierarchical construction of the PES, e.g., three-mode-coupling potentials extrapolated from two-mode-coupling ones using gradients only, here denoted 2gx3, or both gradients and Hessians, here denoted 2hx3, to provide accuracies between the 2M and 3M potentials.³⁶ In addition, yet within the restricted mode-coupling representation, derivative information can be used to interpolate within the same mode-coupling level to provide additional points before a polynomial fitting.³⁶ We shall now describe the interpolation step, how it can be integrated with ADGA, and how a practical extrapolation step can be defined.

2.3.1. Interpolation. In a standard MS interpolation, the interpolated energy value for the point \vec{q} , is obtained as a weighted sum of local Taylor series (one for each expansion point)

$$\tilde{f}(\vec{q}) = \sum_i w_i(\vec{q}) f_i(\vec{q}) \quad (6)$$

$$f_i(\vec{q}) = \sum_{j=0}^{j=m} \left[\frac{1}{j!} ((\vec{q} - \vec{q}_i) \cdot \vec{\nabla}_{\vec{x}})^j f(\vec{x}) \right]_{\vec{x}=\vec{q}_i} \quad (7)$$

where m is the order of the Taylor series and \vec{q}_i are the expansion points. Having usually available only low derivative expansions ($m = 1$ or 2), we investigate in this work whether, in the case of reducing the density of grid, such low-order Taylor series still produce sensible interpolated values. Another important factor for the accuracy of the MS interpolation approach is the weight factors chosen. It is customary to take a function of the inverse of distance as a weight factor. In particular, high powers of the inverse of the distance have been shown to be highly effective in the construction of interpolated values (for the notation used here see ref 36)

$$w_i(\vec{q}) = \frac{\|\vec{q}_i\|^{-2p}}{\sum_j \|\vec{q}_j\|^{-2p}} \quad (8)$$

Here $p = 1, 2, \dots, p = 3N - 6$ has been recommended in MSI interpolations⁴⁶ but should of course be modified according to the dimensionality of the particular mode coupling.

2.3.2. Interpolation Combined with ADGA. During a standard ADGA iteration, the energy information collected from the set of single-point calculations undergoes a polynomial fitting. Especially in the early iterative cycles, the number of single-point calculations, and hence property values, is low and property derivatives are useful for defining a finer grid of points, which combines both calculated and interpolated points to be fitted.

In Figure 1, the strategy for the positioning of the fine grid of points to be interpolated is depicted in the case of a monodimensional surface. For each sector (the interval between two adjacent sampling points x_i and x_{i+1} , blue diamonds) two points to be interpolated are defined at $x_i + a(x_{i+1} - x_i)$ and $x_i + (1.0 - a)(x_{i+1} - x_i)$, where a is a user-defined parameter. For the monodimensional surfaces, two additional interpolation points are added outside the grid boundaries at $x_{\text{first}} - a(x_{\text{second}} - x_{\text{first}})$ and $x_{\text{last}} + a(x_{\text{last}} - x_{text{last-1}})$, respectively.

Although the interpolated points are very useful sources of information, they come with an error which is assumed to increase with the distance to the nearest real point due to the local nature of the low-order Taylor expansion. For that reason, we above introduce only two additional points which by the choice of a can be relatively close to explicitly evaluated points. Second, the i th point is provided with an uncertainty σ_i that has the equivalent effect of a weight during the polynomial fitting. For the explicitly calculated points, a unit σ_i is assumed, whereas for the interpolated points, σ_i is calculated as

$$\sigma_i = \frac{1}{\delta_i} \quad (9)$$

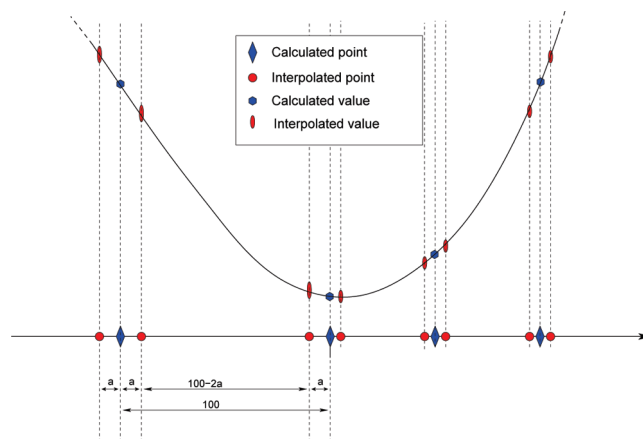


Figure 1. Computed and interpolated points during an ADGA iteration: (blue diamonds) positions of the calculated points; (red dots) positions of the points to be interpolated.

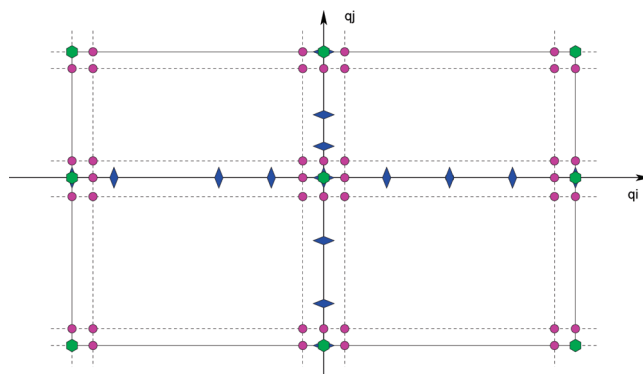


Figure 2. Computed and interpolated points during an ADGA iteration: (green hexagons) positions of the explicitly calculated points for the bidimensional term; (blue diamonds) calculated points on the parental monodimensional surfaces; (magenta dots) points to be interpolated obtained following the a -parameter strategy.

where δ_i depends on the distance between the interpolated point and the calculated points as

$$\delta_i = \sum_k e^{-\lambda d_{ki}/d_{\text{min}}} \quad (10)$$

In eq 10, λ is the decay constant, d_{ki} is the distance between the interpolated point i and the k th calculated point, and d_{min} is the minimum distance between two calculated points in the mode combination. The effect of this is clearly that increasing the distance to all calculated points will increase σ_i such that the point will have less weight in the fitting. This is in some sense a continuous way of expressing that the local expansions can only have a limited trust radius.

For the multidimensional surfaces a straightforward generalization of this scheme is applied as shown in Figure 2 for a bidimensional term (with the exception that there is no point placed outside of the region defined by the monodimensional boundaries). The points explicitly calculated are green (note that the points on the monodimensional surface that define the direct grid of calculated points are also considered to be explicitly calculated), while the points from a lower order mode coupling are marked with blue diamonds. The interpolated points are shown with magenta dots.

For a point on a multidimensional surface, the δ_i used to define the associated uncertainty is computed according to

$$\delta_i = \sum_k \left(\prod_{q \in mc} e^{-\lambda d_{ki}^q / d_{\min}^q} \right) \quad (11)$$

where q is one of the modes present in the mode combination, λ is the decay constant, d_{ki}^q is the q component of distance between the interpolated point i and the k th calculated point, and d_{\min}^q is the minimum distance between two calculated points in the mode combination.

Preliminary tests demonstrated that the performances of the proposed scheme are only weakly dependent on the exact values of the a and λ parameters (if chosen within a reasonable range), and in the following $a = 0.05$ and $\lambda = 8.0$ will be assumed.

For a complete overview of the method, the maximum order of the polynomial entering in the fitting has to be considered. During a standard ADGA iteration, for each direction in the mode coupling, the maximum degree of the fitting polynomial is set to be the larger even number equal to or smaller than n , with n being the number of points determining the direct grid in that direction. The maximum order then increases with more points up to a user-specified maximum order. In the example shown in Figure 2, which matches a first iteration in the ADGA calculation of a two-mode coupling, there are 3 calculation points for each mode. Under these circumstances, the standard procedure of not using interpolation will set up a maximum order of the polynomial equal to 2 for each mode, and hence, the basis for the fitting will include four terms, namely, $q_1 q_1$, $q_1^2 q_2$, $q_1 q_2^2$, and $q_2^2 q_2^2$. In order to account for the additional information provided by the interpolated point, it was found adequate to increase the maximum order of the polynomial to four in the first iteration, whereas in the subsequent iterations the standard rule is used. For the multidimensional case it is in addition customary to set a combined maximum polynomial order (max_order), i.e., $q_1^a q_2^b \in \{q_1^a q_2^b \mid a + b \leq \text{max_order}\}$.

2.3.3. Extrapolation. With a somewhat similar setup, as previously described, one can approximate energy values in a mode coupling with no explicit calculation points except those of the limiting mode couplings obtained by setting one of the coordinates to zero. We denote this as an extrapolation as it is an extrapolation from the grid points of a lower coupling grid calculation (say including up to two-mode couplings) to an estimated function for a higher mode coupling (three-mode couplings). It should be noted, however, that it is still a kind of interpolation from a global view of the function.

In this case, the point to be extrapolated, \vec{q} , is given by the direct product of the grid of points of lower dimensionality. The estimated value of the function at the point to be extrapolated is correspondingly obtained as the weighted sum of the result obtained from the Taylor expansions around the projections of this point to the individual mode-coupling surfaces of one lower dimension.

Detailed equations and a number of graphical illustrations of the procedure are given in ref 36. In Figure 3 the strategy for extrapolation of a bidimensional mode coupling is indicated: the computed grid of points (blue diamonds) on

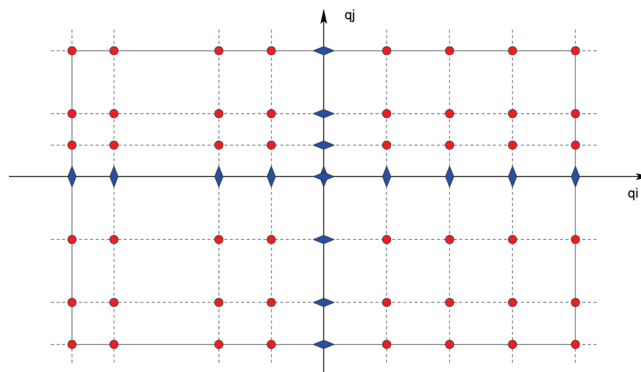


Figure 3. Extrapolation of a bidimensional term: (blue diamonds) positions of the calculated points; (red dots) positions of the points to be extrapolated.

the monodimensional surfaces define a direct product grid of points that will be extrapolated (red dots).

The method, likewise MSI, depends only on an analytical parameter-free weight factor, eq 8, which controls the contributions from each direction of extrapolation. Interestingly, extrapolation is much less dependent on the power of the distance chosen for the weight factors³⁶ and the accuracy of the method is mainly given by the order of the derivatives used in the Taylor series. Should V^{n+1} be represented by an n -variate polynomial, it is easy to prove that the terms of V^{n+1} with powers not higher than the order of derivatives used in the Taylor series are reproduced *exactly*. Other terms are approximated by the aforementioned weighted sum of Taylor expansions. Therefore, when using second-order Taylor expansions one anticipates that pure or mixed linear and quadratic terms (i.e., $q_1 q_2 \dots$, $q_1^2 q_2 \dots$, \dots , $q_1^2 q_2^2 \dots$) are exact in the polynomial representation.

In ref 36 the extrapolation procedure has been successfully applied to a representative set of molecules, showing that the use of first derivatives improves relative to the original PES, V^n , and the use of second derivatives provides a PES with an accuracy comparable to the full calculation.

In the framework of the ADGA, the possibility to extrapolate potential terms from the information collected during construction of the lower dimensionality terms has two applications. The extrapolated PEFs may be used without further refinement to approximate higher mode couplings. Alternatively, the approximated potentials may serve as educated guesses for the “zero”-order $(\rho V)_i$ quantity in the first iteration of the ADGA. The basic extrapolation strategy requires no essential modifications, but there are a few practical differences compared to the previous tests.³⁶ First, the ADGA grid will be small and not regular in contrast to the rich and regular grids for which the extrapolation procedure was tested originally. We expect this to be of minor numerical importance as ADGA per definition should define the grid points to be where there are large contributions and large variations. Furthermore, in accordance with the fact that second-order Taylor expansions provide exact pure and mixed linear and quadratic terms, for each mode in the mode combination a maximum polynomial degree equal to two is adopted in the fitting procedure (recalling that still terms such as $q_1^2 q_2^2 \dots$ are allowed). At this stage, since all the points entering the fitting are extrapolated, all

the weights are identical. Higher order expansion can be requested by the user, but we will throughout use this lower order possibility. This was found satisfactory in preliminary test calculations, and limiting the number of terms in the potential representation will be computationally convenient, especially for larger systems.

2.4. Further Implementational Details: Symmetry and Parallelization. All of the above-mentioned procedures have been implemented in the current version of MidasCpp (Molecular Interactions, Dynamics, and Simulations Chemistry Program Package),⁵⁰ containing both PES and wave function modules. The PES module interacts smoothly with the wave function modules for fast VSCF density calculation during the ADGA calculations and for post-PES vibrational calculations with VSCF, VCI, VMP, VAPT (vibrational autoadjusting perturbation theory),⁵¹ and VCC. The PES module includes interpolation (cubic-splines, modified Shepard), the MS-based extrapolation step, ADGA (with multi-resolution options), as well as the use of point group symmetry and parallelization described very briefly below.

The current version of MidasCpp can exploit point group symmetry to avoid computation of symmetry equivalent structures and generate symmetry pure potentials. The abelian groups are implemented, and implementation takes advantage of the fact that the irreducible representations of the abelian groups are monodimensional. Hence, knowledge of the character table of each group gives the direct effect of the various symmetry operations on the normal coordinates as the effect is multiplication by either +1 or -1. By noting that the normal coordinates are already a set of symmetry-adapted coordinates, it is clear that symmetry-related structures can be found only within the same mode combination. The strategy adopted to exploit molecular symmetry for both energies, properties, and derivatives is as follows. The program analyzes the list of the requested calculations at each ADGA iteration and discards those of symmetry-related structures: only unique structures are computed. Whereas two symmetry-related structures share the same orientation-independent properties (e.g., energy), the rotation matrix between the two structures has to be determined to transform the orientation-dependent properties (e.g., dipole moment components). Derivatives are consistently transformed using the same rotation matrix. The full list of calculations is compiled based on the information in the previous two steps. Furthermore, if the electronic structure program has the capability to recognize and exploit molecular symmetry for a given molecular structure, this feature is used to speed up the corresponding single-point calculation. If the use of symmetry in the single-point calculation requires the program to rotate the molecule into the inertial axes frame and/or the nuclei-specific properties are printed only for symmetry-independent centers, MidasCpp is capable of reconstructing the redundant list of values and reorient the orientation-dependent properties in its coordinate system. In this way MidasCpp can handle several properties such as nuclear magnetic shielding constants, molecular dipole moments (including derivatives), and (hyper)polarizabilities, etc. Finally, we note that symmetry-adapted polynomial functions are used in the fitting and PES representation.

We note that during a specific ADGA iteration, a list of single-point calculations is requested. As the single points are independent, their calculation represents an embarrassingly parallel problem. It is thus possible to setup a very efficient parallelization framework. Our current implementation mimics a master/slaves approach, with the MidasCpp process submitting single-point calculations and collecting the needed results of each calculation. Hybrid approaches are also possible where MidasCpp runs in parallel several electronic structure calculations, each of which runs in parallel.

2.5. On the Relative Cost of the Derivative- vs No-Derivative-Based Calculations. While the use of derivatives is conceptually appealing, derivatives comes with an additional cost. It has previously been discussed how the calculation time in particular modifies the scaling gains obtained by the extrapolation trick.³⁶ However, this was in the context of a fixed and large grid. We now describe how this discussion is modified in important ways in the context of the dynamic ADGA grids, which ultimately will decide if a particular choice of ADGA grid level combined with extrapolation is computationally efficient.

The computational cost of Hessian calculations is, for medium-sized molecules, generally expected to exceed the cost of a energy calculation by a factor $C \times M$. Here, M is the number of modes in the molecule (and for a nonlinear molecule $M = 3N - 6$, and for the purpose of scaling arguments it is not too different from $3N$, with N being the number of atoms). C is some characteristic constant factor relating to the electronic structure method and its implementation, which is anticipated to be on the order of unity. For a gradient calculation, the relative time of calculating the energy and gradient compared to an energy-only calculation is typically a small constant close to 1 for variational wave functions and close to 2 for nonvariational methods such as coupled cluster (CC). Concerning the calculation and use of first derivatives, it is also fair to add that vibrational calculations often require other first-order properties (e.g., dipole moment for calculation of absorption spectra); hence, the gradient is essentially available at very little additional cost.

Clearly, in order to be useful for a reduction of the computational time, the savings in the number of single-point calculations due to the use of derivatives must not to be overwhelmed by the increased cost of each single-point calculation.

If one considers a grid approach where one generates for each mode coupling a direct product grid of points, the cost is dominated by computation of the n -mode couplings where n is the highest coupling level included in the molecular PES. The number of single points needed is then on the order of

$$\binom{M}{n} g_n^n$$

where g_n is the number of single points per direction in generation of the direct grid. If one can extrapolate the higher couplings from the $(n - 1)$ th terms, the required number of single points is on the order of

$$\binom{M}{n-1} g_{n-1}^{(n-1)}$$

For large enough M , this guarantees a reduction in the number of grid points requested on the order of $(M/n)(g_n^n/g_{n-1}^{(n-1)})$. For the gradient extrapolation this is clearly very favorable as the additional cost for computing gradients is small. For the Hessian-based extrapolation the situation is less clear.

Let us for a moment consider the simplest case of the same number of grid points per mode being used in the different mode coupling levels, and thus, $g_n = g_{n-1} = g$, then the reduction in grid points is a factor $(Mg)/(n)$. Taking into account the additional C factor we obtain a ratio in the computational cost of full grid vs Hessian-extrapolated static grid approach as $(g)/(Cn)$. There is a good efficiency gain if the number of grid points per direction is large, but the efficiency gain (if at all a gain) decreases with increasing mode-coupling level n and the factor C . Slightly more general, we can expect a reduction of the total computational cost of a factor $(1/Cn)(g_n^n/g_{n-1}^{(n-1)})$, so there can be significant gain using also Hessians especially for $n = 2$ (thus for the 1M to 2M extrapolation) and when $g_n \approx g_{(n-1)}$.

ADGA was designed to efficiently address construction of mode couplings through a careful selection of the single points needed by analyzing the vibrational wave function density and the strength of the couplings. In agreement with the expectation that the strength of the coupling decreases with increasing mode-coupling level, the number of grid points is not constant anymore. One can say that the effective number of grid points per dimension decreases with n . For ADGA, the effective g_n can be expected to be larger than 10 for $n = 1$ and 2; for systems with modest three-mode couplings g_3 drops significantly below 10 on average, being often as small as 4 with a minimum of 2. The electronic structure program factor C varies of course with different methods and programs, but it should be on the order of unity; we found it to be around 3 in our applications and tests. This in turn means that the Hessian-based extrapolations can be a savings in the computational time when extrapolating the two-mode couplings from the one-mode grid. However, it will often not be competitive for the two- to three-mode extrapolation, if compared with the efficiency of a pure ADGA on higher mode couplings.

In summary, from scaling perspectives it is clear that gradient extrapolations are always useful in the sense of providing additional information at almost no additional cost. Hessian-based extrapolations can be useful, but it depends more on the case, and relative to an efficient ADGA prediction of the higher mode coupling the relative gain in total CPU time will be modest if at all a gain. We will explore in the tests all options at the two- and three-mode coupling level to obtain some insight in the accuracies that are obtainable and, in the light of the above, the accuracy/cost ratios of the different methods.

3. Computational Details

Construction of PESs requires as input vibrational frequencies and normal coordinates. In the first part of the results

section we adopted Hartree–Fock theory in connection with STO-3G basis sets for calculation of the normal coordinates, vibrational frequencies, and single points needed in construction of the PESs. Whereas this level of theory is far from optimal for obtaining highly accurate surfaces, it allows for a systematic investigation of the parameters controlling integration of the adaptive algorithm with the Shepard’s techniques. In addition, analytic gradients and Hessians are available for the ground state energy, whereas first derivatives are used for the dipole moment components.

In the second part of the results section, we aim at an accurate and cost-effective PES construction, using the insight gained with the above benchmarks. At this stage, we employ coupled-cluster singles and doubles with perturbative triples correction⁵² (CCSD(T)) level of theory in connection with triple- ζ quality basis sets (cc-pVTZ)^{53–55} for geometry optimization and harmonic vibrational analysis. Construction of the analytic form of the potential and property surfaces is achieved with a multiresolution scheme:²⁹ monodimensional PEFs are computed at the CCSD(T)/cc-pVTZ level of theory, whereas bidimensional couplings and extrapolated three-dimensional couplings are constructed by means of CCSD/cc-pVTZ or MP2/cc-pVTZ calculations. All electronic structure calculations have been carried out with the CFOUR program.⁵⁶

Concerning the thresholds used in the ADGA procedure (see ref 28 for details), the monodimensional surfaces are converged with $\epsilon_{\text{rel}} = 5 \times 10^{-3}$ and $\epsilon_{\text{abs}} = 5 \times 10^{-7}$, the bidimensional surfaces are converged with $\epsilon_{\text{rel}} = 3 \times 10^{-2}$ and $\epsilon_{\text{abs}} = 3 \times 10^{-6}$, and the three-dimensional surfaces are converged with $\epsilon_{\text{rel}} = 1 \times 10^{-1}$ and $\epsilon_{\text{abs}} = 1 \times 10^{-5}$. These thresholds ensure that the representation of the PEFs are tightly converged.

The maximum polynomial degree used for fitting of the monodimensional and bidimensional surfaces is 12, while for the three-mode PEFs a maximum polynomial degree of 10 is used.

The boundaries of the monodimensional grids are iteratively determined by requiring that 99.9% of the mean density constructed from the three lowest vibrational states for each vibrational mode is included in the boundaries of the monodimensional grids. For the second part of the PES construction, namely, accurate CCSD(T) surfaces, the latter parameter is set to four.

The one-mode vibrational densities used in the ADGA are obtained from VSCF calculations on the ground vibrational states. The VSCF modals are expanded in a set of distributed Gaussians, generated from a basis set density equal to 0.8; the details for this basis set can be found in ref 28. For the correlated vibrational calculation, the VCC[2pt3] method³⁹ is applied. The lowest 8 VSCF modals per mode are retained in the correlated calculations. Coriolis coupling effects are taken into account by approximating the elements of the effective moment of inertia tensor with their equilibrium values and limiting the operator to a 3-mode operator.

Table 1. Methanimine: Potential Energy Surfaces Constructed with and without Derivative Information^a

no. of calcs potential ^d	96		no derivatives in ADGA ^b			derivatives used in ADGA ^c		
			3438		7731	2880		5354
	1M	1hx2M	2M	2hx3M	3M	2M	2hx3M	3M
MAD(2M) ^d	54.9	4.3	0.0			0.3		
MaxAD(2M) ^d	184.8	10.4	0.0			0.7		
MAD(3M) ^e	48.8	10.0	8.8	0.9	0.0	8.6	1.2	0.9
MaxAD(3M) ^e	130.2	65.1	54.7	2.0	0.0	54.0	3.2	3.0

^a Summary of the mean absolute deviation (MAD) and max absolute deviation (MaxAD) computed for the fundamental frequencies with respect to results obtained with a converged potential. The single-point calculations are carried out at the HF/STO-3G level of theory. The vibrational method adopted is VCC[2pt3]. The values are given in cm^{-1} . ^b Derivatives are not used in the 2M and 3M ADGA iterations. (Derivatives are used in construction of the monodimensional terms and in extrapolation of high mode couplings; see text for details.) ^c Interpolation exploited as in Figure 1. ^d Reference data entering the statistical analysis are obtained with the 2M PES calculated with ADGA without the use of derivatives. ^e Reference data entering the statistical analysis are obtained with the 3M PES calculated with ADGA without the use of derivatives.

4. Results and Discussions

4.1. Convergence Benchmark. In this section we explore the properties and accuracy of the higher order PEFs extrapolated by means of the MS technique as well as the speed up in the convergence of the ADGA when derivative information is exploited. As previously described, the spatial extension of the grids sampling the PEFs is determined during construction of the monodimensional terms. In order to obtain a fair comparison, the potentials termed “with derivatives” and “without derivatives” were forced to share the same monodimensional part (i.e., constructed by exploiting derivatives). This ensures that the observed discrepancies are inherent to the PES qualities of the coupling potentials which is the computationally most demanding and not due to subtle differences in the final calculation (e.g., basis set distribution and extent of sampled space).

The first molecule to be addressed is methanimine (CH_2NH), and the corresponding results are reported in Table 1. Initially we consider calculations where MSI is not used during the ADGA for the two- and three-mode grid parts. From the data, it is possible to determine the relative importance of the truncation level in the n -MCR approach. If one compares the fundamental frequencies obtained from a monodimensional approximation to the PES (column 1M) with the data obtained from converged bidimensional potential, the mean absolute deviation (MAD) from the data sets amounts to 54.9 cm^{-1} while the max absolute deviation (MaxAD) is 184.8 cm^{-1} . Significantly smaller numbers are computed from the comparison of the 2M PES with respect to the 3M counterpart (8.8 and 54.7 cm^{-1} , respectively). Although ADGA efficiently lowers the computational cost associated with construction of higher dimension PEFs, the increase in the number of single-point calculations is still significant: construction of the 1M surfaces requires 96 single points, whereas the 2M and 3M surfaces require 3438 and 7731 points, respectively. On the other hand, the correction obtained with inclusion of the extrapolated surfaces is substantial, for instance, the 1hx2M PES (i.e., the PES obtained with the converged 1M potential corrected with 2M terms obtained by extrapolation using up to Hessians) compares well to the standard ADGA 2M potential with a MAD of 4.3 and a MaxAD of 10.4 cm^{-1} . Hence, with respect to complete neglect of two-mode couplings, an order of magnitude smaller discrepancy is achieved with inclusion

of the extrapolated terms. The same conclusion can be drawn if the results from the 2hx3M and full 3M surfaces are compared, only now the absolute size of the deviations is much smaller.

The previous results pertain to the case where the actual ADGA iterations are carried out without the assistance of the derivatives information. The effect of the use of MSI during the ADGA iterations, as depicted in Figure 1, is presented in Table 1. First, with the use of derivatives, the convergence of the ADGA is achieved with fewer single-point evaluations, indicated by the observation that only 2880 and 5354 single-point calculations are needed for the 2M and 3M PESs, respectively. In spite of the reduction in the single energy points required, the accuracy is comparable. The MAD for the fundamental frequencies computed from these potentials, with respect to the counterpart constructed without MS interpolation, is 0.3 and 0.7 cm^{-1} for the 2M and 3M PESs, respectively. Furthermore, it is worth noticing that the reduction in the number of required single-point evaluations (and derivatives) does not affect the efficiency of the extrapolation scheme. This is shown by the statistical analysis of the results obtained on the 2hx3M PES: the MAD(3M) (deviation relative to 3M) is only 0.3 cm^{-1} larger than the counterpart already commented. Considering the number of single-point calculations, the performance of the 2hx3M PES is remarkable. In spite of the fact that it only requires 2880 single-point calculations, it gives rise to fundamental frequencies that compare with a MAD of 1.2 cm^{-1} (MaxAD 3.2 cm^{-1}) to the results achieved using a 3M PES based on 7731 points.

Concerning the relative accuracy of the calculation with and without MSI in the ADGA exploratory calculation pushing the threshold even lower gives results in between the two, indicating that the larger number of points in the calculation without MSI compensates somewhat for the derivative information provided by the MSI.

The second system investigated in this survey is a trisubstituted methane (CHFCIBr). Table 2 shows the corresponding statistical analysis on the computed fundamental frequencies with different approximations for the PES. Although this molecule has the same number of modes (and mode couplings) as the previous one, it is clear that the inherent complexity of the PES is lower, for instance, the MAD between the frequencies computed with a PES

Table 2. CHFCIBr: Potential Energy Surfaces Constructed with or without Derivative Information^a

no. of calcs	90		no derivatives ^b			derivatives ^c		
			2703		4620	2086		3982
	1M	1hx2M	2M	2hx3M	3M	2M	2hx3M	3M
potential ^d								
MAD(2M) ^d	14.6	2.1	0.0			0.0		
MaxAD(2M) ^d	56.4	13.0	0.0			0.1		
MAD(3M) ^e	13.0	3.0	2.1	0.7	0.0	2.2	0.6	0.3
MaxAD(3M) ^e	48.2	15.9	8.2	1.8	0.0	8.2	1.8	0.8

^a Summary of the mean absolute deviation (MAD) and max absolute deviation (MaxAD) computed for the fundamental frequencies with respect to results obtained with a converged potential. The single-point calculations are carried out at the HF/STO-3G level of theory. The vibrational method adopted is VCC[2pt3]. The values are given in cm^{-1} . ^b Derivatives are not used in the 2M and 3M ADGA iterations. (Derivatives are used in construction of the monodimensional terms and in extrapolation of high mode couplings; see text for details.) ^c Interpolation exploited as in Figure 1. ^d Reference data entering the statistical analysis are obtained with the 2M PES calculated with ADGA without the use of derivatives. ^e Reference data entering the statistical analysis are obtained with the 3M PES calculated with ADGA without the use of derivatives.

Table 3. Oxazole: Potential Energy Surfaces Constructed with or without Derivative Information^a

no. of calcs	154		no derivatives ^b			derivatives ^c		
			8686		33 675	6119		16 379
	1M	1hx2M	2M	2hx3M	3M	2M	2hx3M	3M
potential ^d								
MAD(2M) ^d	37.7	5.0	0.0			0.3		
MaxAD(2M) ^d	164.9	22.7	0.0			0.7		
MAD(3M) ^e	36.0	12.2	9.1	2.1	0.0	8.9	2.0	1.1
MaxAD(3M) ^e	114.3	69.1	50.6	11.0	0.0	49.9	11.6	6.1

^a Summary of the mean absolute deviation (MAD) and max absolute deviation (MaxAD) computed for the fundamental frequencies with respect to the results obtained with a converged potential. The single-point calculations are carried out at the HF/STO-3G level of theory. The vibrational method adopted is VCC[2pt3]. The values are given in cm^{-1} . ^b Derivatives are not used in the 2M and 3M ADGA iterations. (Derivatives are used in construction of the monodimensional terms and in extrapolation of high mode couplings, see text for details.) ^c Interpolation exploited as in Figure 1. ^d Reference data entering the statistical analysis are obtained with the 2M PES calculated with ADGA without the use of derivatives. ^e Reference data entering the statistical analysis are obtained with the 3M PES calculated with ADGA without the use of derivatives.

truncated to the 1M terms and the corresponding data from a 2M potential is equal to 14.6 cm^{-1} . The same applies if one considers the 2M and 3M potentials where the discrepancies are 2.1 and 8.2 cm^{-1} for MAD and MaxAD, respectively. This is a clear indication that the strengths of the 2M and 3M mode couplings are quite low. As expected, with such a situation, ADGA is capable of providing converged surfaces using considerably fewer points compared to the methanimine system. The full 2M construction required 2703 single points, while 4620 points are needed for the 3M PES. Focusing on the Shepard extrapolated surface, the good performances are confirmed: the largest 1hx2M – 2M discrepancy amounts to 13.0 cm^{-1} and drops to 1.8 cm^{-1} if one considers the 2hx3M – 3M case. If Shepard interpolation is used during the ADGA iteration the obtained surfaces are virtually equivalent to the previous one (MaxAD(2M), 0.1 cm^{-1} ; MaxAD(3M), 0.8 cm^{-1}) but a considerable savings in terms of the number of required points is achieved: 2086 instead of 2703 single-point calculations are adequate for obtaining convergence for the 2M surfaces.

The last molecule studied in order to evaluate the accuracy and features of the merging of ADGA and MSI interpolation/extrapolation is oxazole ($\text{C}_3\text{H}_3\text{NO}$). This system represents a somewhat larger system than the previous one with its 18 vibrational degrees of freedom. Following the scheme already presented, the statistical analysis of the fundamental frequencies computed with different PES approximations is provided in Table 3. The evolution of the computed fundamental frequencies with the mode–mode coupling in the potential

is similar to the case of methanimine: a strong dependence on the mode-coupling level of the PES is observed. In fact, the computed fundamental frequencies with a 2M or 3M potential differ by as much as 50 cm^{-1} . On the other hand, when extrapolated PEFs are added to the PES, such discrepancies drop by a factor of 4.5. Furthermore, for this molecule the computational savings associated with the use of derivatives is remarkable: a full 3M potential computed with the ADGA approach requires 33 675 single-point calculations, whereas the ADGA+MS approach lowers the amount to 16 379 with an average error of 1.1 cm^{-1} . Finally, it is noteworthy that the 2hx3M potential gives fundamental frequencies with an average error of 2.0 cm^{-1} in spite of a total request of points equal to 6119.

In this section we investigated the performances of the proposed merging of the ADGA for PES construction with the use of both first- and second-derivative information via MS techniques. We showed that (i) a consistent savings in the total number of single-point calculations is achievable and (ii) extrapolation of higher order coupling terms provides an accurate approximation to the fully converged potential. For oxazole, a modest static grid based on 16, 144 (12×12), and 216 ($6 \times 6 \times 6$) points for monodimensional, bidimensional, and tridimensional mode couplings, respectively, requires, after symmetry screening, 135 897 electronic calculations, putting the gain of the combined ADGA/MS approaches in perspective.

As discussed in section 2.5, despite the significant lowering of the number of grid points, the use of second derivatives can be disfavored by the high computational cost of second

Table 4. Use of Gradients in the Construction of Potential Energy Surfaces: Summary of the Mean Absolute Deviation (MAD) and Max Absolute Deviation (MaxAD) Computed for the Fundamental Frequencies with Respect to Results Obtained with Converged Potentials^a

no. of calcs	methanimine					CHFCIBr					oxazole				
	98		3370		7017	90		2691		4739	154		8530		29028
PES	1M	1gx2M	2M	2gx3M	3M	1M	1gx2M	2M	2gx3M	3M	1M	1gx2M	2M	2gx3M	3M
MAD(1M) ^b	0.2					0.1					0.2				
MaxAD(1M) ^b	0.4					0.2					0.4				
MAD(2M) ^c	54.7	56.6	0.1			14.5	19.9	0.0			37.5	37.3	0.3		
MaxAD(2M) ^c	184.5	141.0	0.1			56.3	60.8	0.1			164.5	123.3	1.2		
MAD(3M) ^d			8.8	3.6	1.0			2.1	1.8	0.2			9.3	3.0	1.3
MaxAD(3M) ^d			54.8	9.3	3.0			8.3	5.2	0.7			50.3	19.7	4.3

^a Reference values computed as in Tables 1, 2, and 3. The single-point calculations are carried out at the HF/STO-3G level of theory. The vibrational method adopted is VCC[2pt3]. The values are given in cm^{-1} . ^b The reference data entering the statistical analysis are obtained with the 1M PES calculated with ADGA without the use of derivatives. ^c The reference data entering the statistical analysis are obtained with the 2M PES calculated with ADGA without the use of derivatives. ^d The reference data entering the statistical analysis are obtained with the converged 3M PES calculated with ADGA without the use of derivatives.

derivatives. Thus, while there can be a gain compared to the static grid three-mode calculation, ADGA reduces in effect the number of grid points per dimension of a corresponding explicit three-mode calculation. As a result, the additional savings in the total number of single points obtained by the Hessians in the 2hx3M calculation far from justifies the extra cost of the calculation of the Hessian for the present set of molecules and calculations (i.e., the total time is actually longer).

So far we have discussed the case where, for each single-point calculation, both gradient and Hessian are available. Having illustrated that good accuracy can be obtained and that significant reductions in the number of single-point calculations can be obtained from extrapolation procedures, we will now address the efficiency of our procedure in the case where only gradient information is used. The gradients often come with modest extra effort compared to an energy calculation, while the Hessian requires substantial extra effort. In Table 4, the statistical analysis on the fundamental frequencies computed for various PESs is investigated. Note that the monodimensional PEFs are constructed *ex novo* using only gradients, which results in a negligible difference with respect to the results where both Hessians and gradients are used. Using only gradients in the extrapolations of the two-mode PEFs does not provide significant improvement with respect to the pure 1M potentials and is obviously useless. On the other hand, inclusion of the three-mode extrapolated corrections significantly improves the agreement with the full 3M potentials. The latter observations are in agreement with the results in ref 36. For instance, in the case of methanimine, the MaxAD(3M) drops from 54.8 to 9.3 cm^{-1} when including the extrapolated terms. In the case of oxazole, the MaxAD(3M) residual error is somehow larger (19.7 cm^{-1}); nevertheless, the average error is only 3.0 cm^{-1} . This suggests that the essential part of the 3-mode couplings is well described by the extrapolation procedure, though of course not as accurate as with the extrapolations using up to Hessians. Finally, one notices that the use of gradients in combination with the ADGA provides a small improvement with respect to the number of single points needed for convergence of the PES, but the savings hardly exceeds 10%.

In conclusion, although the accuracy achieved with the use of second derivatives in the MS interpolation/extrapolation

is not fully met if one uses only first derivatives, inclusion of gradient information during construction of a PES is found to be advantageous, especially for extrapolation of the 3-mode coupling corrections, at least for these fairly rigid systems. Thus, the ADGA[2gx3M] potential is certainly a cost-efficient approach and taking into account the much higher cost of Hessians the ADGA[2gx3M] is likely to be more useful in production calculations than ADGA[2hx3M].

4.2. Methanimine and Oxazole: Comparison with Experimental Spectra. The experimental vibrational frequencies of methanimine are taken from refs 57 and 58, while for oxazole, the data in the Computational Chemistry Comparison and Benchmark Database (CCCBDB) are used,⁵⁹ maintained by the National Institute of Standards and Technology, except for the frequencies ν_6 and $\nu_8-\nu_{17}$, which are available in a recent high-resolution infrared study.⁶⁰

In Table 5 we report fundamental frequencies and one combination band calculated with a 3M potential obtained as ADGA[2hx3M] where 1M:CCSD(T)/cc-pVTZ//2hx3M:CCSD/cc-pVTZ. The results presented clearly demonstrate that the ADGA combined with the MS techniques is a robust methodology which provides accurate PESs when based on high-quality single-point calculations. The computed frequencies are found within 8 cm^{-1} from the experimental values, and the statistical analysis of the results gives a MAD 3.7 cm^{-1} , respectively. We note in passing that for ν_2 and ν_3 there is significant mixing of the fundamental with a combination band ($\nu_4 + \nu_6$) and an overtone ($2\nu_5$), respectively. As observed by De Oliveira et al., these mixings are so severe that assignment of any of the observed peaks to match a pure fundamental is inappropriate.⁶¹

In addition, extrapolation of the 3-mode coupling is addressed by means of ADGA[2gx3M] potentials where 1M:CCSD(T)/cc-pVTZ//2gx3M:MP2/cc-pVTZ or 1M:CCSD(T)/cc-pVTZ//2gx3M:CCSD/cc-pVTZ. In all cases a good agreement with the experimental values is found (MAD 3.0–3.5; MaxAD ca. 10.0 cm^{-1}), confirming that the extrapolated 3M PEFs using only first-derivative information are sufficiently accurate to be highly useful. It is worth mentioning that a corresponding vibrational calculation leaving out extrapolated 3-mode couplings gives significantly less accurate results, as evident from the MAD and MaxAD, which amounts to 10.5 and 62.5 cm^{-1} , respectively. In

Table 5. Methanimine: Calculated Frequencies with Respect to the Experimental Counterpart^a

mode	symmetry	experimental ^b	calculated ^c ADGA[2hx3M]	calculated ^d ADGA[2gx3M]	calculated ^e ADGA[2gx3M]
ν_1	A'	3262.6	3265.2 (2.6)	3263.4 (0.8)	3263.4 (0.8)
$\nu_2 - \nu_4\nu_6$	A'	3024.5	3032.3 (7.8)	3024.2 (-0.3)	3027.7 (3.2)
$\nu_4\nu_6 + \nu_2$	A'		2965.8 (-)	2957.9 (-)	2954.9 (-)
$\nu_3 + 2\nu_5$	A'	2914.2	2916.8 (2.6)	2907.9 (-6.3)	2906.0 (-8.2)
$2\nu_5 - \nu_3$	A'	2885.0	2889.4 (4.4)	2885.1 (0.1)	2885.7 (0.7)
ν_4	A'	1638.3	1634.6 (-3.7)	1632.5 (-5.8)	1632.3 (-6.0)
ν_5	A'	1452.0	1451.9 (-0.1)	1449.8 (-2.2)	1449.4 (-2.6)
ν_6	A'	1344.3	1351.1 (6.8)	1346.8 (2.5)	1345.7 (1.4)
ν_7	A'	1058.2	1059.4 (1.2)	1057.4 (-0.8)	1057.0 (-1.2)
ν_8	A''	1127.0	1131.3 (4.3)	1125.6 (-1.4)	1125.8 (-1.2)
ν_9	A''	1060.8	1056.7 (-4.1)	1050.8 (-10.0)	1050.7 (-10.1)
MAD			3.7	3.0	3.5
MaxAD			7.8	10.0	10.1

^a The discrepancy with respect to the experimental values is given in parentheses; Mean absolute deviation (MAD) and maximum absolute deviation (MaxAD) are reported. The values are given in cm^{-1} . ^b Experimental values from ref 59. ^c Vibrational calculation: VCC[2pt3] on a ADGA[2hx3M] potential (1M:CCSD(T)/cc-pVTZ//2hx3M:CCSD/cc-pVTZ see text for details). ^d Vibrational calculation: VCC[2pt3] on a ADGA[2gx3M] potential (1M:CCSD(T)/cc-pVTZ//2gx3M:CCSD/cc-pVTZ see text for details). ^e Vibrational calculation: VCC[2pt3] on a ADGA[2gx3M] potential (1M:CCSD(T)/cc-pVTZ//2gx3M:MP2/cc-pVTZ see text for details).

addition, the physical nature of the converged states can be completely different: In the case of the ν_2 fundamental excitation, the two-mode coupling potential is not capable of providing the nonzero coupling that is required to obtain the significant mixing that is found using the 3M surfaces.

Overall, our results show similar accuracy as the data obtained by Rauhut with vibrational configuration interaction VCI(SDTQ) on a 3M potential based on (large basis set and F12) CCSD(T)⁶² and by De Oliveira et al. with vibrational perturbation theory on a high-quality CCSD(T) quartic force field.⁶¹

For oxazole, the hybrid potential ADGA[2gx3M] where 1M:CCSD(T)/cc-pVTZ//2gx3M:MP2/cc-pVTZ has been constructed. Such a PES requires ca. 170 single-point evaluations at CCSD(T)/cc-pVTZ level of theory for the monodimensional terms, while about 11 600 MP2/cc-pVTZ single-point (including gradient) calculations are needed for the 2gx3M part.

The fundamental vibrational frequencies of oxazole computed at the VCC[2pt3] level of theory are shown in Table 6. The comparison with the experimental data is fairly good, and although the MAD (7.1 cm^{-1}) is slightly larger than in the case of methanimine, the largest error is limited to -15.5 cm^{-1} for the ν_9 frequency. For comparison, it is worth mentioning that the frequencies computed by VCC[2pt3] using the ADGA[1M] PES, constructed with CCSD(T)/cc-pVTZ single-point calculations, have a MaxAD with respect to the experimental results in excess of 108 cm^{-1} . If the extrapolated 3M corrections are excluded from the hybrid PES ADGA [2M] (1M:CCSD(T)/cc-pVTZ//2M:MP2/cc-pVTZ), a maximum discrepancy of 47.7 cm^{-1} is found. Once again, this confirms the importance of the inclusion of the 3M PEFs in the approximated PES, and it corroborates the thesis that the 3M corrections extrapolated by means of gradient information provide a qualitative improvement over the 2M PES.

A summary for the combined effect of using approximated treatment of the triple excitations in the wave function part and approximated 3-mode couplings in the potential is given in Table 7.

Table 6. Oxazole: Calculated Fundamental Frequencies and Comparison with the Experimental Data^a

mode	symmetry	ADGA[2g×3M]	
		exp ^b	calcd ^c
ν_1	A'	3170.0	3174.9 (4.9)
ν_2	A'	3144.0	3158.6 (14.6)
ν_3	A'	3141.0	3142.6 (1.6)
ν_4	A'	1537.0	1534.8 (-2.2)
ν_5	A'	1504.0	1492.4 (-11.6)
ν_6	A'	1329.8	1319.0 (-10.8)
ν_7	A'	1252.0	1239.3 (-12.7)
ν_8	A'	1142.5	1136.7 (-5.8)
ν_9	A'	1091.1	1075.6 (-15.5)
ν_{10}	A'	1081.3	1084.3 (3.0)
ν_{11}	A'	1051.8	1044.2 (-7.6)
ν_{12}	A'	909.3	903.4 (-5.9)
ν_{13}	A'	899.3	892.9 (-6.4)
ν_{14}	A''	858.2	851.7 (-6.5)
ν_{15}	A''	832.0	821.5 (-10.5)
ν_{16}	A''	749.3	747.0 (-2.3)
ν_{17}	A''	646.4	642.1 (-4.3)
ν_{18}	A''	607.0	605.4 (-1.6)
MAD			7.1
MaxAD			15.5

^a The discrepancy with respect to the experimental values is given in parentheses; mean absolute deviation (MAD) and maximum absolute deviation (MaxAD) are reported. The values are given in cm^{-1} . ^b Experimental values from ref 59. ^c Vibrational calculation: VCC[2pt3] on a ADGA[2gx3M] potential (1M:CCSD(T)/cc-pVTZ//2gx3M:MP2/cc-pVTZ; see text for details).

The data confirms that the level of complexity in the construction of the PES and in the vibrational calculation has to match somewhat from an accuracy/efficiency perspective, e.g., the results indicate that for a given potential including up to m -mode coupling, the vibrational coupled cluster parametrization should include at least m levels of excitations to provide a converged result. The most consistent improvement is obtained when both excitation space and potential is improved from the two-mode coupling to three-mode coupling. This observation offers a useful rule of thumb for the definition of a cost-effective set up for vibrational calculations. On the other hand, we previously argued that approximate inclusion of three-mode couplings may be good enough in the potential and in the wave function as also supported by the summarized data. In this respect, combina-

Table 7. Effect of the 3-Mode Couplings in the Wavefunction and PES^a

	VCC[2]	VCC[2pt3]	VCC[3]
	methanimine		
ADGA[2M] ^b	10.5 (63.3)	10.7 (63.7)	10.7 (63.7)
ADGA[2gx3M] ^c	3.9 (14.3)	3.5 (10.1)	3.3 (11.3)
	oxazole		
ADGA[2M] ^b	8.3 (47.7)	8.0 (47.7)	
ADGA[2gx3M] ^c	5.9 (24.1)	7.1 (15.5)	

^a The discrepancy with respect to the experimental values is given; mean absolute deviations (maximum absolute deviation) are reported. The values are given in cm⁻¹. ^b 1M:CCSD(T)/cc-pVTZ//2M:MP2/cc-pVTZ. ^c 1M:CCSD(T)/cc-pVTZ//2gx3M:MP2/cc-pVTZ.

tion of the VCC[2pt3] parametrization of the wave function with a potential including approximate 3-mode couplings appears very promising. We note in passing that the above comments applies for vibrational coupled-cluster response calculations of fundamental vibrations. The excitation space for a vibrational configuration interaction response calculation or similarly, when the VCI excitation space is defined relative to the ground state, must be larger to obtain a similar accuracy. For higher excited states and combination bands higher excitation levels are generally expected to be more important also for VCC calculations.

5. Summary

We described the combined use of the adaptive density-guided approach and the MS interpolation/extrapolation techniques in construction of potential energy surfaces for vibrational calculations. The procedure has been applied to three molecules (methanimine, trisubstituted methane, and oxazole) in exploratory calculations and in calculations aiming at comparison with experiment. We have shown that the ADGA in concert with MS interpolation and extrapolation and a multiresolution approach provides a cost-effective route to access high-quality PES for use in explicit anharmonic vibrational calculations. Comparison with experimental data for methanimine and oxazole using ADGA[2gx3M] and ADGA[2hx3M] potentials constructed on the basis of CCSD(T), CCSD, and MP2 calculations give encouraging results.

It was shown that purely extrapolated higher order mode-combination terms capture a large part of the contribution to the fully converged surfaces. A few more detailed conclusions are in place: The accuracy of the two-mode couplings from the 1M grids is remarkable if Hessians are available but not useful when only gradients are used in the extrapolation. On the other hand, in the extrapolation from two- to three-mode couplings very good results are found in either case. From a computational point of view, it seems that the cost associated with calculation of the second derivatives makes the extrapolation using Hessians for three and higher mode couplings too costly to be competitive in the present setup. The use of gradients, on the other hand, is cost efficient and provides a substantial improvement relative to complete neglect of three-mode couplings. Taking into account both accuracy and efficiency the following practical hierarchy of PES construction protocols are suggested for further studies in the context of ADGA: ADGA[1M],

ADGA[1hx2M], ADGA[2M], ADGA[2gx3M], ADGA[3M], ADGA[3gx4M], ... In addition, we also anticipate that the gradient-based extrapolated representation of high mode couplings can be used in future studies using screening techniques of some sort, e.g., the extrapolated surfaces are used as they are for most couplings, while those few couplings that were estimated to be particularly important are further refined by explicit construction.

Certainly, the possibility to obtain three-mode couplings with only two-mode grids is a unique opportunity especially if one considers that gradient information is available with very little extra cost when computing first-order properties (e.g., dipole moments) besides the energy. For systems whose size usually precludes a robust ab initio treatment including three-mode couplings the combined use of, e.g., ADGA[2gx3M] and VCC[2pt3] as done here may turn out to be the only realistic option. Thus, it is hoped that this work can pave the way for more accurate calculations on larger molecules, where calculations have been limited to approaches including only up to two-mode couplings.

Acknowledgment. This work was supported by the Lundbeck Foundation, the Danish National Research Foundation, the Danish Center for Scientific Computing (DCSC), and EUROHORCs through a EURYI award.

References

- (1) Bowman, J. M. *J. Chem. Phys.* **1978**, *68*, 608.
- (2) Bowman, J. M. *Acc. Chem. Res.* **1986**, *19*, 202.
- (3) Gerber, R. B.; Ratner, M. A. *Adv. Chem. Phys.* **1988**, *70*, 97.
- (4) Hansen, M. B.; Sparta, M.; Seidler, P.; Christiansen, O.; Toffoli, D. *J. Chem. Theory Comput.* **2010**, *6*, 235.
- (5) Norris, L. S.; Ratner, M. A.; Roitberg, A. E.; Gerber, R. B. *J. Chem. Phys.* **1996**, *105*, 11261.
- (6) Jung, J. O.; Gerber, R. B. *J. Chem. Phys.* **1996**, *105*, 10332.
- (7) Chaban, G. M.; Jung, J. O.; Gerber, R. B. *J. Chem. Phys.* **1999**, *111*, 1823.
- (8) Christiansen, O. *J. Chem. Phys.* **2003**, *119*, 5773.
- (9) Matsunaga, N.; Chaban, G. M.; Gerber, R. B. *J. Chem. Phys.* **2002**, *117*, 3541.
- (10) Yagi, K.; Hirata, S.; Hirao, K. *J. Chem. Phys.* **2007**, *127*, 034111.
- (11) Bowman, J. M.; Christoffel, K.; Tobin, F. *J. Phys. Chem.* **1979**, *83*, 905.
- (12) Christoffel, K. M.; Bowman, J. M. *Chem. Phys. Lett.* **1982**, *85*, 220.
- (13) Carter, S.; Bowman, J. M.; Handy, N. C. *Theor. Chem. Acc.* **1998**, *100*, 191.
- (14) Christiansen, O. *J. Chem. Phys.* **2004**, *120*, 2149.
- (15) Rauhut, G. *J. Chem. Phys.* **2004**, *121*, 9313.
- (16) Begue, D.; Gohaud, N.; Pouchan, C.; Cassam-Chenai, P.; Lievin, J. *J. Chem. Phys.* **2007**, *127*, 164115.
- (17) Christiansen, O. *J. Chem. Phys.* **2005**, *122*, 194105.
- (18) Seidler, P.; Christiansen, O. *J. Chem. Phys.* **2007**, *126*, 204101.

- (19) Carter, S.; Culik, S. J.; Bowman, J. M. *J. Chem. Phys.* **1997**, *107*, 10458.
- (20) Gerber, R.; Jung, J. In *The vibrational self-consistent field approach and extensions: Method and applications to spectroscopy of large molecules and clusters*; Jensen, P., Bunker, P. R., Eds.; Wiley: Chichester, 2000; pp 365–390.
- (21) Bowman, J. M.; Carter, S.; Huang, X. C. *Int. Rev. Phys. Chem.* **2003**, *22*, 533.
- (22) Benoit, D. M. *J. Chem. Phys.* **2004**, *120*, 562, and references therein.
- (23) Kongsted, J.; Christiansen, O. *J. Chem. Phys.* **2006**, *125*, 124108.
- (24) Yagi, K.; Hirata, S.; Hirao, K. *Theor. Chem. Acc.* **2007**, *118*, 681.
- (25) Toffoli, D.; Kongsted, J.; Christiansen, O. *J. Chem. Phys.* **2007**, *127*, 204106.
- (26) Scribano, Y.; Benoit, D. *J. Chem. Phys.* **2007**, *127*, 164118.
- (27) Hirata, S.; Yagi, K.; Perera, S.; Yamazaki, S.; Hirao, K. *J. Chem. Phys.* **2008**, *128*, 214305.
- (28) Sparta, M.; Toffoli, D.; Christiansen, O. *Theor. Chem. Acc.* **2009**, *123*, 413.
- (29) Sparta, M.; Høyvik, I.-M.; Toffoli, D.; Christiansen, O. *J. Phys. Chem. A* **2009**, *113*, 8712.
- (30) Jakowski, J.; Sumner, I.; Iyengar, S. S. *J. Chem. Theory Comput.* **2006**, *2*, 1203.
- (31) Sumner, I.; Iyengar, S. S. *J. Phys. Chem. A* **2007**, *111*, 10313.
- (32) Manzhos, S.; Carrington, T. *J. Chem. Phys.* **2006**, *125*, 084109.
- (33) Manzhos, S.; Carrington, T., Jr. *J. Chem. Phys.* **2008**, *129*, 224104.
- (34) Dawes, R.; Thompson, D. L.; Guo, Y.; Wagner, A. F.; Minkoff, M. *J. Chem. Phys.* **2007**, *126*, 184108.
- (35) Dawes, R.; Thompson, D. L.; Wagner, A. F.; Minkoff, M. *J. Chem. Phys.* **2008**, *128*, 084107.
- (36) Matito, E.; Toffoli, D.; Christiansen, O. *J. Chem. Phys.* **2009**, *130*, 134104.
- (37) Rauhut, G.; Hartke, B. *J. Chem. Phys.* **2009**, *131*, 014108.
- (38) Seidler, P.; Christiansen, O. *J. Chem. Phys.* **2009**, *131*, 234109.
- (39) Seidler, P.; Matito, E.; Christiansen, O. *J. Chem. Phys.* **2009**, *131*, 034115.
- (40) Watson, J. K. G. *Mol. Phys.* **1968**, *15*, 479.
- (41) Except for the fact that the multidimensional grids are not allowed to extend beyond the boundaries defined by the corresponding monodimensional ones.
- (42) Hrenar, T.; Werner, H.; Rauhut, G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3123.
- (43) Pfluger, K.; Paulus, M.; Jagiella, S.; Burkert, T.; Rauhut, G. *Theor. Chem. Acc.* **2005**, *114*, 327.
- (44) Rodriguez-Garcia, V.; Hirata, S.; Yagi, K.; Hirao, K.; Taketsugu, T.; Schweigert, I.; Tasumi, M. *J. Chem. Phys.* **2007**, *126*, 124303.
- (45) Yagi, K.; Taketsugu, T.; Hirao, K. *J. Chem. Phys.* **2002**, *116*, 3963.
- (46) Yagi, K.; Oyanagi, C.; Taketsugu, T.; Hirao, K. *J. Chem. Phys.* **2003**, *118*, 1653.
- (47) Oyanagi, C.; Yagi, K.; Taketsugu, T.; Hirao, K. *J. Chem. Phys.* **2006**, *124*, 064311.
- (48) Evenhuis, C.; Manthe, U. *J. Chem. Phys.* **2008**, *129*, 024104.
- (49) Carbonniere, P.; Begue, D.; Dargelos, A.; Pouchan, C. *Chem. Phys.* **2004**, *300*, 41.
- (50) MidasCpp (Molecular Interactions, dynamics and simulation Chemistry program package in C++), 2007; <http://www.chem.au.dk/midas>.
- (51) Matito, E.; Barroso, J. M.; Besalú, E.; Christiansen, O.; Luis, J. M. *Theor. Chem. Acc.* **2009**, *123*, 41.
- (52) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Headgordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.
- (53) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.
- (54) Kendall, R.; Dunning, T.; Harrison, R. *J. Chem. Phys.* **1992**, *96*, 6769.
- (55) Woon, D.; Dunning, T. *J. Chem. Phys.* **1993**, *98*, 1358.
- (56) Stanton, J.; Gauss, J.; Harding, M.; Szalay, P. *CFOUR, Coupled-Cluster techniques for Computational Chemistry, a quantum-chemical program package*; 2009; see <http://www.cfour.de>.
- (57) Halonen, L.; Duxbury, G. *J. Chem. Phys.* **1985**, *83*, 2091.
- (58) Halonen, L.; Duxbury, G. *Chem. Phys. Lett.* **1985**, *118*, 246.
- (59) Johnson III, R. D. *NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101 Release 14*, 2006; <http://srdata.nist.gov/cccbdb>.
- (60) Hegelund, F.; Larsen, R. W.; Palmer, M. H. *J. Mol. Spectrosc.* **2007**, *241*, 26.
- (61) De Oliveira, G.; Martin, J.; Silwal, I.; Liebman, J. J. *Comput. Chem.* **2001**, *22*, 1297.
- (62) Rauhut, G.; Knizia, G.; Werner, H. *J. Chem. Phys.* **2009**, *130*, 054105.

JCTC

Journal of Chemical Theory and Computation

A Theoretical Study of Brominated Porphycenes: Electronic Spectra and Intersystem Spin–Orbit Coupling

Angelo Domenico Quartarolo, Sandro Giuseppe Chiodo, and Nino Russo*

Dipartimento di Chimica and Centro di Calcolo ad Alte Prestazioni per Elaborazioni Parallele e Distribuite—Centro d’Eccellenza MIUR, Università della Calabria, I-87030 Arcavacata di Rende, Italy

Received June 1, 2010

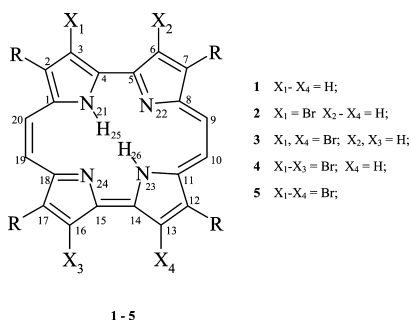
Abstract: In this paper, a time-dependent density functional theoretical study (TDDFT) has been carried out for brominated 2,7,12,17-tetra-*n*-propylporphycenes. Their potential therapeutic use in photodynamic therapy (PDT), a noninvasive medical treatment of cancer diseases, is due to the strong absorbance in the red part of the visible spectrum and the presence of heavy atoms (bromine). The prediction of electronic spectra for photosensitizer molecules can be a valuable tool in the design of drugs for application in PDT. Singlet and triplet vertical excitation energies have been calculated by means of the nonempirical hybrid functional PBE0 in conjunction with a split valence basis set (SVP), on previously optimized, at the density functional level of theory, ground state geometries. In particular, the quantum-chemical simulation of their absorption electronic spectra, both *in vacuo* and in solvent environments (dichloromethane and bromobenzene), has evidenced the red shift maxima wavelengths for the Q bands (or lower energy bands) with an increasing number of bromine atoms, in agreement with experimental results. The mean absolute deviation for the Q-electronic bands is about 0.3 eV. Calculated vertical triplet energies are between 1.04 (for tetra-brominated derivative) and 1.20 eV (for dibrominated derivative). The influence of bromine atoms on intersystem spin crossing has been investigated by applying a computational code which calculates spin–orbit matrix elements between singlet and triplet excited state wave functions weighted by the TDDFT transition coefficients.

1. Introduction

Porphycene is a nonporphyrin tetrapyrrolic macrocycle and the first constitutional isomer of porphyrin reported in the literature, having an 18 π -electron planar conjugated system.¹ This compound was first synthesized by Vogel and co-workers in 1986 by means of the McMurry coupling of bipyrrrole dialdehydes.^{2,3} Although the synthesis of porphycene is generally low-yielding, its unique optical features and in particular the strong absorption in the far-red part of the visible region have attracted chemical researchers to the synthesis of new derivatives for application in the medical area of photodynamic therapy (PDT).⁴ PDT is a noninvasive technique for the treatment of different kinds of tumors (e.g.,

early lung, breast, and prostate cancers), age-related macular degeneration (AMD), and skin diseases.^{5–8} The combination of a laser light source, a photosensitizer drug dose, previously injected in the human body, and the presence of molecular oxygen $^3\text{O}_2$ ($^3\Sigma_g^-$), naturally present in human tissues, can induce apoptosis and/or necrosis of tumoral cells as a result of *in locu* generated cytotoxic oxygen species.^{9,10} The mechanism of generation of singlet oxygen $^1\text{O}_2$ ($a^1\Delta_g$) as the key cytotoxic agent, which is usually referred to as the type II PDT mechanism, involves a cascade of photochemical steps.^{11–13} First, the photosensitizer in its ground state S_0 is excited to the first excited state S_1 state by irradiation with a wavelength, preferably in the so-called therapeutic window between 600–900 nm, that should overlap the maximum absorption wavelength of the sensitizer.¹⁴ Red-shifted absorption wavelengths are preferred since they better penetrate

* To whom correspondence should be addressed. E-mail: nrusso@unical.it.

Scheme 1. Molecular Structures and Atom Numbering

human tissues, allowing the treatment of deeper tumors. In the type II mechanism, the S_1 state of the sensitizer decays to the first triplet excited state (T_1), through a radiationless transition or intersystem spin crossing process (isc), and the energy gain is then transferred to molecular oxygen, forming the highly reactive species 1O_2 . The latter step requires an activation energy of about 0.98 eV, which corresponds to the $^3\Sigma_g^- \rightarrow ^1\Delta_g$ electronic transition for molecular oxygen.¹⁵ The efficiency of a photosensitizer as a PDT drug mainly depends on its electronic absorption spectrum features, such as maximum wavelength positions and intensities, and on the ability to generate 1O_2 . The latter factor is measured by the singlet oxygen quantum yield (Φ_Δ) of the sensitizer and is related to the electron transfer process. Other structural factors, like the photosensitizer amphiphilic character or the possibility to form aggregates in aqueous solution which decrease the photodynamic action, play an important role in the PDT drug design. The attempt to increase the value of Φ_Δ can be searched for by introducing heavy atoms, such as halogens (Br, I) or transition metal atoms in the molecular structure. The presence of a heavy atom (high atomic number) causes the mixing of pure electronic states of different spin multiplicities (the so-called heavy atom or spin-orbit coupling effect), increasing the rate of isc between S_1 and T_1 states ($S_1 \rightarrow T_1$ radiationless transition).¹⁶⁻¹⁸ When the triplet lifetime is sufficiently long-lived (typically on the order of few microseconds) in order to avoid deactivation by solvent molecular collisions, the energy transfer from the T_1 photosensitizer state to 3O_2 is favored, and consequently Φ_Δ increases. Currently, photosensitizers approved for clinical use belong to the porphyrin-like class of molecules.^{19,20} Photophrin, a mixture of hematoporphyrin monomers, dimers, and oligomers, was the first accepted PDT drug for the treatment of early stage lung cancer.²¹ Other synthetic photosensitizers like *m*-tetrahydroxyphenylchlorin (mTHPC, Foscan)²² and lutetium texaphyrin (Lutrin)^{23,24} have been accepted for clinical use. However, also, the synthesis of new nonporphyrin potential PDT drug (e.g., phenotiazinium, tetra-aryl-azadipyrromethenes, or hypericin derivatives) represents an important growing research field.²⁵⁻²⁹ In this paper, we will present a theoretical study of the structures and photophysical properties of brominated 2,7,12,17-tetra-*n*-propylporphycenes (Scheme 1, structures **2–5**) along with the corresponding unsubstituted molecule (Scheme 1, structure **1**). The chemical synthesis of the brominated compounds, described in previous papers, provides the formation of mono-, di-, tri-, and tetra-brominated porphycenes ac-

ording to the amount of bromine added to compound **1**.³⁰⁻³³ The experimental study by Shimakoshi et al. of the photo-physical properties (absorption, fluorescence, and phosphorescence spectra as well as their quantum yields) revealed the dependence of these properties on the number of bromine atoms.³³ In particular, the rate of isc (k_{isc}), which measures the efficiency of the $S_1 \rightarrow T_1$ radiationless transition, is at a maximum for the tetra-brominated porphycene and increases with the number of bromine atoms, confirming the role of the heavy atom effect on these compounds. The experimental values of Φ_Δ for the brominated derivatives range from 0.49 (**5**) to 0.95 (**2**) and are higher than the corresponding value of the unsubstituted structure **1** ($\Phi_\Delta = 0.36$).^{33,34} Theoretical calculations can usually predict and rationalize electronic spectra and wavelength shift depending upon the nature of the substituent. This fact can be useful for the molecular design of red-shifted PDT photosensitizers, taking advantage from the prediction of absorption electronic spectra by means of time-dependent density functional theory (TDDFT).³⁵ In the past decade, this theoretical methodology for electronic excited states has become an efficient and routine tool for predicting electronic spectra even for large molecules.³⁶⁻³⁸ Moreover, in the computational approach for the study of electronic excited states, it is also possible to have theoretical insights about the isc mechanism. In fact it is well-known from the application of perturbation theory to radiationless transitions that the $S_1 \rightarrow T_1$ transition depends quadratically on the corresponding matrix element of the spin-orbit quantum operator H_{so} . For this purpose, a computational code has been developed in our lab, for the calculation of the H_{so} matrix elements between two reference wave functions with different spin multiplicities. In our case, we are interested particularly in the S_1 and T_1 electronic excited states. In a first approximation, the knowledge of these elements can be correlated to the constant rate k_{isc} , giving some preliminary hints about the photosensitizer's ability to populate the triplet state and act as a PDT drug.

2. Theoretical Approach for Spin-Orbit Contributions

One particular application of our recently developed^{39,40} method is an occasionally useful way to evaluate the SO matrix elements between the S_i and T_j states. In the following, we shortly describe the strategy, actually in use by us, for calculating these contributions. A good hint is furnished by this expression:

$$\langle S_i | H_{SO} | T_j \rangle = \sum_l^{N_{S_i}} \sum_m^{N_{T_j}} C_{il}^S C_{jm}^T \langle \Psi_{il}^S | H_{SO} | \Psi_{jm}^T \rangle \quad (1)$$

where Ψ^S and Ψ^T are the singlet and triplet state wave functions, respectively, arising from one-electron vertical excitations performed over the ground state (S_0) electronic configuration, C_{il}^S and C_{jm}^T are the coefficients of the l th one-electron singlet transition and m th one-electron triplet transition belonging to the S_i and T_j states, respectively. N_{S_i} and N_{T_j} are the number of transitions defining the i th and j th singlet and triplet states, respectively. H_{SO} is the full Breit-Pauli operator.⁴¹ More advanced discussion about the

approach used to compute the SO matrix elements between the Ψ^S and Ψ^T wave functions can be found elsewhere.^{38,39} However, the used approximation that is entered in the expression (eq 1) is based on the TD-DFT assignment ansatz,³⁴ which has been shown to be exact to linear order for some matrix elements between ground and excited states.^{42,43} But to achieve accuracy up to linear order, the second order response of the density has to be taken into account.⁴⁴ However, this ansatz is widely used for matrix elements between excited states with satisfactory results.⁴⁵

Notice that each set of the coefficients C_{ii}^S and C_{jm}^T , of which their squared values symbolically are reported in Tables 4 and 6 as c^2 , comes from TD-DFT calculations.

Not as well appreciated as it ought to be is the fact that, despite the size of the molecules investigated, this procedure uses the full and exact Breit–Pauli operator instead of an approximate one. But the expression (eq 1) could not necessarily be the most efficient or fastest executing one, due to the large number of terms that could be enclosed. Thus, one can, in principle, easily find a good approximation by simply truncating the expression (eq 1) to a few terms, those belonging to larger values of the weighted coefficients, making it easier to evaluate. In this work, we have not used the full set of coefficients and the corresponding transitions defining the singlet, Ψ^S , and triplet, Ψ^T , wave functions but the main configurations belonging to the higher values of their squared coefficient, c^2 , as listed in Tables 4 and 6.

It is worth emphasizing that the strategy for finding SO matrix elements between S_i and T_j states assumes that singlet, Ψ^S , and triplet, Ψ^T , wave functions are built from molecular orbitals (MO) of the ground state S_0 , without invoking all the machinery beyond the separate optimization of these wave functions. Their electronic configurations are defined according to the TD-DFT results by simply promoting one electron from an occupied orbital to an unoccupied one.

Open shell singlets are obtained as a combination of $\alpha\beta$ and $\beta\alpha$ wave functions:

$$2^{-1/2}[\alpha\beta - \beta\alpha] \quad (2)$$

and, concerning the triplets, as $\alpha\alpha$ and $\beta\beta$ wave functions and, as before, by a combination of $\alpha\beta$ and $\beta\alpha$ ones:

$$\alpha\alpha, 2^{-1/2}[\alpha\beta + \beta\alpha], \beta\beta \quad (3)$$

Supposing there is *a priori* knowledge about previous works concerning our SO basic method,³⁹ often it is a workable procedure to biorthogonalize the MOs of different states coming from separate optimizations. But here, occasionally, we have switched off the biorthogonalization, since, as hinted, the MOs of the singlet, Ψ^S , and triplet, Ψ^T , wave functions, being unaltered from the S_0 ground state wave function, are already biorthogonals. Nevertheless, in doing so, it is needed, very often, to reorder the MOs in such a way that the discoincident orbitals, the orbitals involved in the coupling mechanism, must be shifted to external ones, as HOMO orbitals (without changing the MOs' occupation of the singlet, Ψ^S , and triplet, Ψ^T , wave functions), as required in the implementation of MOLSOC code.^{39,40} But, due to this operation, it is absolutely necessary

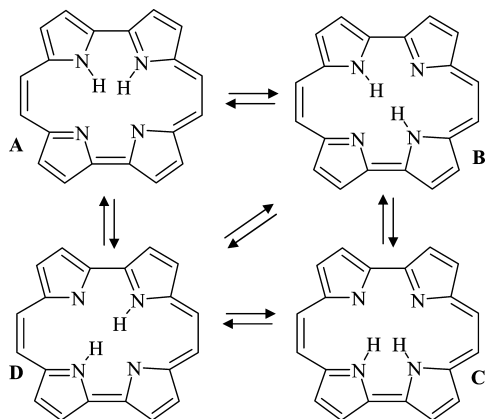
to assign the right sign for the wave function, because of the antisymmetrization principle.

3. Computational Details

Structure optimizations and excitation energies were calculated by means of the TURBOMOLE quantum chemistry program (version V5.10).⁴⁶ Gas-phase ground state geometry optimizations were carried out at the density functional level of theory. For that purpose, the nonempirical PBE0 hybrid functional was employed, which includes a fixed amount (1/4) of the exact Hartree–Fock exchange energy to the gradient-corrected PBE exchange–correlation functional.^{47,48} For structure optimization calculations, the Stuttgart effective-core quasi-relativistic pseudopotential (SDD, 28 core electrons) and the corresponding optimized basis set for valence electrons (ecp-28-mwb: (6s6p1d)/[3s3p1d]) were assigned to Br atoms.⁴⁹ The standard split valence basis set, with one set of polarization functions added, was used for C, N (SVP: (7s4p1d)/[3s2p1d]), and H (SVP: (4s1p)/[2s1p]).⁵⁰ Pseudopotential and basis sets were taken from the TURBOMOLE basis set library. A vibrational frequency analysis performed at the same level of theory confirmed the stationary points found as potential energy surface minima.⁵¹ Single point energy calculations, for the *cis* and *trans* optimized tautomers of Scheme 1, were done with the recently developed double-hybrid functional B2PLYP in conjunction with the resolution of identity approximation and the polarized valence triple- ζ basis set (TZVP).^{52,53} The B2-PLYP functional gives high accuracy for noncovalent interactions with a mean absolute deviation between 0.2 and 0.3 kcal/mol and can better assess the most stable tautomer.⁵⁴ Linear response properties (singlet and triplet excitation energies) were calculated on the ground state equilibrium structures, by means of time-dependent density functional theory (TD-DFT)⁵⁵ and the same basis set (SVP) for all atoms (Br, N, C, H). The simulation of the UV–vis electronic spectra band shape was made by convolution of the first 20 singlet excitation energy roots with Gaussian functions having a constant full-width at half-maximum of 0.2 eV, using the SWizard program.⁵⁶ Bulk solvation effects on ground state geometries and excitation energies were included by means of the conductor-like polarizable model (COSMO).⁵⁷ For this reason, the dielectric constant of dichloromethane ($\epsilon = 8.93$) with a solvent radius of 2.27 Å was manually set, while optimized atomic radii and other default parameters for the cavity construction were taken from the COSMO module. For the calculation of triplet energies, bromobenzene ($\epsilon = 5.4$) has also been considered as a solvent, doing single point energy calculations on the gas phase optimized structures. Some calculations of excitation energies have also been performed with two recently developed hybrid meta-GGA functionals (BMK and M06–2X)^{58,59} and the GAUSSIAN 03 program.⁶⁰

SO matrix elements have been evaluated using the MolSOC^{39,40} code interfaced with the TURBOMOLE program package.⁴⁶ The *mos* file of TURBOMOLE containing the coefficients of the MOs, the basis set file, and the input file, containing a route section (specifying the keywords), a charge and multiplicity section, and a geometry section are necessary for carrying out SO calculation with MolSOC. As

Scheme 2. Tautomeric Equilibria between *trans* (**B** and **D**) and *cis* (**A** and **C**) Configurations in Porphycene



mentioned above, the MOs of the Ψ^S and Ψ^T wave functions (expression 1) are those of the ground state, S_0 . These have been optimized at the PBE0/SVP level.

4. Results and Discussion

The following discussion is divided into four parts containing (a) the molecular structure and energetic aspects of brominated porphycenes (section 4.1), (b) one-electron absorption electronic spectra, (c) triplet energies (section 4.2), and (d) the efficiency of the $S_n \rightarrow T_n$ intersystem spin crossing mechanism as derived from spin-orbit matrix elements analysis (section 4.3).

4.1. Ground State Structure Properties. Porphycenes display a reduced inner cavity in comparison to porphyrins, and for that reason the inner pyrrolic hydrogens can undergo faster ground-state tautomerization between the adjacent (ethylenic bridged) pyrrole rings (see Scheme 2).⁶¹ This process is also favored by the formation of strong intramolecular hydrogen bonds (as for example between $N_{21}-H_{25} \cdots N_{24}$ atoms of Scheme 1). The proposed mechanism of tautomerization in porphyrins, as deduced by nuclear magnetic resonance (NMR) studies, implies the interconversion between the *trans* and *cis* conformations of the four inner hydrogens, through a sequence of one-step processes.⁶² The thermal activated hydrogen transfer between opposite pyrrole rings, leading to *trans* to *cis* conversion, can be followed by the formation of the other *trans* conformation or a reversal process to the initial *trans* structure. An equivalent representation of this process, describing the interconversion between tautomeric forms in porphycene, is reported in Scheme 2. The extent of the interconversion mechanism mainly depends on the energetic barriers among the different tautomeric forms. In order to assess the most stable tautomer to adopt for all subsequent response property calculations, an energetic analysis was performed for the *cis* and *trans* tautomeric forms of compounds **1–5**. All four tautomers (both *trans* and *cis* species) are found to be local energy minima. The absolute (atomic units) and relative calculated energies (kcal/mol), with respect to the most stable tautomer found within the mono-, di-, tri-, and tetrabrominated porphycene series (**2–5**) as well as for the tetra-propylporphycene (**1**), are reported in Table 1. From gas-phase PBE0 calculations, the most stable tautomer form was found to

be, in all cases, the *trans* configuration. The energetic trend was also supported by single point energy calculations with the B2PLYP/TZVP approach. Notwithstanding the stability of *trans* molecular structures, the difference between *trans* and *cis* tautomers is lower than 4 kcal/mol. So the energetic barrier of these compounds can be easily overcome, in particular in liquid media, through the energy transfer mechanism due to molecular collisions. A potential energy profile *in vacuo*, relative to the different tautomers of monobrominated porphycene **2**, has been depicted in Figure 2. The connection between *trans* and *cis* ground state equilibrium structures goes through the formation of transition states, which have been found for the *trans* **D** and **B** tautomers to the *cis* **A** and **C** (TS_{D-A} , TS_{D-C} , TS_{B-A} , and TS_{B-C} molecular structures in Figure 1). The transition state structures correspond to a hydrogen-transfer process between the two pyrrolic rings connected by the ethylenic bridge. The interconversion barrier energies, between the *trans* **D** structure to the *cis* **A** and **C** forms, are respectively 4.0 and 3.4 kcal/mol with the corresponding N–H stretching imaginary frequencies of $1030i$ and $1060i$ cm^{-1} . For the transition state TS_{D-A} (likewise for TS_{D-C}), the equilibrium bond lengths found between the exchanged hydrogen and pyrrolic nitrogens are slightly asymmetrical, being respectively 1.235 and 1.331 Å ($N_{21}-H_{25}$ and $N_{24}-H_{25}$ distances of Scheme 1). For *trans* **B** structure interconversion to *cis* **A** and **C**, the energy barriers are respectively 2.66 and 2.99 kcal/mol with imaginary frequencies of $1072i$ and $1112i$ cm^{-1} . The low energetic barriers found are compatible with the simultaneous presence in solution of the *cis* and *trans* forms, with the latter being the dominant tautomeric form. From fluorescence polarization experiments on bare porphycene, the *trans* configuration resulted to be dominant in both S_0 and S_1 states.⁶¹ This technique also allows the determination of tautomerization rates through the analysis of emission anisotropy values. For porphycene, in the S_1 state, a double-hydrogen transfer or *trans–trans* conversion has been observed as the interconversion step and is faster than the *cis–trans* mechanism.⁶³ Double-hydrogen transfer (or tunneling) is also valid for the ground state, as demonstrated in a previous study at cryogenic temperatures for 9-acetoxy-2,7,12,17-tetra-*n*-propylporphycene.⁶⁴

The approach that can be used to explain the tunneling effect depends mainly on three factors: (a) the energy difference between the ground state equilibrium structures (*trans*), which can interconvert through the double-hydrogen transfer reaction, (b) the energetic transition state barrier height, and (c) the distance of pyrrolic hydrogen between the two *trans* structures (or interminimal distance). Low values of each of these factors can promote the tunneling. For the above-mentioned 9-acetoxy-2,7,12,17-tetra-*n*-propylporphycene molecule, lying in the ground state, factors a and b have been estimated to be respectively less than 180 cm^{-1} (0.5 kcal/mol) and 1820 cm^{-1} (5.2 kcal/mol). So despite the asymmetric molecular structure for the two *trans* tautomers, due to the acetoxy substituent, these low values allow hydrogen atoms to tunnel directly through the double-minimum potential energy in competition with the *trans* to *cis* tautomerization step. In the case of porphycene derivative

Table 1. Absolute (hartree) and Relative Electronic Energies (kcal/mol) for *trans* and *cis* Tautomeric Forms of Compounds 1–5, Calculated at the PBE0/(SVP-SDD) and B2PLYP/TZVP Levels of Theory^a

molecule		tautomer 1 ^b	tautomer 2 ^b	tautomer 3 ^c	tautomer 4 ^c
1	PBE0/SVP	-1458.595004 (0.01)	-1458.595017 (0.00)	-1458.592111 (1.82)	-1458.592111 (1.82)
	B2PLYP/TZVP ^d	-1460.247910 (0.0)	-1460.247894 (0.01)	-1460.244038 (2.42)	-1460.244033 (2.42)
2	PBE0/SVP	-1472.128805 (1.526)	-1472.131124 (0.00)	-1472.126145 (3.12)	-1472.127534 (2.25)
	B2PLYP/TZVP	-4033.401372 (1.42)	-4033.403637 (0.00)	-4033.397691 (3.73)	-4033.398977 (2.92)
3	PBE0/SVP	-1485.666347 (0.00)	-1485.661807 (2.85)	-1485.660664 (3.57)	-1485.660664 (3.57)
	B2PLYP/TZVP	-6606.556697 (0.00)	-6606.552295 (2.76)	-6606.549926 (4.25)	-6606.549927 (4.25)
4	PBE0/SVP	-1499.183369 (1.55)	-1499.185844 (0.00)	-1499.179813 (3.78)	-1499.181718 (2.59)
	B2PLYP/TZVP	-9179.696892 (1.44)	-9179.699194 (0.00)	-9179.692054 (4.48)	-9179.694702 (2.82)
5	PBE0/SVP	-1512.703398 (0.00)	-1512.703249 (0.01)	-1512.698968 (2.78)	-1512.698862 (2.75)
	B2PLYP/TZVP	-11752.837861 (0.00)	-11752.837658 (0.13)	-11752.832882 (3.12)	-11752.832710 (3.23)

^a Energy minima are denoted in parentheses in bold character. ^b *trans* configuration. ^c *cis* configuration. ^d Single point energy calculations at PBE0/SVP (H, C, N)-SDD (Br) optimized geometries.

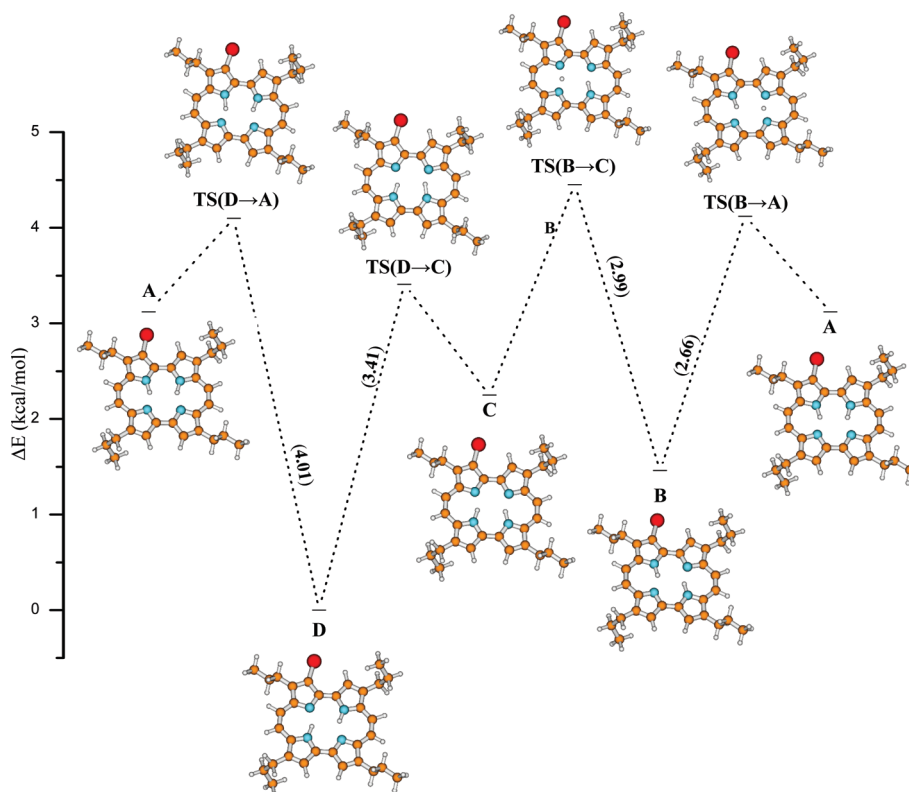


Figure 1. Energy profile and molecular structures (minima and transition states) for the interconversion between the tautomeric forms A–D of Scheme 1 for molecule 2. In parentheses are reported the energy barriers (kcal/mol) between *trans* and *cis* tautomers. The most stable tautomer (structure D) is taken as the zero-energy reference.

2, the evaluation of the energetic barrier for the transition state corresponding to the double-hydrogen transfer can give insights into the possibility of tunneling for this class of brominated porphycenes. Some attempts to find for compound 2 this hypothetical transition state with the two inner hydrogens equidistant from the pyrrolic nitrogens have not been successful. This fact is probably due to the asymmetric chemical environment for the inner hydrogens, which cause

the starting geometry to revert to the $\text{TS}_{\text{D}-\text{C}}$ or $\text{TS}_{\text{B}-\text{C}}$ transition state final geometries. A different strategy adopted for finding an approximate energy value for this hypothetical transition state was to perform a potential energy surface scan versus the two N–H distances. These distances were increased simultaneously by 0.1 Å, starting from the *trans* equilibrium ground state geometry of tautomer D. During the scan energy calculation, pyrrolic nitrogen and hydrogen Cartesian coord-

Table 2. Bond Lengths (Å) and Valence and Dihedral Angles (deg) for Brominated Porphycenes^a

	2		3		4		5	
	calcd	$\Delta^{\text{exp-calcd}}$	calcd	$\Delta^{\text{exp-calcd}}$	calcd	$\Delta^{\text{exp-calcd}}$	calcd	$\Delta^{\text{exp-calcd}}$
Bond Length								
C ₃ -X ₁	1.837	(0.013)			1.827	(0.033)	1.826	(0.05)
C ₆ -X ₂			1.836	(0.039)	1.832	(-0.029)	1.832	(0.042)
C ₁₆ -X ₃			1.836	(0.039)	1.836	(-0.033)	1.832	(0.035)
C ₁₃ -X ₄							1.826	(0.054)
C ₄ -C ₅	1.415	(0.009)	1.413	(-0.005)	1.424	(-0.005)	1.420	(0.006)
C ₁₉ -C ₂₀	1.391	(-0.006)	1.391	(0.002)	1.386	(-0.032)	1.387	(-0.008)
C ₂₀ -C ₁	1.411	(-0.001)	1.403	(0.001)	1.406	(0.01)	1.403	(0.003)
C ₁ -C ₂	1.457	(0.015)	1.443	(-0.006)	1.432	(-0.022)	1.432	(0.007)
C ₂ -C ₃	1.368	(0.013)	1.378	(-0.014)	1.381	(-0.022)	1.380	(-0.009)
N ₁ -C ₁	1.355	(-0.014)	1.360	(0.006)	1.361	(0.013)	1.359	(0.014)
N ₁ -C ₄	1.346	(0.023)	1.357	(0.012)	1.363	(-0.009)	1.363	(-0.007)
MAD ^b	0.01		0.01		0.02		0.02	
rms ^b	0.01		0.02		0.02(5)		0.03	
Bend Angle								
C ₁ -N ₁ -C ₄	107.91	(-0.44)	110.42	(-0.35)	111.33	(-4.04)	111.31	(-3.3)
C ₁₉ -C ₂₀ -C ₁	132.21	(-0.79)	131.15	(-0.27)	131.77	(0.12)	131.01	(1.12)
Dihedral Angles								
C ₃ -C ₄ -C ₅ -C ₆	-0.137	(0.5)	-0.080	(0.7)	-0.707	(3.5)	-20.857	(6.0)
C ₁₃ -C ₁₄ -C ₁₅ -C ₁₆	0.078	(0.6)	0.075	(0.7)	21.863	(17.7)	-20.879	(2.0)

^a In parentheses are reported the deviations ($\Delta^{\text{exp-calcd}}$) from X-ray experimental data. ^b Absolute mean deviation (MAD) and root-mean-square deviation (rms).

dinates were kept fixed while the remaining parameters were fully optimized. The energy versus N-H coordinate plot, reported in the Supporting Information (Figure S1, S14), gave an approximate high energy barrier of about 16 kcal/mol, which is greater than the energy required for the *trans-trans* interconversion mechanism by a two-step transfer mechanism. Similar results have been obtained for compound **5** (Figures S2, Supporting Information), which has a more symmetrical chemical environment and a lower *trans-trans* energy gap. In this case, the energy barrier height is even higher (about 30 kcal/mol).

The energetic stability of the *trans* conformations in comparison to that of the *cis* can be mainly ascribed to different geometrical parameters inside the inner cavity, which in turn can mutually influence the energy contribution of intramolecular hydrogen bonds. For instance, in *trans* conformations, the inner hydrogen distance is slightly greater than in the *cis* by about 0.02 Å; similarly the increase of the opposite nitrogen distance implies an overall minor sterical hindrance inside the cavity. It can be noted from Table 1 that in compounds **1** and **5**, for which inner hydrogen atoms show a similar chemical environment by symmetry, the *trans* tautomers have about the same electronic energy, and the same holds for the *cis* forms. In the same way, the presence of bromine substituents and in particular the proximity to pyrrolic hydrogens can account for the small energy difference between each *trans* or *cis* tautomeric pair by the electron-withdrawing effect, which induces a different electronic charge distribution on the molecular system. A list of some geometrical parameters (bond lengths, valence and torsional angles) for the most energetically stable tautomer of each porphycene derivative (**2-5**) is presented in Table 2. For bond lengths, the agreement with the experimental data has been evaluated by means of the absolute mean deviation (MAD) and the root mean squared deviation (rms) values. For both statistical descriptors, the

bond length deviation for all compounds ranges between 0.01 and 0.03 Å, showing a good agreement with experimental data. Bend and torsional angles are generally described with an accuracy within 6°, except for compound **4**, where the steric repulsion between vicinal bromine atoms is described theoretically by a torsion displacement between adjacent pyrrolic rings by about 22° (angle C₁₃-C₁₄-C₁₅-C₁₆). On the other hand, crystallographic measurements gave an angle of *ca.* 4°, due to the formation of intermolecular π -stacking interactions which are responsible for the coplanarity between the rings.³³ In the present discussion, we have implicitly omitted the existence of the two other possible *cis* tautomers having the inner pyrrolic hydrogens connected on the same side of the ethylenic bridges (bonds C₁₉-C₂₀ and C₉-C₁₀ in Scheme 1). In fact, from preliminary optimization calculations on compound **2**, their relative energies lie well above the *trans* stable tautomeric forms by about 90 kcal/mol. For that reason, their influence on molecular properties, with a good approximation, can be neglected in comparison with the more stable *trans* and *cis* structures reported in Scheme 2.

4.2. Bromine Effect on Electronic Absorption Spectra.

The electronic spectra of porphyrin and porphyrin-like systems (e.g., chlorine and bacteriochlorin) can be interpreted by means of the so-called four-orbital model proposed in the early 1960s by Gouterman and co-workers and based on semiempirical theoretical calculations.^{65,66} The main features of the electronic bands in the UV-visible region of the spectrum are described as excitation electronic transitions between the occupied (HOMO and next-HOMO, hereafter called H and H-1) and the unoccupied (LUMO and next-LUMO, or L and L+1) molecular frontier orbitals. For the free-base porphyrin, the electronic spectrum consists of a series of absorption bands between 500 and 600 nm (Q bands) of weak intensity (molar absorptivity $\epsilon \sim 10\,000\text{ M}^{-1}\text{ cm}^{-1}$) and a more intense band near the UV region at about 400

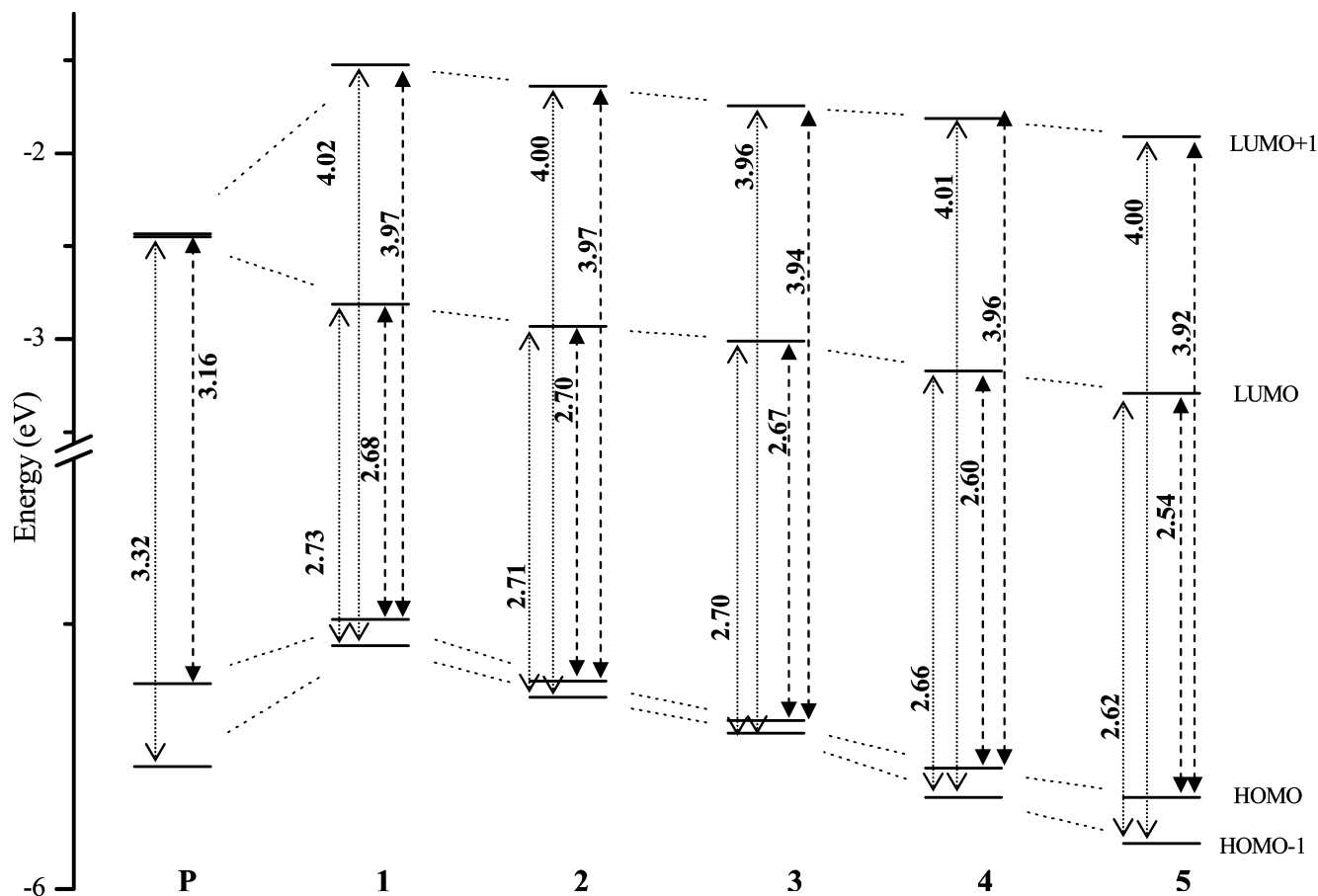


Figure 2. Frontier molecular orbital energy diagram for porphyrin (**P**) and molecules **1–5**. Gap energies (eV) between H–1 to L/L+1 (dot arrow) and H to L/L+1 (dash arrow) are also shown.

nm (called B or Soret bands, $\epsilon \sim 200\,000\text{ M}^{-1}\text{ cm}^{-1}$).^{6,65} Other electronic bands are present at higher energies (between 3.65 and 5.50 eV) and are classified according to Platt's nomenclature as N, L, and M bands.⁶⁷ The proposed explanation for the weak intensity of the electronic Q bands (denoted as Q_x and Q_y) in porphyrin is related to the equal weight of the two transition configurations which compose each Q band and to the near-degeneracy of the orbital energies of the final states (next-LUMO and LUMO for **P** in Figure 2).⁶⁸ For example, the low-energy-lying Q band (Q_x , experimental energy in gas phase at 1.98 eV)⁶⁹ is described, from theoretical calculations, by similar transition coefficient weights and two distinct electronic transitions: the next-HOMO to next-LUMO and the HOMO to LUMO electronic transition. The similar but different signs of the contribution to the transition dipole moment cause its *quasi*-vanishing or weak intensity of the Q bands. A similar argument holds for the intensity of the next-excitation energy band or Q_y band (experimental gas-phase excitation energy at 2.42 eV),⁶⁹ whereas for the B bands (B_x and B_y), the two transition configurations do not cancel transition moments and strengthen the intensity of the band. In porphycene, the molecular symmetry is lower than in porphyrin (C_{2h} vs D_{2h}) and the energy splitting between the L+1 and L molecular orbitals causes the red-shift wavelength absorption and a different weight for the two Q-band electronic transitions (see for example compound **1** in Figure 2). As a consequence, the reciprocal cancellation effect on the intensity, which

determines the weakness of the porphyrin Q bands, is less effective in porphycene, and the result is a stronger absorption Q band with a red-shift maximum absorption ($\lambda_{\text{max}} = 630\text{ nm}$, 1.97 eV; $\epsilon \sim 50\,000\text{ M}^{-1}\text{ cm}^{-1}$ in benzene).³ The same spectroscopic features (red-shift wavelength and strong absorption) found for porphycene are also present in its brominated derivatives **1–5** (see Figure 1), for which, in the following, a strict comparison between experimental and calculated data is made in order to elucidate some aspects of their electronic spectra. Before discussing the theoretical results for the most stable tautomer for each porphycene derivative, a brief comparison among the absorption electronic spectra of the different tautomers (**A–D** structures in Scheme 2) of compound **2** will be made. The choice of the monobromo-substituted porphycene as a test case has been made on the basis of two factors: the introduction of at least one bromine substituent with respect to porphycene derivative **1** and at the same time the low computational cost. The influence of the basis set choice (all electron or pseudopotential basis set for bromine atom) on excitation energies is shown in Table 3, where are reported the most relevant electronic bands in the visible region (Q and B bands). Among tautomers **A–D**, the excitation energy difference for each electronic band is generally within 0.01 eV, taking into account also the different basis sets for bromine atoms. The partial conclusion that can be deduced from this result is that the tautomerization mechanism, at least for absorption electronic spectra, does not drastically influence excitation

Table 3. Comparison between Excitation Energies ΔE in eV (Q and B bands) between the *trans* (**A**, **B**) and *cis* (**C**, **D**) Tautomers of Monobromo-Substituted Porphycene **2** (in parentheses are reported the oscillator strengths)

	A		B		C		D	
	SVP ^a	SVP/SDD ^b	SVP ^a	SVP/SDD ^b	SVP ^a	SVP/SDD ^b	SVP ^a	SVP/SDD ^b
Q ₁	2.24 (0.1825)	2.24 (0.1782)	2.23 (0.1643)	2.24 (0.1686)	2.24 (0.1852)	2.24 (0.1817)	2.24 (0.1743)	2.25 (0.1753)
Q ₂	2.35 (0.2614)	2.36 (0.2654)	2.35 (0.2732)	2.36 (0.2729)	2.37 (0.2413)	2.37 (0.2445)	2.33 (0.2316)	2.34 (0.2381)
B ₁	3.76 (1.0411)	3.76 (1.0447)	3.75 (0.9361)	3.76 (0.9507)	3.75 (0.9209)	3.75 (0.9287)	3.74 (0.8699)	3.75 (0.8990)
B ₂	3.87 (1.0380)	3.88 (1.0388)	3.88 (1.1311)	3.89 (1.1119)	3.86 (1.0648)	3.86 (1.0608)	3.87 (1.0311)	3.87 (1.0538)

^a SVP all-electron basis set for all atoms. ^b SVP basis set for H, C, and N atoms and Stuttgart pseudopotential (SDD) plus optimized basis set for valence electrons for Br.

energies. So the approximation of considering the most stable tautomer for each compound as the dominant model structure present in solution, without accounting for statistically averaged tautomer populations, can offer just enough theoretical information for the electronic spectrum interpretation. The main results for the most stable tautomeric forms of compounds **1–5** have been reported in Table 3, including excitation energies, main transition configurations, and oscillator strengths. In particular, the results focus on Q-band trends by changing the number of bromine atoms at positions 3, 6, 13 and 16 (see Scheme 1). Since experimental absorption maxima were measured in dichloromethane, calculated bulk solvation shifts on excitation energies were also included in Table 3. It has to be noted that in the experimental spectra a third absorption peak appears in the Q-band region. From TDDFT as well as the symmetry-adapted cluster configuration interaction method (SAC-CI), there is no evidence for this electronic transition, and in previous work, it has been assigned as a Q₁ side vibrational band.^{70,71} For the parent compound 2,7,15,19-tetra-*n*-propylporphycene (**1**) of brominated molecules **2–5**, the calculated Q bands show a slight bathochromic wavelength red-shift in comparison to the free base porphycene (**Pc**). From *in vacuo* calculations, this difference is estimated to be 10 and 6 nm, respectively, for the Q₁ and Q₂ bands. This result is the same order of magnitude as the corresponding experimental shift (5 and 4 nm), although the absorption measurements are referred to different solvents (benzene for **Pc** and dichloromethane for derivative **1**). The basic feature that emerges from the experimental absorption spectra is the red shift for the wavelength absorption of Q₁ and Q₂ bands with an increasing number of bromine atom substitutions (positions 3, 6, 13, and 16 of Scheme 1). The same findings come from theoretical calculations; in fact, both Q₁ and Q₂ bands increase in wavelength by about 37 and 25 nm, respectively, on going from compound **2** to **5**. Excitation energies, from TDDFT calculations, expressed in electronvolt units, are systematically overestimated, and the mean absolute deviation from the experimental values, for both Q₁ and Q₂ bands, is 0.3 eV, a result that is consistent with the error generally found for the hybrid TDDFT method. The oscillator strengths *f*, comprised between 0.15 and 0.3 units, are stronger than the corresponding values for porphyrin calculated at the same level of theory, but less intense than the B bands which are found to be greater by about 1 order of magnitude (*f* ~ 1.0). Molecular frontier orbitals (H–1, H, L, and L+1)

are mainly responsible for the electronic transitions which compose Q bands and are all $\pi\pi^*$ in character, as can be shown by plotting the relative molecular orbital isodensity surfaces (see Supporting Information, S14–S18). The main transition configuration for porphycene derivatives **1–5** does not follow the same transition configuration scheme shown in Table 4 for the free base porphycene **Pc**. In that case (**Pc**), the Q₁ electronic band resulted from H–1→L (78%) and H→L+1 (13%) electronic transitions, while the Q₂ band resulted from H→L (81%) and H–1→L+1 (13%). On the other hand, the Q₁ band of derivative **1** shows a non-negligible contribution from H→L (~8%), and the Q₂ band main configuration also includes the electronic transition from H–1 to L (~7%). At the same time for derivatives **4** and **5**, the four-orbital model is no longer completely valid since the Q₁ band of **4** and Q₂ band of **5** have minor contributions from inner orbitals (respectively H–2 and H–3). Theoretical calculations with the SAC-CI method have evidenced the failure of the four-orbital model also for the B bands of porphycene and porphyrin isomers, by proposing an alternative five-orbital model for the interpretation of their electronic spectra.⁷⁰ The Q₁ and Q₂ wavelength red shifts can be rationalized on the basis of their proper transition configuration and *in vacuo* absolute molecular orbital energies (Figure 2 and Supporting Information, S19). For instance, the Q₁ band of both **1** and **2** is mainly composed of H–1→L electronic transition by about 76% (**1**) and 82% (**2**), respectively (Table 4). The energy difference between H–1 and L slightly decreases by 0.02 eV (see Figure 2) on going from **1** to **2**, explaining the Q₁ bathochromic shift. For the derivative **3**, the energy difference between H and L molecular orbitals, with the corresponding transition being responsible by 80% for the Q₁ band, further decreases by 0.03 eV. For derivatives **4** and **5**, also considering the second electronic transition H→L, which shows a considerable weight of about 20%, the orbital energy difference decreases following the same trend as previously shown and confirming the overall Q₁ band red-shift effect from compounds **1** to **5**. The energy stabilization of L and L+1 orbitals, which decreases the energy gaps between H and L (L+1), has a greater importance for the Q-band red shift in comparison to that of H and H–1 molecular orbitals. In a similar fashion, the above interpretation scheme can be applied for explaining the Q₂ band red shift, taking into account all relevant electronic transitions contributing to that band. For example, for derivative **5**, the energy difference corresponding to the

Table 4. TDDFT Excitation Energies ΔE (eV, nm), Oscillator Strengths f , and Transition Main Configuration for Porphycene (**Pc**) and Compounds **1–5**^a

molecule	state	TDDFT						SAC-CI ^b		Expt ^b	
		ΔE		main configuration	c^2	f	ΔE^{solv}	ΔE	ΔE		
Pc	1 ¹ A	2.30,	540.1	H-1→L	0.785	0.1503	2.31	1.62	1.97	629	
				H→L+1	0.135						
	2 ¹ A	2.40,	517.2	H→L	0.812	0.2202	2.41	1.86	2.08	596	
1	1 ¹ A	2.25,	551.8	H-1→L+1	0.126	0.1649	2.25			1.96	633
				H-1→L	0.757						
				H→L+1	0.121						
				H→L	0.082						
	2 ¹ A	2.37,	523.8	H→L	0.763	0.2721	2.36		2.06	601	
				H-1→L+1	0.134						
				H-1→L	0.066						
2	1 ¹ A	2.23,	555.9	H-1→L	0.824	0.1625	2.25		1.93	641	
				H→L+1	0.110						
	2 ¹ 1	2.35,	528.6	H→L	0.840	0.2726	2.36		2.04	607	
				H-1→L+1	0.140						
3	1 ¹ A	2.21,	562.3	H→L	0.806	0.1548	2.22		1.91	649	
				H-1→L+1	0.125						
	2 ¹ A	2.34,	530.4	H-1→L	0.836	0.2729	2.36		2.02	613	
				H→L+1	0.135						
4	1 ¹ A	2.16,	573.9	H-1→L	0.605	0.1555	2.19			1.86	665
				H→L	0.202						
				H→L+2	0.082						
				H-2→L	0.044						
				H→L	0.648						
	2 ¹ A	2.29,	540.4	H→L	0.223	0.3007	2.31		1.98	625	
				H-1→L	0.223						
				H-1→L+1	0.088						
5	1 ¹ A	2.10,	589.4	H-1→L	0.556	0.1509	2.13			1.81	684
				H→L	0.250						
				H-3→L	0.089						
				H→L+1	0.071						
				H→L	0.613						
	2 ¹ A	2.26,	549.6	H→L	0.265	0.3168	2.28		1.94	640	
				H-1→L	0.265						
				H-1→L+1	0.076						

^b Ref 60. ^b In dichloromethane, ref 33, except for **Pc** in benzene (ref 3). ^a Excitation energies (eV) in dichloromethane are denoted as ΔE^{solv} .

first electronic transition of **Q**₂ (H-1→L) is lower by 0.02 eV in comparison to that of compound **4**. At the same time for the second electronic transition (H→L, transition weight of 25%), the energy difference is greater by 0.12 eV, the overall effect being the wavelength red shift on the resulting band from **1** to **5**. This effect can also be appreciated from the simulation of the electronic spectra, obtained by overlapping with Gaussian functions for each electronic transition, showing the resulting **Q**₁ and **Q**₂ maximum absorption wavelengths (see Figure 3). The introduction of bulk solvation effects through the C-PCM method does not improve the agreement between calculated and experimental excitation energies. Moreover, **Q** bands appear red-shifted in wavelength only for compounds **2–5**, while the calculated excitation energies between **1** and **2** are identical. The performance of new hybrid meta-GGA functionals (BMK and M06-2X) on electronic excitation energies was also investigated, and the results have been reported in Table 7. Since the mean average deviation (MAD) for both **Q**₁ and **Q**₂ bands is about 0.3 eV, these two new functionals, for the investigated systems, do not significantly improve PBE0 results.

Triplet Energies. Photosensitizer triplet energy is an important photophysical parameter in photodynamic therapy. A value above 0.98 eV is a prerequisite for energy transfer from the photosensitizer T₁ state to molecular oxygen in the type II PDT mechanism.¹⁵ Singlet oxygen quantum yield

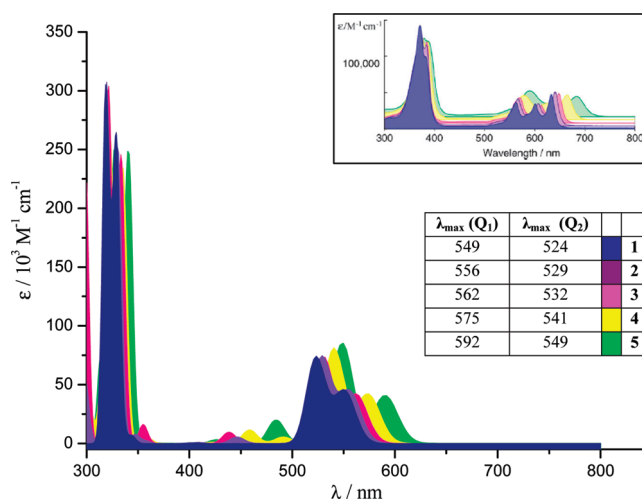


Figure 3. *In vacuo* simulated electronic spectra for molecules **1–5** and relative maxima absorption wavelength for **Q**₁ and **Q**₂ electronic bands. The half-height bandwidth was set to 0.2 eV. In the upper panel are reproduced for comparison the experimental spectra. Half-height bandwidths were set to 0.2 eV.

measurements reveal the efficiency of this mechanism and indirectly demonstrate that the above condition is fulfilled when singlet oxygen is experimentally detected and quantified, for example, through the peak intensity in the O₂

Table 5. Singlet ΔE ($S_0 \rightarrow S_n$, $n = 1-2$) and Triplet ΔE ($T_1 \rightarrow T_n$, $n = 1-2$) Vertical Excitation Energies and Oscillator Strengths f for Free Base Porphycene (**Pc**) and Derivatives **1-5**^a

molecule	state	BMK ^b				M06-2X ^b			
		ΔE ($S_0 \rightarrow S_n$) (eV, nm)		f	$\Delta E(T_1 \rightarrow T_n)$ (eV)	ΔE ($S_0 \rightarrow S_n$) (eV, nm)		f	$\Delta E(T_1 \rightarrow T_n)$ (eV)
Pc	1	2.31	536.8	0.1803	1.10	2.26	548.8	0.1863	1.05
	2	2.42	512.5	0.2573	1.38	2.38	521.7	0.2667	1.30
1	1	2.27	545.2	0.1968	1.16	2.23	556.2	0.2107	1.12
	2	2.39	517.9	0.3102	1.40	2.35	527.6	0.3130	1.34
2	1	2.26	549.8	0.1987	1.17	2.21	561.4	0.2114	1.14
	2	2.38	521.3	0.3138	1.37	2.34	530.2	0.3222	1.31
3	1	2.23	556.2	0.1918	1.21	2.18	568.0	0.2024	1.18
	2	2.37	522.5	0.3138	1.34	2.33	531.3	0.3265	1.27
4	1	2.19	566.0	0.2044	1.11	2.15	578.0	0.2180	1.07
	2	2.33	532.4	0.3436	1.33	2.29	541.4	0.3571	1.27
5	1	2.14	580.7	0.2048	1.03	2.08	593.5	0.2193	0.98
	2	2.29	541.4	0.3583	1.32	2.25	550.7	0.3738	1.25
MAD ^c	1	0.33			0.07	0.28			0.11
	2	0.32				0.30			

^a The values are calculated in vacuo by means of BMK and M06-2X hybrid meta-GGA functionals. ^b PBE0 optimized geometry. ^c Mean absolute deviation (MAD), in eV, for singlet ($n = 1, 2$) and triplet states ($n = 1$).

Table 6. Calculated Vertical Triplet Energies E_T (eV) *in vacuo* and in Dichloromethane ($E_T^{\text{solv(I)}}$) and Bromobenzene ($E_T^{\text{solv(II)}}$), with the Main Transition Configuration and Their Coefficients (c^2)^a

molecule	state	E_T	$E_T^{\text{solv(I)}}$	$E_T^{\text{solv(II)}}$	E_T^{adiab}	main configuration ^b	c^2	exp.
1	T ₁	1.16	1.19	1.19	0.73	H→L	0.930	1.27
		1.41	1.41	1.41		H-1→L	0.966	
	T ₂	2.41	2.41	2.42	H-1→L+1	0.361		
					H-3→L	0.332		
	T ₃				H→L+1	0.213		
					H-1→L+1	0.548		
					H-3→L	0.218		
					H→L+1	0.151		
2	T ₁	1.17	1.20	1.20	0.77	H→L	0.849	1.25
		1.38	1.38	1.38		H-1→L	0.094	
	T ₂	1.38	1.38	1.38	H-1→L	0.882		
					H→L	0.101		
	T ₃	2.31	2.34	2.35	H-2→L	0.843		
					H→L+1	0.054		
					H-2→L+2	0.028		
					H-1→L+1	0.391		
T ₄	2.42	2.42	2.43	H-3→L	0.265			
				H→L+1	0.193			
				H-2→L	0.059			
				H-1→L	0.930			
3	T ₁	1.20	1.22	1.22	0.81	H→L	0.975	1.24
		1.35	1.35	1.35		H-1→L	0.930	
	T ₂	2.30	2.35	2.32	H-2→L	0.924		
		2.31	2.37	2.34	H-3→L	0.788		
	T ₃				H→L+1	0.069		
					H-1→L+1	0.052		
					H→L	0.813		
					H-1→L	0.134		
T ₄	1.11	1.13	1.13	0.77	H-1→L	0.840	1.18	
	1.34	1.33	1.34	H→L	0.143			
	2.10	2.21	2.15	H-2→L	0.908			
	2.23	2.28	2.25	H-3→L	0.878			
5	T ₁	1.04	1.05	1.06	0.62	H→L+1	0.035	1.03-1.13 ^c
		1.33	1.32	1.32		H→L	0.875	
	T ₂	2.04	2.13	2.08	H-1→L	0.075		
	T ₃	2.08	2.16	2.11	H-1→L	0.902		
MAD ^d		0.07			0.46	H-2→L	0.930	
						H-3→L	0.904	

^a Adiabatic triplet energies E_T^{adiab} for the first excited state and experimental phosphorescence (eV) are also given. ^b *In vacuo*. ^c Low intensity peak. ^d Mean absolute deviation (MAD), in eV, for the vertical and adiabatic T₁ state.

phosphorescence spectrum. The theoretical evaluation of the triplet energies ($S_0 \leftarrow T_1$), in the framework of the TDDFT approach, can be a valuable tool in the design of a new PDT photosensitizer. For the molecular systems just present in

the literature, as is our case study, the comparison with experimental data can support or limit the applicability of a theoretical protocol. The first four TDDFT vertical triplet energies *in vacuo* have been reported in Table 6, together

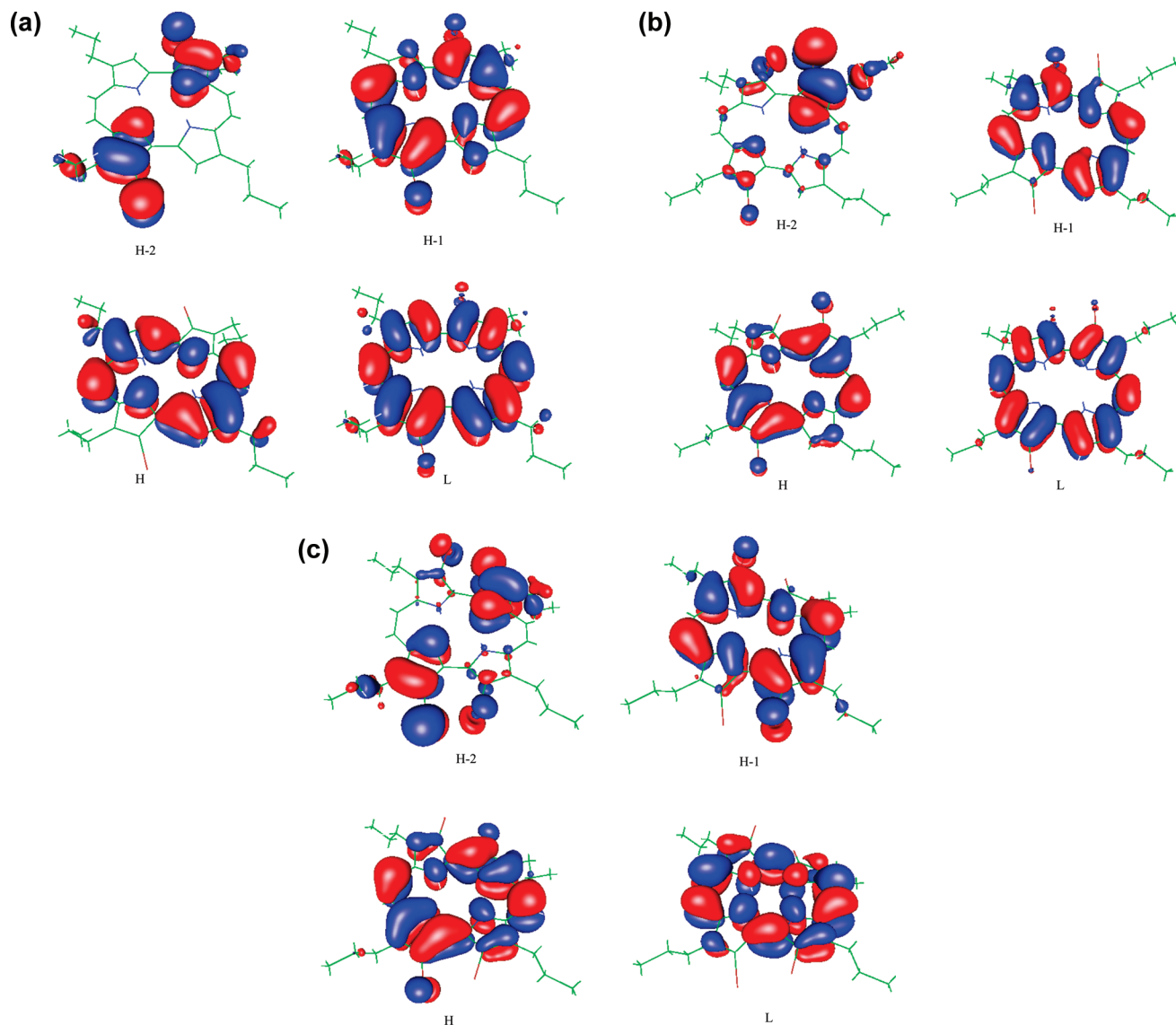


Figure 4. Graphical representation of the highest occupied and lowest unoccupied MOs for the (a) molecule **3** tautomer **1**, (b) molecule **4**, and (c) molecule **5**, obtained from the geometry of the ground states, S_0 , at the PBE0 level.

with the main transition configuration and in solvent (dichloromethane and bromobenzene) values. Experimental triplet energies for derivatives **1–5**, as obtained from phosphorescence spectra in degassed bromobenzene, are between 1.27 and 1.03 eV.³³ All these compounds have shown appreciable singlet oxygen quantum yield Φ_Δ (between 0.36 and 0.95), a fact that confirms the strict relation between this parameter and photosensitizer triplet energies greater than 0.98 eV. The TDDFT vertical triplet energies obtained in this study cover a range between 1.20 (**3**) and 1.04 eV (**5**), with a maximum deviation of 0.1 eV from the experimental counterparts. Since phosphorescence spectra of compounds **1–5** refer to an emission process at the excited state relaxed geometry, it is more convenient to calculate excitation energies from the first triplet optimized geometry, in order to get a more realistic comparison with experimental data. *In vacuo* adiabatic excitation energies for the first triplet excited state of **1–5**, obtained in such a way, are reported as E_T^{adiab} in Table 6 and are further underestimated with respect to vertical energies with a mean absolute deviation from the experimental value of 0.46 eV. From theoretical calculations,

a strict correlation between the number of bromine atoms and vertical (and adiabatic) triplet energy values has not been found in going from derivative **1** to **5**, as is the case for the experimental results where the triplet energy decreases with the number of bromine atoms. Vertical triplet energies calculated in dichloromethane and bromobenzene are almost identical and similar to *in vacuo* results. In this case, the energy deviation ranges between 0.01 and 0.03 eV. The introduction of bulk solvation effects slightly improves the agreement with experimental triplet energies (Table 6). As for the case of calculated singlet excitation energies, the mean absolute deviation for vertical triplet energies, calculated by BMK and M02–6X exchange–correlation functionals (Table 5), is comparable to that of PBE0, being respectively 0.07 and 0.1 eV.

4.3. Spin–Orbit Matrix Elements. The key concept in order to understand the trend of the SO matrix elements of the molecules investigated lies in the nature of the MOs involved in the coupling mechanism arising from the one-electron singlet and triplet transition defining the S_i and T_j states. These MOs are shown in Figure 4. Instead, the Table

Table 7. Spin–Orbit Matrix Elements (cm^{-1}) between Singlet and Triplet Excited States of Porphycene Derivatives 1–5 Computed Using the Geometry of the S_0 Ground State

		3					
		1	2	Tau1	Tau2	4	5
$\langle S_1 H_{so} T_1\rangle$	x	0.2	0.3	2×10^{-2}	0.1	6.3	4.6
	y	0.2	0.9	0.2	1.5	10.8	9.9
	z	2×10^{-2}	0.5	1×10^{-2}	1.6	2.3	0.3
$\langle S_1 H_{so} T_2\rangle$	x	4×10^{-2}	0.1	2×10^{-3}	3×10^{-2}	2.5	1.3
	y	5×10^{-2}	0.3	1×10^{-4}	0.9	6.8	7.8
	z	8×10^{-3}	0.2	1×10^{-4}	0.8	4.0	0.4
$\langle S_1 H_{so} T_3\rangle$	x	6×10^{-2}	0.2	1×10^{-4}	6×10^{-2}	7.8	30.4
	y	8×10^{-2}	1.3	2×10^{-3}	3.4	33.5	0.3
	z	0.1	1.6	6×10^{-3}	2.1	11.9	4.3
$\langle S_1 H_{so} T_4\rangle$	x	0.2	0.3	8×10^{-2}	2×10^{-4}	5.4	4.1
	y	0.1	1.6	1.2	3×10^{-3}	2.2	6.5
	z	7×10^{-2}	2.1	2.9	2×10^{-3}	6.1	1.0
$\langle S_2 H_{so} T_1\rangle$	x	3×10^{-2}	9×10^{-2}	4×10^{-3}	3×10^{-2}	0.2	0.4
	y	5×10^{-2}	0.3	3×10^{-4}	0.8	0.8	0.9
	z	7×10^{-3}	0.2	4×10^{-4}	0.7	1.2	0.2
$\langle S_2 H_{so} T_2\rangle$	x	0.2	0.3	2×10^{-2}	0.1	0.4	0.8
	y	0.2	0.9	0.2	1.5	4.8	2.0
	z	2×10^{-2}	0.5	5×10^{-3}	1.6	2.7	0.7
$\langle S_2 H_{so} T_3\rangle$	x	0.1	0.2	9×10^{-4}	6×10^{-2}	26.4	17.2
	y	0.2	0.9	4×10^{-4}	1.9	21.5	7.0
	z	3×10^{-2}	0.3	4×10^{-3}	2.0	0.7	2.9
$\langle S_2 H_{so} T_4\rangle$	x	5×10^{-2}	9×10^{-2}	0.3	5×10^{-5}	0.2	10.4
	y	0.1	0.3	2.6	3×10^{-3}	2.7	25.7
	z	3×10^{-2}	0.3	10.5	1×10^{-3}	0.8	1.4

7 listing of the SO matrix elements between S_j and T_j states, irrespective of their sign, gives several details about this matter. By analyzing the results in Table 7, we might expect an increase of the SO values as the number of substituted bromine atoms becomes larger, due to the dependence of the relativistic effects on the size of the heteroatoms. This is true for some matrix elements like the $\langle S_1|H_{so}|T_1\rangle$ when the tautomer 2 (**Tau2**) is taken into account. Nevertheless, it matters a lot whether, on the whole, this behavior is not respected, since this means that not all sites of substitution could affect the SO contributions in the same way.

What is more evident is, on the whole, the hole belonging to tautomer 1 (**Tau1**) of compound 3, which clearly indicates that there is no benefit, as far as the SO contributions are concerned, when for this molecule the heteroatom double substitution is involved. In this case, the smallest values occur when the larger contributions to the summation above involve the same discoincident orbitals or when these are mainly located in different regions of the molecule. As an example, for $\langle S_1|H_{so}|T_2\rangle$, the coupling involves mainly the β H and the α H with coefficients of 0.90 and 0.99, respectively. Analogous considerations are held for the $\langle S_2|H_{so}|T_1\rangle$ matrix elements where the larger involved orbitals are the β H–1 and the α H–1 with coefficients of -0.91 and 0.96 , respectively. The highest values belong to the $\langle S_1|H_{so}|T_3\rangle$ matrix elements of molecules 4 and 5, respectively. Obviously, there is a more clear correlation to this fact, since, here, the MOs involved in the coupling mechanism show also a p atomic orbital contribution located in the same bromine atom. However, this is not the only reason why the SO matrix elements are enhanced. For the $\langle S_1|H_{so}|T_3\rangle$ matrix elements of compound 4, the larger involved orbitals in the coupling mechanism are β H–2 and α H–1 with coefficients of 0.78 and -0.95 , respectively, and β H–2 and α H with

coefficients of 0.45 and -0.95 , respectively. For the $\langle S_1|H_{so}|T_3\rangle$ matrix elements of compound 5, the orbitals are β H–2 and α H–1 with coefficients of 0.75 and -0.96 and β H–2 and α H with coefficients of 0.50 and -0.96 , respectively. Note from Figure 4b and c that in both cases the atomic orbitals of at least one of the bromines are not the same p orbitals, or p orbitals with almost the same orientation. Just these findings enlarge unconditionally the values of the SO matrix elements. Instead, concerning the $\langle S_2|H_{so}|T_3\rangle$ of tautomer 1 (**Tau1**) of molecule 3, where larger weight is given from the coupling between the β H–2 and the α H–1 orbitals, there are still atomic orbitals centered on the heteroatom site (see Figure 4a) but with the same orientation of the p orbital involved. The best orbital orientation for SO mixing is when the two p orbitals are at 90° with respect to one another. In order to generate angular momentum, an orbital jump (as an example of the $p_z \rightarrow p_x$ type) is required.

In conclusion, for most of the matrix elements, the trend is determined by the substitution of bromine atoms. On the whole and with the exception of some case, the values of the matrix elements increase with the number of heteroatoms.

Conclusions

Ground state structures and electronic absorption spectra of 2,7,12,17-tetra-*n*-tetrapropyl porphycene and its brominated derivatives have been theoretically investigated by means of DFT and TDDFT methods. From our study, the following conclusion can be drawn:

The reduced internal cavity allows a tautomerization mechanism for the inner pyrrole hydrogens with the formation of *cis* and *trans* conformers. The most energetically stable structures are found to be the *trans* structures, though

the gas-phase energy barrier for the *trans* to *cis* interconversion is very low (within 4 kcal/mol for the monobrominated porphycene).

Optimized structure parameters, for the most stable tautomer of derivatives **1–5**, are in good agreement with available crystallographic data. The increasing number of bromine atoms tends to twist the torsional angle between adjacent pyrrole rings (about 20° for the tetrabrominated compound) in order to minimize sterical repulsion.

The calculated electronic spectra Q bands are red-shifted on going from the mono- to tetra-brominated porphycene derivative, as found from experimental spectra. In particular, the *in vacuo* calculated Q₁ and Q₂ maximum wavelengths are shifted respectively by 43 and 35 nm for tetrabrominated compound with respect to the unsubstituted case (compound **1**). Solvent shifts on excitation energies, as obtained from the C-PCM model, showed little difference in comparison with the calculated gas-phase ones.

The computed spin-orbit matrix elements between S_i and T_j electronic states at the ground state optimized geometry tend to increase, with the exception of some case, with the number of bromine atoms, in qualitative agreement with the experimental intersystem spin crossing rate constant trend. One exception is made by the tribrominated porphycene derivative (compound **3**), for which it is necessary to take into account also the presence of other *trans* tautomers in order to rationalize the experimental trend.

Acknowledgment. Financial support from the Università degli Studi della Calabria and MIUR (PRIN 2008) is gratefully acknowledged.

Supporting Information Available: *In vacuo* optimized structures and absolute energies (atomic units) for *trans* and *cis* tautomeric forms of compounds **1–5**. For the most stable tautomers, the optimized Cartesian coordinates, molecular orbital isodensity surfaces and frontier molecular orbital energies are also given. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Sánchez-García, D.; Sessler, J. L. *Chem. Soc. Rev.* **2008**, *37*, 215–232.
- (2) Aramendía, P. F.; Redmond, R. W.; Nonell, S.; Schuster, W.; Braslavsky, S. E.; Schaffner, K.; Vogel, E. *Photochem. Photobiol.* **1986**, *44*, 555–559.
- (3) Vogel, E.; Köcher, M.; Schmickler, H.; Lex, J. *Angew. Chem., Int. Ed. Engl.* **1986**, *25*, 257–259.
- (4) Stockert, J. C.; Cañete, M.; Juarranz, A.; Villanueva, A.; Horobin, R. W.; Borrell, J. I.; Teixidó, J.; Nonell, S. *Curr. Med. Chem.* **2007**, *14*, 997–1026.
- (5) Dougherty, T. J.; Gomer, C. J.; Henderson, B. W.; Jori, G.; Kessel, D.; Korbek, M.; Moan, J.; Peng, Q. *J. Natl. Cancer Inst.* **1998**, *90*, 889–905.
- (6) Bonnett, R. In *Chemical Aspects of Photodynamic Therapy*; Gordon & Breach Science Publishers: Amsterdam, 2000; pp 1–289.
- (7) Juzeniene, A.; Peng, Q.; Moan, J. *J. Photochem. Photobiol. Sci.* **2007**, *6*, 1234–1245.
- (8) Schuitmaker, J. J.; Baas, P.; van Leengoed, H. H. L. M.; van der Meulen, F. W.; Star, W. M.; van Znadwijk, N. *J. Photochem. Photobiol. B* **1996**, *34*, 3–12.
- (9) Buytaert, E.; Dewaele, M.; Agostinis, P. *Biochim. Biophys. Acta* **2007**, *1776*, 86–107.
- (10) Oleinick, N. L.; Morris, R. L.; Belichenko, I. *Photochem. Photobiol. Sci.* **2002**, *1*, 1–21.
- (11) DeRosa, M. C.; Crutchley, R. J. *Coord. Chem. Rev.* **2002**, *233–234*, 351–371.
- (12) Schweitzer, C.; Schmidt, R. *Chem. Rev.* **2003**, *103*, 1685–1757.
- (13) Schmidt, R. *Photochem. Photobiol.* **2006**, *82*, 1161–1177.
- (14) Dougherty, T. J.; MacDonald, I. J. *J. Porphyrins Phthalocyanines* **2001**, *5*, 105–129.
- (15) Herzberg, G. In *Spectra of Diatomic Molecules*, 2nd ed.; Van Nostrand Reinhold: New York, 1950; pp 344–346.
- (16) Darmanyan, A. P.; Foote, C. J. *Phys. Chem.* **1992**, *96*, 3723–3728.
- (17) Turro, N. J. In *Modern Molecular Photochemistry*; Benjamin: Menlo Park, 1978; pp 153–198.
- (18) Atkins, P. W. In *Molecular Quantum Mechanics*; Oxford University Press: New York, 1989; pp 319–343.
- (19) Nyman, E. S.; Hynninen, P. H. *J. Photochem. Photobiol. B: Biol.* **2004**, *73*, 1–28.
- (20) Sternberg, E. D.; Dolphin, D.; Bruckner, C. *Tetrahedron* **1998**, *54*, 4151–4202.
- (21) Lipson, R. L.; Baldes, E. J. *Arch. Dermatol.* **1960**, *82*, 508–516.
- (22) Ronn, A. M.; Nouri, M.; Lofgren, L. A.; Steinberg, B. M.; Westerborn, A.; Windahl, T.; Shikowitz, M. J.; Abramson, A. L. *Lasers Med. Sci.* **1993**, *11*, 267–272.
- (23) Young, S. W.; Woodburn, K. W.; Wright, M.; Mody, T. D.; Fan, Q.; Sessler, J. L.; Dow, C.; Miller, R. A. *Photochem. Photobiol.* **1996**, *63*, 892–897.
- (24) Hsi, R. A.; Kapatkin, A.; Strandberg, J.; Zhu, T.; Vulcan, T.; Solonenko, M.; Rodriguez, C.; Chang, J.; Saunders, M.; Mason, N.; Hahn, S. *Clin. Cancer Res.* **2001**, *7*, 651–660.
- (25) O'Connor, A. E.; William, M.; Gallagher, W. M.; Byrne, A. T. *Photochem. Photobiol.* **2009**, *85*, 1053–1074.
- (26) Wainwright, M. *Chem. Soc. Rev.* **1996**, *25*, 351–359.
- (27) New, O. M.; Dolphin, D. *Eur. J. Org. Chem.* **2009**, *16*, 2675–2686.
- (28) Gorman, A.; Killoran, J.; O'Shea, C.; Kenna, T.; Gallagher, W. M.; O'Shea, D. F. *J. Am. Chem. Soc.* **2004**, *126*, 10619–10631.
- (29) Roelants, M.; Lackner, B.; Waser, M.; Falk, H.; Agostinis, P.; Van Poppeland, H.; De Witte, P. A. M. *Photochem. Photobiol. Sci.* **2009**, *8*, 822–829.
- (30) Will, S.; Rahbar, A.; Schmickler, H.; Lex, J.; Vogel, E. *Angew. Chem., Int. Ed. Engl.* **1990**, *29*, 1390–1393.
- (31) Aritome, I.; Shimakoshi, H.; Hisaeda, Y. *Acta Crystallogr., Sect. C* **2002**, *58*, 563–564.
- (32) Baba, T.; Shimakoshi, H.; Aritome, I.; Hisaeda, Y. *Chem. Lett.* **2004**, *33*, 906–907.
- (33) Shimakoshi, H.; Baba, T.; Iseki, Y.; Aritome, I.; Endo, A.; Adachib, C.; Hisaeda, Y. *Chem. Commun.* **2008**, 2882–2884.

- (34) Braslavsky, S. E.; Müller, M.; Mártire, D. O.; Pörting, S.; Bertolotti, S. G.; Chakravorti, S.; Koç-Weier, G.; Knipp, B.; Schaffner, K. *J. Photochem. Photobiol. B: Biol.* **1997**, *40*, 191–198.
- (35) Casida, M. E. In *Recent Developments and Applications in Density-Functional Theory*; Seminario, J. M., Ed.; Elsevier: Amsterdam, 1996; pp 155–192.
- (36) Werschnik, J.; Gross, E. K. U.; Burke, K. *J. Chem. Phys.* **2005**, *123*, 62206–62206(9).
- (37) Elliott, P.; Burke, K.; Furche, F. In *Recent Advances in Density Functional Methods*; Lipkowitz, K. B., Cundari, T. R., Eds.; Wiley: Hoboken, NJ, 2009; Vol. 26, pp 91–165.
- (38) Jacquemin, D.; Wathelet, V.; Perpète, E. A.; Carlo Adamo, C. *J. Chem. Theory Comput.* **2009**, *5*, 2420–2435, No. 9.
- (39) Chiodo, S.; Russo, N. *J. Comput. Chem.* **2008**, *29*, 912.
- (40) Chiodo, S. G.; Russo, N. *J. Comput. Chem.* **2009**, *30*, 832.
- (41) Bethe, H. A.; Salpeter, E. E. *Quantum Mechanics of the One and Two Electron Atoms*; Plenum: New York, 1977.
- (42) Chernyak, V.; Mukamel, S. *J. Chem. Phys.* **2000**, *112*, 3572.
- (43) Send, R.; Furche, F. *J. Chem. Phys.* **2010**, *132*, 044107.
- (44) Furche, F. *J. Chem. Phys.* **2001**, *114*, 5982.
- (45) Tavernelli, I.; Tapavicza, E.; Rothlisberger, U. *J. Chem. Phys.* **2009**, *130*, 124107.
- (46) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, M.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.
- (47) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (48) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982–9985.
- (49) Schwerdtfeger, P.; Dolg, M.; Schwarz, W. H. E.; Bowmaker, G. A.; Boyd, P. D. W. *J. Chem. Phys.* **1989**, *91*, 1762–1774.
- (50) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571–2577.
- (51) Deglmann, P.; Furche, F. *J. Chem. Phys.* **2002**, *117*, 9535–four pages.
- (52) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 34108.
- (53) Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397–3406.
- (54) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143.
- (55) Bauernschmitt, R.; Ahlrichs, R. *Chem. Phys. Lett.* **1996**, *256*, 454–464.
- (56) Gorelsky, S. I. SWizard program, revision 4.6, <http://www.sg-chem.net/>.
- (57) Klamt, A.; Schüürmann, G. *J. Chem. Soc. Perkin Trans.2* **1993**, *5*, 799–805.
- (58) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405–16.
- (59) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.
- (60) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; 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.; Ochtersky, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision B.05; Gaussian, Inc.: Wallingford, CT, 2004.
- (61) Waluk, J. *Acc. Chem. Res.* **2006**, *39*, 945–952.
- (62) Braun, J.; Schlabach, M.; Wehrle, B.; Köcher, M.; Vogel, E.; Limbach, H. H. *J. Am. Chem. Soc.* **1996**, *118*, 7231–7232.
- (63) Gil, M.; Waluk, J. *J. Am. Chem. Soc.* **2007**, *129*, 1335–1341.
- (64) Gil, M.; Jasny, J.; Vogel, E.; Waluk, J. *Chem. Phys. Lett.* **2000**, *323*, 534–541.
- (65) Gouterman, M. *J. Mol. Spectrosc.* **1961**, *6*, 138–163.
- (66) Gouterman, M.; Wagnière, G.; Snyder, L. *J. Mol. Spectrosc.* **1963**, *11*, 108–127.
- (67) Platt, J. R. In *Radiation Biology*; Hollaender, A., Ed.; McGraw-Hill: New York, 1956; Vol. 3, pp 71–123.
- (68) Toyota, K.; Hasegawa, J.; Nakatsuji, H. *Chem. Phys. Lett.* **1996**, *250*, 437–442.
- (69) Edwards, L.; Dolphyn, D. H.; Gouterman, M.; Adler, A. D. *THEOCHEM* **1997**, *401*, 301–314.
- (70) Hasegawa, J.; Takata, K.; Miyahara, T.; Neya, S.; Frisch, M.; Nakatsuji, H. *J. Chem. Phys. A* **2005**, *109*, 3187–3200.
- (71) Waluk, J.; Müller, M.; Swiderek, P.; Köcher, M.; Vogel, E.; Hohlneicher, G.; Michl, J. *J. Am. Chem. Soc.* **1991**, *113*, 5511–5527.

Structure, Stabilities, Thermodynamic Properties, and IR Spectra of Acetylene Clusters $(C_2H_2)_{n=2-5}$

S. Karthikeyan, Han Myoung Lee, and Kwang S. Kim*

Center for Superfunctional Materials, Department of Chemistry, Pohang University of Science and Technology, San 31, Hyojadong, Namgu, Pohang 790-784, Korea

Received June 12, 2010

Abstract: There are no clear conclusions over the structures of the acetylene clusters. In this regard, we have carried out high-level calculations for acetylene clusters $(C_2H_2)_{2-5}$ using dispersion-corrected density functional theory (DFT-D), Møller–Plesset second-order perturbation theory (MP2); and coupled-cluster theory with single, double, and perturbative triple excitations [CCSD(T)] at the complete basis set limit. The lowest energy structure of the acetylene dimer has a T-shaped structure of C_{2v} symmetry, but it is nearly isoenergetic to the displaced stacked structure of C_{2h} symmetry. We find that the structure shows the quantum statistical distribution for configurations between the T-shaped and displaced stacked structures for which the average angle ($\langle \hat{\theta} \rangle$) between two acetylene molecules would be $53-78^\circ$, close to the T-shaped structure. The trimer has a triangular structure of C_{3h} symmetry. The tetramer has two lowest energy isomers of S_4 and C_{2h} symmetry in zero-point energy (ZPE)–uncorrected energy (ΔE_e), but one lowest energy isomer of C_{2v} symmetry in ZPE-corrected energy (ΔE_0). For the pentamer, the global minimum structure is C_1 symmetry with eight sets of T-type π –H interactions and a set of π – π interactions. Our high-level *ab initio* calculations are consistent with available experimental data.

Introduction

In recent years, there has been much interest in the structure and properties of weakly bound complexes,^{1–4} because of their ubiquitous role in diverse fields including molecular clusters,^{5–10} biomolecular structures,^{11–13} supramolecular chemistry,^{14–16} and self-assembled nanostructures.^{17–19} In particular, to understand the aromatic π –H and π – π interactions,^{20–33} the aromatic dimers including the benzene dimer have been studied extensively.^{34–39} In addition, aliphatic π interactions have also been studied.^{40–59}

For the acetylenic π interactions, it is necessary to investigate the acetylene clusters. Some selected structures are shown in Figures 1–4. For the acetylene dimer $[(C_2H_2)_2]$, Pendley and Ewing reported five bands using Fourier transform infrared spectroscopy.⁴⁴ These bands are consistent with the staggered structure (S-shaped with C_{2h} symmetry) proposed by Sakai et al.⁴⁵ The free-jet infrared absorption spectroscopy study of Ohshima et al.⁴⁶ showed that the

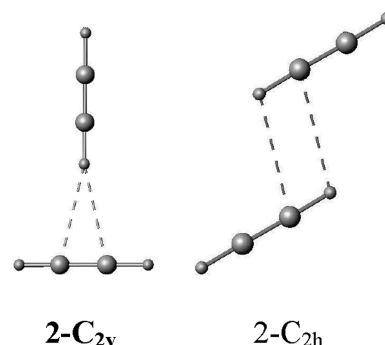


Figure 1. Low-energy structures of $(C_2H_2)_2$.

acetylene dimer is a T-shaped hydrogen-bonded structure of C_{2v} symmetry. Dykstra and Shuler,⁵³ Alberts et al.,⁵⁴ Bone and Handy,⁵⁵ Hobza et al.,⁵⁶ Karpfen,^{57,58} and Brenner and Millie⁵⁹ proposed the T-shaped structure with C_{2v} symmetry. On the other hand, Prichard et al.⁴⁷ reported that the acetylene dimer is a twisted T-shaped hydrogen-bonded structure, but not the C_{2v} symmetry structure. Thus, there is no clear conclusion for the acetylene dimer structure.

* Corresponding author e-mail: kim@postech.ac.kr.

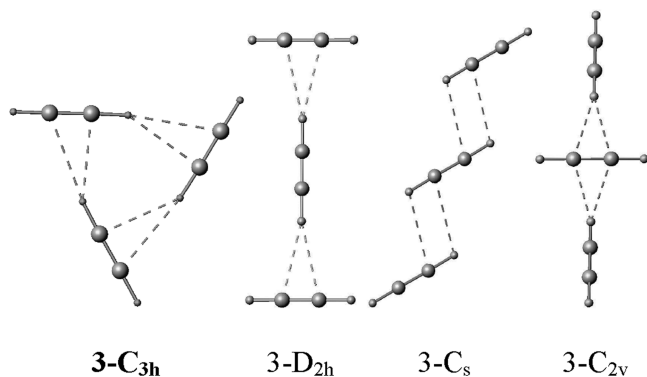


Figure 2. Low-energy structures of $(\text{C}_2\text{H}_2)_3$.

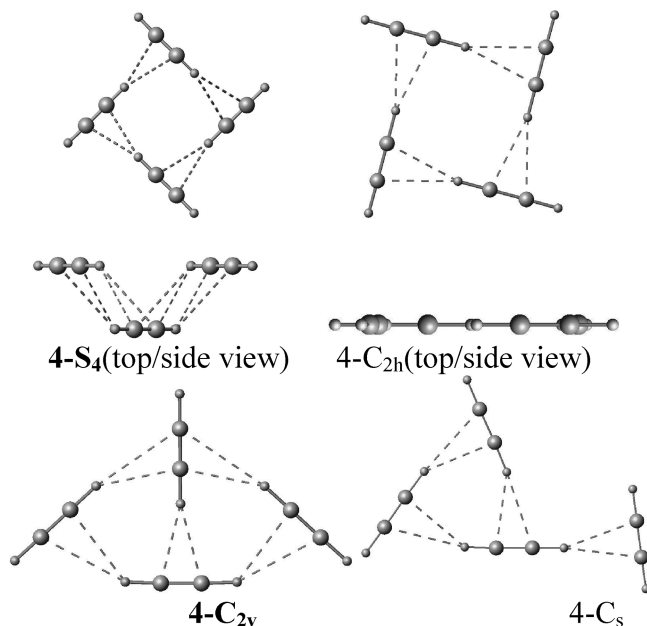


Figure 3. Low-energy structures of $(\text{C}_2\text{H}_2)_4$.

For the acetylene trimer, Dykstra and Shuler,^{41,53} Alberts et al.,⁵⁴ Brenner and Millie,⁵⁹ Prichard et al.,^{47,48} and Bone et al.⁴⁹ reported the C_{3h} symmetry structure. In the case of the acetylene tetramer, Bryant et al.⁵⁰ and Dykstra and Shuler⁴¹ proposed the S_4 symmetry structure on the basis of their infrared C–H stretching spectra and the Molecular Mechanic Cluster (MMC) method, respectively. On the other hand, Bone et al.⁵¹ proposed the cyclic structure of C_{4h} symmetry. For the acetylene pentamer, Bone et al.⁵¹ reported that the global minimum energy structure is the C_{2h} symmetry structure based on the MP2/DPZ level of theory. Yu et al.⁵² assumed the acetylene cyclic structure of C_{5h} symmetry based on the Hartree–Fock level of theory. On the other hand, Dykstra and Shuler⁵³ reported that the global minimum structure for the acetylene pentamer is C_1 symmetry. Furthermore, it is not clear whether a new more stable structure could be found for the acetylene pentamer because of its structural complexity.

In this regard, a more accurate theoretical investigation is required. We have carried out DFT-D, MP2, and CCSD(T) calculations. In order to obtain the concrete conclusion, we have focused our attention on the following: (a) binding energy at high levels of theory, (b) ZPE correction, (c)

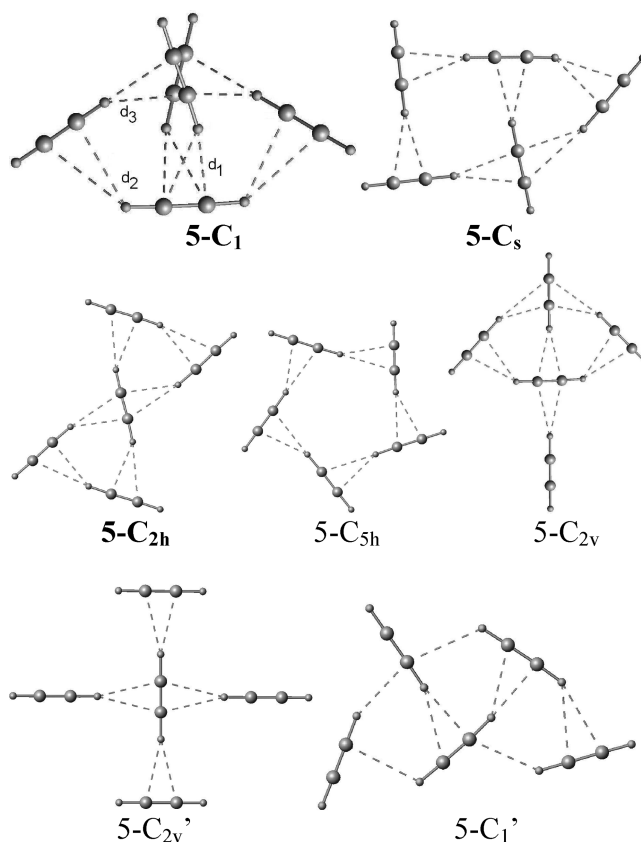


Figure 4. Low-energy structures of $(\text{C}_2\text{H}_2)_5$.

complete basis set (CBS) limit, and (d) comparison of the predicted spectra with the available experimental data.^{46–50}

Computational Method

A detailed conformation search was followed by a complete geometry optimization at the DFT-D(M06-2X)⁶⁰ and MP2 levels of theory. Then, frequency calculations were carried out for several low-energy isomers at the M06-2X/aug-cc-pVDZ level. All atoms were treated with the aug-cc-pVDZ and aug-cc-pVTZ basis sets (which will be abbreviated as aVDZ and aVTZ, respectively). The MP2/aug-cc-pVTZ and CCSD(T)/aug-cc-pVDZ energies were obtained using the single-point energy calculations on the MP2/aug-cc-pVDZ geometries. The basis set superposition error (BSSE) correction was made after geometry optimization. The calculations were performed with the Gaussian 03 suite of programs.⁶¹ The molecular structures were drawn with the Posmol package.⁶²

We estimated the MP2/CBS binding energies on the basis of the extrapolation scheme, which exploits the fact that the electron correlation error is proportional to N^{-3} for aug-cc-pVNZ basis sets.^{63,64} The CCSD(T)/CBS energies were estimated by assuming that the difference in binding energies between MP2/aug-cc-pVDZ and MP2/CBS calculations is similar to that between CCSD(T)/aug-cc-pVDZ and CCSD(T)/CBS calculations.⁶⁵ The spectral features of acetylene clusters $(\text{C}_2\text{H}_2)_{2-5}$ were investigated at the M06-2X/aug-cc-pVDZ and MP2/aug-cc-pVDZ levels of theory. In calculating the CCSD(T)/CBS ΔE_0 , ΔH_r (enthalpy at room temperature), and ΔG_{50} (free energies at 50 K, 1 atm), we employed the

MP2/aug-cc-pVDZ frequencies. The structure and binding energies of M06-2X tend to be more reliable than the BSSE-uncorrected MP2/aug-cc-pVDZ, because BSSE-uncorrected MP2/aug-cc-pVDZ overestimates binding energies.^{7,66} The MP2/aug-cc-pVDZ frequencies are similar to the M06-2X/aug-cc-pVDZ values but tend to be overestimated because of BSSE. In this regard, the vibrational stretching frequencies will be discussed on the basis of the M06-2X/aug-cc-pVDZ results. Since the CCSD(T)/CBS results are the most reliable, our discussion will be based on the CCSD(T)/CBS results, unless otherwise specified.

To facilitate the comparison of the calculated frequencies with experimental frequencies, the DFT-D theoretical harmonic frequencies are scaled by a constant factor. When the MP2 predicted harmonic frequencies are compared with the experimental frequencies, higher vibrational frequencies tend to be overestimated while lower frequencies do not. Thus, the MP2 frequencies can often be exponentially scaled^{67–69} as $\nu_i^s = \nu_i e^{-\alpha \nu_i}$, where ν_i^s and ν_i are scaled and unscaled frequencies corresponding to the vibrational mode *i*. The exponent α , a single parameter, was chosen to optimally fit the theoretical acetylene harmonic frequencies with the experimental frequencies. This method scales down lower vibrational frequencies slightly less than higher frequencies; thus, the experimentally scaled MP2 frequencies often turned out to be better than the constant-scaled values. Since the unscaled M06-2X/aug-cc-pVDZ and MP2/aug-cc-pVDZ values for the asymmetric C–H stretching and bending C–C–H frequencies of C₂H₂ are 3492 and 774 cm⁻¹ and 3431 and 734 cm⁻¹, as compared with the experimental values (3287 and 729 cm⁻¹),⁴⁴ the scale factor 0.942 is used for M06-2X/aug-cc-pVDZ and the value of α is chosen as 0.0000125 for MP2/aug-cc-pVDZ.

Results and Discussion

Important low-lying energy structures of the acetylene clusters (C₂H₂)_{2–5} are shown in Figures 1–4. This binding energies and thermodynamics properties at the M06-2X/aug-cc-pVDZ level are in Table 1, and the binding energies and thermodynamics properties at the MP2/aug-cc-pVDZ, MP2/aug-cc-pVTZ, MP2/CBS, CCSD(T)/aug-cc-pVDZ, and CCSD(T)/CBS levels are listed in Table 2, where ΔE_e , ΔE_0 , ΔH_r , and $\Delta G(50\text{ K})$ are the BSSE-corrected thermodynamic quantities. The selected geometrical parameters of 2-C_{2v}, 2-C_{2h}, 3-C_{3h}, 4-S₄, 4-C_{2v}, and 5-C₁ are given in Table 3. Acetylene clusters are named “Num-Sym”, where “Num” is the number acetylene monomer in the cluster, and “Sym” shows the symmetry of the cluster. If there are two clusters for the same symmetry, a prime is used for the structure with the less stable energy. The M06-2X and MP2 vibrational frequencies of selected acetylene clusters are in Figure 5 and the Supporting Information (Table S1).

For the acetylene dimer, the T-shaped 2-C_{2v} isomer and the displaced-stacked 2-C_{2h} isomer are nearly isoenergetic. The T-shaped 2-C_{2v} is more stable in ΔE_e by 1.01, 0.59, and 0.68 kJ/mol at the M06-2X/aug-cc-pVDZ, MP2/CBS, and CCSD(T)/CBS levels, respectively. The 2-C_{2h} isomer (displaced-stacked structure) is the transition state of the 2-C_{2v} isomer with one imaginary frequency (45 cm⁻¹ for

Table 1. DFT-D(M06-2X) Interaction Energies and Thermodynamic Quantities (kJ/mol) for Low-Energy Structures of the Acetylene Clusters (C₂H₂)_{n=2–5}^a

name	M06-2X/aVDZ			
	$-\Delta E_e$	$-\Delta E_0$	$-\Delta H$	$-\Delta G(50\text{ K})$
2-C _{2v}	5.72	3.64	2.87	0.89
2-C _{2h} (1)	4.71	(3.74)	4.82	(0.86)
3-C _{3h}	17.05	11.73	9.98	4.70
3-D _{2h} (1)	11.15	(7.35)	6.70	(0.41)
3-C _s (4)	9.28	(8.06)	13.24	(1.03)
3-C _{2v}	11.18	6.57	3.90	0.02
4-S ₄	26.14	17.65	14.69	6.45
4-C _{2h}	26.17	17.59	14.66	6.17
4-C _{2v}	24.37	19.05	17.03	7.52
4-C _s	23.18	15.81	11.93	5.43
5-C ₁	33.98	26.72	27.83	9.90
5-C _s	34.85	22.24	18.98	6.00
5-C _{2h}	34.75	24.61	19.36	8.87
5-C ₁ '(2)	30.68	(22.13)	20.81	(6.23)
5-C _{3h} (2)	32.29	(21.98)	21.74	(4.68)
5-C _{2v} (1)	30.89	(21.42)	18.12	(5.44)
5-C _{2v} '(3)	27.41	(18.17)	19.33	(1.94)

^a If the structure is not a minimum, the numbers of imaginary frequencies are given in parentheses after the structure name. In these cases, the ZPE and thermal energy corrections are not reliable enough, so these value for $-\Delta E_0$ and $-\Delta G$ are given in parentheses. The most stable isomers are given in boldface.

M06-2X/aug-cc-pVDZ, 29 cm⁻¹ for BSSE-uncorrected MP2/aug-cc-pVDZ, 34 cm⁻¹ for BSSE-corrected MP2/aug-cc-pVDZ). However, the energy difference in ΔE_e between 2-C_{2v} and 2-C_{2h} is too small; thus, the stability in ZPE-corrected energy (ΔE_0) could be changed. In a simple approximation that the low imaginary frequency is neglected for ZPE correction because this mode could be replaced by an internal translational or rotational mode of monomers in the cluster, the 2-C_{2h} isomer is more stable than the 2-C_{2v} isomer by 0.10, 0.42, and 0.33 kJ/mol at the M06-2X/aVDZ, MP2/CBS, and CCSD(T)/CBS levels, respectively. This indicates that the two isomers have a flat potential surface, and so the structure would show the quantum statistical distribution of configurations ($\theta = 0–90^\circ$) between the T-shaped ($\theta = 90^\circ$) and displaced stacked ($\theta = 0^\circ$) structures, as in the case of the C_{2v} vs C_s benzene–water cluster conformation,⁷⁰ water dimer with an excess electron,⁷¹ or the linear vs bent HCCN conformation.⁷²

For the 2-C_{2v} isomer, the predicted distance between two centers of mass of each monomer unit (*R*) is 4.42 Å and 4.44 Å at the M06-2X/aug-cc-pVDZ and BSSE-corrected MP2/aug-cc-pVDZ levels, respectively, in reasonable agreement with the experimental value 4.41 Å.⁵⁶ For the 2-C_{2h} isomer, the predicted *R* is 4.27 Å and 4.18 Å at the M06-2X/aug-cc-pVDZ and BSSE-corrected MP2/aug-cc-pVDZ levels, respectively, which is slightly shorter but still close to the experimental value 4.41 Å.⁵⁶ Thus, the structure having the quantum statistical distribution of the 2-C_{2v} to 2-C_{2h} configurations should be closer to the 2-C_{2v} configuration. The vertical distance between two acetylene structures of 2-C_{2h} is 2.75 Å, which is much shorter than the stacking distances (~3.4 Å) between stacking aromatic rings in organic crystals,⁷³ multiwalled carbon nanotube layers,⁷⁴ or DNA stacks.^{75,76} The calculated C≡C and C–H bond distances of the acetylene dimer (2-C_{2v}) are 1.201 and 1.066

Table 2. MP2 and CCSD(T) Interaction Energies and Thermodynamic Quantities (kJ/mol) for Low Energy Structures of the Acetylene Clusters $(C_2H_2)_{n=2-5}$ ^a

MP2/aVDZ; MP2/aVTZ; [MP2/CBS]				
name	$-\Delta E_0$	$-\Delta E_0$	$-\Delta H$	$-\Delta G(50\text{ K})$
2- C_{2v}	5.76; 6.65; [7.02]	2.87; 3.75; [4.13]	2.52; 3.41; [3.78]	0.07; 0.88; [1.26]
2- $C_{2h}(1)$	5.41; 6.13; [6.43]	3.53; 4.25; [4.55]	5.10; 5.82; [6.12]	0.54; 1.25; [1.55]
3- C_{3h}	17.94; 20.62; [21.74]	10.77; 13.44; [14.56]	10.10; 12.78; [13.90]	3.06; 5.73; [6.86]
3- $D_{2h}(1)$	11.22; 12.99; [13.74]	5.84; 7.61; [8.36]	5.97; 7.74; [8.49]	-1.37; 0.40; [1.15]
3- $C_2(2)$	10.99; 12.49; [13.12]	7.38; 8.88; [9.51]	9.14; 10.64; [11.27]	1.09; 2.59; [3.22]
3- C_{2v}	10.75; 12.51; [13.26]	5.05; 6.82; [7.56]	0.72; 4.76; [5.51]	-1.69; 0.07; [0.82]
4- S_4	27.09; 31.01; [32.66]	16.41; 20.33; [21.98]	14.49; 18.41; [20.07]	4.71; 8.63; [10.28]
4- C_{2h}	26.97; 30.83; [32.46]	15.85; 19.71; [21.34]	14.13; 18.00; [19.62]	4.51; 8.38; [10.00]
4- C_{2v}	26.58; 30.63; [32.33]	16.60; 20.64; [22.34]	14.50; 18.54; [20.24]	4.51; 8.56; [10.26]
4- $C_s(1)$	24.33; 27.94; [29.46]	14.62; 18.22; [19.74]	14.51; 18.11; [19.63]	3.19; 6.80; [8.32]
5- C_1	38.98; 44.77; [47.21]	27.76; 33.55; [35.98]	23.73; 29.52; [31.96]	11.01; 16.80; [19.24]
5- C_s	37.32; 42.81; [45.12]	22.72; 28.21; [30.52]	20.08; 25.57; [27.89]	6.31; 11.81; [14.12]
5- C_{2h}	36.36; 41.61; [43.82]	22.03; 27.28; [29.49]	19.28; 24.53; [26.74]	5.60; 10.85; [13.06]
5- C_1'	33.83; 39.17; [41.42]	20.50; 25.84; [28.09]	17.18; 22.52; [24.77]	4.13; 9.48; [11.73]
5- $C_{5h}(2)$	33.20; 37.96; [39.96]	19.63; 24.39; [26.40]	21.21; 25.97; [27.98]	2.03; 6.79; [8.79]
5- C_{2v}	32.64; 37.63; [39.74]	19.43; 24.42; [26.52]	15.87; 20.86; [22.96]	3.22; 8.21; [10.31]
5- $C_{2v}'(3)$	27.89; 32.00; [33.74]	16.28; 20.40; [22.13]	18.61; 22.73; [24.46]	-0.35; 3.77; [5.50]

CCSD(T)/aVDZ; [CCSD(T)/CBS]				
name	$-\Delta E_0$	$-\Delta E_0$	$-\Delta H$	$-\Delta G(50\text{ K})$
2- C_{2v}	5.05; [6.30]	2.15; [3.41]	1.81; [3.06]	-0.72; [0.54]
2- C_{2h}	4.60; [5.62]	2.72; [3.74]	4.29; [5.31]	-0.27; [0.74]
3- C_{3h}	15.56; [19.36]	8.34; [12.18]	7.72; [11.52]	0.68; [4.48]
3- D_{2h}	9.79; [12.31]	4.41; [6.93]	4.54; [7.06]	-2.80; [-0.28]
3- C_s	9.38; [11.51]	5.76; [7.89]	7.53; [9.66]	-0.53; [1.60]
4- S_4	23.50; [29.08]	12.82; [18.40]	10.91; [16.48]	1.12; [6.70]
4- C_{2h}	23.57; [29.06]	12.45; [17.94]	10.73; [16.22]	1.11; [6.60]
4- C_{2v}	22.63; [28.37]	12.64; [18.39]	10.54; [16.28]	0.56; [6.30]
4- C_s	21.07; [26.20]	11.36; [16.48]	11.25; [16.37]	-0.07; [5.06]
5- C_1	30.81; [39.03]	19.59; [27.81]	15.56; [23.78]	2.84; [11.07]
5- C_s	32.25; [40.06]	17.66; [25.46]	15.02; [22.83]	1.25; [9.06]
5- C_{2h}	31.52; [38.98]	17.19; [24.65]	14.45; [21.91]	0.76; [8.22]
5- C_1'	28.18; [35.78]	14.85; [22.44]	11.53; [19.12]	-1.52; [6.08]
5- C_{5h}	29.05; [35.82]	15.48; [22.25]	17.06; [23.83]	-2.12; [4.64]
5- C_{2v}	27.78; [34.87]	14.57; [21.66]	11.01; [18.10]	-1.64; [5.45]

^a The BSSE corrections are made. CCSD(T)/CBS energies were estimated by applying the correction term (the difference between MP2/aVDZ and CCSD(T)/aVDZ energies) to the MP2/CBS interaction energies, which were obtained with the extrapolation scheme utilizing the electron correlation error proportional to N^{-3} for the aug-cc-pVNZ basis set. In the CCSD(T)/CBS energies, the MP2/aVDZ thermal energies were used. If the structure is not a minimum, the numbers of imaginary frequencies are given in parentheses after the structure name. In these cases, the ZPE and thermal energy corrections are not reliable enough. The most stable isomers are given in boldface.

Table 3. Selected Geometrical Parameters of 2- C_{2v} , 2- C_{2h} , 3- C_{3h} , 4- S_4 , 4- C_{2v} , and 5- C_1 at the M06-2X/aVDZ (and MP2/aVDZ [Geometry-Optimized BSSE-Corrected MP2/aVDZ]) Levels

	2- C_{2v}	2- C_{2h}	3- C_{3h}	4- S_4
d_{CH}	2.81(2.67[2.81])	3.03(2.90[2.90])	2.76(2.60)	2.73(2.58)
d_{CM-CM}	4.42(4.30[4.44])	4.27(4.18[4.18])	4.37(4.25)	4.34(4.24)
$\theta(\angle HC_{cm}H)$	90	0	60	90

	4- C_{2v}	5- C_1
d_{CH}	2.62(2.53)	2.92(2.79)
d_{CM-CM}	4.22(4.15)	4.19(4.09)
$\theta(\angle HC_{cm}H)$	90	94

Å at the M06-2X/aug-cc-pVDZ level and 1.232 and 1.076 Å at the MP2/aug-cc-pVDZ level, which are close to the experimental values 1.203 and 1.062 Å. The acetylene dimer is predicted to be stable without dissociation below 55 K at 1 atm.

From the microwave and infrared spectra, Prichard et al.⁴⁷ reported that the structure of acetylene dimer is a twisted T-shaped hydrogen-bonded structure but not of the C_{2v}

symmetry. One of the acetylene monomers was twisted by $\theta = 63^\circ$ from the center of mass. The distance R is 4.38 Å, and the electric dipole moment of the dimer is 0.28 D. Most of MP2 calculations predict that the global minimum of the acetylene dimer is the T-shaped structure with C_{2v} symmetry. Alberts et al.⁵⁴ reported the binding energy of 6.90 kJ/mol. Bone and Handy⁵⁵ reported the C_{2v} symmetry structure with a distance R of 4.34 Å and a binding energy of 5.69 kJ/mol

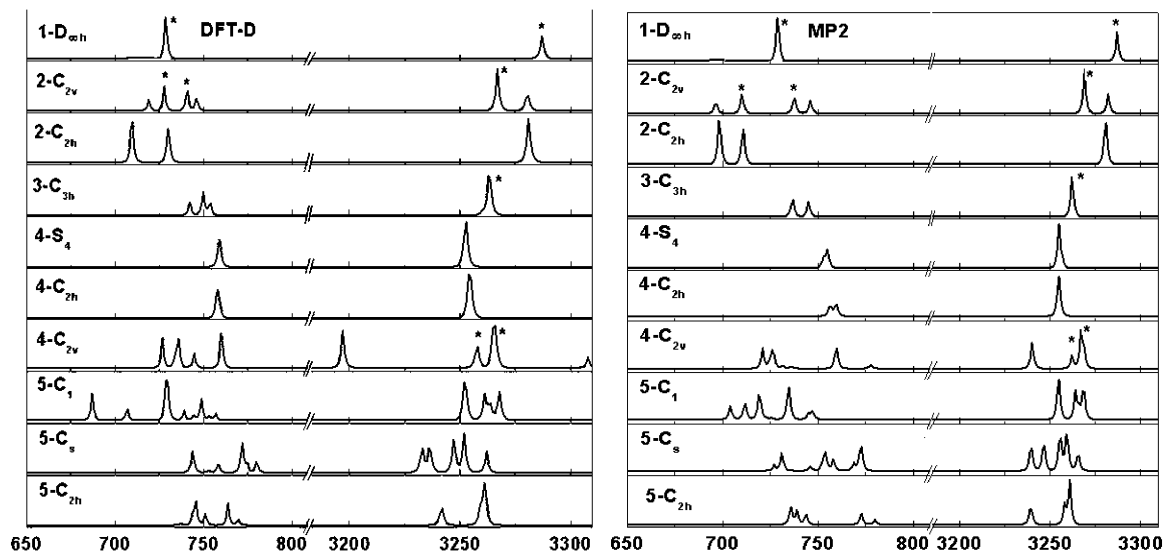


Figure 5. DFT-D/M06-2X and MP2 predicted vibrational spectra of $(\text{C}_2\text{H}_2)_{n=2-5}$. The M06-2X values are scaled by 0.942, while the exponential scale factor α is chosen as 0.000125 for the MP2/aug-cc-pVDZ values. The experimental frequencies are marked as asterisks.

at the MP2/TZ2P level. Hobza et al.⁵⁶ reported the C_{2v} structure with a distance R of 4.49 Å and a binding energy of 4.49 kJ/mol at the MP2/[DZ+(2df,2p)] level. Yu et al.⁵² and Karpfen⁵⁷ predicted that the C_{2v} structure (distance R : 4.32 Å) has a binding energy of 9.20 and 5.74 kJ/mol, respectively. Dykstra and Shuler⁵³ predicted that the T-shaped structure with C_{2v} symmetry has a distance R of 4.36 Å and a binding energy of 6.34 kJ/mol.

We have calculated the binding energy and thermodynamic quantities of the twisted acetylene dimer, but it is slightly less stable than the $2-C_{2v}$ structure. During the full optimization of the twisted T-shaped acetylene dimer, it becomes the T-shaped structure of $2-C_{2v}$. Thus, here we show that this structure is reconciled by considering the quantum statistical distribution of the $2-C_{2v}$ to $2-C_{2h}$ configurations for which the $2-C_{2v}$ is much more populated than the $2-C_{2h}$.

The study of the accurate quantum statistical distribution would be a challenging subject because it requires the full potential surface of the configurations. Here, we report the results. The MP2/CBS ΔE_c potential energy surface (PES) is presented in Figure 6a where the minimum potential well is at $\theta = 90^\circ$, while the MP2/CBS $\Delta E_0'$ PES is in Figure 6b, where the minimum potential well is at $\theta = 63^\circ$, which happens to be the same with the experimental value. $\Delta E_0'$ indicates the ZPE-corrected energy excluding only the ZPE along the C_{2v} - C_{2h} transition pathway eigenmode. The ΔE_c potential energy well is nearly harmonic at the minimum point ($\theta = 90^\circ$) where the MP2/aug-cc-pVDZ harmonic vibrational frequency is 39 cm^{-1} . On this potential surface, the anharmonicity-corrected fundamental vibrational frequency is evaluated to be 44 cm^{-1} ($E_1 - E_0$). The anharmonic vibrational frequency on the $\Delta E_0'$ potential energy well is evaluated as $\Delta E'_{1-0} = 49 \text{ cm}^{-1}$. From the wave functions at the vibrational ground state (Ψ_0 and Ψ_0' corresponding to ΔE and $\Delta E'$, respectively), the average angle ($\langle \theta \rangle$) is 78° on the ΔE_c PES and 53° on the $\Delta E_0'$ PES. Since ΔE PES and $\Delta E'$ PES would be considered the limiting cases without and with ZPEs for all the other eigenmodes of the PES, the

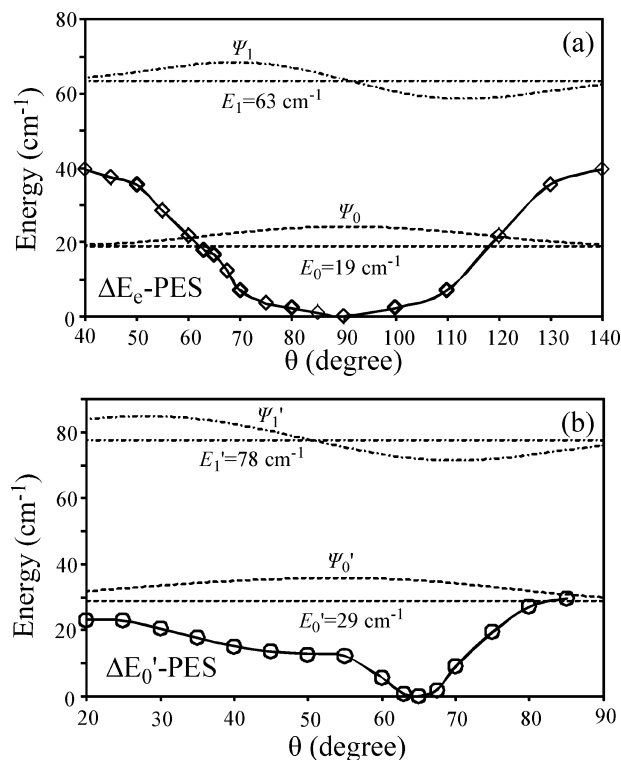


Figure 6. Anharmonic vibrational energies and wave functions for the MP2/CBS ΔE_c (a) and $\Delta E_0'$ (b) potential energy profiles with respect to the intermolecular angle (θ) of the acetylene dimer.

realistic value of $\langle \theta \rangle$ would be between the two cases; i.e., $\langle \theta \rangle$ would likely be 53 – 78° , in agreement with the experimentally observed value ($\theta = 63^\circ$). Although this study is based on the MP2/CBS PES, the results based on CCSD(T)/CBS would be similar because the difference between the two cases is small, as noted in Table 2.

The predicted M06-2X/aug-cc-pVDZ bending frequencies for the $2-C_{2v}$ isomer appearing at 728 and 741 cm^{-1} [average value of 735 cm^{-1}] is consistent with the experimental

frequency at 737 cm^{-1} . This band corresponds to the out-of-phase in-plane coupling of HCCH and is red-shifted by 1 and 12 cm^{-1} [average value of 6 cm^{-1}] as compared to the monomer frequencies. The predicted asymmetric stretching frequencies appear at 3267 and 3281 cm^{-1} [average value of 3274 cm^{-1}], which are 20 and 6 cm^{-1} [average value of 13 cm^{-1}] red-shifted from the acetylene monomer frequencies, in good agreement with the experimental frequency^{46,47} of 3273 cm^{-1} , which is blue-shifted by 6 cm^{-1} .

When we compare the acetylene dimer with the benzene dimer, we find some interesting results. In the case of benzene dimer, the C–H bond of the proton donor becomes shorter, showing a blue shift, whose unusual features were discussed by Hobza and co-workers.⁷⁷ However, in the case of the acetylene dimer, it shows the opposite trend, with a slightly increased bond length showing a red shift by $\sim 20\text{ cm}^{-1}$, as in a normal hydrogen bond.

The acetylene dimer in a nonlinear configuration by van der Waals interaction gives four fundamental frequencies. The remaining stretching frequencies may be represented as in-phase and out-of-phase couplings between acetylene monomer bonds. At the M06-2X/aug-cc-pVDZ level, the two strong peaks of the bending and asymmetric stretching frequencies for the $2\text{-}C_{2h}$ isomer appear at 730 and 3281 cm^{-1} , which are slightly smaller and larger than the experimental frequencies of 737 and 3273 cm^{-1} , respectively. This shows that the $2\text{-}C_{2v}$ configuration ($\theta = 90^\circ$) is more quantum statistically populated than the $2\text{-}C_{2h}$ configuration ($\theta = 0^\circ$), in agreement with the fact that the experimental value is 63° . The dipole moment of the $2\text{-}C_{2v}$ configuration ($\theta = 90^\circ$) is 0.36 D, while that of the $2\text{-}C_{2h}$ configuration ($\theta = 0^\circ$) is 0 D. In consideration of the quantum statistical distribution, the average dipole moment is near the experimental⁴⁷ value of 0.28 D.

In the case of the acetylene trimer (C_2H_2)₃, the cyclic structure of C_{3h} symmetry is much more stable than those of the other isomers because the C_{3h} structure has three sets of slightly slanted T-shaped $\pi\text{-H}$ interactions, while other structure has only two sets of $\pi\text{-H}$ interactions. The interaction energy of the acetylene trimer (C_{3h} isomer) is more than the sum of three single H– π (i.e., C_{2v} interaction energy) interaction energies, due to the relay effect which enhances the positive charge of H by 0.01 au (from 0.24 to 0.25 au) and the negative charge of the neighboring C by 0.03 au (from -0.26 to -0.23 a.u.). The $3\text{-}D_{2h}$ isomer ($\pi\text{-H}\text{-}\pi$ interaction type) is the transition state of the $3\text{-}C_{3h}$ isomer. To maximize the H– π interaction, the acetylene trimer has to be twisted. Similar structures were reported by Bone et al.,⁴⁹ Yu et al.,⁵² Dykstra and Shuler,⁵³ Alberts et al.,⁵⁴ and Brenner and Millie.⁵⁹ The distance from the C_3 symmetry axis to the center of mass of each monomer is reported to be 0.247, 2.668, 2.478, and 2.460 Å by Bone et al.,⁴⁹ Yu et al.,⁵² Dykstra and Shuler,⁵³ and Alberts et al.,⁵⁴ respectively. Our predicted distance is 2.526 Å, in good agreement with the experimental value of 2.514 Å. The predicted asymmetric vibrational stretching frequency at 3264 cm^{-1} for $3\text{-}C_{3h}$ is in close agreement with the experimental vibrational frequency at 3265.6 cm^{-1} . The predicted bending frequencies appear at 742, 750, and 754 cm^{-1} .

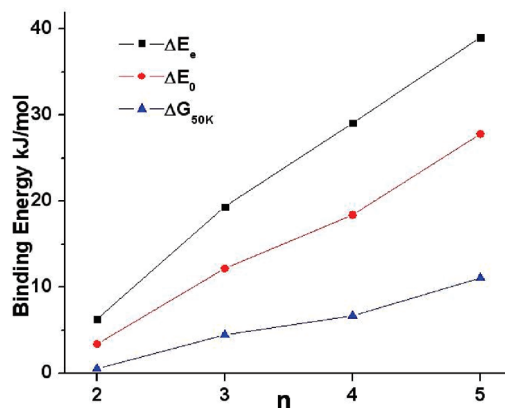


Figure 7. Thermodynamic properties of the low-energy structure of the $(\text{C}_2\text{H}_2)_{n=2-5}$ clusters.

In the case of the acetylene tetramer (C_2H_2)₄, one might expect that the structure is a planar configuration with C_{4h} symmetry, based on the previous discussion on the acetylene dimer and trimer in which more $\pi\text{-H}$ interaction bonds are formed between acetylene monomer units. We have carried out an extensive search for the most stable isomer. Previously reported structures ($4\text{-}S_4$ and $4\text{-}C_{4h}$)^{50,51} were also considered for comparison. Among many isomers, the lowest energy isomer in ΔE_e is the square shaped $4\text{-}C_{2h}$ and $4\text{-}S_4$ structures, which have four sets of $\pi\text{-H}$ interactions, while the lowest energy isomer in ΔE_0 is the bitriangular shaped $4\text{-}C_{2v}$ structure which has five sets of $\pi\text{-H}$ interactions. The $4\text{-}C_s$ isomer is the transition state of the $4\text{-}C_{2v}$ isomer. The asymmetric stretching vibrational frequencies of the $4\text{-}C_{2v}$ isomer (3258 , 3266 cm^{-1} ; average value: 3262 cm^{-1}) are consistent with the experimental value 3262 cm^{-1} .

In the case of the acetylene pentamer (C_2H_2)₅, Bone et al.⁵¹ proposed the “8”-shaped structure formed by two triangles ($5\text{-}C_{2h}$) based on the MP2/DPZ level. Dykstra and Shuler⁴¹ reported that the global minimum energy structure for the acetylene pentamer is $5\text{-}C_1$ on the basis of the MMC method, which has eight T-shaped interactions. We have done an extensive search to find the most stable structure. We find two isomers ($5\text{-}C_1$ and $5\text{-}C_s$) which are more stable than the previously reported structures ($5\text{-}C_{5h}$, $5\text{-}C_{2h}$). The $5\text{-}C_s$ is the most stable in ΔE_e , while the $5\text{-}C_1$ is the most stable in ΔE_0 . The $5\text{-}C_1$ isomer contains eight sets of $\pi\text{-H}$ interactions and a set of $\pi\text{-}\pi$ interactions. For the most stable $5\text{-}C_1$ structure, the predicted asymmetric stretching frequencies appear at 3252, 3261, 3264, and 3268 cm^{-1} , and the bending frequencies appear at 727, 736, and 760 cm^{-1} . These predicted vibrational stretching frequencies will be useful for experimentalists to identify the most stable acetylene pentamer. We note that the pentamer structure based on the MMC method by Dykstra and Shuler is consistent with our *ab initio* predicted global minimum structure ($5\text{-}C_1$).

Figure 7 shows the plot of the binding energies (ΔE_e , ΔE_0 , ΔG_{50K}) with respect to the cluster size n . These values almost linearly increase as the number of monomer

units increases. The dimer is stable below 55 K at 1 atm, while the pentamer is stable below 70 K 1 atm.

Conclusion

We have studied the geometrical isomers, energies, thermodynamic properties, and IR spectra of acetylene clusters $(C_2H_2)_{n=2-5}$. We have clarified the lowest energy structure of the $(C_2H_2)_{n=2-5}$ clusters. According to the CCSD(T)/CBS level of theory, the T-shape acetylene dimer of C_{2v} symmetry is the most stable in ΔE_e , but the displaced-stacked dimer of C_{2h} symmetry is also as stable as the $2-C_{2v}$ structure within 0.3 kJ/mol. This leads to the quantum statistical distribution of configuration over θ from 90° ($2-C_{2v}$ configuration) through 0° ($2-C_{2h}$ configuration), resulting in an average value of the angle $|\theta| = 53-78^\circ$, in good agreement with the experimental angle of $\theta = 63^\circ$. For the acetylene trimer, the cyclic $3-C_{3h}$ isomer is the most stable. In the case of the acetylene tetramer, the isomers $4-S_4$ and $4-C_{2h}$ are isoenergetic in ΔE_e , but the $4-C_{2v}$ isomer is the most stable in ΔE_0 and at nonzero temperatures. For the acetylene pentamer, we find two new structures ($5-C_1$ and $5-C_s$), which are more stable than the previously reported global minimum structure ($5-C_{5h}$ and $5-C_{2h}$). Although the $5-C_s$ structure is the most stable in ΔE_e , the $5-C_1$ structure is the most stable in ΔE_0 and at nonzero temperatures. These acetylene clusters are predicted to be stable below temperatures of 55–70 K at 1 atm. High-level *ab initio* calculated results are consistent with available experimental results. We also note that the structures predicted by Dykstra and Shuler⁴¹ based on the MMC method are mostly consistent with those predicted by high-level *ab initio* calculations.

Acknowledgment. This work was supported by NRF (WCU, R32-2008-000-10180-0; EPB Center, 2009-0063312, GRL, National Honor Scientist Program) and KISTI (KSC-2008-K08-0002).

Supporting Information Available: Vibrational frequencies (Table S1), binding energies and selected distances (Table S2), and rotational constants and dipole moment (μ) (Table S3) of $(C_2H_2)_{n=2-5}$ are available free of charge via the Internet at <http://pubs.acs.org/>.

References

- Chenoweth, K.; Dykstra, C. D. *Theor. Chem. Acc.* **2003**, *110*, 100.
- Hobza, P.; Selzle, H. L.; Schlag, E. W. *Chem. Rev.* **1994**, *94*, 1767.
- Brutschy, B. *Chem. Rev.* **2000**, *100*, 3891.
- Kim, K. S.; Tarakeshwar, P.; Lee, J. Y. *Chem. Rev.* **2000**, *100*, 4145.
- Rezac, J.; Fanfrlik, J.; Salahub, D.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1749.
- Maity, S.; Patwari, G. N.; Karthikeyan, S.; Kim, K. S. *Phys. Chem. Chem. Phys.* **2010**, *12*, 6150.
- Guin, M.; Patwari, G. N.; Karthikeyan, S.; Kim, K. S. *Phys. Chem. Chem. Phys.* **2009**, *11*, 11207.
- Vaupel, S.; Brutschy, B.; Tarakeshwar, P.; Kim, K. S. *J. Am. Chem. Soc.* **2006**, *128*, 5416.
- Tarakeshwar, P.; Kim, K. S.; Brutschy, B. *J. Chem. Phys.* **2001**, *114*, 1295.
- Dykstra, C. E.; Lisy, J. M. *THEOCHEM* **2000**, *500*, 375.
- Burley, S. K.; Petsko, G. A. *Science* **1985**, *229*, 23.
- Cerny, J.; Kabelac, M.; Hobza, P. *J. Am. Chem. Soc.* **2008**, *130*, 16055.
- Sponer, J.; Riley, K. E.; Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2595.
- Singh, N. J.; Lee, H. M.; Hwang, I.-C.; Kim, K. S. *Supramol. Chem.* **2007**, *19*, 321.
- Singh, N. J.; Lee, H. M.; Suh, S. B.; Kim, K. S. *Pure Appl. Chem.* **2007**, *79*, 1057.
- Ren, T.; Jin, Y.; Kim, K. S.; Kim, D. H. *J. Biomol. Struct. Dyn.* **1997**, *15*, 401.
- Hoeben, F. J. M.; Jonkheijm, P.; Meijer, E. W.; Schenning, A. P. H. *Chem. Rev.* **2005**, *105*, 1491.
- 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.
- 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.
- Pitonak, M.; Neogrady, P.; Rezac, J.; Jurecka, P.; Urban, M.; Hobza, P. *J. Chem. Theory Comput.* **2008**, *4*, 1829.
- Hohenstein, E. G.; Sherrill, C. D. *J. Phys. Chem. A* **2009**, *113*, 878.
- Tsuzuki, S.; Honda, K.; Fujii, A.; Uchimaru, T.; Mikami, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2860.
- Piacenza, M.; Grimme, S. *J. Am. Chem. Soc.* **2005**, *127*, 14841.
- Lee, E. C.; Hong, B. H.; Lee, J. Y.; Kim, J. C.; Kim, D.; Kim, Y.; Tarakeshwar, P.; Kim, K. S. *J. Am. Chem. Soc.* **2005**, *127*, 4530.
- Sponer, J.; Jurecka, P.; Hobza, P. *J. Am. Chem. Soc.* **2004**, *126*, 10142.
- Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095.
- Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. *J. Am. Chem. Soc.* **2002**, *124*, 10887.
- Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2000**, *122*, 3746.
- Hong, B. H.; Lee, J. Y.; Cho, S. J.; Yun, S.; Kim, K. S. *J. Org. Chem.* **1999**, *64*, 5661.
- Tarakeshwar, P.; Lee, S. J.; Lee, J. Y.; Kim, K. S. *J. Chem. Phys.* **1998**, *108*, 7217.
- Grimme, S. *Angew. Chem., Int. Ed.* **2008**, *47*, 3430.
- DiStasio, R. A., Jr.; Helden, G. V.; Steele, R. P.; Head-Gordon, M. *Chem. Phys. Lett.* **2007**, *437*, 277.
- Janowski, T.; Pulay, P. *Chem. Phys. Lett.* **2007**, *447*, 27.
- Hunter, C. A.; Sanders, J. K. M. *J. Am. Chem. Soc.* **1990**, *112*, 5525.
- Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Phys. Chem.* **1996**, *100*, 18790.

- (36) Sinnokrot, O. M.; Sherrill, D. C. *J. Phys. Chem. A* **2006**, *110*, 10656.
- (37) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 104.
- (38) Lee, E. C.; Kim, D.; Jureka, P.; Tarakeshwar, P.; Hobza, P.; Kim, K. S. *J. Phys. Chem. A* **2007**, *111*, 3446.
- (39) Singh, N. J.; Min, S. K.; Kim, D. Y.; Kim, K. S. *J. Chem. Theory Comput.* **2009**, *5*, 515.
- (40) Tarakeshwar, P.; Choi, H. S.; Kim, K. S. *J. Am. Chem. Soc.* **2001**, *123*, 3323.
- (41) Shuler, K.; Dykstra, C. E. *J. Phys. Chem. A* **2000**, *104*, 11522.
- (42) Fraser, G. T.; Suenram, R. D.; Lovas, F. J.; Pine, A. S.; Hougen, J. T.; Lafferty, W. J.; Muentner, J. S. *J. Chem. Phys.* **1988**, *89*, 6028.
- (43) Matsumurak, K.; Lovas, F. J.; Suenram, R. D. *J. Mol. Spectrosc.* **1991**, *150*, 576.
- (44) Pendley, R. D.; Ewing, G. E. *J. Chem. Phys.* **1983**, *78*, 3531.
- (45) Sakai, K.; Koide, A.; Kihara, T. *Chem. Phys. Lett.* **1977**, *47*, 416.
- (46) Ohshima, Y.; Matsumoto, Y.; Takami, M.; Kuchitsu, K. *Chem. Phys. Lett.* **1988**, *147*, 1.
- (47) Prichard, D. G.; Nandi, R. N.; Muentner, J. S. *J. Chem. Phys.* **1988**, *89*, 115.
- (48) Prichard, D.; Muentner, J. S.; Howard, B. J. *Chem. Phys. Lett.* **1987**, *135*, 9.
- (49) Bone, R. G. A.; Murray, C. W.; Amos, R. D.; Handy, N. C. *Chem. Phys. Lett.* **1989**, *161*, 166.
- (50) Bryant, G. W.; Eggers, D. F.; Watts, R. O. *Chem. Phys. Lett.* **1988**, *151*, 309.
- (51) Bone, R. G. A.; Amos, R. D.; Handy, N. C. *J. Chem. Soc., Faraday Trans.* **1990**, *86*, 1931.
- (52) Yu, J.; Shujun, S.; Bloor, J. E. *J. Phys. Chem.* **1990**, *94*, 5589.
- (53) Shuler, K.; Dykstra, C. E. *J. Phys. Chem. A* **2000**, *104*, 4562.
- (54) Alberts, I. L.; Rowlands, T. W.; Handy, N. C. *J. Chem. Phys.* **1988**, *88*, 3811.
- (55) Bone, R. G. A.; Handy, N. C. *Theor. Chim. Acta.* **1990**, *78*, 133.
- (56) Hobza, P.; Selzle, H. L.; Schlag, E. W. *Collect. Czech. Chem. Commun.* **1992**, *57*, 1186.
- (57) Karpfen, A. *J. Phys. Chem. A* **1999**, *103*, 11431.
- (58) Karpfen, A. *J. Phys. Chem. A* **1998**, *102*, 9286.
- (59) Brenner, V.; Millie, P. Z. *Phys. D.* **1994**, *30*, 327.
- (60) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (61) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, V.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.
- (62) Lee, S. J.; Chung, H. Y.; Kim, K. S.; Bull, *Korean Chem. Soc.* **2004**, *25*, 1061.
- (63) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639.
- (64) Min, S. K.; Lee, E. C.; Lee, H. M.; Kim, D. Y.; Kim, D.; Kim, K. S. *J. Comput. Chem.* **2008**, *29*, 1208.
- (65) Császár, A. G.; Allen, W. D.; Schaefer III, H. F. *J. Chem. Phys.* **1998**, *108*, 9751.
- (66) Kolar, M.; Hobza, P. *J. Phys. Chem. A* **2007**, *111*, 5851.
- (67) Lee, J. Y.; Hahn, O.; Lee, S. J.; Choi, H. S.; Mhin, B. J.; Lee, M. S.; Kim, K. S. *J. Phys. Chem.* **1995**, *99*, 2262.
- (68) Lee, J. Y.; Hahn, O.; Lee, S. J.; Choi, H. S.; Shim, H.; Mhin, B. J.; Kim, K. S. *J. Phys. Chem.* **1995**, *99*, 1913.
- (69) Kolaski, M.; Lee, H. M.; Choi, Y. C.; Kim, K. S.; Tarakeshwar, P.; Miller, D. J.; Lisy, J. M. *J. Chem. Phys.* **2007**, *126*, 074302.
- (70) Kim, K. S.; Lee, J. Y.; Choi, H. S.; Kim, J.; Jang, J. H. *Chem. Phys. Lett.* **1997**, *265*, 497.
- (71) Kim, J.; Lee, J. Y.; Oh, K. S.; Park, J. M.; Lee, S.; Kim, K. S. *Phys. Rev. A* **1999**, *59*, R930–933.
- (72) Kim, K. S.; Schaefer III, H. F.; Radom, L.; Pople, J. A.; Binkley, J. S. *J. Am. Chem. Soc.* **1983**, *105*, 4148.
- (73) 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.
- (74) Hong, H.; Small, J. P.; Purewal, M. S.; Mullokandov, A.; Sfeir, M. Y.; Wang, F.; Lee, J. Y.; Heinz, T. F.; Brus, L. E.; Kim, P.; Kim, K. S. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 14155.
- (75) Kim, K. S.; Clementi, E. *J. Am. Chem. Soc.* **1985**, *107*, 227.
- (76) McGaughey, G. B.; Gagne, M.; Rappe, A. K. *J. Biol. Chem.* **1998**, *273*, 15458.
- (77) Hobza, P.; Spirko, V.; Selzle, H. L.; Schlag, E. W. *J. Chem. Phys. A* **1998**, *102*, 2501.

Calculation of Dipole Transition Matrix Elements and Expectation Values by Vibrational Coupled Cluster Method

Subrata Banik and Sourav Pal*

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

M. Durga Prasad

School of Chemistry, University of Hyderabad, Hyderabad 500046, India

Received June 30, 2010

Abstract: An effective operator approach based on the coupled cluster method is described and applied to calculate vibrational expectation values and absolute transition matrix elements. Coupled cluster linear response theory (CCLRT) is used to calculate excited states. The convergence pattern of these properties with the rank of the excitation operator is studied. The method is applied to a water molecule. Arponen-type double similarity transformation in extended coupled cluster (ECCM) framework is also used to generate an effective operator, and the convergence pattern of these properties is compared to the normal coupled cluster (NCCM) approach. It is found that the coupled cluster method provides an accurate description of these quantities for low lying vibrational excited states. The ECCM provides a significant improvement for the calculation of the transition matrix elements.

I. Introduction

An accurate description of anharmonic molecular vibration is often necessary to account for the experimental results obtained from modern high resolution techniques of molecular spectroscopy. Several methods have been discussed in the literature over the past three decades. The vibrational self-consistence method (VSCF)^{1–5} and its generalizations to multiconfigurational reference functions (VMCSCF)^{6–10} has been developed and used extensively by several authors. The vibrational configuration interaction method (VCI)^{11–15} has also been developed and used for small molecules. The dimension of VCI matrix increases exponentially with the number of degrees of freedom. This makes VCI difficult to apply for large systems. Vibrational Moller–Plesset perturbation theory (VMP)^{16–18} has been used for the calculation of the vibrational spectra of many systems. Canonical van Vleck perturbation theory^{19–23} has been developed and applied extensively.

Recently, some attempts have been made to describe anharmonic molecular vibrations by the coupled-cluster

method (CCM).^{24–36} The CCM has been established as one of the most accurate techniques for the description of the many body systems.^{37–47} In this method, the ground-state wave function of a many body system is decomposed into a reference function and an exponential wave operator. The exact ground-state function in CCM is

$$|\psi_g\rangle = \exp(S)|\phi_{\text{ref}}\rangle \quad (1)$$

where $|\phi_{\text{ref}}\rangle$ is the reference wave function. The cluster operator S consists of connected singles, doubles, up to n -tuple excitation operators. The cluster matrix elements are determined from the equation

$$\langle\phi_c|e^{-S}He^S|\phi_{\text{ref}}\rangle = 0 \quad (2)$$

where $|\phi_c\rangle$ are the excited states. There are two advantages in the coupled cluster approach. First, the method is size consistent by virtue of the exponential ansatz. Second, again due to the exponential structure of the wave operator, the resulting wave function and energy are highly accurate in an approximate calculation even with a low order truncation of the cluster operator.

* Corresponding author e-mail: s.pal@ncl.res.in.

There are two approaches to construct the Fock space required for the coupled cluster calculations for molecular vibrations. The first method is the basis set representation in which the Fock space is constructed as a union of all k -particle Hilbert spaces constructed as the tensor products of basis functions of the appropriate degrees of freedom. This is the route followed mainly by Christiansen and co-workers.^{24–30} The second approach is to construct the Fock space using harmonic oscillator (HO) ladder operators acting on an appropriate vacuum state.^{31–36} In this representation, the cluster operators are given by

$$S = \sum S_{n_1 n_2 \dots} a_1^{\dagger n_1} a_1^{\dagger n_2} \dots \quad (3)$$

The vacuum state is a variationally optimized multidimensional Gaussian state, and a_i^\dagger/a_i are the usual creation/annihilation operators of the harmonic oscillator algebra defined with respect to the vacuum state. We use this representation. Because the ladder a_i^\dagger/a_i satisfy the canonical commutation relations, we term this as the bosonic representation of the CCM.

There are two routes for calculation of excited-state energies in the coupled cluster approach. In the first approach, variously called the coupled cluster linear response theory (CCLRT) or the coupled cluster equation of motion method (EOMCC),^{48–51} the excited-state energies are obtained as the eigenvalues of the similarity transformed effective Hamiltonian H_{eff}^N

$$H_{\text{eff}}^N = e^{-S} H e^S \quad (4)$$

The second approach uses a multireference coupled cluster theory tailored to describe the excited states directly.^{52–56} All of the vibrational applications to date have used the CCLRT to obtain the excited-state energies.^{28,30,34–36} In an earlier work, we had studied the convergence trends of the CCM in the bosonic representation³⁶ in terms of the rank (the maximum number of creation operators used to define the truncation in eq 3) of the cluster operator.

In this work, we turn our attention to the reliability of the CCM approach for the calculation of properties other than energies. Specifically, we study the convergence pattern of the CCM approach to the calculation of expectation values and transition matrix elements of the dipole moment operator. This is the first implementation of the coupled cluster method to study the expectation values and transition matrix elements in the context of molecular vibration. To the extent of our knowledge, no calculation on transition matrix elements is reported in the literature using the coupled cluster method even in electronic structure theory. The convergence pattern of the expectation values and the transition matrix elements are studied as a function of the rank of the excitation operator.

The similarity transformation of the Hamiltonian (eq 4) lies at the heart of the CCM. Consequently, the H_{eff} of eq 4 is not manifestly Hermitian. While this poses no problem in an exact calculation, in an approximate calculation where the basis set is truncated, the effective Hamiltonian can, and does on occasion, develop complex eigenvalues. One possibility of eliminating such complex eigenvalues is to use a unitary wave operator. The resulting equations, however,

generate an infinite series on the left-hand side of eq 2 and thus are subject to uncontrolled approximation.

An approximate way of treating such complex eigenvalues is to use a second similarity transformation inspired by the work of Arponen.^{57–59} The ground-state wave function is parametrized as

$$|\psi_g\rangle = e^S e^{-\sigma} |\phi_{\text{ref}}\rangle \quad (5)$$

Here, the generator of the second similarity transformation σ consists of the de-excitation operator alone. The effective Hamiltonian, H_{eff} , is now hermitized up to first order. We noted in our earlier study³⁶ that this modification of the wave operator eliminates some of the complex eigenvalues. According to the Lie algebraic decoupling theorem, the equation of motion for S is decoupled from the σ matrix elements in the exact limit.^{60,61} To distinguish the two approaches, we term them as normal coupled cluster method (NCCM) and extended coupled cluster method (ECCM) in the spirit of Arponen. The second goal of the present study is to see whether the ECCM approach offers any additional advantages over the NCCM approach for the calculations of expectation values and transition matrix elements.

The rest of this Article is organized as follows. In the next section, we describe the essential aspects of the calculation of expectation values and transition matrix elements from the CCM perspective. We have applied the formalism to water molecule and its isotopomers using an ab initio potential energy surface and dipole moment surface to understand the convergence properties of these quantities with respect to the truncation in the excitation operators. These results are presented in section III.

II. Theory

Within the Born–Oppenheimer approximation, the vibrational Hamiltonian for nonrotating molecules is given by

$$H = \sum_i \frac{P_i^2}{2} + V(Q) + V_c + V_w \quad (6)$$

Here, Q_i and P_i represent the mass weighted normal coordinates and their conjugate momenta. $V(Q)$ is the potential energy function. This is often approximated by a quartic polynomial in the Taylor series expansion

$$V = \frac{1}{2} \sum_i \omega_i^2 Q_i^2 + \sum_{i \leq j \leq k} f_{ijk} Q_i Q_j Q_k + \sum_{i \leq j \leq k \leq l} f_{ijkl} Q_i Q_j Q_k Q_l \quad (7)$$

V_c and V_w are the Coriolis coupling and the Watson's mass term, respectively.⁶² The formulation of CCM approach for molecular anharmonic vibration requires three steps. In the first step, Hartee approximation is invoked for the ground state. In this, a multi-dimensional Gaussian ansatz

$$\psi = N \exp[-(\sum_i \omega_i (Q_i - Q_i^0)^2/2)] \quad (8)$$

is optimized variationally with respect to ω_i and Q_i^0 . This optimized Gaussian function acts as vacuum state $|0\rangle$ for the

construction of Fock space of CCM. The harmonic oscillator ladder operators a_i^\dagger and a_i are defined with respect to this vacuum state

$$a_i = \sqrt{\frac{\omega_i}{2}} \left(Q_i - Q_i^0 + \frac{1}{\omega_i d(Q_i - Q_i^0)} \right) \quad (9)$$

$$a_i^\dagger = \sqrt{\frac{\omega_i}{2}} \left(Q_i - Q_i^0 - \frac{1}{\omega_i d(Q_i - Q_i^0)} \right) \quad (10)$$

The Hamiltonian is written in terms of these ladder operators. By definition, the optimized Hartree product satisfies the relation

$$a_i |0\rangle = 0 \quad (11)$$

In the second step, the ground-state wave function is parametrized as

$$|\psi_g\rangle = e^S |0\rangle \quad (12)$$

The cluster operator is expanded as

$$S = \sum_i s_i a_i^\dagger + \sum_{i < j} s_{ij} a_i^\dagger a_j^\dagger + \sum_{i < j < k} s_{ijk} a_i^\dagger a_j^\dagger a_k^\dagger + \dots \quad (13)$$

The working equations for coupled cluster ground-state energy and cluster matrix elements are given by

$$\langle 0 | H_{\text{eff}}^N | 0 \rangle = E_g \quad (14)$$

$$\langle e | H_{\text{eff}}^N | 0 \rangle = 0 \quad (15)$$

Here,

$$H_{\text{eff}}^N = e^{-S} H e^S \quad (16)$$

Equation 15 represents a set of coupled nonlinear equations that has to be solved iteratively. The detailed procedure to solve this set of equations is described in ref 36.

In the last step, we invoke the CCLRT for the descriptions of excited states. The excited-state wave function is written as

$$|\psi_e\rangle = e^{\Omega} |0\rangle \quad (17)$$

Here, Ω is a linear excitation operator, which is given by

$$\Omega = \sum_i \Omega_i a_i^\dagger + \sum_{i < j} \Omega_{ij} a_i^\dagger a_j^\dagger + \sum_{i < j < k} \Omega_{ijk} a_i^\dagger a_j^\dagger a_k^\dagger + \dots \quad (18)$$

The working equation for CCLRT to get excitation energies is

$$[H_{\text{eff}}^N, \Omega] |0\rangle = \Delta E \Omega |0\rangle \quad (19)$$

We now turn to the calculation of expectation values transition matrix elements. A straightforward approach based on the CCM ansatz for the expectation values leads to a nonterminating series^{63,64}

$$\begin{aligned} \langle \hat{O} \rangle &= \frac{\langle 0 | \exp(S^\dagger) \hat{O} \exp(S) | 0 \rangle}{\langle 0 | \exp(S^\dagger) \exp(S) | 0 \rangle} \\ &= \langle 0 | \exp(S^\dagger) \hat{O} \exp(S) | 0 \rangle_L \end{aligned} \quad (20)$$

making it impractical for the numerical work. Prasad⁶⁰ has earlier suggested an alternative approach for the calculation of expectation values and transition matrix elements within the CCM framework that bypasses the need to evaluate such infinite series. Here, it is recognized that because the CCM approach involves the construction and diagonalization of an effective Hamiltonian via the similarity transformation in eq 16, it is possible to relate the left and right eigenvectors of H_{eff} to the eigenvectors of the original Hamiltonian.

$$H_{\text{eff}} |R_i\rangle = E_i |R_i\rangle \quad (21a)$$

$$\langle L_i | H_{\text{eff}} = \langle L_i | E_i \quad (21b)$$

$$|\psi_i\rangle = N_i e^S |R_i\rangle \quad (22a)$$

$$\langle \psi_i | = \langle L_i | e^{-S} M_i \quad (22b)$$

By choosing the normalization constants M_i and N_i such that $M_i N_i \langle L_i | R_i \rangle = 1$, the expectation value of any arbitrary operator O is given by

$$\langle O \rangle = \langle \psi_i | O | \psi_i \rangle = \langle L_i | O_{\text{eff}} | R_i \rangle \quad (23)$$

where

$$O_{\text{eff}} = e^{-S} O e^S \quad (24)$$

These equations are identical to the equations derived by the Z -vector^{65,66} or λ -vector⁶⁷ formalism by earlier workers because all of these methods use a linearly parametrized left vector to calculate the expectation values. Similarly, the transition matrix elements between two states $|\psi_i\rangle$ and $|\psi_j\rangle$ are given by

$$|\langle \psi_i | O | \psi_j \rangle|^2 = \langle L_i | O_{\text{eff}} | R_j \rangle \langle L_j | O_{\text{eff}} | R_i \rangle \quad (25)$$

and the phase of the transition matrix element $\phi(O_{ij} = |O_{ij}| e^{i\phi})$ is given by

$$\phi = \frac{1}{2} \text{Im} [\ln \langle L_i | O_{\text{eff}} | R_j \rangle / \langle L_j | O_{\text{eff}} | R_i \rangle] \quad (26)$$

We use this approach for the calculation of the expectation values and transition matrix elements of the dipole operator.

The structure of the equations remains unchanged in the case of ECCM. The only difference in the case of the ECCM is that the effective operators defined in eq 24 are replaced by

$$O_{\text{eff}} = e^\sigma e^{-S} O e^S e^{-\sigma} \quad (27)$$

and the σ matrix elements are given by

$$\langle \phi_{\text{ref}} | e^\sigma e^{-S} H e^S e^{-\sigma} | \phi_e \rangle = 0 \quad (28)$$

As mentioned in the Introduction, the equations for the S -matrix elements are decoupled from the σ matrix elements

in the exact limit. We assume that this holds even in approximate calculations and solve eqs 15 and 28 sequentially.

III. Results and Discussion

We have applied the above-discussed methodology of section II to study the vibrational corrections to dipole moments of different vibrational states and transition matrix elements between ground state and several excited states of water molecule and isotopic variants HDO and D₂O. Over the years, there have been extensive studies on the vibrational spectra of water molecule.^{68–70} It is an archetypical local mode molecule because of the large mass disparity between oxygen and hydrogen atoms. Moreover, the low barrier of inversion makes it highly anharmonic. Consequently, it is a very good test molecule for any theoretical method based on a normal coordinate system. There are several accurate quartic ab initio potential energy surfaces (PES) available for these systems in the literature. However, there are very few dipole moment surfaces (DMS) reported in the literature. We have taken both PES and DMS from ref 23. In addition to the calculation of the potential energy surface and dipole moment surface, these authors made extensive calculations to the dipole moment expectation values and transition matrix elements using perturbation theory. We choose both PES and DMS based on CISD calculations using STO basis for applying our methodology to H₂O, HDO, and D₂O molecules. Although the potentials presented here are old, we chose these for consistency between PES and DMS in terms of basis set and method used in electronic structure calculations. The PES does not contain Coriolis coupling terms. Because the goal of the present work is to study the reliability of the effective operator approach based on the coupled cluster linear response theory rather than attaining experimental accuracy, we compared our results to converged full CI results. As is well-known, the quartic force field provides a poor description for the H₂O molecule.⁷¹ Consequently, it does not give numbers that can be compared to experimental data even with full CI level. A higher order expansion in the potential is required to match experimental values. The present methodology can be easily expanded for higher order potential functions. It will only add more terms in eq 15. Among these three molecules, we found that the deviations between CCM and converged full CI are maximum in the case of the H₂O molecule. Here, we discuss the results for the H₂O molecule.

A. NCCM-Based Calculation. In our earlier work,³⁶ we presented extensive calculations on the convergence of state energies with respect to the variation of the rank of both cluster operator S and excitation operator Ω from four boson to six boson level in NCCM and CCLRRT, respectively. In two illustrative examples of formaldehyde and water, we found that both the ground-state and the excited-state energies have converged with respect to cluster operator by S_4 in NCCM. However, in some cases, the results were not converged even with six boson rank of excitation operator Ω . On the basis of this, in the present work we study the convergence pattern of the dipole operator expectation values and transition matrix elements with respect to rank of excitation operator Ω only. In all calculations, we kept the

Table 1. Variation of Expectation Values of Dipole Moment of H₂O with Varying Excitation Operator from Four Boson to Six Boson^a

state	4 boson	5 boson	6 boson	full CI	PT2 ^b
000	0.88	0.88	0.88	0.87	0.90
010	-2.09	-2.10	-2.10	-2.09	-1.93
020	-5.39	-5.62	-5.71	-5.87	4.80
100	2.26	2.26	2.26	2.26	2.32
030	-7.07	-9.17	-10.36	-11.43	-7.71
110	-0.31	-0.28	-0.29	-0.26	-0.47
120	-2.44	-2.81	-2.82	-3.41	-3.30
200	3.24	3.38	3.36	3.42	3.69
002	6.57	6.65	6.65	6.87	7.12
210	0.25	0.54	0.67	1.41	0.95
012	1.70	2.06	2.07	4.53	4.35
300	3.04	4.20	4.58	4.34	5.02
102	4.75	6.98	7.72	7.00	8.38
001	3.93	3.93	3.92	3.93	4.03
011	1.28	1.34	1.34	1.40	1.23
021	-1.72	-1.64	-1.55	-1.66	-1.61
101	4.67	4.84	4.83	4.88	5.37
111	1.32	2.35	2.63	2.60	2.61
201	3.90	5.48	5.87	5.52	6.66
003	6.83	9.13	9.59	9.75	10.17

^aThe tabulated values are the vibrational corrections to the dipole moments ($\mu_{vv} - \mu_e$). Units are in 10^{-2} debye. ^bReference 23.

cluster operator of NCCM fixed at six boson level. We compare our results with the converged full CI results. For full CI, we used 10–18–10 harmonic oscillator basis. Comparisons have also been made with the second-order perturbation results of ref 23.

1. Expectation Values of Dipole Operator. In Table 1, we present the variation of expectation values of dipole operator with respect to truncation levels of excitation operator Ω , keeping the cluster operator fixed at six boson level. The values presented in the table are vibrational corrections to the dipole moment. The Z axis is taken as the molecular axis. The states with maximum three quanta excitations are reported here. We find that for the ground state and fundamentals, CCM results are in excellent agreement with the converged full CI results. The values are converged even with as low as four boson excitation operator. For all states with two quanta of excitations, the dipole moment expectation values are very close to the full CI values. In 200, 002, and 011 states, the values are nearly converged with the rank of excitation operator. In case of 020, 030, 003 states, we find that dipole moment is monotonically converging with respect to excitation operator but has not saturated even at six boson level. For most of the three quantum states, the error is about 5% except for three states (120, 210, and 012) for which the maximum error is as high as 50%. As can be seen, the expectation values are not converged with respect to Ω even at the six boson level. We find dramatic improvements in the dipole moment expectation values on increasing the rank of excitation operator from four boson to five boson to six boson in some cases. For example, the dipole moment of 030 state changes from -7.07×10^{-2} to -10.36×10^{-2} debye from four boson to six boson rank of excitation operator. The converged full CI value for this state is -11.43×10^{-2} debye. As we noted in our earlier work,³⁶ lower lying states like fundamentals, first overtones, etc., are well represented by the CCLRRT method because of its

Table 2. Variation of Absolute Transition Matrix Elements of H₂O with Varying Excitation Operator from Four Boson to Six Boson^a

state	4 boson	5 boson	6 boson	full CI	PT2 ^b
010	14.75	14.75	14.76	14.75	14.6
020	0.58	0.57	0.57	0.73	0.93
100	3.59	3.59	3.59	3.59	3.50
030	0.16	0.18	0.19	0.07	
110	0.10	0.10	0.10	0.06	0.22
200	0.44	0.44	0.44	0.35	0.44
002	0.08	0.07	0.08	0.01	0.08
012	0.13	0.13	0.13	0.02	
300	0.02	0.01	0.03	0.09	
102	0.11	0.11	0.11	0.03	
001	6.27	6.27	6.27	6.26	6.60
011	1.62	1.62	1.62	1.64	3.10
021	0.06	0.06	0.05	0.07	
101	0.94	0.94	0.94	0.78	1.10
111	0.34	0.35	0.35	0.32	
201	0.08	0.10	0.11	0.07	
003	0.14	0.14	0.13	0.07	

^a Values greater than 0.01 are reported. ^b Reference 23.

bivariational nature. However, truncation of the linear excitation operator at six boson level does not describe the wave functions of higher states adequately. The convergence pattern of states energies also reflects these improper descriptions of the higher excited-state wave functions.

In Table 1, we have compared the CCM dipole moment values with the second-order perturbation theory results presented in ref 23 also. We found that for almost all states the CCM results are better than the second-order perturbation results.

2. Transition Matrix Elements. The absolute values of the transition matrix elements of H₂O from the ground state to different excited states are presented in Table 2. Like state energies and dipole moment expectation values, here also we find excellent agreement between converged full CI and CCM with as low as four boson excitation operator for the fundamentals. Even for lower lying two quanta excited state 011 and three quanta excited state 111, we find that converged full CI values are reached by NCCM with four boson excitation operator. For two quanta states, the results are converged with the truncation of excitation operator, and they are close to converged full CI values except in the cases where the transition matrix element is very small. In our earlier study on energetics,³⁶ we found that the ground-state ket vector is well represented by as low as four body cluster operator S_4 . So the errors in transition matrix elements must be due to an inadequate description of the ground-state bra vector within the NCCM approach. In the NCCM approach, the ground-state bra vector is linearly parametrized. Parametrizing the ground-state bra vector by an exponential ansatz as is done in the ECCM approach should improve the transition matrix elements.

Finally, NCCM generally gives a better description for transition matrix elements than does the second-order perturbation theory.

B. ECCM-Based Calculation. In this section, we compare the results of different levels of truncation of Ω operator after Arponen-type double similarity transformation. In all calculations, both cluster operator S and σ are kept at the six boson level.

Table 3. Variation of Expectation Values of Dipole Moment of H₂O with Varying Excitation Operator from Four Boson to Six Boson after Arponen-Type Double Similarity Transformation^a

state	4 boson	5 boson	6 boson	full CI	PT2 ^b
000	0.88	0.88	0.88	0.87	0.90
010	-2.03	-2.04	-2.04	-2.09	-1.93
020	-5.46	-5.58	-5.60	-5.87	-4.80
100	2.25	2.25	2.25	2.26	2.32
030	-7.10	-9.42	-10.18	-11.43	-7.71
110	-0.28	-0.29	-0.29	-0.26	-0.47
120	-2.44	-2.76	-2.87	-3.41	-3.30
200	3.37	3.37	3.38	3.42	3.69
002	6.62	6.64	6.64	6.87	7.12
210	0.26	0.69	0.73	1.41	0.95
012	1.71	2.11	2.18	4.53	4.35
300	3.03	4.41	4.49	4.34	5.02
102	4.66	7.19	7.69	7.00	8.38
001	3.90	3.90	3.90	3.93	4.03
011	1.32	1.35	1.35	1.40	1.23
021	-1.71	-1.64	-1.62	-1.66	-1.61
101	4.79	4.82	4.83	4.88	5.37
111	1.33	2.51	2.63	2.60	2.61
201	4.04	5.70	6.01	5.52	6.66
003	6.65	9.24	9.44	9.75	10.17

^a The tabulated values are the vibrational corrections to the dipole moments ($\mu_{vv} - \mu_e$). Units are in 10^{-2} debye. ^b Reference 23.

Table 4. Variation of Absolute Transition Matrix Elements of H₂O after Arponen-Type Double Similarity Transformation with Varying Excitation Operator from Four Boson to Six Boson^a

state	4 boson	5 boson	6 boson	full CI	PT2 ^b
010	14.81	14.81	14.81	14.75	14.6
020	0.75	0.75	0.75	0.73	0.93
100	3.63	3.63	3.63	3.59	3.50
030	0.02	0.02	0.02	0.07	
110	0.04	0.04	0.04	0.06	0.22
200	0.34	0.34	0.34	0.35	0.44
002	0.01	0.01	0.01	0.01	0.08
012	0.03	0.03	0.03	0.02	
300	0.09	0.09	0.09	0.09	
102	0.00	0.02	0.02	0.03	
001	6.32	6.32	6.32	6.26	6.60
011	1.61	1.61	1.61	1.64	3.10
021	0.08	0.08	0.08	0.07	
101	0.78	0.78	0.78	0.78	1.10
111	0.32	0.32	0.32	0.32	
201	0.07	0.08	0.09	0.07	
003	0.10	0.08	0.08	0.07	

^a Values greater than 0.01 are reported. ^b Reference 23.

1. Dipole Moment Expectation Values. We present the variation of dipole moment expectation values for H₂O with different levels of truncation of excitation operator with ECCM in Table 3. Like energetics of the states, here also we find the improvement due to double similarity transformation over NCCM is marginal.

2. Transition Matrix Elements. Variation of the transition matrix elements with rank of excitation operator for H₂O is given in Table 4. Here, we find significant improvements due to the double similarity transformation of ECCM approach over NCCM. For fundamentals, the NCCM results are very close to converged full CI. So improvements due to ECCM over NCCM are marginal. Beyond the fundamental states, we find dramatic improvements with ECCM-based

calculations. For example, with the NCCM-based method, the converged transition matrix element value for the 020 state is 0.57×10^{-2} debye, whereas the full CI value is 0.73×10^{-2} debye. With the ECCM-based calculation, it improves to 0.75×10^{-2} debye. Similarly, for the 002 state, the full CI value is 0.01×10^{-2} debye. NCCM-based calculation gives 0.08×10^{-2} debye, whereas ECCM gives the exact full CI value. Similarly, for 300, 101, 111 states, we find exact full CI values with the ECCM-based method. In all of the cases, the errors by the ECCM-based method are negligible. As we stated earlier, in the NCCM-based method, the ground-state bra vector is not properly described. In the ECCM, the ground-state bra vector is parametrized with an exponential operator. This makes the ECCM approach significantly superior over NCCM in calculating transition matrix elements.

We find a similar convergence pattern of expectation values and transition matrix elements in HDO and D₂O molecules. For low energy states, particularly states with two quanta excitations (in some cases with three quanta excitations, e.g., 030, 300, etc.), the results are converged with respect to full CI even with four boson operator rank. Some higher energy excited values have not reached the converged full CI values, but they are monotonically converging toward full CI results. Like the H₂O molecule, ECCM does not give any significant improvement over the NCCM approach in calculating the expectation values. However, transition matrix elements are better represented by the ECCM-based approach than by the NCCM-based approach.

IV. Conclusion

In this work, we presented an effective operator approach within the framework of CCM to calculate expectation values of operators and absolute transition matrix elements. We conclude that these properties can be calculated very accurately using CCLRT. We studied the convergence pattern of these properties with respect to truncations of excitation operator in CCLRT. We found that for fundamentals and most of the states with two quanta excitations, these properties are converged with the rank of excitation operator and reached full CI limit by Ω_4 . For higher states, the values tend to approach full CI values on going from four boson to six boson rank of excitation operator.

Next, we turn to the utility of Arponen-type double similarity transformation. We found that the ECCM does not offer any significant advantage over NCCM as far as state energies and expectation values are concerned. However, the story is quite different in case of transition matrix elements. Here, the ECCM fares far better than the NCCM, particularly when the transition matrix elements are small.

The CCLRT approach with a low rank excitation operator does not appear to be suitable for the description of highly excited states. As the number of quanta of excitation in a molecule increases, the wave function samples a larger region of coordinate space and, consequently, is affected to a greater extent by the anharmonicity. This has some intriguing consequences on the wave functions. For example, in a system described by quartic potential, the centroid of the wave functions would move away from the origin in the

energy regime dominated by cubic terms of the potential, but would return toward the origin as the quartic term becomes significant at higher energies. Thus, a proper description of the shifting of the wave function centroids and changes in their effective frequencies is necessary to describe such states. The CCLRT, with its linear structure, is perhaps not the best way to parametrize such changes. A multireference CCM for the excited states that describes the shifts in the centroids and frequencies in a state-specific manner might provide a better description. Efforts in this direction are in progress and will be presented in due course.

Acknowledgment. S.B. acknowledges financial support from the University Grant Commission, India. S.P. acknowledges partial financial support from the SSB grant of CSIR, India and the J. C. Bose fellowship grant of DST, India. Facilities at the Centre of Excellence in Scientific Computing at NCL are acknowledged. M.D.P. acknowledges support from UGC, India in the form of CAS to the School of Chemistry and DST, India for use of the HPCF facility at the University of Hyderabad.

References

- (1) Carney, D. C.; Sprandel, L. L.; Kern, C. W. *Adv. Chem. Phys.* **1978**, *37*, 305.
- (2) Bowman, J. M. *J. Chem. Phys.* **1978**, *68*, 608.
- (3) Bowman, J. M. *J. Chem. Phys.* **1986**, *19*, 202.
- (4) Christoffel, K. M.; Bowman, J. M. *Chem. Phys. Lett.* **1985**, *85*, 220.
- (5) Culot, F.; Liévin, J. *Phys. Scr.* **1992**, *46*, 502.
- (6) Culot, F.; Liévin, J. *Theor. Chim. Acta* **1994**, *89*, 227.
- (7) Culot, F.; Laruelle, F.; Liévin, J. *Theor. Chim. Acta* **1995**, *92*, 221.
- (8) Heislbetz, H.; Rauhut, G. *J. Chem. Phys.* **2010**, *132*, 124102.
- (9) Drukker, K.; Hammes-Schiffer, S. *J. Chem. Phys.* **1997**, *107*, 363.
- (10) Webb, S. P.; Hammes-Schiffer, S. *J. Chem. Phys.* **2000**, *113*, 5214.
- (11) Christoffel, K. M.; Bowman, J. M. *Chem. Phys. Lett.* **1982**, *85*, 220.
- (12) Carter, S.; Handy, N. C. *Comput. Phys. Rep.* **1986**, *5*, 117.
- (13) Carter, S.; Meyer, W. *J. Chem. Phys.* **1990**, *93*, 8902.
- (14) Tennyson, J. *Comput. Phys. Rep.* **1986**, *4*, 1.
- (15) Henderson, J. R.; Tennyson, J.; Sutcliffe, B. T. *J. Chem. Phys.* **1992**, *96*, 2426.
- (16) Jung, J. O.; Gerber, R. B. *J. Chem. Phys.* **1996**, *105*, 10332.
- (17) Chaban, G.; Jung, J. O.; Gerber, R. B. *J. Chem. Phys.* **1999**, *111*, 1823.
- (18) Chirstiansen, O. *J. Chem. Phys.* **2003**, *119*, 5773.
- (19) Sibert, E. L. *J. Chem. Phys.* **1988**, *88*, 4378.
- (20) McCoy, A. B.; Sibert, E. L. *J. Chem. Phys.* **1991**, *95*, 3476.
- (21) McCoy, A. B.; Sibert, E. L. *J. Chem. Phys.* **1991**, *95*, 3488.
- (22) McCoy, A. B.; Sibert, E. L. *J. Chem. Phys.* **1990**, *92*, 1893.
- (23) Ermler, W. C.; Rosenberg, B. J.; Shavitt, I. In *Comparison of Ab Initio Quantum Chemistry with Experiment for Small*

- Molecules*; Bartlett, R. J., Ed.; Reidel: Dordrecht, 1985; p 171.
- (24) Sree Latha, G.; Prasad, M. D. *Chem. Phys. Lett.* **1995**, *241*, 215.
- (25) Christiansen, O. *J. Chem. Phys.* **2004**, *120*, 2140.
- (26) Christiansen, O. *J. Chem. Phys.* **2004**, *120*, 2149.
- (27) Christiansen, O. *J. Chem. Phys.* **2005**, *122*, 194105.
- (28) Seidler, P.; Christiansen, O. *J. Chem. Phys.* **2007**, *126*, 204101.
- (29) Christiansen, O. *Theor. Chem. Acc.* **2006**, *116*, 106.
- (30) Seidler, P.; Mattito, E.; Christiansen, O. *J. Chem. Phys.* **2009**, *131*, 034115.
- (31) Madhavi Sastry, G.; Durga Prasad, M. *Theor. Chim. Acta* **1994**, *89*, 193.
- (32) Madhavi Sastry, G.; Durga Prasad, M. *Chem. Phys. Lett.* **1994**, *228*, 213.
- (33) Sree Latha, G.; Prasad, M. D. *J. Chem. Phys.* **1996**, *105*, 2972.
- (34) Prasad, M. D. *Indian J. Chem.* **2000**, *39A*, 196.
- (35) Nagalakshmi, V.; Lakshminarayana, V.; Sumithra, G.; Prasad, M. D. *Chem. Phys. Lett.* **1994**, *217*, 279.
- (36) Banik, S.; Pal, S.; Prasad, M. D. *J. Chem. Phys.* **2008**, *129*, 134111.
- (37) Cizek, J. *Adv. Chem. Phys.* **1969**, *14*, 35.
- (38) Cizek, J. *J. Chem. Phys.* **1966**, *45*, 4256.
- (39) Bartlett, R. J. *Annu. Rev. Phys. Chem.* **1981**, *32*, 359.
- (40) Bartlett, R. J.; Musial, M. *Rev. Mod. Phys.* **2007**, *79*, 291.
- (41) Cederbaum, L. S.; Alon, O. E.; Streltsov, A. I. *Phys. Rev. A* **2006**, *73*, 043609.
- (42) Farnell, D. J. J.; Zinke, R.; Schulenburg, J.; Richter, J. *J. Phys. C* **2009**, *21*, 406002.
- (43) Dean, D. J.; Gour, J. R.; Hagen, G.; Hjorth-Jensen, M.; Kowalski, K.; Papenbrock, T.; Piecuch, P.; Wloch, M. *Nucl. Phys. A* **2005**, *752*, 299.
- (44) Hsue, C. S.; Chern, J. L. *Phys. Rev. D* **1984**, *29*, 643.
- (45) Bishop, R. F.; Flynn, M. F. *Phys. Rev. A* **1988**, *38*, 2211.
- (46) Bishop, R. F.; Bosca, M. C.; Flynn, M. F. *Phys. Rev. A* **1989**, *40*, 3484.
- (47) Bishop, R. F.; Bosca, M. C.; Flynn, M. F. *Phys. Lett. A* **1988**, *132*, 420.
- (48) Monkhorst, H. J. *Int. J. Quantum Chem., Symp.* **1977**, *11*, 421.
- (49) Mukherjee, D.; Mukherjee, P. K. *Chem. Phys.* **1979**, *39*, 325.
- (50) Comeau, D. C.; Bartlett, R. J. *Chem. Phys. Lett.* **1993**, *207*, 414.
- (51) Emrich, K. *Nucl. Phys. A* **1981**, *351*, 379.
- (52) Mukherjee, D.; Pal, S. *Adv. Quantum Chem.* **1989**, *20*, 291.
- (53) Pal, S.; Rittby, M.; Bartlett, R. J.; Sinha, D.; Mukherjee, D. *J. Chem. Phys.* **1988**, *88*, 4357.
- (54) Vaval, N.; Pal, S. *Chem. Phys. Lett.* **1999**, *300*, 125.
- (55) Lindgren, I.; Mukherjee, D. *Phys. Rep.* **1987**, *151*, 93.
- (56) Jeziorski, B.; Monkhorst, H. J. *Phys. Rev. A* **1981**, *24*, 1668.
- (57) Arponen, J. *Ann. Phys.* **1983**, *151*, 311.
- (58) Arponen, J.; Bishop, R. F.; Pajanne, E. *Phys. Rev. A* **1987**, *36*, 2519.
- (59) Arponen, J.; Bishop, R. F.; Pajanne, E. *Phys. Rev. A* **1987**, *36*, 2539.
- (60) Prasad, M. D. *Theor. Chim. Acta* **1994**, *88*, 383.
- (61) Latha, S. L.; Prasad, M. D. *Theor. Chim. Acta* **1993**, *86*, 511.
- (62) Watson, J. K. G. *Mol. Phys.* **1968**, *15*, 479.
- (63) Pal, S.; Prasad, M. D.; Mukherjee, D. *Theor. Chim. Acta* **1983**, *62*, 523.
- (64) Noga, J.; Urban, M. *Theor. Chim. Acta* **1988**, *73*, 291.
- (65) Salter, E. A.; Trucks, G.; Bartlett, R. J. *J. Chem. Phys.* **1989**, *90*, 1752.
- (66) Pal, S.; Ghosh, K. B. *Curr. Sci.* **1992**, *63*, 667.
- (67) Christiansen, O.; Jørgensen, P.; Hättig, C. *Int. J. Quantum Chem.* **1998**, *68*, 1.
- (68) Lodi, L.; Tolchenov, R. N.; Tennyson, J.; Lynus-Gray, A. E.; Shirin, S. V.; Zovok, N. F.; Polyanksy, O. L.; Csaszar, A. G.; van Stralen, J. N. P.; Visscher, L. *J. Chem. Phys.* **2008**, *128*, 44304.
- (69) Csaszar, A. G.; Czako, G.; Furtenbacher, T.; Tennyson, J.; Szalay, V.; Shirin, S. V.; Zovok, N. F.; Polyanksy, O. L. *J. Chem. Phys.* **2005**, *1282*, 214305.
- (70) Jorgensen, U. G.; Jensen, P. *J. Mol. Spectrosc.* **1993**, *161*, 219.
- (71) Csaszar, A. G.; Mills, I. M. *Spectrochim. Acta, Part A* **1997**, *53*, 1101.

CT1003669

Effects of Discrete Charge Clustering in Simulations of Charged Interfaces

John M. A. Grime* and Malek O. Khan

Department of Physical and Analytical Chemistry, Physical Chemistry, Uppsala University, Uppsala, Sweden

Received January 7, 2010

Abstract: A system of counterions between charged surfaces is investigated, with the surfaces represented by uniform charged planes and three different arrangements of discrete surface charges - an equispaced grid and two different clustered arrangements. The behaviors of a series of systems with identical net surface charge density are examined, with particular emphasis placed on the long ranged corrections via the method of “charged slabs” and the effects of the simulation cell size. Marked differences are observed in counterion distributions and the osmotic pressure dependent on the particular representation of the charged surfaces; the uniformly charged surfaces and equispaced grids of discrete charge behave in a broadly similar manner, but the clustered systems display a pronounced decrease in osmotic pressure as the simulation size is increased. The influence of the long ranged correction is shown to be minimal for all but the very smallest of system sizes.

1. Introduction

The study of charged interfaces is of interest in a variety of diverse fields, for example electrochemistry,¹ pharmaceutical research,² and membrane biology.^{3–5} Perhaps the most common traditional models of such interfaces use the Poisson–Boltzmann approximation for the treatment of electrostatic interactions, a “mean field” approach which predicts a purely repulsive interaction between like-charged interfaces. The absence of counterion correlations in such a model fail to predict the existence of attractive regimes,^{6,7} however, and hence can be of less general applicability where length scales are on the order of nanometres. Such microscopic effects may be significant in the study of a variety of biochemical processes such as vesicle/liposome aggregation and fusion⁴ as well as for colloidal stability.⁷ Charged interfaces are conventionally approximated in simulation as uniformly charged surfaces,^{6–10} but at the microscopic level surface charge exists as discrete packets. It is therefore of interest to examine any differences in the energy and pressure produced in model systems featuring different representations of an identical net surface charge density, and it is reasonable to assume that the ability of the discrete charges to collocate into clusters will affect the surface–surface interactions as

well as the counterion distributions. In order to treat these effects accurately, it is essential to correctly treat the long-range electrostatics.

Naji and Podgornik¹¹ have demonstrated how two surfaces, with a quenched charge disorder, have an attractive component to the force between the two surfaces. Mamasakhlisov et al.¹² extended this field theory approach and state that partial annealing of mobile wall ions, which move on a slower time scale than the counterions, enlarges the attractive interwall force. Using both a field-theoretical approach and Monte Carlo simulations, Fleck and Netz,^{13,14} have shown that surface charge disorder leads to counterions being more attracted to the surface in comparison with ordered surface charges. Studying a slightly different problem, Naydenov et al.¹⁵ show theoretically that two different charged wall species can form finite domains and that the resulting domains depend both on short-range attractive forces and electrostatics (via the salt concentration).

We present certain differences in the properties of a system of counterions between like-charged surfaces, modeled using Monte Carlo simulations of both uniformly charged surfaces and surfaces bearing discrete charges at various levels of surface charge clustering. Particular attention is paid to the energy of a counterion as a function of the z coordinate (perpendicular to the charged surfaces) and the osmotic

* Corresponding author e-mail: john.grime@fki.uu.se.

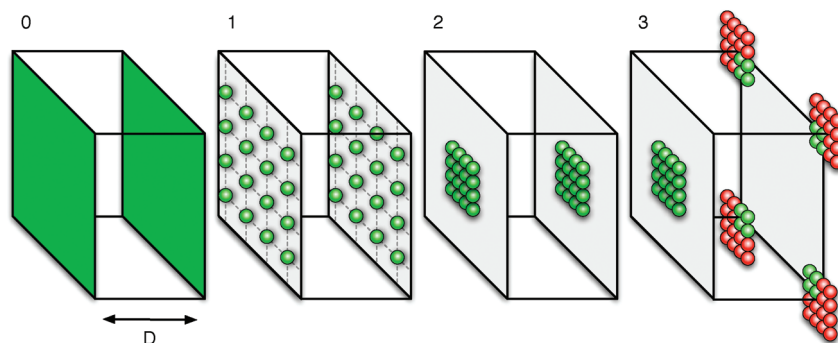


Figure 1. Illustration of the four system types studied here. From left to right; uniformly charged planes (system 0, planes shown in green), equidistantly spaced discrete wall charges (system 1, wall charges shown in green), clustered discrete wall charges directly opposite (system 2), clustered discrete wall charges with maximum separation between clusters on opposing walls (system 3, periodic wall charges shown in red to emphasize periodic boundaries). Counterions not shown.

pressure with respect to the potential system size dependence for a commonly used simulation technique.

Four types of systems are considered, as shown in Figure 1. Briefly, these consist of two parallel planes of uniform surface density σ (system 0), a lattice arrangement of equispaced discrete wall charges (system 1) and two arrangements of discrete wall charges condensed into square clusters positioned directly opposite one another on the two walls (system 2) and with the maximum spacing possible in the simulation cell (system 3).

2. Method

Monte Carlo simulations for counterions contained between a pair of uniformly charged surfaces are well established.^{6–10} The energy of the system is described as a combination of interionic effects (including Coulomb interactions and a hard sphere potential), ion-wall interactions, and a long ranged correction to account for the relatively slow decay of the electrostatic potential and the limited system size

$$U = U_q + U_{\text{HS}} + U_w + U_{\text{LRC}} \quad (1)$$

Here the interionic potentials U_q and U_{HS} are sums over the N interacting particles

$$U_q = \sum_{i=1}^{N-1} \sum_{j=i+1}^N \frac{q_i q_j e^2}{4\pi\epsilon_0\epsilon_1 r_{ij}} \quad (2)$$

$$U_{\text{HS}}(i,j) = \sum_{i=1}^{N-1} \sum_{j=i+1}^N \begin{cases} \infty & \text{where } r_{ij} < r_{\text{HS}} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

with q_1 and q_2 being the charges on the ions (in elementary charge units, e), and ϵ_1 is the permittivity of the medium relative to that of vacuum (ϵ_0). The separation between particles i and j is r_{ij} , and any overlap of the hard spheres surrounding the charges (i.e. $r_{ij} < r_{\text{HS}}$) produces an infinite energy penalty to ensure such a situation does not occur during simulation. We consider only single-valence ions with a hard sphere radius of 1 Å. The choice of $\epsilon_1 = 78.5$ is made to approximate an aqueous solution, giving the Bjerrum length, $l_B = e^2/(4\pi\epsilon_0\epsilon_1 k_B T) \approx 7.1$ Å at the simulated temperature of 300 K.

Ion-wall interactions are simply those of a point charge and an infinite plane of uniform surface charge density σ

$$U_w = \sum_{i=1}^N -\frac{\sigma q_i e}{2\epsilon_0\epsilon_1} \left(\left| \frac{D}{2} + z_i \right| + \left| \frac{D}{2} - z_i \right| \right) \quad (4)$$

Note that here we assume the plane at $z = 0$ bisects the two charged walls; hence, the origin lies in the center of the simulation cell, and $z = \pm(D)/2$ are the positions of the charged wall planes. The energy of the interaction between the two walls themselves is constant during simulation and is therefore ignored for the purposes of the Monte Carlo procedure. Where discretely charged walls are present, the energy expression becomes somewhat simpler - the counterion/wall interactions are already considered explicitly via the Coulomb pair potential (in the minimum image convention) and its long ranged correction, and hence U_w above is ignored.

To ensure a flat plane of closest approach for the counterions, we add an excluded region of volume extending to twice the ion hard sphere radius in front of each wall. This also provides identical accessible volume to the counterions for comparable systems with either uniform or discretely charged walls.

Finally, the long ranged correction to the electrostatic energy is performed under the “charged planes” formulation of Valleau et al.⁷ All discrete charges in the system (counterions and discrete wall charges, where present) contribute to an average charge density profile on the axis perpendicular to the wall planes. This average distribution forms the charge density on a series of charged slabs (infinite in the plane parallel to the walls) arranged in a stack between the wall planes. The long ranged correction to the electrostatic energy for a discrete charge is then the interaction with this series of infinite slabs (cf. eq 4) minus the interaction of the charge with the finite regions of each slab which lie inside the simulation cell

$$U_{\text{LRC}} = \sum_{i=1}^N \sum_{s=1}^S -\frac{\sigma q_i e}{2\epsilon_0\epsilon_1} \left(|s_i - z_i| + \frac{W}{2\tau} f(|s_i - z_i|/W) \right) \quad (5)$$

Table 1. Charged Wall Dimension W as a Function of L for the Systems Studied

L	1	2	3	4	5	6	7	8	9
$W, \text{\AA}$	12.66	25.32	37.97	50.63	63.29	75.95	88.60	101.26	113.92

where W is the x dimension of the simulation cell on the wall plane (assumed to be square, so x and y dimensions are both W) and

$$f(z) = 4 \ln\left(\frac{0.5 + r_1}{r_2}\right) - 4z \left(\sin^{-1} \left[\frac{r_2^2 + 0.5 \times r_1}{0.5 \times r_2 + r_1 r_2} \right] + \tan^{-1} \left[\frac{1}{2z} - \frac{\pi}{2} \right] \right) \quad (6)$$

with

$$r_1 = (0.5 + z^2)^{1/2} \quad (7)$$

$$r_2 = (0.25 + z^2)^{1/2} \quad (8)$$

Using this procedure, the average charge distribution normal to the wall planes (and hence the long ranged contribution to the electrostatic energy) may be generated self-consistently.

We adopt the convention of describing the wall charges with the parameter L ; for example, $L = 4$ describes an set of 4×4 wall charges which lie on the wall plane. By changing L , we can investigate the size dependence of the systems at a constant total surface charge density provided we ensure appropriate simulation cell dimensions and enforce charge neutrality with a suitable number of counterions (with the net counterion concentration also held constant across simulations). A series of systems with $1 \leq L \leq 9$ is considered, and for the case of system type 0 (see Figure 1) the walls are represented by uniformly charged planes of the appropriate area which ensures charge neutrality and dimensions identical to a discretely charged system of equal L and σ . The wall dimensions in the simulation cell which correspond to the studied values of L for $\sigma = 0.1 \text{ Cm}^{-2}$ are listed in Table 1. In all cases, the separation D between the planes of closest approach for a counterion to the charged walls is 15 \AA , more than twice the Bjerrum length for a system of this nature.

The osmotic pressure is measured at the midplane of the system,^{6,7,16} where the counterion concentration varies slowly. Such an approach helps to reduce uncertainty in the kinetic contribution to the pressure, which may be poorly defined in the regions of rapidly changing concentration adjacent to the charged planes; such an artifact can lead to difficulties where the pressure is evaluated according to, for example, the contact theorem.^{6,17} We calculate the pressure according to the formulation of Valleau et al.,⁷ with certain changes; Valleau et al. simplified the expression by noting that the charge on each plane exactly balanced, on average, the charge from the counterions in one-half of the simulation cell, and hence certain components of the pressure cancel for uniformly charged walls. This is not the case for systems with discrete wall charges, and hence we use the full pressure expression with no simplifying assumptions in each simula-

tion. This treatment was confirmed to reproduce the values of the simplified expression⁷ for systems with uniformly charged walls. The long ranged correction to the electrostatic energy for each system studied was generated using 2.5×10^7 Monte Carlo cycles, with a cycle defined as n attempted moves of a randomly selected counterion, with n being the number of counterions in the system. All data used herein were produced from an additional 2.5×10^7 Monte Carlo cycle.

3. Results and Discussion

3.1. Uniformly Charged Walls. We first examine the size dependence of certain observables in systems with uniformly charged walls. The counterion charge density profiles and mean total energy of a counterion as a function of the z coordinate are shown in Figure 2, where we observe that the results are effectively independent of system size. As the energy contribution from an infinite, uniformly charged plane and a point charge is linear in distance (eq 4), and the counterions are contained between symmetrical surfaces with identical surface charge density, this charge distribution is effectively a result of the counterion correlations as the energy due to the walls is constant for all counterion positions. Any value of σ will produce the same counterion charge distributions given the same systems size and counterion concentration, although such systems are unphysical where a net charge exists.

The total osmotic pressure, however, shows a strong size dependence for small L as can be seen in Figure 2. The value of P is very much lower for $L = 1$ compared to $L = 2$ ($\approx 460 \text{ mM}$ vs $\approx 560 \text{ mM}$), for example. As L increases, we observe a fast convergence onto a stable value of $P \approx 575 \text{ mM}$. This demonstrates the robust nature of such a representation, even at relatively small systems sizes.

3.2. Discrete Wall Charges on an Equispaced Grid.

Although uniformly charged surfaces are both convenient for analytical theory and computationally efficient in models of charged interfaces, at the microscopic level such an interface features discrete packets of charge from embedded ions or partial charges in the wall material. We therefore examine the effects of replacing the uniformly charged surfaces with regular grids of discrete charges (see system 1, Figure 1). All the parameters are otherwise identical to those considered previously.

Figure 3 shows the counterion charge distribution for a series of systems with discretely charged walls of $1 \leq L \leq 9$. Somewhat surprisingly the results are very comparable, even for small L , with slight differences becoming apparent only in the regions very close to the charged walls. The results are similar to those of the uniformly charged walls (see Figure 2) albeit with slightly higher concentrations immediately adjacent to the wall which reflects the stronger attraction of the counterions to the discrete wall charges at close ranges. This difference is rather subtle, in agreement with the findings of Fleck and Netz^{13,14} who noted that, where wall charge exists in regularly spaced packets, the effects on the counterion distribution - specifically the tendency for counterions to aggregate closer to the charged

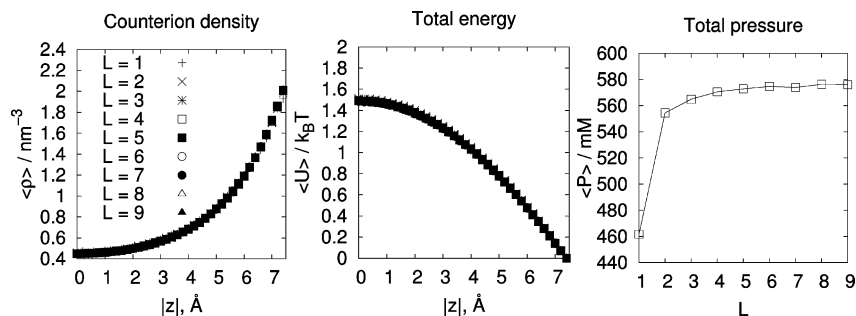


Figure 2. Counterion distributions perpendicular to the interfaces, total energy of a counterion as a function of z position, and total pressure for the systems with uniformly charged walls (system 0, see Figure 1). Energy given as excess over that at the plane of closest approach to the wall, $|z| = 7.5 \text{ Å}$. Histogram resolution is 5 bins per Å, with the data point plotted at the center of each histogram bin. Error bars are too small to be shown clearly.

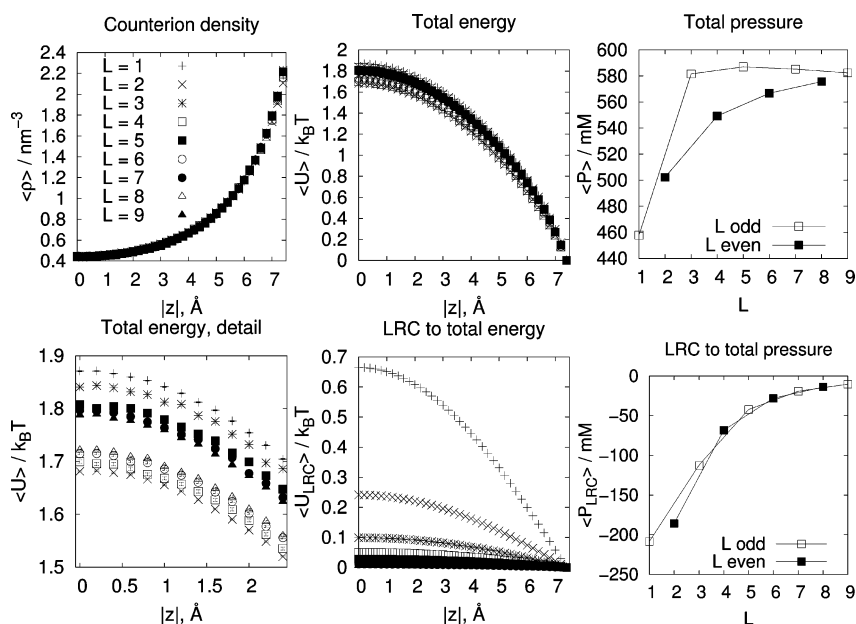


Figure 3. Simulation results for systems with equidistantly spaced discrete wall charges (system 1, see Figure 1). Data are as in Figure 2, with the addition of a detailed plot of the counterion energy in the midplane region and the long ranged contributions to the total energy and pressure. Energies are given as excess over that at the plane of closest approach to the wall, $|z| = 7.5 \text{ Å}$. Error bars are too small to be shown clearly.

wall - is much reduced compared to systems with disordered arrangements of wall charge.

As the interactions between the counterions and the wall charges are modeled using a combination of a short ranged Coulomb potential (in the minimum image convention) and a long ranged correction to the electrostatic energy, care must be taken to avoid asymmetry due to different numbers of discrete wall charges on the x and y plane either side of a counterion. Where L is an even number, we can encounter pronounced asymmetry in the number of wall charges which are explicitly considered on either side of the counterion on the x and y axes. In the limit of large L this effect will vanish, but for relatively small systems the effects can be significant. Where L is an odd number, however, we expect this asymmetry to be reduced as the counterion can preferentially locate next to any of the discrete wall charges, and the odd L allows the minimum image convention to provide an equal number of wall charges in both directions on the x and y axes.

This effect can be seen in the total energy of a counterion as a function of its z position (Figure 3), where there are two main groups of results; systems where L is odd, and systems where L is even. Wall charge arrangements with odd L appear to slightly overestimate the mean energy of a counterion as a function of z coordinate, whereas the even L systems appear to slightly underestimate the mean energy. As L increases, the mean energy for both odd and even L converges onto an intermediate value.

The long ranged correction to the total energy of a counterion between discretely charged walls as a function of z position is also shown in Figure 3. As expected, smaller systems require a larger LRC - but the magnitude of this correction decays rapidly as L increases. For $L \geq 4$, the long ranged correction contributes, on average, less than 3% of the total energy as a function of counterion z position.

A similar pattern is seen in the osmotic pressure (Figure 3); here P appears to have effectively converged for odd values of $L \geq 3$, whereas convergence for even values of L

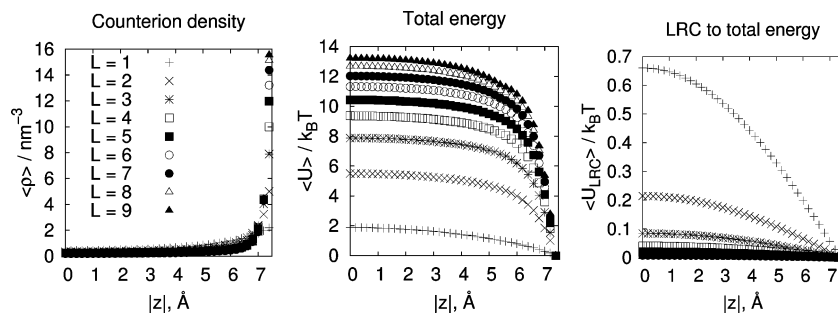


Figure 4. Charge distributions perpendicular to the interfaces, total energy of a counterion as a function of z position, and the contribution of the long ranged correction to the total energy for the systems with clustered discretely charged walls (system 2, see 1). Energy given as excess over that at the plane of closest approach to the wall, $|z| = 7.5 \text{ \AA}$. The simulation results for system 3 are almost indistinguishable from these data. Error bars are too small to be shown clearly.

requires a larger system. The long ranged correction to the pressure decays rapidly in L , as did the long ranged correction to the energy. The total pressure converges onto a value which is very similar to that of the uniformly charged walls (ca. 580 mM at $L = 9$).

3.3. Discrete Wall Charges and the Effects of Clustering. We have so far considered a charged interface as described by both a uniformly charged surface and a regular lattice of discrete charges. These models produce very comparable counterion distributions perpendicular to the interface, and there is rapid convergence onto similar pressure as L increases. However, not all surfaces can be adequately described in these terms; an interface with a particular mean σ may consist of pockets of relatively high and low surface charge density. Interfaces of equal net σ as modeled by uniform charge densities and regularly spaced discrete charges can produce qualitatively and quantitatively different counterion concentration profiles, for example in the case of rodlike counterions,¹⁷ and the variable clustering of discrete surface charges may further affect the energy and pressure in a given system.

A quantitative discussion of the surface charge clustering found in various experimental systems is problematic due to the difficulties of directly measuring these aspects of an interface. Although there is experimental evidence to suggest the presence of charged domain formation in, for example, lipid membrane systems,^{18–21} quantitative analysis of the size of the charged domains formed has proven to be difficult. Even though the existence of such domains is well recognized in vitro, there remains some debate as to their presence for in vivo membrane systems.²¹

We now examine the extreme situation of discrete wall charges condensing into a single cluster with high local surface charge density but the same net σ and system composition as those discussed previously. The two model arrangements of clustered discrete surface charge are depicted in Figure 1 as systems 2 and 3, and here wall charges are staggered on the x and y axes with spacing equal to the impenetrable diameter of the wall charges (2 \AA) to form square arrangements of clustered discrete charge.

For this aspect of the study, further attention is given to the effects of the long ranged electrostatic correction via the charged slabs method. As this correction is calculated from the average charge distribution in the system via the

counterion concentration profiles, any significant changes to this distribution via clustering of the wall charges (with a subsequent shift in average counterion density) has the potential to affect the properties of the system. We therefore simulate each of systems 1, 2, and 3 (see Figure 1) using not only the LRC appropriate to the system itself but also the LRC of the other two discretely charged system types (of equal size and net σ) to examine the significance of the LRC.

We adopt the notation LRC_n to describe the long ranged correction from system n (see Figure 1). There are 9 sets of simulations performed, as each of the three arrangements of discrete wall charges are simulated using the three different long ranged corrections for every value of L .

The simulation results for system types 2 and 3 (see Figure 1) are extremely similar, and hence only the counterion density profiles, energy, and LRC for system 2 are plotted in Figure 4. Both clustered systems display very different properties when compared to the systems with uniformly charged walls (Figure 2) and unclustered discretely charged walls (Figure 3). As L increases, we are no longer simply adding additional repeating sections to the simulation cell as was the case for the uniformly charged and unclustered discretely charged wall systems. Instead, the models represent what are effectively different types of interface, which can be seen in the pronounced changes of the counterion density profiles and energy as L increases. This is readily explained in terms of the rapid increase in the local surface charge density of the clustered region of the walls relative to the expansion of the overall wall dimensions in the simulation cell; this effect is particularly noticeable for large L , where we observe a density of $\rho \geq 10 \text{ e nm}^{-3}$ adjacent to the clustered walls compared to $\rho \approx 2 \text{ e nm}^{-3}$ for the uniformly charged and unclustered discretely charged walls. Such behavior demonstrates the utility of measuring pressures at the midplane of the system, as the large changes in counterion concentration make techniques such as the contact theorem numerically problematic at the plane of closest approach to the charged surfaces.

In all cases, providing the LRC from a different discretely charged system of equal size and net σ does not noticeably alter the counterion charge distributions perpendicular to the charged walls, and nor is there a discernible effect on the energy of a counterion as a function of z (data not shown).

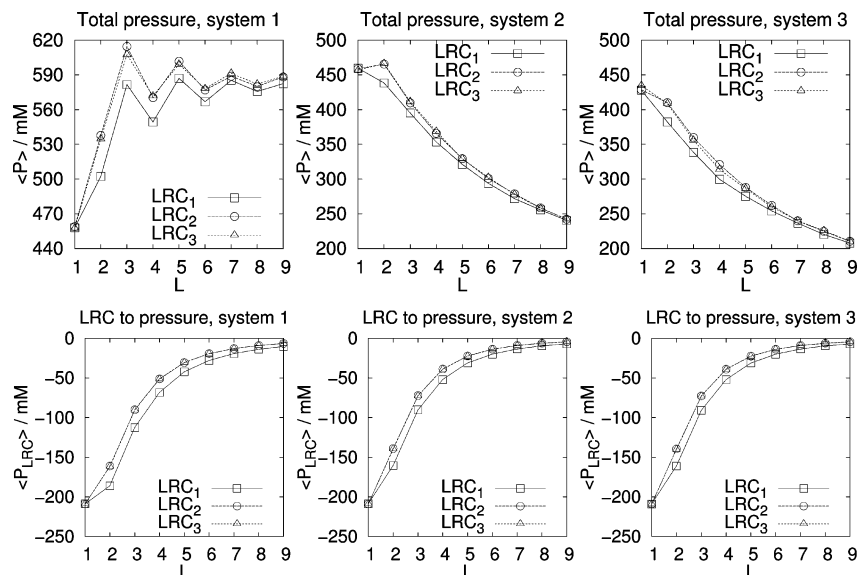


Figure 5. Total and long ranged contribution to P for the three discretely charged wall systems. Compare total pressure data for system 1 to Figure 3, where this pressure is separated into even and odd L . Error bars are too small to be shown clearly.

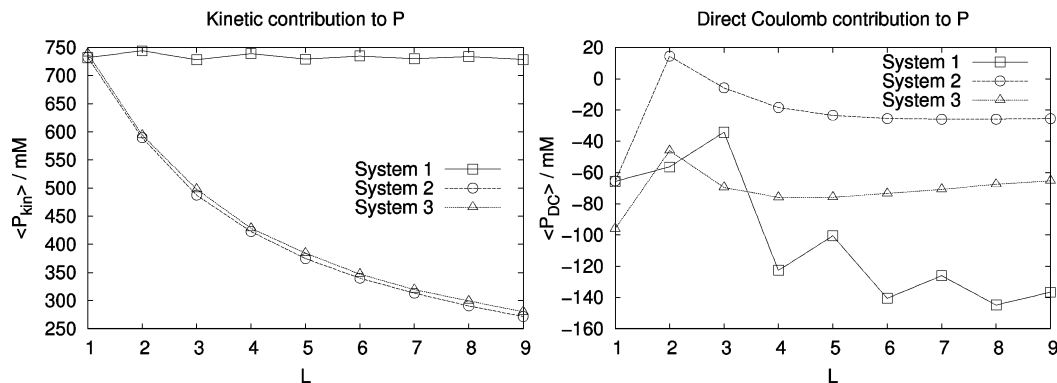


Figure 6. Individual components of the osmotic pressure for systems 1, 2, and 3 (Figure 1). In each case, the “native” long ranged correction appropriate to the system is used. Error bars are too small to be shown clearly.

Figure 5 displays the recorded pressures for systems 1, 2, and 3 for the values of L studied here. For system 1, where the wall charges are arranged on an equispaced grid, P quickly increases in L to values which oscillate around the approximate plateau value of P for the uniformly charged systems (Figure 2), with the magnitude of the oscillations decreasing as L increases. This effectively demonstrates the convergence of P onto a similar value to that of system 0, showing that these two representations are basically equivalent provided the system is sufficiently large (although this is not necessarily valid for high electrostatic coupling¹⁶). In systems 2 and 3, where the wall charges are densely clustered, P behaves in an entirely different manner - essentially *decreasing* as a function of L .

The completely different trend in P as a function of L for the clustered systems cannot be explained by the LRC to the pressure, which is seen to be very similar in each case (Figure 5). Figure 6 shows the kinetic and direct Coulomb contributions to the total pressure, and we observe that the reduction in P is largely a result of the marked decrease in the kinetic contribution to the osmotic pressure as a function of L in systems with clustered wall charges. This is an expected effect of the very large counterion concentrations

in the vicinity of the clustered wall charges, with a conjugate depletion of counterions in the middle of the systems where the kinetic contribution to P is measured (compare, for example, the counterion densities in Figure 3 and Figure 4). Although the equispaced grid arrangement of wall charges in system 1 leads to a smaller direct Coulomb pressure contribution with increasing L , this reduction is small compared to the differences in kinetic pressure we see between system 1 and systems 2 and 3. For both the kinetic and direct Coulomb contributions to the osmotic pressure for system 1, we again see the characteristic oscillations, which decrease in magnitude with increasing L . The measured hard sphere contribution to the pressure is orders of magnitude smaller than the other contributions and hence is ignored in this discussion.

The total pressures measured for system 3 are consistently somewhat lower than those of system 2; again, the LRC contributions to the pressure are almost identical, and the answer may be found in Figure 6 where we see that although the kinetic contribution to P for system 3 is consistently slightly higher than that of system 2, the direct Coulomb pressure found in system 3 is markedly lower as the repulsion between like-charged wall ions on the opposing surfaces is

decreased as the distance between the clusters on the x , y plane increases. This produces a net P which is marginally lower for the systems with misaligned wall clusters compared to clusters which are positioned directly opposite one another.

4. Conclusion

Charged interfaces of a particular net surface charge density σ may be represented in models as uniformly charged planes, discretely charged walls with equidistant charge spacing, or walls with locally clustered discrete wall charges. Although the behaviors of the former two systems are known to differ under certain circumstances, here we show that the addition of local clustering of the discrete wall charges can itself change the nature of the system entirely via large differences in the counterion distributions and an enhanced sensitivity of the energy and osmotic pressure to the size of the simulation cell, even in the presence of long ranged corrections to the electrostatic energy.

The long ranged correction to the electrostatic energy and pressure become rapidly less significant as a fraction of the total energy and pressure as the system increases in size. We show that despite the sensitivity of the long ranged corrections used here to the average charge profile of the system, in many cases a precomputed charge distribution will suffice to reproduce the essential behaviors regardless of the level of wall charge clustering in the system used to precompute the long ranged correction and the actual system of interest.

Notably, we observe a trend of rapidly decreasing osmotic pressure as a function of system size for clustered arrangements of wall charges. The osmotic pressure for such clustered arrangements can be $\approx 1/3$ of that measured from equidistantly spaced discrete wall charges or uniformly charged surfaces for the range of systems studied. We observe that these changes are largely the result of the electrostatic interaction within the simulation cell and that the influence of the long ranged corrections to the energy and pressure decay rapidly as a function of system size for the simulation methods used.

Acknowledgment. We acknowledge the high performance computational (and storage) capacity allocated through the Swedish National Infrastructure for Computing (SNIC) on resources at the National Supercomputer Centre (NSC), alongside the UPPMAX high performance computational resources provided by Uppsala University under Project SNIC s00109-20. M.O.K. also acknowledges an Ingvar Carlsson grant from The Swedish Foundation for Strategic Research.

References

- (1) Piedade, J. A. P.; Mano, M.; de Lima, M. C. P.; Oretskaya, T. S.; Oliveira-Brett, A. M. *Biosens. Bioelectron.* **2004**, *20*, 975–984.
- (2) Carla, M.; Cuomo, M.; Arcangeli, A.; Olivotto, M. *Biophys. J.* **1995**, *68*, 2615–2621.
- (3) Düzgünes, N.; Nir, S.; Wilschut, J.; Bentz, J.; Newton, C.; Portis, A.; Papahadjopoulos, D. *J. Membr. Biol.* **1981**, *59*, 115–125.
- (4) Ekerdt, R.; Papahadjopoulos, D. *Proc. Nat. Acad. Sci. U.S.A.* **1982**, *79*, 2273–2277.
- (5) Evans, E.; Kukan, B. *Biophys. J.* **1983**, *44*, 255–260.
- (6) Guldbbrand, L.; Jönsson, B.; Wennerström, H.; Linse, P. *J. Chem. Phys.* **1984**, *80*, 2221–2228.
- (7) Valleau, J. P.; Ivkov, R.; Torrie, G. M. *J. Chem. Phys.* **1991**, *95*, 2221–2228.
- (8) Jönsson, B.; Wennerström, H.; Halle, B. *J. Phys. Chem.* **1980**, *84*, 2179–2185.
- (9) Urbanija, J.; Bohinc, K.; Bellen, A.; Maset, S.; Igljč, A.; Kralj-Igljč, V.; Kumar, P. B. S. *J. Chem. Phys.* **2008**, *129*, 1051015–1011055.
- (10) May, S.; Igljč, A.; Reščič, J.; Maset, S.; Bohinc, K. *J. Phys. Chem. B* **2008**, *112*, 1685–1692.
- (11) Naji, A.; Podgornik, R. *Phys. Rev. E* **2010**, *72*, 041402–1–041402–11.
- (12) Mamasakhlisov, Y. S.; Naji, A.; Podgornik, R. *J. Stat. Phys.* **2008**, *133*, 659–681.
- (13) Fleck, C. C.; Netz, R. R. *Europhys. Lett.* **2005**, *70*, 341–347.
- (14) Fleck, C. C.; Netz, R. R. *Eur. Phys. J. E* **2007**, *22*, 261–273.
- (15) Naydenov, A.; Pincus, P. A.; Safran, S. A. *Langmuir* **2007**, *23*, 12016–12023.
- (16) Khan, M. O.; Petris, S.; Chan, D. Y. C. *J. Chem. Phys.* **2005**, *122*, 1047051–1047057.
- (17) Grime, J. M. A.; Khan, M. O.; Bohinc, K. *Langmuir* 2010In press.
- (18) Huang, J.; Swanson, J. E.; Dibble, A. R. G.; Hinderliter, A. K.; Feingenson, G. W. *Biophys. J.* **1993**, *64*, 413–425.
- (19) Hinderliter, A. K.; Huang, J.; Feingenson, G. W. *Biophys. J.* **1994**, *67*, 1906–1911.
- (20) Ahn, T.; Yun, C.-H. *J. Biochem.* **1998**, *124*, 622–627.
- (21) Allender, D. W.; Schick, M. *Biophys. J.* **2006**, *91*, 2928–2935.

CT100009M

Toward a Coarse Graining/All Atoms Force Field (CG/AA) from a Multiscale Optimization Method: An Application to the MCM-41 Mesoporous Silicates

A. Ghoufi,^{*,†} D. Morineau,[†] R. Lefort,[†] and P. Malfreyt[‡]

Institut de Physique de Rennes, UMR 6251 CNRS, Université de Rennes 1, France

Thermodynamique et Interactions Moléculaires, UMR CNRS 6272, Université Blaise Pascal, France

Received March 29, 2010

Abstract: Many interesting physical phenomena occur on length and time scales that are not accessible by atomistic molecular simulations. By introducing a coarse graining of the degrees of freedom, coarse-grained (CG) models allow the study of larger scale systems for longer times. Coarse-grained force fields have been mostly derived for large molecules, including polymeric materials and proteins. By contrast, there exist no satisfactory CG potentials for mesostructured porous solid materials in the literature. This issue has become critical among a growing number of studies on confinement effects on fluid properties, which require both long time and large scale simulations and the conservation of a sufficient level of atomistic description to account for interfacial phenomena. In this paper, we present a general multiscale procedure to derive a hybrid coarse grained/all atoms force field CG/AA model for mesoporous systems. The method is applied to mesostructured MCM-41 molecular sieves, while the parameters of the mesoscopic interaction potentials are obtained and validated from the computation of the adsorption isotherm of methanol by grand canonical molecular dynamic simulation.

1. Introduction

Considerable effort has been devoted to the study of molecular fluids confined in mesoporous/nanoporous solids.^{1–4} Confined systems exhibit very original features, in terms of structural arrangements, molecular dynamics, and phase transitions (freezing, melting, capillary condensation, and mesomorphic transitions), which cannot be simply understood from the properties of the bulk system.^{5–10} According to the abundant literature on the matter, it has been recognized that the properties of confined systems are not simply related to the typical size of the confining medium but are intimately related to the details of the porous morphology and the structure and chemistry of the pore surface.^{11–15} This significantly complicates the interpretation of experimental results and their comparison with theoretical

predictions based on simple pore models (structureless pore, simple geometry). Molecular simulations of fluids adsorbed in realistic pores offer unique possibilities to connect some macroscopic properties to a microscopic description of the physical phenomena at play in nanoconfined phases. As a result, molecular simulations have become widespread in the literature and have become a powerful method to investigate mesoporous confinement effects.^{1,4} Fully atomistic simulations, which allow a realistic account of the details of the confining medium and the surface interaction, are also very time-consuming. This usually becomes a serious drawback to investigating long-time adsorption and relaxation processes, which are usual in confined systems as well as those with large pore sizes (e.g., 10 nm and more). To overcome these difficulties, coarse-grained force fields CG¹⁶ and mesoscopic methods—such as dissipative particles dynamics^{17,18}—have been designed for biological¹⁹ and polymeric systems.²⁰ Much benefit could be gained from the development of mesoscopic methods for porous solid materials,

* Corresponding author e-mail: aziz.ghoufi@univ-rennes1.fr.

[†] Université de Rennes 1.

[‡] Université Blaise Pascal.

which surprisingly have not been given full attention.²¹ A realistic CG potential for mesoporous solids should necessarily incorporate a part of modeling at the atomistic level in order to describe the specific interfacial interactions, which play a fundamental role in confinement phenomena. The situation shares some similarities with the question of folding proteins, which has been treated with hybrid potential models.^{16,22} However, the approaches derived in the latter case cannot be simply transferred to the case of solid systems. Another approach proposed by Dupuis et al.²¹ extends the quasi-continuum method to treat the dynamics of crystalline solids at a constant temperature. This method is based upon the calculation of the potential of mean force. Its accuracy is determined by the sampling of the solid atoms' positions in the molecular dynamics (MD) simulations. This method is not designed for the determination of the CG potential for porous solids, which are generally approximated as rigid systems in numerical studies of adsorption. Alternatively, Dubbeldam et al.²³ used an iterative search for Lennard-Jones parameters to reproduce the inflection points in isotherms of adsorption. This full microscopic method does not allow one to obtain the intrinsic parameters of porous solids because we only obtain the crossed (framework/adsorbate) interactions parameters. Then, the obtained parameters are not transferable, and the method must be applied for each adsorbate.

In the present paper, we introduce a hybrid coarse grained/all atoms force field CG/AA for porous materials. We detail a general procedure to determine the mesoscopic potential parameters using a multiscale method based on the computation of the isotherm of adsorption by molecular simulations. The method proposed here is developed in close connection with atomistic models, but it is based upon a very different approach from iterative Boltzmann inversion,²⁴ inverted Monte Carlo schemes,^{25,26} or force matching methods.²⁷ Recently, Das and Andersen²⁸ used a force matching method based on a multiscale approach and the computation of the potential of mean force for solutes. They propose new basis functions for the variational calculation. This method is quite accurate, provided that the phase space is correctly sampled, but the grid potential does not provide a general force field. Here, we present a coarse grained force field resulting from an optimization procedure of energetic parameters to model the porous solid. This method is based on the calculation and optimization of the macroscopic properties, allowing one to provide the intrinsic parameters of porous material. Our CG/AA force field can be easily combined with the CG or AA solvent/adsorbate. Besides, we do not need to compute the grid of potential as it is done in the force matching approach.²⁸ The goal of our study is to provide a simple method that can be used to derive computationally fast and practical models of porous solids. Thereby, we report a CG/AA force field for MCM-41 types of mesoporous silicates modeled by SiO_4 and SiO_3OH units, the interaction parameters of which are obtained from the isotherm of adsorption of methanol. Already, there exist some simplifications to decrease the computational time as the rigid consideration. However, at variance with the usual AA rigid framework model (decreasing of the intramolecular of degree of

freedom), our CG/AA rigid description allows one to decrease the number of centers of force.

We opt for a highly hydrated MCM-41 because it is one of the most studied forms of silicate used in molecular simulations, and thus it can be considered as a reference system. This system is especially suited for a validation of our method by comparing some adsorption and dynamic properties between the AA and the CG/AA models. In addition, we apply our method for another MCM-41 with a lower hydration level, which provides a more realistic description of some experimental porous materials. The improvement provided by this approach allows the simulation of large scale mesoporous systems while keeping the possibility of later tuning the nature of the surface interaction, which is a currently debated issue in a number of experimental studies. The properties of confined methanol have already been addressed in numerical and experimental studies.^{29–33} They have provided us with useful groundwork for the development and the validation of the CG/AA force field. The first milestone of this study stands on the methodological side and consists of the validation of the coarse-graining approach, by comparing adsorption and dynamics results between the AA and the CG/AA models.

2. The Simulation Models

2.1. The Atomistic Model. *Highly Hydrated (HH) MCM-41:* $\rho_{\text{OH}} = 7.5 \text{ OH/nm}^2$. Simulations of confined methanol have been performed by building a realistic model of the porous silicate used experimentally. This is required to account for the complexity of the liquid–substrate interactions and confinement effects. Since the geometry of the porous MCM-41 is properly characterized in terms of channels of a definite section, it has been possible to produce comparable conditions of confinement. We derived an atomic description of the silicate starting from an equilibrium structure of amorphous silica within a cubic cell of 36 \AA on a side provided by Vink and Barkema.³⁴ Then, we applied a procedure proposed by Brodka and Zerda³⁵ to consider a realistic porosity within the amorphous silica. We first generate a cavity along the z axis of the silica cell by removing the atoms within a cylinder of diameter (D) 24 \AA . From their coordination numbers, we distinguished bridging oxygens (O_b) bonded to two silicon atoms from nonbridging oxygens (O_{nb}) bonded to only one silicon and bonded to one hydrogen atom (H_{nb}). An iterative procedure of atom (O and Si) removal was applied until only tetra-coordinated silicon atoms, bonded to a maximum of two O_{nb} 's, were present in the structure. Finally, nonbridging oxygens were saturated with hydrogen atoms to form surface hydroxyl groups. Although the silica matrix was subsequently kept rigid, rotation around the Si–O bond of the hydroxyl groups was allowed. In the MC procedure, we used a trial rotation move of the H atom around the Si–O bond. The parameters of the bending potential ($U_\theta = k_\theta(\theta - \theta_0)^2$) are $k_\theta = 284.37 \text{ kJ mol}^{-1} \text{ rad}^{-2}$ and $\theta_0 = 118^\circ$ where k_θ is the constant of force and θ_0 the equilibrium angle. This procedure leads to a realistic description of the irregular inner surface of the porous silicate and of the interfacial interactions between

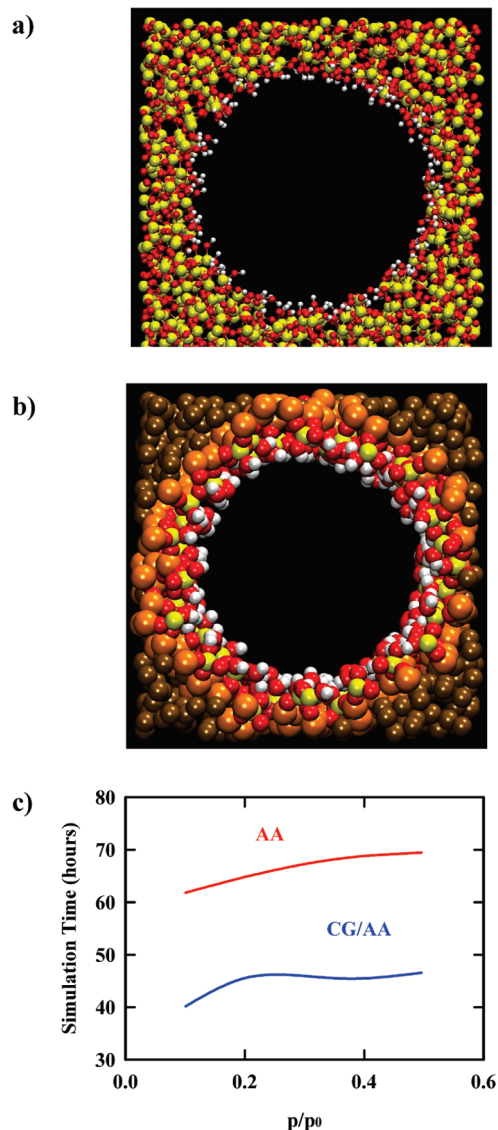


Figure 1. AA (a) and CG/AA (b) description of MCM-41. Red indicates the oxygen atoms. The hydrogen positions are in white. Yellow are the silicon atoms, orange the SiO_{4-x} CG beads, and brown the SiO_4 CG beads. (c) Times of simulation for both models as a function of the reduced pressure. The simulations were performed using a time step of 0.002 ps to sample 2 ns, e.g., 10^6 MD steps.

the fluid and the matrix. The inner surface coverage of silanol groups was about 7.5 per square-nanometer, which is comparable to previous models of surface silica and correspond to highly hydrated actual silica (HH MCM-41).^{12,35} The resulting pore morphology is shown in the snapshots in Figure 1a.

For AA and CG/AA models, the total intermolecular potential is a sum of the electrostatic and Lennard-Jones interactions (eq 1).

$$U = \sum_i^{N-1} \sum_a^{n_i} \sum_{j=i+1}^N \sum_b^{n_j} \left[\frac{q_a q_b}{4\pi\epsilon_0 r_{iajb}} + 4\epsilon_{ab} \left(\left(\frac{\sigma_{ab}}{r_{iajb}} \right)^{12} - \left(\frac{\sigma_{ab}}{r_{iajb}} \right)^6 \right) \right] \quad (1)$$

In eq 1, q_i is the charge of the i atom, ϵ_{ab} and σ_{ab} are the different Lennard-Jones parameters calculated from Lorentz–

Table 1. Force Field Parameters of MCM-41 Resulting from the Optimization Procedure^a

	σ (Å)	ϵ (kJ mol ⁻¹)	q (u.e)
MCM-41			
AA Force Field			
H _{nb}	0.000	0.000	0.206
O _b	2.700	1.912	-0.6349
O _{nb}	3.000	1.912	-0.5399
Si	0.000	0.000	1.2739
CG/AA Force Field			
H _{nb}	0.000	0.000	0.206
O _b	2.700	1.622	-0.6349
O _{nb}	2.700	1.622	-0.5399
Si	0.000	0.000	1.2739
SiO ₃	4.500	0.832	0.950
SiO ₂	4.500	0.832	0.638
SiO	4.500	0.832	0.320
SiO ₄	5.500	2.411	0.0054
CH ₃ OH			
CH ₃	3.750	0.815	0.265
O _H	3.020	0.773	-0.700
H _o	0.000	0.000	0.435

^a For SiO_{4-x} , the charge depends on the number x of oxygen atoms of the silicon's first coordination shell that are modelled by the (AA) description. The geometric characteristics of methanol are given in Ref. 36.

Berthelot mixing rules ($\sigma_{ab} = (\sigma_{aa} + \sigma_{bb})/2$, $\epsilon_{ab} = \sqrt{(\epsilon_{aa}\epsilon_{bb})}$). Note that Lorentz–Berthelot mixing rules have been also applied for the CG model. r_{iajb} is the distance between atom a of molecule i and b of j . n_i is the number of particles in the i group. We consider the adsorption of CH_3OH in the gas phase for its simplicity. To model CH_3OH , we used the TRAPPE force field developed by Siepmann et al.³⁶ (Table 1), which was validated on the liquid–vapor diagram of phase.

Weakly Hydrated (WH) MCM-41: $\rho_{\text{OH}} = 3.4 \text{ OH/nm}^2$. A porous framework with a lower density of silanol groups ($\rho_{\text{OH}} = 3.4/\text{nm}^2$), lately denoted as the weakly hydrated (WH) MCM-41, was built using the HH MCM-41 as a starting structure. Dehydration of the pore surface was obtained by removing the OH group and the oxygen atom engaged together in a hydrogen bond between randomly selected couples of adjacent silanol groups. An explicit chemical bond was then created so that the remaining oxygen bridges the two surface silicon atoms considered. We repeated this iterative procedure until the targeted surface density of the silanols groups was achieved. The obtained structure was allowed to relax during MD simulations of duration $t = 5$ ns, with a harmonic description of the Si–O bond ($U = k_0(r - r_0)^2$) with an equilibrium distance $r_0 = 1.58 \text{ Å}$ and a force constant $k_0 = 2000 \text{ kJ mol}^{-1}$. The total residual charge introduced by this dehydration process (+1.1 u.e.) has been homogeneously redistributed over the interaction sites of the system. It corresponds to a tiny increase by $\delta q \sim 0.0001$ u.e. of the charge assigned to each electrostatic site.

The resulting WH porous structure presents a less hydrophilic character with respect to the HH MCM-41. The radial density profiles of the two types of empty MCM-41 (cf. Figure 2a) agree with a comparable average value of the pore radius of about 12 Å. Direct insight into the inner pore

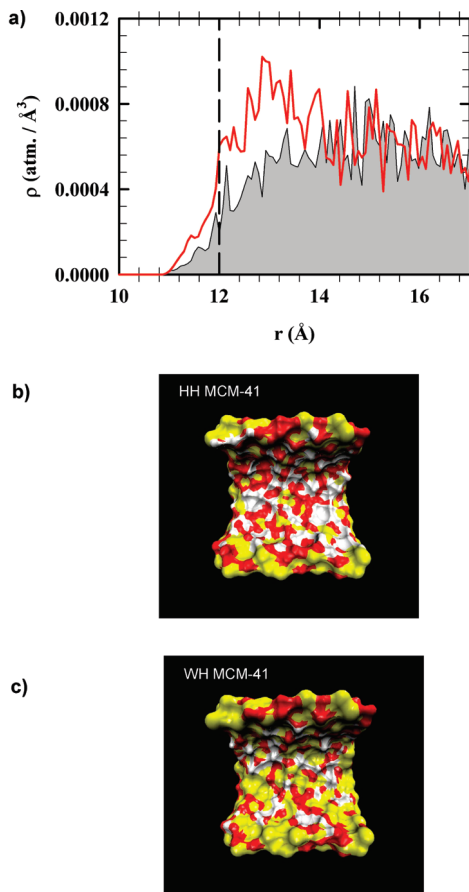


Figure 2. (a) Radial density profile of MCM-41 for HH (red) and WH (gray zone) types of MCM-41. Connolly's surface of HH (b) and WH (c) MCM-41. Red, white, and yellow colors indicate the oxygen, hydrogen, and silicon atoms, respectively.

surface coverage of the two types of materials is provided from the surface Connolly diagrams displayed in Figure 2b and c.

2.2. Derivation of the Coarse-Grained Force Field.

SiO_3OH silanol groups located on the internal pore surface confer a hydrophilic character to the hydrated porous silicates and can promote the formation of hydrogen bonds (Hb) with associating fluids, which strongly affect the structural organization and the dynamics of the adsorbates. Interfacial hydrogen bonds and hydrophilic interactions are some fundamental aspects of the physics of confined phases. Hence, silanol groups must be considered as a foremost ingredient of any reliable force field which aims at modeling porous silicates. The multiscale coarse-graining strategy that we have adopted provides a hybrid force field where the surface atoms and the hydrogen bonds are described at the atomistic level, whereas a coarse-grained description is used for the groups of atoms far away from the internal surface. To define the CG beads, we decompose the framework in three parts according to a layered description depicted in Figure 1b. The CG appellation is preferred to the one of united atoms (UA) because this force field does not include explicit representation of nonpolar hydrogen atoms and only polar hydrogens are included in the force field definition. Additionally, our description is in line with the CG statistical physics. The first layer (N_1) corresponds to the silanol

SiO_3OH groups treated with an atomistic description. In this atomistic description, distinction is made between bridging oxygens (O_b), which are bonded to two silicon atoms and nonbridging oxygens (O_{nb}) bonded to only one silicon and one silanol hydrogen atom. The second layer (N_2) corresponds to the nonsilanol groups bonded to $x = 1, 2,$ or 3 oxygen atoms of N_1 and which were treated as CG beads, denoted SiO_{4-x} . In this case, one bead stands for a coarse-grained description of the force field, which corresponds to one silicon atom and half the interaction of the $4 - x$ bridging oxygens in the atomistic model. Different values of the force field parameters are obtained depending on the coordination number x , as explained later. The second type of CG bead, denoted N_3 , is introduced for the remaining SiO_4 groups, which are connected to N_2 and/or N_3 beads. In the latter case, the bead stands for the atomistic interaction arising from one silicon atom and half the interaction arising from the oxygens within the 4-fold coordination shell. Figure 1b shows a picture of this arrangement. According to this description, the surface is covered by $-\text{SiOH}$ and $\text{Si}-\text{O}-\text{Si}$ to get a physical picture of the atomistic description. As the electrostatic field is independent of the description level, the mesoscopic charge of each bead is the sum of the AA charges of atoms participating in one CG particle. Nevertheless, one should pay great attention to the definition of the CG charges of SiO_{4-x} beads, which make the link between the coarse-grained and atomistic parts of the sample. According to the definition, the silicon atom of a SiO_{4-x} bead is connected to $x = 1, 2,$ or 3 oxygens, which is/are explicitly treated by an atomistic AA force field. Therefore, the net mesoscopic charge reflecting the silicon and half the interaction from the oxygens not treated by the AA model is $q_{\text{Si}}^{\text{AA}} + (4 - x)/(2)q_{\text{O}}^{\text{AA}}$. The calculation of q_{SiO_4} is deduced by dividing the residual charge by the total number of silicons in the layer (N_3) so as to satisfy electric charge neutrality. The charges of the different sites of the coarse-grained model are summarized in Table 1. We opt for the simplest degree of coarse-grained (1 CG unit = SiO_x ($x = 1, 4$)). However, it is possible to undertake a study using an higher degree of CG, but our main point was to test and validate our multiscale optimization.

Once the CG beads are defined, we opt as a center of force the center of mass of each CG unit. The Lennard-Jones parameters (σ^{CG} and ϵ^{CG}) have been obtained and refined from the simulation of the isotherm of adsorption of methanol at $T = 300$ K, the results of which are detailed in the next section. Adsorption quantities have been calculated from the simulations of the CG and AA models and have been used as inputs in the merit function (F ; eq 2). The parameters of the CG model have been optimized by minimization of the value of F according to a procedure discussed in ref 37.

$$F = \frac{1}{n} \sum_{i=1}^n \frac{[f_i^{\text{CG/AA}} - f_i^{\text{AA}}]}{s_i^2} \quad (2)$$

In eq 2, s_i is the estimated statistical uncertainty and $s_i^2 = (s_i^{\text{CG/AA}})^2 + (s_i^{\text{AA}})^2$, f_i^{AA} , and $f_i^{\text{CG/AA}}$ are the values of the i th physical property on n , which were calculated by simulation of the AA and CG/AA models, respectively. In the present

case, $n = 2$, and the two properties considered were the enthalpy of adsorption $\Delta_r H_{\text{ads}}$ and the adsorbed amount n_{ads} . The Lennard-Jones parameters have been optimized for SiO_{4-x} , SiO_4 , O_b , O_{nb} , and H_{nb} , whereas they are identical to the all atoms force field for the silicon atoms. The choice to keep the AA parameters for silicon atoms in the coarse-grained description of silica has been justified by the fact that they do not vary significantly if they are considered as free parameters in the refinement procedure (about 0.2% change). This is most probably a consequence of the distant location of Si from the inner surface with regard to other species (H_b , O_b , and O_{nb}), which reduce their influence on the diffusion and adsorption processes. The minimum condition of F is that every partial derivative must be zero, which means solving eq 3. Equation 3 is a function of $(\partial f_i^{\text{CG/AA}})/(\partial y_k)$ where y_k is one of the p different optimized potential parameters.

$$\sum_{i=1}^n \left[\frac{f_i^{\text{CG/AA}}(y_j^o) - f_i^{\text{AA}} + \sum_{k=1}^p \frac{\partial f_i^{\text{CG/AA}}}{\partial y_k} \Delta y_k}{s_i^2} \right] \frac{\partial f_i^{\text{CG/AA}}}{\partial y_j} = 0, \quad j = 1, \dots, p \quad (3)$$

$\Delta y_k = y_k - y_k^o$ is the difference between the refined value y_k and the initial value y_k^o of the k th potential parameter. In our study, $n = 2$ and $p = 10$. The partial derivatives $(\partial f_i^{\text{CG/AA}})/(\partial y_k)$ are calculated using the fluctuation relation³⁸ as indicated in relation 4.

$$\frac{\partial \langle X \rangle}{\partial y_k} = \left\langle \frac{\partial X}{\partial y_k} \right\rangle - \beta \left(\left\langle X \frac{\partial U}{\partial y_k} \right\rangle - \langle X \rangle \left\langle \frac{\partial U}{\partial y_k} \right\rangle \right) \quad (4)$$

$\langle \dots \rangle$ indicates the ensemble average in the canonical statistical ensemble. We give the final expression of $(\partial \langle \Delta_r H_{\text{ads}} \rangle)/(\partial y_k)$ and $(\partial \langle n_{\text{ads}} \rangle)/(\partial y_k)$ in the appendix. The result of our optimization is given in Table 1. For Lennard-Jones parameters, we provide the intrinsic terms (aa) while the crossed terms (ab) are calculated using the Lorentz–Berthelot combining rules ($\sigma_{ab}^{\text{CG}} = (\sigma_{aa}^{\text{CG}} + \sigma_{bb}^{\text{CG}})/2$; $\epsilon_{ab}^{\text{CG}} = \sqrt{(\epsilon_{aa}^{\text{CG}} \epsilon_{bb}^{\text{CG}})}$). Smaller values of σ (i.e., more attractive) were obtained for the O_{nb} and O_b parameters with respect to the AA description. This can be attributed to some changes in the local environment interaction, which effectively counterbalance the highly repulsive contribution of the mesoscopic beads. For the Lennard-Jones (LJ) parameters, different initial values were used to account for the microscopic difference between each unit. The optimized final values are very close to each other, indicating a weak dispersive/repulsive interaction between the adsorbant and the framework, which can be safely averaged over the atoms constituting each unit.

First, we applied the same procedure to derive the CG/AA model corresponding to the WH type of MCM-41 from the WH AA framework. Moreover, a CG/AA model of the WH type of MCM-41 was obtained after dehydration and thermal relaxation of the HH CG/AA matrix, following the procedure described in section 2.1. These two procedures provide very similar structures and a residual charge, which does not differ by more than 0.2%.

3. Simulation Methods

The isotherms of adsorption were computed using grand canonical molecular dynamic (GCMD) simulations combined with an explicit reservoir of gas.³⁹ We used the full insertion/deletion⁴⁰ trial move to model an open system. Additionally, by comparison with the fractional⁴¹ particle insertion/deletion, we obtain the right frequency of insertion. Indeed, the fractional method allows for getting stable dynamics, whereas in the full approach, if the frequency of insertion is higher or the MD move is too low, an alteration of the dynamics is found. In contrast to the GCMC simulations, the kinetic energy is included in the partition function (eq 5). Hence, the GCMD method will allow us to model a dynamical description of the confinement and adsorption process.

$$Q_{\mu VT} = \frac{1}{h^{3N} N!} \int \text{d}\mathbf{r}^N \text{d}\mathbf{p}^N \exp(-(\beta[U(\mathbf{r}^N) + K(\mathbf{p}^N) - \mu N]) \quad (5)$$

In eq 5, U and K are the potential and kinetic energies respectively, μ is the chemical potential, \mathbf{p}^N the momentum vector, \mathbf{r}^N the positions, β is the reverse temperature ($1/(k_B T)$), where k_B is the constant of Boltzman, N the total number of molecules, and h is Planck's constant. The initial kinetic energy of the inserted molecules was calculated from the Maxwell–Boltzmann distribution. The expression of the probability of acceptance of the deletion/insertion trial move is given in ref 41. We used a modified DL_POLY package⁴² to compute the isotherms of adsorption using the grand canonical molecular dynamic (GCMD) simulations. We used the Ewald summation for the computation of electrostatic interactions, and the long-range corrections were applied from $r = 12$ Å. For GCMD, the full insertion/deletion trial move was attempted every 100 configurations to reach the mechanical and thermal equilibrium at $T = 300$ K. The simulations were performed using a time step of 0.002 ps over an acquisition phase of 2 ns. The equilibration time was fixed to 5 ns to allow stabilization of the amount adsorbed, energy, temperature, and pressure. Here, we study a larger size of pore rather than the long time properties, given that the size effect in the confinement is problematic. A reduction of the number of interactions to be calculated is one straightforward strategy to access longer time scale than those usual atomistic models. We believe that this second strategy can equally illustrate the capability of the approach. It has been preferred because it allows one to investigate another pore diameter which was not reachable to an atomistic description in a reasonable time. Changing the pore size from micro to mesoporous diameters is indeed a crucial issue in exploring the physics of confinement. Size effects have been widely investigated experimentally for a variety of pore sizes, including diameters as large as some tens of nanometers (with SBA-15 silicates for instance) where departure from the bulk is initiated. This is a range of pore sizes where improvements in molecular simulation are needed. The AA model of MCM-41 for $D = 25$ Å and 50 Å implies 2012 and 11384 atoms, respectively. The same porous systems described by the CG model require 1322 and 7480 sites, respectively. One consequence of the reduction

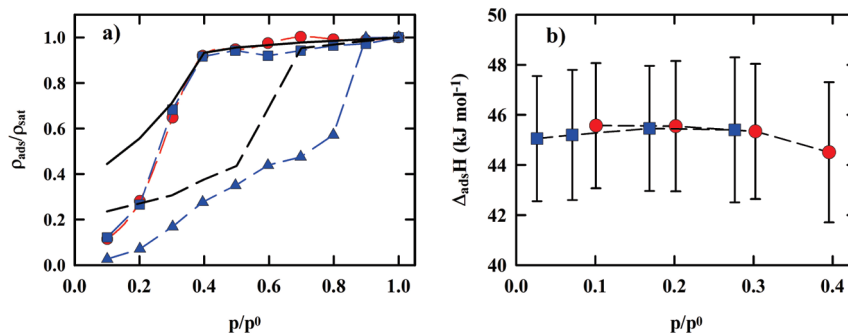


Figure 3. Isotherm (a) and enthalpy (b) of adsorption of methanol in MCM-41 for a pore diameter (D) of 24 Å for AA (●) and CG/AA force fields (■). In figure a, black dashed and solid lines are the experimental data for $D = 24 \text{ \AA}$ and $D = 50 \text{ \AA}$, respectively. ▲ corresponds to the computed isotherm for $D = 50 \text{ \AA}$ using the CG/AA model. ρ is the number of molecules per unit of volume.

in the number of interaction sites induced by coarse-graining is an increase in computation time (by a factor of about 1.5 in the present cases). Figure 1c shows obviously the increase in time obtained with the CG/AA model in relation to the AA description. Molecular dynamics were run on an Intel Core 2 quad core 2.66 GHz CPU gigascale workstation. From parallelized molecular dynamics simulations, the time gained is very impressive. However, from Monte Carlo simulations, the computational time gain is less convincing, and then the CG description and the rigid considerations can decrease it. At saturation, the CPU times of sequential Monte Carlo simulations for CG/AA and AA model descriptions are 61.1 and 30.2 h, respectively. This gain can be increased from a higher coarse-graining level. Combination of the CG/AA description and parallel MD simulations is an open way to the possible extension of simulation studies to larger time and length scales. Petascale architectures are a very promising technology for exploring the large-sized systems. Indeed, the MD simulations are 1000 faster, which allows for reaching the microsecond time scale. We think that the combination of the petascale machines and the CG description will allow the exploration of very large length and time scales. Then, it is important to develop the methodologies as presented in this work to decrease the complexity of the studied systems. Indeed, the size and the complexity of these nearly increase with the power of the computer. Combination of the CG/AA model with petascale architectures will improve the exploring and understanding of the physics properties under confinement.

The coarse-graining procedure keeps the surface roughness and mesostructure (related to SANS) unchanged with respect to the initial AA model. Some better insights into the microstructure and surface roughness could be beneficial to the description of some materials of current interest, such as porous silicon or SBA-15. In these cases, quenched disorder induced by large surface roughness as well as microporosity inside the wall have been discussed mainly from the experimental side and in a limited number of promising simulation works. These studies would go beyond the simple case of MCM-41 addressed here but are of major interest.

4. Results and Discussion

4.1. Highly Hydrated MCM-41. We report in Figure 2a and b the reduced enthalpy of adsorption $\Delta_{\text{ads}} H$ (eq A1 of the Appendix) and the adsorbed amount, n_{ads} , as function of the reduced pressure for the AA and CG/AA force fields.

We obtain a perfect agreement between the two models for a pore diameter of $D = 2.4 \text{ nm}$ for the two physical properties. This validates the coarse-graining method used to describe the mesoscopic interactions between the porous framework and the adsorbate. This allows us to take advantage of the gain in computation speed provided by the coarse-grained description. For instance, we have performed CG/AA simulations of a larger pore size ($D = 5 \text{ nm}$). We observe a significant shift of the pressure of capillary condensation for the two pore sizes from about $P = 0.3$ to $P = 0.8$. This can be attributed to the different density of silanol at the inner surface. This pore size effect on the adsorption isotherm is in qualitative agreement with experimental results obtained for the same pore size and shown as filled symbols in Figure 2a.⁴⁴ However the agreement between simulations and experiments is not fully quantitative, especially the capillary condensation which occurs around $P = 0.6$ for the largest pore in the experimental case. Although, this would mean that the AA interaction parameters could probably be improved, it does not affect our conclusions about the validity of the coarse-graining procedure. As a hint for future studies, it is most probable that the adsorption properties are sensitive to small variations of the nature of the hydrophilic surfaces of the MCM-41 used in the different experiments—in terms of silanol density, for instance. This is a crucial parameter that can be tuned in the simulation models to monitor its influence on the physics of confined materials. A critical check of the coarse-grained model with respect to interfacial interaction and liquid structure is provided by the number of hydrogen bonds and the radial density profiles shown in Figure 3a and b, respectively.

Hydrogen bonds between methanol and silanol groups ($\text{Me}-\text{HO}\cdots\text{HO}-\text{Si}$ and $\text{Me}-\text{OH}\cdots(\text{OH})\text{Si}$) have been defined according to commonly used geometrical criteria.⁴⁵ They have been calculated as a function of pressure and presented in Figure 4a after normalization to one methanol or one silanol. At the lowest pressure, about one H bond

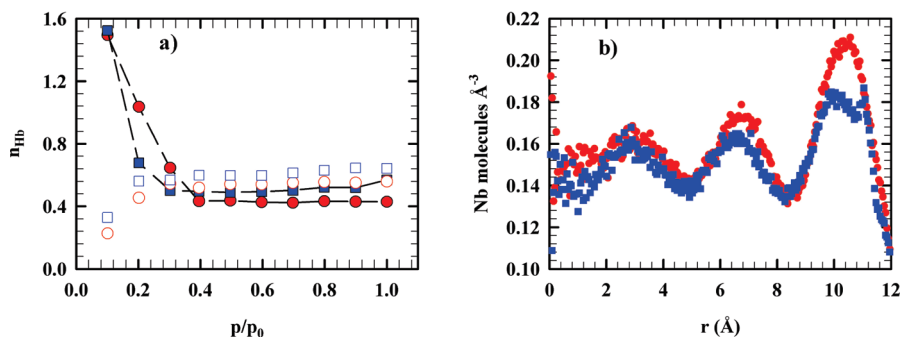


Figure 4. (a) Number of hydrogen bonds between the hydrogen atom of methanol and oxygen atom of the silanol located at the internal surface (and the oxygen atom of methanol and the hydrogen atom of silanol) per silanol (full symbols) and per methanol (empty symbols). The circle and square symbols represent the AA and CG/AA models, respectively, at $p/p_0 = 1$. (b) Radial distributions of methanol in the pore ($D = 24 \text{ \AA}$) using two models: AA (●) and CG/AA (■).

per methanol is formed, which means that in the regime of low coverage, methanol molecules are essentially adsorbed on silanol sites. In the gas phase, the competition between methanol/methanol and silanol/methanol interactions mostly favors surface adsorption.

This average number decreases on increasing the loading, since multiple layers are formed, which leads to an increasing number of methanol molecules far from the interface and the occurrence of more methanol–methanol interactions. In the dense liquid phase, the competition between the different types of interactions is balanced, and they equally contribute to the enthalpy of the system. At complete loading (264 molecules), the number of H bonds per silanol saturates to only about 0.6. This incomplete saturation of surface H-bonded sites means that steric hindrance near the pore wall probably prevents surface silanols from forming more H bonds. These features are equally observed for the AA and CG/AA models, suggesting that the interfacial fluid/framework interactions are correctly described. We obtain the same concordance for the radial profiles of methanol in the pore (Figure 4b). The CG model nicely captures the details of the layering organization at ambient pressure as well as the structural organization at the interface, which are key features of the confinement effect on fluid properties.³¹ The difference in radial density between the AA and CG/AA models near the walls is due to the difference in the adsorbed amount.

As for dynamic properties, we provide in Figure 4a the 3D average translational diffusion coefficient D_t . It is obtained from the time evolution of the isotropic mean-square displacement (MSD) of the molecular center-of-mass according to eq 6. In this equation, N_0 is the number of the origin time (t_0), t is the MD time, r the vector position of the centre of mass (com) and N the number of the particles.

Same values of D_t are obtained within the statistical uncertainties for the two models, which gives support to the coarse-graining procedure with respect to the dynamical behavior too. With increasing pressure, the value of D_t decreases systematically, which can be related to the increase of the average density of the confined phase during filling and the collapse of the free liquid–gas interface.

$$D_t = \lim_{t \rightarrow \infty} \frac{\langle \sum_{t_0}^{N_0} \sum_{i=1}^N |r_{\text{com},i}(t + t_0) - r_{\text{com},i}(t)|^2 \rangle}{6NN_0t} \quad (6)$$

In order to get better insight into the translational dynamics, we compare in Figure 4b the MSD along the three directions of space, x , y , and z , and their isotropic average for the two models at complete filling, $p/p_0 = 1$.

A crossover is observed around $t = 100 \text{ ps}$ in the time variation of the MSD measured along the z direction of the pore axis (Figure 5b). It corresponds to a change of the translation motion from subdiffusive at short times to fully diffusive at longer times (i.e., $\text{MSD}(t) \sim t$ [blue line]), which is a typical feature of the dynamics of dense liquids. At variance in the two directions perpendicular to the pore axis (x and y), the short-time subdiffusive regime is not followed by a diffusive regime. On the contrary, the MSD bends and reaches a plateau value for times longer than 1 ns. This anisotropy in the transport reflects the effects of the unidirectional spatial confinement, which primarily constrains molecule motions in the x and y directions. Experimental and numerical signatures of similar low-dimensional diffusion have been reported in the literature for different sorts of confined materials.^{46–48} Again, the details of these dynamical features are well reproduced by the CG/AA model.

Finally, we checked the transferability of the model, which is one crucial criterion for the prediction of thermodynamic properties. We computed the isotherm of adsorption of water in silica for a pore size $D = 2.4 \text{ nm}$ using the TIP4P2005 model for water⁴⁸ and the AA and CG/AA force fields discussed previously for the mesoporous silicate. Figure 6 shows excellent agreement between the adsorbed quantities obtained for the two different models.

Furthermore, a comparable agreement is obtained for the enthalpy of adsorption⁴³ at low pressures, which is $45 \pm 1.1 \text{ kJ mol}^{-1}$ for AA and $46 \pm 0.9 \text{ kJ mol}^{-1}$ for CG/AA. This successful test with respect to the adsorption isotherm and an energetic quantity shows the high transferability of the CG/AA force field to some other molecular adsorbates.

4.2. Weakly Hydrated MCM-41. Figure 7a shows the isotherms of adsorption of AA and CG/AA force fields of the weakly hydrated porous silicates. It exhibits a good

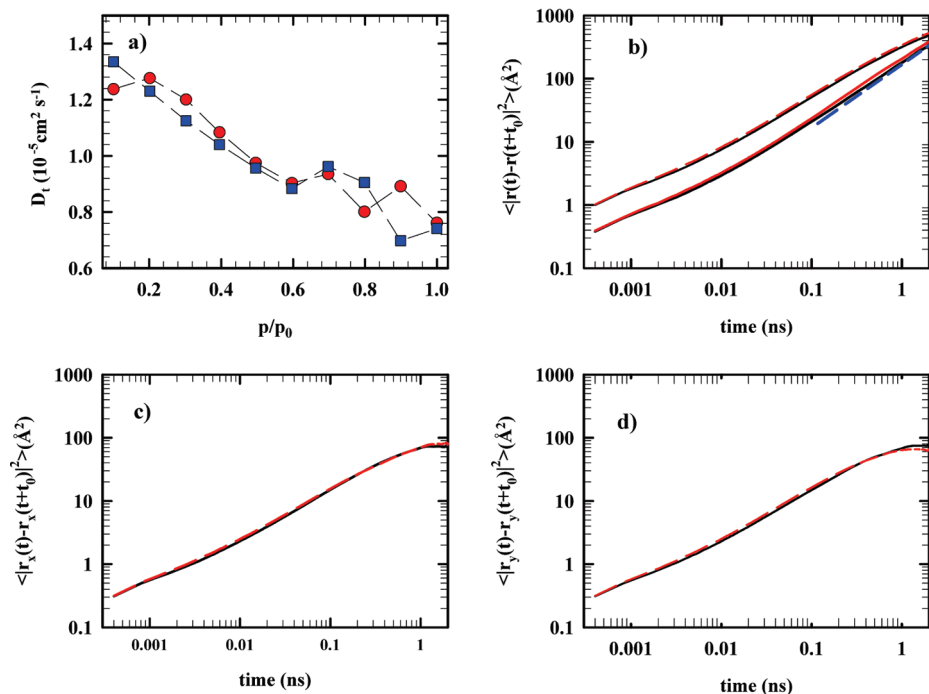


Figure 5. (a) Translational diffusion of methanol using two models: AA (●) and CG/AA (■) as function of the reduced pressure. Mean square displacement (MSD) of methanol into the pore ($D = 24 \text{ \AA}$) according to the x (c), y (d), and z and total components (b). The red dashed line represents the CG/AA model, while the black solid line corresponds to the AA model. Parts b, c, and d are represented in logarithmic scale. In part b, we reported the curve (blue dashed line) of $y = Dt$ as a guide for the eye.

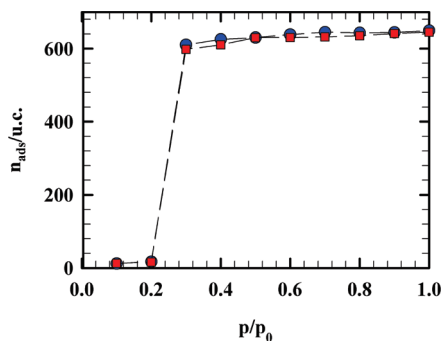


Figure 6. Isotherm of adsorption of water in a pore of diameter $D = 24 \text{ \AA}$, using two models: AA (●) and CG/AA (■).

accordance between the two models, which further validates our multiscale optimization method. Figure 7b provides a comparison between the isotherms of adsorption of methanol in WH and HH MCM-41, where direct effects of the hydrophilic character of the surface on the adsorption properties can be observed. We have considered three different regions of the adsorption curves, labeled A, B, and C in Figure 7a. In region A, which corresponds to low pressure, we have a higher adsorbed amount for WH MCM-41 than for HH. This trend is inverted in region B, since the capillary filling is shifted to higher pressure for the WH material. Finally, in region C, a slightly higher adsorbed quantity is reached for the WH form at complete loading.

A greater adsorbed amount in regions A and C, for the WH material with respect to the HH, can be related to the greater accessible surface and porous volume of the former material, respectively. The adsorption mechanism in region A mostly corresponds to the pore surface coverage by

methanol. The adsorbed amount is therefore sensitive to the accessible surface (S_{acc}), which has become larger for the WH material after removal of a part of the surface silanol groups. To get a more quantitative depiction, we calculate the accessible surface area (usually denoted ASA in the literature), which corresponds to the area traced out by the center of a probe molecule as the probe is rolled across the surface of the framework atoms. In practice, the accessible surface area is obtained from a simple Monte Carlo integration where the probe sphere is randomly inserted around the surface of each framework atom in turn and tested for overlap.³⁸ The fraction of probes that do not overlap with other framework atoms is used to calculate the accessible surface area. The probe should be chosen to correspond to the size of the adsorbate of interest, i.e., 3.9 \AA for the methanol molecule. According to this procedure, we obtain $S_{\text{acc}}(\text{HH}) = 2840.03 \text{ \AA}^2$ and $S_{\text{acc}}(\text{LH}) = 3078.09 \text{ \AA}^2$, which is consistent with the different adsorbed amount at low pressure controlled by the state of the surface. For similar reasons, the slightly larger adsorbed amount at complete filling (region C) for the WH materials can be related to a larger accessible volume (linked to a larger S_{acc}) induced by silanol group removal during surface dehydration.

The inversion in region B is related to a shift to higher pressure of the capillary condensation for the WH material. The pressure of capillary condensation, which corresponds to a rather steep increase of the adsorbed amount at intermediate pressure, is closely related to the value of the fluid–pore interaction as well as the pore size. A weaker fluid–pore interaction is expected for methanol when the surface hydrophobicity is tuned from high (for the HH MCM-41) to moderate (for the WH MCM-41). On the microscopic

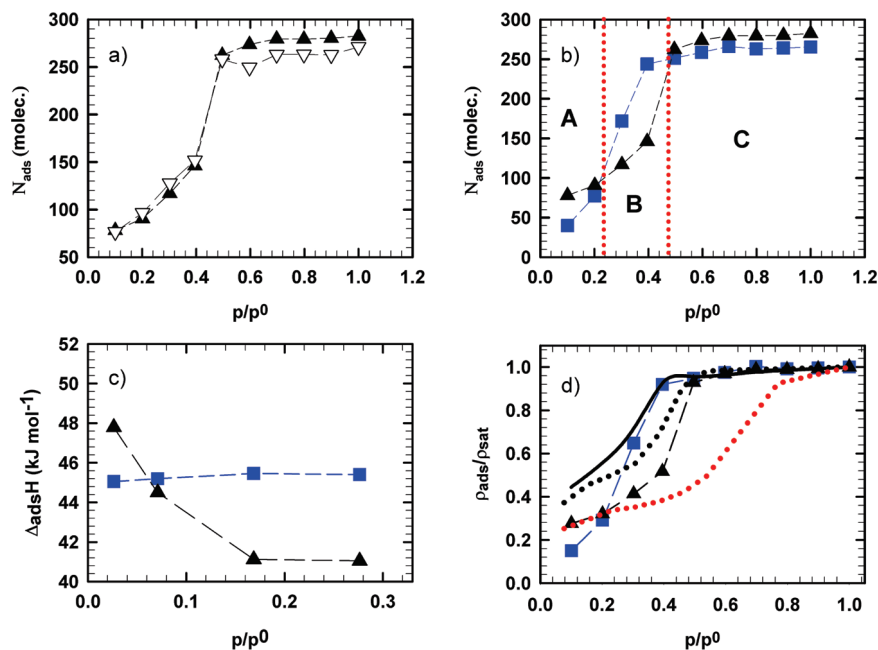


Figure 7. (a) Isotherm of the adsorption of methanol in a pore of the weakly hydrated porous silicates with diameter $D = 24 \text{ \AA}$, using two models: AA (▲) and CG/AA (▽). Isotherm (b) and enthalpy (c) of the adsorption of methanol in a pore of diameter $D = 24 \text{ \AA}$ using the AA force field for WH (▲) and HH (■) MCM-41. (d) Isotherm of the adsorption of methanol in a pore of diameter $D = 24 \text{ \AA}$ using the AA force field for WH (▲) and HH (■) MCM-41. Experimental isotherms for $D = 24.1^{44} \text{ \AA}$ (solid line), $D = 21 \text{ \AA}$ (dotted line), and $D = 28 \text{ \AA}$ (red dotted line) taken from ref 50.

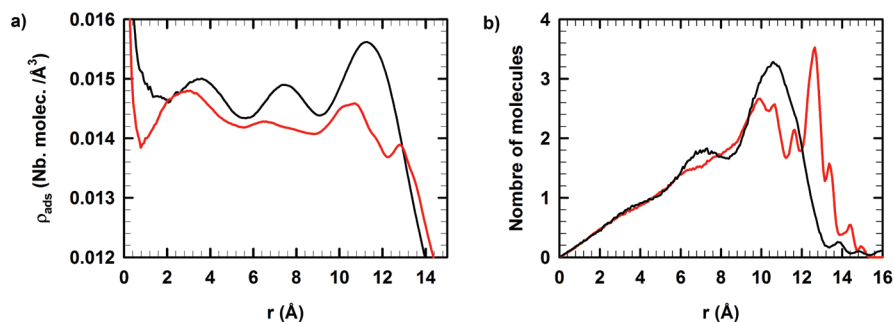


Figure 8. (a) Radial density of the methanol for HH (black line) and WH (red line) MCM-41. (b) Number of methanol molecules as a function of the distance from the center of the pore.

scale, this phenomenon can be related to the lower tendency of methanol to form an interfacial network of hydrogen bonds with the WH silicate. This interpretation is confirmed by the calculation of the number of hydrogen bonds by silanol groups in the two forms (~ 0.4 in HH and ~ 0.2 in WH at complete loading). This behavior is also reflected by the variation of the adsorption enthalpy (Figure 6c) where we observe a similar crossover from regions A and B for the two types of porous materials. Figure 6d shows a fair agreement between our simulation results for the HH form of MCM-41 and the experimental isotherm of adsorption from Carrot et al. for $D = 24 \text{ \AA}$ ³³ while the WH isotherm is shifted to higher pressure. It would suggest that the highly hydrated form considered in this case provides a more realistic description of the experimental sample. However, if we consider the more recent work on materials where a low silanol density was reported ($2\text{--}3 \text{ OH/nm}^2$), it appears that our simulated adsorption curve for the WH MCM-41 with diameter $D = 24 \text{ \AA}$ lies nicely between the two experimental curves obtained for adjacent values of diameter

$D = 21$ and 28 \AA .^{49,50} This can be considered as a valuable test of the validity of our derivation of a CG/AA force field for MCM-41.

Finally, we compare the radial density profile of methanol for WH and HH MCM-41 at saturation in order to analyze the effect of silanol density on the layered structure in the confined liquid. As discussed in the previous section, methanol confined in the HH form exhibits a strongly layered structure, as shown in Figure 8a. This radial structuring of the fluid is weaker in the WH porous silicate: the amplitude of the modulation of density is smaller, and the peaks are broader. In addition, the density profile of the contact layer ($r > 8 \text{ \AA}$) exhibits two maxima. Layering is induced by the enhancement of methanol order from the solid surface that propagates toward the inner pore. This effect is expected to be smaller for the WH materials, because it induces a smaller number of interfacial H bonds between methanol and silanol groups. In addition, the WH surface presents sites with different hydrophilic character, which is at variance with the HH surface that is smooth and uniformly covered by silanols.

In the WH matrix, interfacial molecules can therefore feel a broader variety of local environments. It can induce different types of preferred configurations for the molecules located at the surface, as the shape of the density profile within the contact layer would suggest. This phenomenon further weakens the surface ordering effect.

5. Conclusion

Molecular simulations can provide unique contributions to a better understanding of the mesoporous confinement effect on fluid properties. They have been constantly improved by appropriate methodological developments.

Indeed, the physics of confined materials generally requires extending the investigation to long length and time scales. This can introduce strong limitations to the capability of all atoms descriptions, which are very time-demanding. On the other hand, full mesoscopic approaches are deficient in a sufficient microscopic account of the solid/fluid interfacial interactions.

For that purpose, we developed a general and multiscale procedure to design a hybrid coarse grained/all atom force field that gathers the benefits of both approaches: it includes an atomistic description of the inner pore surface, including a realistic account of silanols and a coarse-grained description of the overall porous framework. We presented a derivation method based on the optimization of the CG parameters with respect to macroscopic properties, i.e., the adsorbed amount and the enthalpy of adsorption calculated by AA and CG simulations. We have fully validated our approach in the case of methanol adsorbed in a MCM-41 material. This coarse-graining method is fast and versatile. We have shown a good reproduction of both static and dynamical properties, in terms of adsorbed quantity, isosteric heat of adsorption, interfacial structure, and translation coefficient. Besides, we have shown that GC/AA provides promising openings for simulation of much larger systems, as illustrated with a pore of diameter 5 nm and is fully transferable to other adsorbates, such as water. We have also shown that the GC/AA can address the currently debated question of the effect of the hydrophilic character of the porous materials on the adsorption properties and the liquid structure. Indeed, it fully accounts for the presence of surface silanol groups, the number of which can be tuned as shown in our comparative study for two MCM-41's with different hydration levels. Finally, this model represents a very interesting and powerful intermediate model between all microscopic and mesoscopic descriptions. This method can be easily transferred to the other porous materials with the specific sites of interactions such as metal organic framework (MOF), covalent organic framework, or polymer coordination from a higher degree of coarse graining. Additionally, other macroscopic properties can be used in the derivation process.

Appendix

The calculation of $(\partial\Delta_n H_{\text{ads}})/(\partial y)$ with $y = \varepsilon$ and σ implies a reminder of the definition of $\Delta_n H_{\text{ads}}$ (eq A1).

$$\Delta_n H_{\text{ads}} = \frac{\langle n_{\text{ads}} H \rangle - \langle n_{\text{ads}} \rangle \langle H \rangle}{\langle n_{\text{ads}}^2 \rangle - \langle n_{\text{ads}} \rangle^2} \quad (\text{A1})$$

In eq A1 n_{ads} is the amount adsorbed and H the total Hamiltonian of the system composed of potential and kinetic terms. For more clarity, we omitted the dependence with respect to \mathbf{r} and \mathbf{p} in the Hamiltonian (H). The derivation with respect to y provides eq A2.

$$\frac{\partial \Delta_n H_{\text{ads}}}{\partial y} = \frac{1}{(\langle n^2 \rangle - \langle n \rangle^2)} \left[\left(\frac{\partial \langle nH \rangle}{\partial y} - \langle H \rangle \frac{\partial \langle n \rangle}{\partial y} - \langle n \rangle \frac{\partial \langle H \rangle}{\partial y} \right) (\langle n^2 \rangle - \langle n \rangle^2) - \left(\langle nH \rangle - \langle n \rangle \langle H \rangle \right) \left(\frac{\partial \langle n^2 \rangle}{\partial y} - 2 \langle n \rangle \frac{\partial \langle n \rangle}{\partial y} \right) \right] \quad (\text{A2})$$

For more clarity we omit the subscript ads. The terms $(\partial \langle nH \rangle)/(\partial y)$, $n(\partial \langle H \rangle)/(\partial y)$, $(\partial \langle n \rangle)/(\partial y)$, $(\partial \langle n^2 \rangle)/(\partial y)$, and $(\partial \langle H \rangle)/(\partial y)$ have been determined by following relations while $(\partial \langle H \rangle)/(\partial y)$ is evaluated by a first-order finite difference.

$$\frac{\partial \langle nH \rangle}{\partial y} = \left\langle n \frac{\partial H}{\partial y} \right\rangle - \beta \left(\left\langle (nH) \frac{\partial H}{\partial y} \right\rangle - \langle nH \rangle \left\langle \frac{\partial H}{\partial y} \right\rangle \right) \quad (\text{A3})$$

$$\frac{\partial \langle n \rangle}{\partial y} = -\beta \left(\left\langle n \frac{\partial H}{\partial y} \right\rangle - \langle n \rangle \left\langle \frac{\partial H}{\partial y} \right\rangle \right) \quad (\text{A4})$$

$$\frac{\partial \langle n^2 \rangle}{\partial y} = -\beta \left(\left\langle (n^2) \frac{\partial H}{\partial y} \right\rangle - \langle n^2 \rangle \left\langle \frac{\partial H}{\partial y} \right\rangle \right) \quad (\text{A5})$$

$$\frac{\partial \langle H \rangle}{\partial y} = -\beta \left(\left\langle (H) \frac{\partial H}{\partial y} \right\rangle - \langle H \rangle \left\langle \frac{\partial H}{\partial y} \right\rangle \right) \quad (\text{A6})$$

References

- (1) Gelb, L. D.; Gubbins, K. E.; Radhakrishnan, R.; Sliwinski-Bartkowiak, M. *Rep. Prog. Phys.* **1999**, *62*, 1573.
- (2) Christenson, H. K. *J. Phys. Cond. Mat.* **2001**, *13*, R95–R133.
- (3) Alcoutlabi, M.; McKenna, G. B. *J. Phys.: Cond. Matter* **2005**, *17*, R461.
- (4) Alba-Simionesco, C.; Coasne, B.; Dossch, G.; Dudziak, G.; Gubbins, K. E.; Radhakrishnan, R.; Sliwinski-Bartkowiak, M. *J. Phys.: Condens. Matter* **2006**, *18*, R15.
- (5) Klein, J.; Kumacheva, E. *Science* **1995**, *269*, 816.
- (6) Granick, S. *Science* **1991**, *253*, 1374.
- (7) Bellini, T.; Radzihovsky, L.; Toner, J.; Clark, N. A. *Science* **2001**, *294*, 1074.
- (8) Coasne, B.; Jain, S. K.; Gubbins, K. *Phys. Rev. Lett.* **2006**, *97*, 105702.
- (9) Guégan, R.; Morineau, D.; Lefort, R.; Moréac, A.; Béziel, W.; Guendouz, M.; Zanotti, J.-M.; Frick, B. *J. Chem. Phys.* **2007**, *126*, 1064902.
- (10) Lefort, R.; Morineau, D.; Guégan, R.; Guendouz, M.; Zanotti, J.-M.; Frick, B. *Phys. Rev. E* **2008**, *78*, 040701(R).
- (11) Coasne, B.; Hung, F. R.; Pellenq, R. J.-M.; Siperstein, F. R.; Gubbins, K. E. *Langmuir* **2006**, *22*, 194.
- (12) Puibasset, J.; Pellenq, R. J.-M. *J. Chem. Phys.* **2005**, *122*, 094704.
- (13) Coasne, B.; Renzo, F. D.; Galarneau, A.; Pellenq, R. J.-M. *Langmuir* **2008**, *24*, 7285.
- (14) Guégan, R.; Morineau, D.; Loverdo, C.; Béziel, W. *Phys. Rev. E.* **2006**, *73*, 011707.

- (15) Kityk, A. V.; Wolff, M.; Knorr, K.; Morineau, D.; Lefort, R.; Huber, P. *Phys. Rev. Lett.* **2008**, *101*, 187801.
- (16) Neri, M.; Anselmi, C.; Cascella, M.; Maritan, A.; Carloni, P. *Phys. Rev. Lett.* **2005**, *95*, 218102.
- (17) Hoogerbrugge, P. J.; Koelman, J. *Eur. Phys. Lett.* **1992**, *19*, 155.
- (18) Flekkøy, E. G.; Coveney, P. V. *Phys. Rev. Lett.* **1999**, (83), 1775.
- (19) Chen, N.-Y.; Su, Z.-Y.; Mou, C.-Y. *Phys. Rev. Lett.* **2006**, *96*, 078103.
- (20) Detcheverry, F. A.; Pike, D. Q.; Nealey, P. F.; Müller, M.; Pablo, J. J. d. *Phys. Rev. Lett.* **2009**, *102*, 197801.
- (21) Dupuis, L. M.; Tadmor, E. B.; Miller, R. E.; Phillips, R. *Phys. Rev. Lett.* **2005**, *95*, 060202.
- (22) Fabritiis, G. D.; Delgado-Buscalioni, R.; Coveney, P. V. *Phys. Rev. Lett.* **2006**, *97*, 134501.
- (23) Dubbeldam, D.; Calero, S.; Vlugt, T. J. H.; Krishna, R.; Maesen, T. L. M.; Beerdsen, E.; Smith, B. *Phys. Rev. Lett.* **2004**, *93*, 088302.
- (24) Ashbaugh, H. S.; Patel, H. A.; Kumar, S. K.; Garde, S. *J. Chem. Phys.* **2005**, *122* (10), 104908.
- (25) Elezgaray, J.; Laguerre, M. A. *Comput. Phys. Commun.* **2006**, *175* (4), 264.
- (26) Lyubatsev, A. P. *Eur. Biophys. J. Biosophys. Lett.* **2005**, *35* (1), 5361.
- (27) Ivzvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109* (7), 2469.
- (28) Das, A. C.; Andersen, H. *J. Chem. Phys.* **2009**, *131*, 034102.
- (29) Gupta, N. M.; Kumar, D.; Kamble, V. S. S.; Mitra; Mukhopadhyay, R.; Kartha, V. B. *J. Phys. Chem. B* **2006**, *110*, 4815.
- (30) Takahara, S.; Kittaka, S.; Mori, T. Y.; Kuroda; Takamuku, T.; Yamaguchi, T. *J. Phys. Chem. C* **2008**, *112*, 14385.
- (31) Guégan, R.; Morineau, D.; Alba-Simionesco, C. *Chem. Phys.* **2005**, *317*, 236.
- (32) Morineau, D.; Guégan, R.; Xia, Y.; Alba-Simionesco, C. *J. Chem. Phys.* **2004**, *121*, 1466.
- (33) Ribeiro Carrot, M. M. L.; Candeias, A. J. E.; Carrot, P. J. M.; Ravikovitch, P. I.; Neimark, A. V. *Microporous Mesoporous Mater.* **2001**, *47*, 323–337.
- (34) Vink, R. L. C.; Barkema, G. T. *Phys. Rev. B* **2003**, *67*, 245201.
- (35) Bródka, A.; Zerda, T. W. *J. Chem. Phys.* **1996**, *104*, 6319.
- (36) Chen, B.; Potoff, J. J.; Siepmann, J. I. *J. Phys. Chem. B* **2001**, *105*, 2569.
- (37) Bourasseau, E.; Haboudou, M.; Boutin, A.; Fuchs, A. H.; Ungerer, P. *J. Chem. Phys.* **2003**, *118*, 3020.
- (38) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: New York, 1987.
- (39) Chempath, S.; Clark, L.; Snurr, R. *J. Chem. Phys.* **2003**, *118*, 7635.
- (40) Lupkowski, M.; Swol, F. v. *J. Chem. Phys.* **1991**, *95*, 1995.
- (41) Boinepalli, S.; Attard, P. *J. Chem. Phys.* **2003**, *119*, 12769.
- (42) Forester, T. R.; Smith, W. *DLPOLY CCP5 Program Library*; Daresbury Lab.: Cheshire, U. K., 1994.
- (43) Nicholson, D.; Parsonage, N. G. *Computer Simulation and the Statistical Mechanics of Adsorption*; Academic Press: London, 1982.
- (44) Ghoufi, A.; Morel, J. P.; Desrosiers, N. M.; Malfreyt, P. *J. Phys. Chem. B* **2005**, *109*, 23579.
- (45) Busselez, R.; Lefort, R.; Ji, Q.; Affouard, F.; Morineau, D. *Phys. Chem. Chem. Phys.* **2009**, *11*, 11127.
- (46) Malikova, N.; S.; Longeville, S.; Zanotti, J.-M.; Dubois, E.; Marry, V.; Turq, P.; Ollivier, J. *Phys. Rev. Lett.* **2008**, *101*, 265901.
- (47) Scheidler, P.; Kob, W.; Binder, K. *J. Phys. Chem. B* **2004**, *108*, 6673.
- (48) Vega, C.; de Miguel, E. *J. Chem. Phys.* **2007**, *126*, 154707.
- (49) Takamuku, T.; Maruyama, H.; Kittaka, S.; Takahara, S.; Yamaguchi, T. *J. Phys. Chem. B* **2005**, *109*, 892.
- (50) Yamagushi, T.; Yoshida, K.; Smirnov, P.; Takamuku, T.; Kittaba, S.; Takahara, S.; Kuroda, Y.; Bellissent-Funel, M. C. *Eur. Phys. J.* **2007**, *141*, 19–27.

CT100169R

Defining Condensed Phase Reactive Force Fields from *ab Initio* Molecular Dynamics Simulations: The Case of the Hydrated Excess Proton

Chris Knight,[†] C. Mark Maupin,[‡] Sergei Izvekov,[‡] and Gregory A. Voth^{*,†,‡}

Department of Chemistry, James Franck Institute, and Computation Institute, University of Chicago, 5735 South Ellis Avenue, Chicago, Illinois 60637, United States, and Center for Biophysical Modeling and Simulation and Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States

Received August 9, 2010

Abstract: In this report, a general methodology is presented for the parametrization of a reactive force field using data from a condensed phase *ab initio* molecular dynamics (AIMD) simulation. This algorithm allows for the creation of an empirical reactive force field that accurately reproduces the underlying *ab initio* reactive surface while providing the ability to achieve long-time statistical sampling for large systems not possible with AIMD alone. In this work, a model for the hydrated excess proton is constructed where the hydronium cation and proton hopping portions of the model are statistically force-matched to the results of Car–Parrinello Molecular Dynamics (CPMD) simulations. The flexible nature of the algorithm also allows for the use of the more accurate classical simple point-charge flexible water (SPC/Fw) model to describe the water–water interactions while utilizing the *ab initio* data to create an overall multistate molecular dynamics (MS-MD) reactive model of the hydrated excess proton in water. The resulting empirical model for the system qualitatively reproduces thermodynamic and dynamic properties calculated from the *ab initio* simulation while being in good agreement with experimental results and previously developed multistate empirical valence bond (MS-EVB) models. The present methodology, therefore, bridges the AIMD technique with the MS-MD modeling of reactive events, while incorporating key strengths of both.

1. Introduction

The simulation of certain classes of chemical reactions and charge transport processes in the condensed phase is a challenging problem. This is because of the large system sizes necessary to eliminate system size dependent artifacts from long-ranged electrostatics as well as the long time scales necessary to accurately simulate competing or intertwined dynamical processes such as diffusion. While simple ion transport is difficult in its own right, the hydrated excess proton, which is a delocalized net positive charge defect spanning over multiple water molecules, is considered one of the most difficult ions to simulate due to the variable

bonding topology that accompanies transport via the Grotthuss mechanism.^{1,2} For modeling of complex chemical processes in the condensed phase such as the hydrated excess proton, one can utilize *ab initio* molecular dynamics (AIMD) simulations (refs 3–5 and those cited therein). However, the computational cost of AIMD simulations typically limits the system size and simulation time length when applied to condensed phase systems. In addition, one is also currently limited by the accuracy of density functionals, at least until high-level Schrödinger wave function methods, such as MP2,^{6,7} become tractable for condensed phase simulations. Therefore, it is advantageous to develop a general procedure for accurately parametrizing empirical reactive force fields that allow for the study of complex chemical processes, i.e., the making and breaking of chemical bonds, and that can reproduce thermodynamic and dynamic properties of an

* Corresponding author e-mail: gavoith@uchicago.edu.

[†] University of Chicago.

[‡] University of Utah.

AIMD simulation while also extending both the length and time scale of the simulation to probe more complex phenomena.

In typical molecular mechanics force fields, a single bonding topology is assumed throughout the entire simulation, which precludes the possibility of modeling reactive events that rely upon the breaking and formation of chemical bonds. To address this fixed bonding topology deficiency, reactive molecular dynamics methodologies that explicitly handle variable coordination and bonding topologies have been utilized to study various chemically reactive systems of interest.^{8,9} The multistate molecular dynamics (MS-MD) approach, of which the multistate empirical valence bond (MS-EVB) methodology is an example,^{10–14} is a reactive simulation method that dynamically changes the underlying bonding topology of a reactive species in response to the environment. This dynamic evolution of the bonding topology is achieved by constructing a linear combination of several bonding topologies describing various “states” of the acceptor/donor species; the relative weights of which are determined “on the fly” during the course of the simulation. It is the dynamic nature of this reactive force field that allows for an accurate description of bond making/breaking processes, such as those involved in the transport of the hydrated excess proton and hydroxide ions.^{2,5,13–15}

The original EVB approach^{16–19} was used as an interpolation scheme to describe the reaction path for chemical reactions between well-defined reactant–product pairs. In this original method, a Hamiltonian matrix, commonly 2×2 , is constructed where all elements are expressed as functions of geometric coordinates of the system. The diagonal elements of the Hamiltonian matrix correspond to the diabatic states (bonding topologies) representing the donor and acceptor in the chemical reaction of interest. The off-diagonal elements describe the coupling between two diabatic states (donor and acceptor) and provide a mechanism for transitions, allowing for the donor and acceptor to switch identities. This scheme is general enough to be able to describe various types of chemical reactions, such as intramolecular isomerization/rearrangement²⁰ and nucleophilic substitution.²¹ Once the reactant and product states are identified, the Hamiltonian matrix is populated (diagonal and off-diagonal components) and subsequently diagonalized. Upon diagonalization, the coefficients of the lowest energy eigenvector are used to calculate the forces using the Hellman–Feynman theorem and the system is propagated in time using Newtonian equations of motion. The same basic idea with certain extensions has been implemented in the multiconfigurational molecular mechanics approach.^{22,23}

The multistate MD generalization of the EVB idea was to dynamically include bonding topologies spanning multiple molecules and larger effective distances (MS-EVB).^{13,14} For each simulation step, the multiple diabatic states, used to construct the Hamiltonian matrix, are determined by identifying all possible donor–acceptor pairs proximal to an initial donor–acceptor. The dimension of this Hamiltonian can be kept relatively modest using geometric criteria, such as bond distances and angles, to “weed out” those bonding topologies that are likely to be energetically unfavorable. For

the case of the hydrated excess proton in bulk water, it is found to be sufficient to include ~ 30 states (up to and including the third solvation shell of the excess proton),²⁴ while only a few states have significant weight for channel-type systems.²⁵ By allowing the MS-MD algorithm the freedom to consider all of these possible states, one can properly model the excess proton charge defect delocalization over several molecules adjacent to the donor–acceptor species while accounting for the dynamical rearrangement of the chemical bond topology.

The MS-MD (or specifically MS-EVB) models developed previously have been parametrized against high-level ab initio gas phase calculations^{12,26,27} and experimental data. These fitting procedures only provide a partial route for defining an accurate condensed phase potential energy surface for the reactive system. It is, thus, desirable to develop new, robust, and efficient algorithms for the parametrization of accurate reactive MS-MD force fields based instead on explicit condensed phase AIMD simulations. As detailed earlier and below, the MS-MD methodology then enables simulations of large systems over the long time scales necessary to adequately sample the evolving environment, whether it is the fluctuating hydrogen bond (H-bond) rearrangements in aqueous solutions or the slow protein conformational changes in biomolecular environments.

Force matching (FM) algorithms provide a means by which the forces on atoms calculated from some method with a high level treatment of interactions, typically a computationally expensive method such as accurate ab initio calculations, can be used to construct and parametrize an empirical model that best reproduces the original forces.^{28–31} The result of a force matching calculation is an empirical force field that is capable of accurately reproducing the physical properties of the original method used to build the training set of configurations. When the interactions are expressed as a sum of pairwise functions that are linear with respect to the model parameters,^{30,31} one is able to take advantage of linear least-squares fitting procedures to determine the model parameters that best reproduce the original forces. This latter approach dramatically simplifies the FM algorithm for complex condensed phase systems.^{30,31} By choosing these effective interactions to be interpolated as cubic spline functions (third order polynomials), one can also take advantage of the favorable property of continuity of the force and its first derivative.

In this work, a force matching algorithm is developed and applied to construct a reactive force field (in this case for aqueous proton transport) using a training set of ab initio data from a condensed phase simulation. Fitting a model to data from a condensed phase simulation is advantageous due to issues related to transferability of potentials parametrized with gas phase calculations. The method presented here also allows the flexibility to force match only those portions of the model that one wants. Instead of going through the procedure of parametrizing all interactions from a training set of AIMD simulation data, one can simply use a previously parametrized model. For example, as discussed below, the previously parametrized simple point-charge flexible water (SPC/Fw) model is used to describe water–water interactions

instead of force matching a water model based on the AIMD simulation.³⁰

In the next section, the AIMD simulations that were utilized for the force matching procedure are discussed along with the procedure for the MS-EVB simulations. The use of ab initio simulations in the construction of a nonreactive hydronium–water model and reactive force-matched MS-EVB model (FM-MS-EVB), which describes the dynamic proton hopping events, is then discussed. In the Results and Discussion, the results of the newly created FM-MS-EVB model, the underlying AIMD simulation, and the original MS-EVB3 model of Wu et al.²⁴ are compared. The Article is concluded with a discussion regarding the future outlook for the present force matching algorithm and its application to complex condensed phase reactions.

2. Methods and Development

2.1. Condensed Phase ab Initio MD Simulations. To generate data to parametrize the reactive force field, AIMD simulations were conducted on a system containing 128 water molecules plus an excess proton. The simulation cell was cubic with side lengths measuring 15.66 Å. The initial configuration for the simulation was taken from an equilibrated MD simulation of bulk water using a force-matched water model.³⁰ An excess proton was added, and the system was equilibrated for 8 ps using the Car–Parrinello Molecular Dynamics (CPMD) simulation method.³² The CPMD software package (v3.5) was used for all AIMD simulations.^{32,33} The HCTH/120 density functional³⁴ was used because it has been shown to give an improved description of experimental liquid water as compared to other density functionals.^{34–36} The Kohn–Sham orbitals were expanded in a plane wave basis set with a cutoff of 80 Ry, and Troullier–Martins pseudopotentials³⁷ were used. The equations of motion were integrated with a time step of 3 au (0.073 fs), and the fictitious electron mass was set at 340 au. For hydrogen atoms, it was shown that fictitious electron masses around this value lead to stable CPMD simulations and a proper separation of the electronic and nuclear degrees of freedom.^{35,38} During this equilibration phase, the temperature was maintained at 300 K by rescaling the velocities. From the latter portion of this trajectory, two configurations were chosen at random as the initial configurations for simulations in the microcanonical (constant NVE) ensemble. Forces on all atoms in the two NVE ensemble simulations, 27 and 35 ps in length, respectively, were then used as input to force matching algorithms to parametrize both a nonreactive and reactive hydronium–water model. Results from these simulations were discussed in an earlier publication.³⁹

2.2. MS-EVB Simulations. The MS-EVB simulations were run with systems containing 216 water molecules plus an excess proton. The simulation cells were cubic with side lengths measuring 18.62 Å. The equations of motion for the FM-MS-EVB simulations were integrated with a time step of 0.5 fs using the leapfrog algorithm in a modified version of the DL_POLY (v2) software package.⁴⁰ Simulations with the nonreactive hydronium–water model and the MS-EVB3 model were run in a modified version of the LAMMPS

software package using the velocity verlet algorithm and a time step of 1.0 fs.^{41,42} The Ewald method was used to calculate the long-range electrostatics of the system. For each parameter set, as discussed below, the system was equilibrated using a Nose–Hoover thermostat⁴³ with a 0.5 ps relaxation time to maintain the temperature at 298.15 K. Once equilibrated, the final configuration from a simulation was then used to initialize a subsequent simulation in the constant NVE ensemble from which physical properties were calculated.

2.3. Force Matching the Nonreactive Hydronium Model. A reactive model describing transport of the excess proton in bulk water was constructed in several stages. First, a classical nonreactive hydronium cation model solvated by water was developed. This model was used to parametrize the diabatic states of the resulting reactive FM-MS-EVB model. The intramolecular model for the hydronium cation, H₃O⁺, was based on the MS-EVB2 hydronium model¹² with the Morse bonding potential refitted to gas phase ab initio calculations. The hydronium OH stretch was modeled with a Morse oscillator potential, eq 1, with parameters $D_e = 143.81$ kcal/mol, $\alpha = 1.7$ Å⁻¹, and $r_0 = 1.0$ Å.

$$V(r) = D_e[(1 - e^{-\alpha(r-r_0)})^2 - 1] \quad (1)$$

A harmonic bending potential, eq 2 below, was used for the hydronium HOH angle with parameters $k_b = 73.269$ kcal/mol rad⁻¹ and $\theta_0 = 116^\circ$.

$$V(\theta) = \frac{k_b}{2}(\theta - \theta_0)^2 \quad (2)$$

Charges were fixed at $-0.5e$ and $+0.5e$ for the oxygen and hydrogen atoms of the hydronium cation, respectively. In ab initio simulations of water with gradient-corrected functionals, nonergodic sampling has been observed at temperatures below 330 K.³⁶ The melting temperature for water using the PBE and BLYP density functionals was also determined to be ~ 100 K higher than the experimental value, indicating that one is actually sampling a supercooled phase at a density of 1.0 g/cm³ and temperatures near 300 K.⁴⁴ Instead of force matching a water model based on the CPMD trajectories, the more accurate SPC/Fw⁴⁵ water model was, therefore, used to describe the water–water interactions. This is a flexible point-charge model that has been shown to reproduce a number of thermodynamic and dynamic properties as compared to experiment.⁴⁵ The remaining nonbonded interactions between the hydronium cation and water molecules were obtained from a force matching calculation using data from the ab initio trajectories.

The ab initio simulation data used for the force matching calculation was selected from configurations of the CPMD trajectory where the hydronium and its first solvation shell closely resembled the Eigen cation, H₉O₄⁺, which donates three strong H-bonds, as opposed to the Zundel cation, H₅O₂⁺. The Eigen cation was selected by identifying the configurations that correspond to a “special pair” coordinate $\delta_{\text{O1x}} > 0.2$ Å.⁴⁶ The δ_{O1x} coordinate is a measure of the difference between the H-bond and covalent bond distances for a specific hydronium–water pair, as illustrated in Figure

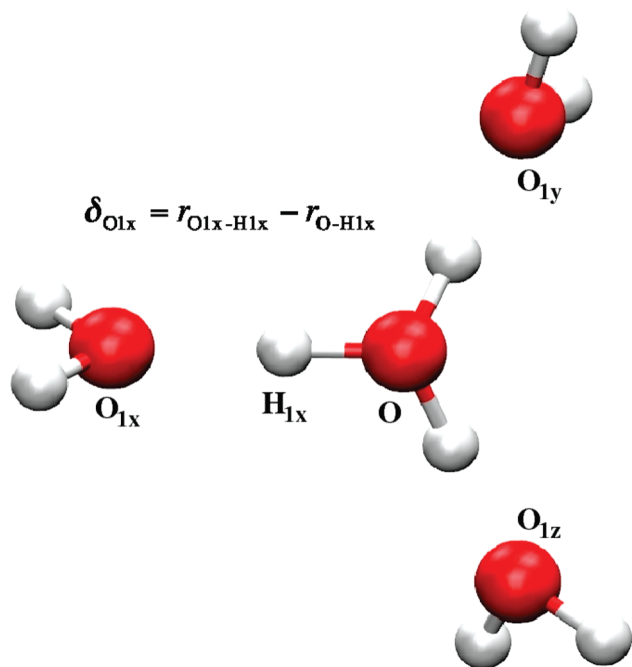


Figure 1. Hydronium cation and three H-bond accepting waters in the first solvation shell. The three water molecules are ordered according to the value of the delta parameter ($\delta_{O1x} < \delta_{O1y} < \delta_{O1z}$) for each H-bond defined as the H-bond distance minus the covalent bond distance.

1. The “O1x” labeled water with the smallest δ value identifies the water molecule that is the most likely candidate to accept a proton from the hydronium, thereby facilitating proton transfer through the Grotthuss mechanism.⁴⁶ This subset of configurations with $\delta_{O1x} > 0.2$ Å was then used as the training set for the force matching calculation³¹ that derived a set of effective pairwise potentials for the short-

ranged nonbonded interactions between the atoms of a hydronium cation and those of a water molecule.

With this method, tabulated potentials expressed as a pairwise sum of piecewise spline functions were obtained for interactions between the atoms of the hydronium cation and water molecules. The forces determined from the hydronium intramolecular model, SPC/Fw water model, and the electrostatic interactions between the hydronium cation and water molecules, calculated using Ewald summation and metallic boundary conditions,⁴⁷ were first subtracted from the ab initio forces. The remaining contribution to the ab initio force on each atom was then used to fit the hydronium–water pairwise potentials for all oxygen and hydrogen interatomic pairs. Each pairwise potential was represented by a spline curve on a grid with a spacing of 0.1 au (~ 0.05 Å). The spline force curve and its first derivative were constrained to be zero at the cutoff distance for the pairwise interactions, $r_{\text{cut}} = 9.3$ Å. At short interatomic separations that were not sampled in the simulations, the force was set to a constant (the potential was linearly extrapolated) starting at values of the “core” cutoff r_{core} : 2.275 Å for OH–OW, 2.169 Å for OH–HW, 0.9525 Å for HH–OW, and 1.5345 Å for HH–HW interatomic potentials. The resulting force curves are shown as the solid black lines in Figure 2 for each interatomic pair of atoms. The first letter of the atom label defines the element, and the second letter indicates the molecule type. For example, “OH” denotes the oxygen of the hydronium cation while “HW” defines the hydrogen of a water molecule. The original spline tabulated force curves were then fit to a polynomial series to smooth out roughness due to limited statistical sampling of interatomic separations from the CPMD simulations. A least-squares minimization was used, and the fitted

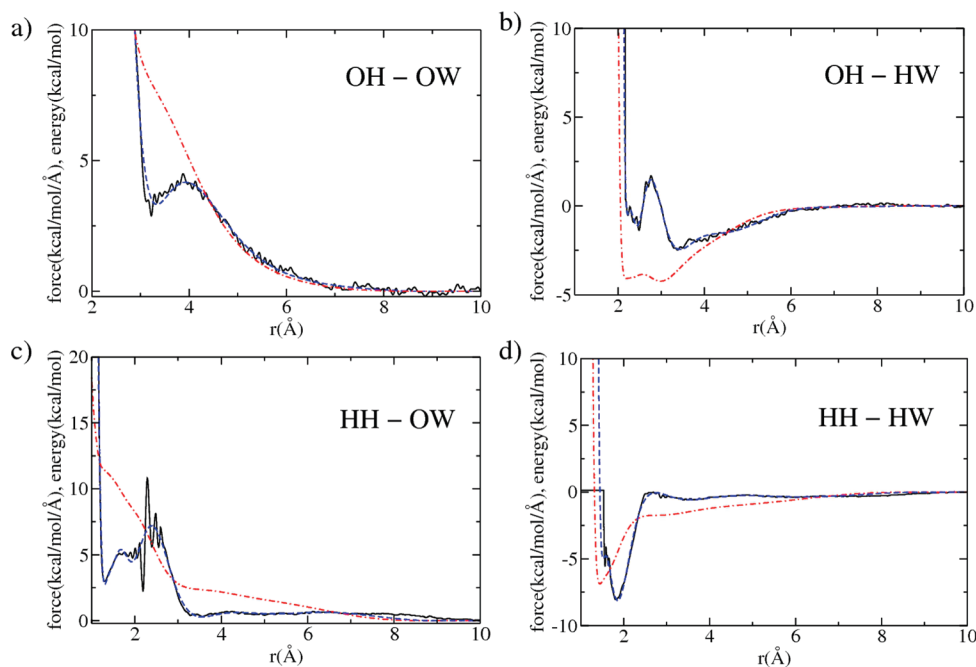


Figure 2. Pairwise forces and potentials for the hydronium–water interactions as a function of distance obtained using forces from the CPMD trajectories as input to a force matching calculation. In each plot, the solid black line is the original spline tabulated force curve as a function of interatomic distance, the dashed blue line is the smoothed force curve, and the dashed-dotted red line is the potential obtained by integrating the smoothed force.

Table 1. Coefficients from a Least Squares Fit of the Original Spline Tabulated Force Curves Generated from the Force Matching Calculation for Each Hydronium–Water Interatomic Pair^a

order	OH–OW	OH–HW	HH–OW	HH–HW
0	-17.75179440	-245.77772117	765.85959888	-20.12874479
1	39.99415051	512.74997783	-1471.08432228	30.95649268
2	-32.19721078	-470.26747909	1298.96366415	-32.77387486
3	25.71041189	447.91419530	-1051.04732640	23.76357441
4	-18.85323611	-391.49346573	779.48446553	-18.42840895
5	12.38193946	345.09130440	-523.30121181	12.73151116
6	-6.09658945	-288.91345247	315.98084927	-6.71330105
7	3.13297827	231.82017615	-168.82081489	3.74134680
8	-1.07030405	-181.81971455	77.78832254	-1.48022490
9	0.14352517	134.02526896	-29.96910784	0.22780730
10		-96.50667993	8.76791113	
11		65.83383950	-1.61090267	
12		-41.22269725	0.13169694	
13		24.56642398		
14		-13.17741630		
15		6.82225054		
16		-2.86081399		
17		0.54515725		

^a The resulting curves are shown in Figure 2.

curves were subject to constraints that the force and its first derivative are zero at the cutoff, $r_{\text{cut}} = 9.3 \text{ \AA}$. Also, the values for the core cutoff were extended to shorter distances to enable smooth extrapolation in this region: 1.5 \AA for OH–OW, 1.5 \AA for OH–HW, 0.7 \AA for HH–OW, and 1.0 \AA for HH–HW interatomic potentials.

In previous work, rapid convergence of the least-squares fitting of force curves was found when a truncated expansion of Chebyshev polynomials of the first kind as a function of the inverse interatomic distance was used to approximate the original spline tabulated data for the force:³¹

$$f(r) = \sum_{n=0}^N A_n T_n(\bar{r}) \quad (3)$$

where $T_n(\bar{r})$ is an n th order Chebyshev polynomial, A_n is a least-squares coefficient, and

$$\bar{r} = \frac{2/r - 1/r_{\text{core}} - 1/r_{\text{cut}}}{1/r_{\text{core}} - 1/r_{\text{cut}}} \quad (4)$$

For each curve, the highest order term included in the fit was the lowest one that yielded an overall good fit and for which the inclusion of higher order terms did not significantly reduce the fitting residual. The least-squares coefficients for the effective short-ranged nonbonded force for each interatomic pair are shown in Table 1. The fitted force curves and the resulting potential curves are shown in Figure 2. Results from simulations with the nonreactive hydronium model are discussed below and compared with results from the CPMD simulations and the MS-EVB3 model. The nonreactive hydronium model was also used to describe the diabatic states in the full FM-MS-EVB model, the parametrization of which is discussed in the next subsection.

2.4. Force Matching the Reactive MS-EVB Model. The FM-MS-EVB model developed here uses the state search algorithm and underlying functional forms of the recently developed MS-EVB3 model for proton transport in water.²⁴ The functions taken from the MS-EVB3 model describe the

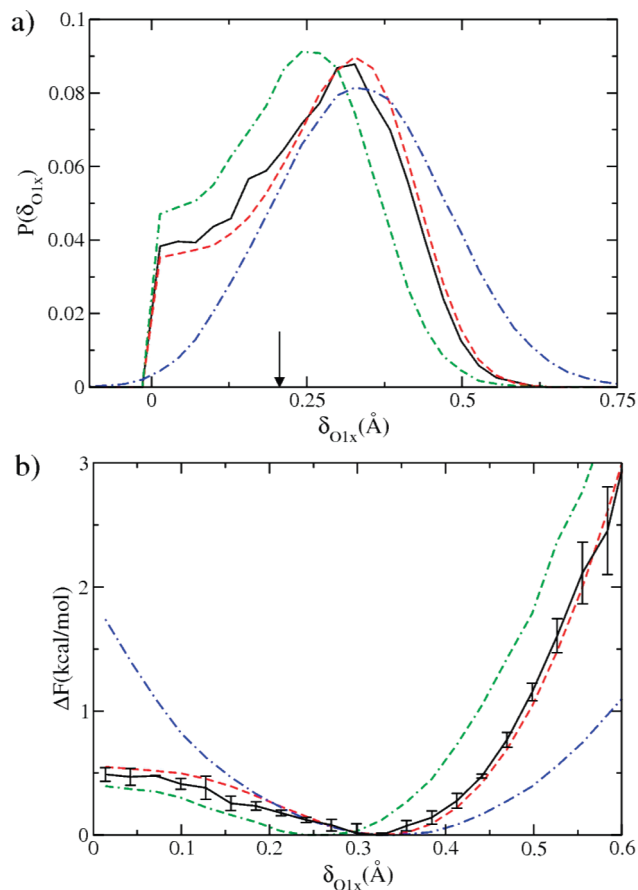


Figure 3. Probability (a) and potential of mean force (b) along the δ_{O1x} coordinate. The solid black lines were calculated from the CPMD trajectories. The dashed red lines are from the FM-MS-EVB model; the dashed-dotted blue lines are from the nonreactive hydronium model, and the double-dashed-dotted green lines are from the MS-EVB3 model. Configurations with $\delta_{\text{O1x}} > 0.2$, as indicated by the arrow, were included in the force matching calculation to determine the hydronium–water pairwise interactions for the nonreactive model. For the CPMD potential of mean force, the error bars indicate one standard deviation. Typical standard deviations for the FM-MS-EVB curve are 0.005–0.01 kcal/mol.

off-diagonal coupling of states which are overall attractive and promote successful chemical reactions. To compensate an overestimated attraction at short interatomic separations and prevent unphysical configurations from being sampled, the underlying nonreactive model is supplemented with a set of short-ranged repulsive interactions for the description of the diabatic states in the MS-EVB simulations. Using forces from the ab initio trajectories, a new set of MS-EVB3 model parameters are determined via the force matching algorithm. To ensure a sufficient sampling of all possible excess proton solvation structures, 4000 configurations were randomly sampled from the ab initio trajectories with weights determined from the δ_{O1x} probability distribution (Figure 3a). A residual, χ^2 , was then constructed as a function of the forces on all atoms in the system,

$$\chi^2 = \frac{1}{3N_{\text{C}}N_{\text{A}}} \sum_{j=1}^{N_{\text{C}}} \sum_{i=1}^{N_{\text{A}}} w(r_{ij}) |\mathbf{F}_{ij} - \mathbf{F}_{ij}^{\text{CPMD}}|^2 \quad (5)$$

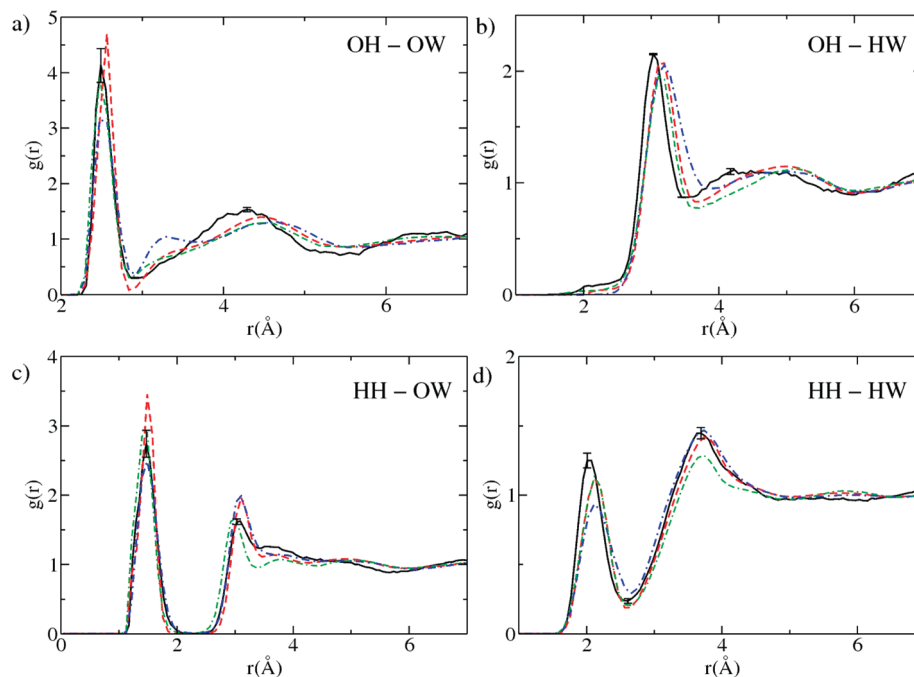


Figure 4. Radial distribution functions between atoms of the hydronium cation and water molecules. In each plot, the solid black line is calculated from the CPMD trajectories, the dashed red line is obtained from the FM-MS-EVB model, the dashed-dotted blue line is for the nonreactive hydronium model, and the double-dashed-dotted green line is for the MS-EVB3 model. Error bars for the first two peaks of the CPMD curves indicate one standard deviation.

where N_C is the number of configurations, N_A is the number of atoms, $\mathbf{F}_{ij}^{\text{CPMD}}$ are the atomic forces from the CPMD simulation, and \mathbf{F}_{ij} are the FM-MS-EVB model forces for a given set of model parameters. In eq 5, a simple weighting function, $w(r)$, was used to ensure that those atoms proximal to the excess proton carried the largest weight in the calculated residual:

$$w(r) = \begin{cases} 1 & r \leq r_w \\ e^{-\alpha(r-r_w)} & r > r_w \end{cases} \quad (6)$$

where $r_w = 3.0 \text{ \AA}$ and $\alpha = 1.54 \text{ \AA}^{-1}$. The parameters for the weighting function were chosen on the basis of the peak positions in the radial distribution functions (RDFs) calculated from the ab initio trajectories, shown in Figure 4. Those atoms within the first solvation shell of the excess proton have full weight, and those atoms beyond the second solvation shell have reduced weights. The use of the weighting function prevented the atoms far from the excess proton, where the main contribution to the atomic forces is due to the underlying water model, from contributing significantly to the residual. It should be noted that the MS-EVB parameter determination from the minimization of the residual in eq 5 is a nonlinear optimization, which is different from the linear algorithm described previously.^{30,31}

In total, the reactive portion of the FM-MS-EVB model to be parametrized requires the determination of 24 parameters. The Nelder–Mead downhill simplex algorithm⁴⁸ was used for the simultaneous optimization of all model parameters to minimize the nonlinear residual in eq 5. Twenty sets of optimizations were carried out, and the initial solution vectors for each one were chosen by randomly sampling parameter values within specified ranges. These ranges, along

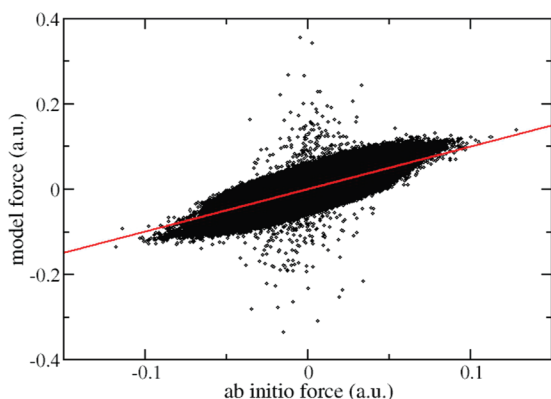
with quadratic restraints, were employed to prevent the sampling of wildly unphysical solutions. For example, these restraints prevented significant sampling of a positive partial charge on oxygen atoms and exceptionally large values ($>10 \text{ \AA}$) for parameters that corresponded to nearest-neighbor oxygen–oxygen distances. Each of these separate optimizations were continued as necessary until the termination criterion (relative difference between residual of best and worst solution vectors) was less than 10^{-4} . The final results from these calculations were then used as the starting point for a new series of calculations in which the initial solution vectors were chosen as random perturbations from the previous solution. This process was repeated until further optimization lead to no significant changes in the model parameters. For all models found in this manner, molecular dynamics simulations were used to calculate physical properties, such as RDFs and potentials of mean force along the δ_{O1x} coordinate, which were compared against results from the ab initio trajectories. From this final set of solutions, the best model was identified as the one that had the best agreement with the physical properties calculated from the CPMD trajectories. The parameters for this model are listed in Table 2.

A comparison of the forces from the ab initio trajectories and those calculated with the FM-MS-EVB model is shown in Figure 5. For most data points, it can be seen that the deviations are a small fraction of the range of sampled forces. With this model, the average root-mean-squared-deviation (rmsd) for all components of the total forces on all atoms was 0.02 au. The maximum deviation in any component of a total force was 0.36 au with 0.3% of forces having deviations larger than 0.05 au. Of all the outliers in Figure 5, with deviations larger than 0.05 au, only 10% of those

Table 2. Optimized Force Field Parameters for the FM-MS-EVB Model Obtained by Minimizing the Residual in Eq 5^a

B	3.3732 kcal/mol	γ	0.3628 Å ⁻²
b	8.3494 Å ⁻¹	P	1.1066
d_{OO}^0	2.1165 Å	k	7.5707 Å ⁻²
b'	8.1908 Å ⁻²	D_{OO}	2.7128 Å
C	11.8746 kcal/mol	β	12.4834 Å ⁻¹
c	9.6546 Å ⁻¹	R_{OO}^0	3.1366 Å
d_{OH}^0	0.4073 Å	P'	4.2352
$V_{\text{const}}^{\text{ij}}$	-11.1268 kcal/mol	α	13.7322 Å ⁻¹
q_{O}^{ex}	-0.0500 e	r_{OO}^0	2.0545 Å
q_{H}^{ex}	0.0167 e		
$q_{\text{H}^+}^{\text{ex}}$	0.0332 e		

^a The switching ranges for the smooth-cutoff function for the V_{OO}^0 and V_{HO}^0 short-ranged repulsions are 2.414–2.610 and 1.335–1.536 Å, respectively. The parameters and corresponding equations are defined in the original MS-EVB3 publication.²⁴

**Figure 5.** Comparison of total forces from the CPMD simulation calculated with the MS-EVB model for all atoms. A line of slope unity is shown to indicate where points would lie for perfect agreement. As discussed in the text, the majority of outliers are associated with water molecules beyond the first solvation shell of the excess proton.

involve components of an atomic force on the excess proton or atoms within the first solvation shell.

3. Results and Discussion

The FM-MS-EVB model was used to generate ten independent trajectories for 3 ns each in the constant NVE ensemble. The equations of motion for these FM-MS-EVB simulations were integrated with a time step of 0.5 fs in a modified version of the DL_POLY (v2) software package.⁴⁰ The MS-EVB3 state search algorithm was used. The average temperature over all simulations was 300.6 ± 1.4 K. The energy drift observed in the FM-MS-EVB simulations was 11.4 ± 2.0 kcal/mol per ns. This is smaller than the drift in the MS-EVB2 model but three times larger than the drift observed in the MS-EVB3 model, 13.1 and 3.4 kcal/mol per ns, respectively.²⁴ The increase in the drift of total energy as compared to the MS-EVB3 model is due in part to the hydronium–water pairwise interaction potentials which extend beyond the second solvation shell (unlike in the MS-EVB3 model) and the MS-EVB3 state search algorithm. An improved state search algorithm (work currently in progress) would need to be insensitive to the range of interaction

Table 3. Positions (Å) and Heights of the First Maximum and Minimum for the OH–OW and OH–HW Hydronium–Water RDFs^a

OH–OW	r_{max}	g_{max}	r_{min}	g_{min}	n
CPMD	2.49	4.16	2.91	0.30	3.1
FM nonreactive	2.48	3.15	2.93	0.36	3.0
FM-MS-EVB	2.57	4.68	2.84	0.09	3.0
MS-EVB3	2.48	3.87	2.84	0.31	3.0
OH–HW	r_{max}	g_{max}	r_{min}	g_{min}	n
CPMD	3.02	2.12	3.47	0.87	9.4
FM nonreactive	3.20	2.05	3.92	0.95	15.6
FM-MS-EVB	3.20	2.06	3.65	0.83	11.2
MS-EVB3	3.11	1.95	3.74	0.77	11.4

^a Integrated coordination numbers (n) for the first peak are also given.

potentials used in the model and the addition/removal of MS-EVB states over the course of a simulation to eliminate any drift in the total energy and ensure a proper sampling of ensemble distributions. As discussed below, good agreement was found when comparing results with the CPMD simulations for a number of equilibrium and dynamic properties. A comparison of the calculated properties for the CPMD, FM-MS-EVB, and MS-EVB3 models is discussed next.

The hydronium–water RDFs calculated with both the force-matched nonreactive and FM-MS-EVB models are shown in Figure 4. For all curves, there is general agreement with respect to the peak positions and heights between the CPMD simulations and the force-matched models. Deviations can be seen in the OH–OW RDF, Figure 4a, where there is a small, enhanced depletion, near 2.9 Å, between the first and second peaks in the FM-MS-EVB model compared to the CPMD results. This depletion is partially a result of waters accepting H-bonds from the hydronium that are somewhat too strong (short). The contribution to the RDF from these waters, particularly the waters labeled “O1z”, is shifted closer to the first peak. This shift in density has the effect of increasing the height of the first peak and shifting the peak maximum toward larger separations, ~ 0.1 Å, compared to the RDF calculated from the CPMD simulation (Table 3). This effect can also be seen in the first peak of the HH–OW RDF (Figure 4c) where again there is a slight shift of density toward shorter distances, increasing the height of the first peak. The integrated coordination number for the first peak of the OH–OW RDF for CPMD and the force-matched models are listed in Table 3. In line with the small shifts in density, the coordination numbers calculated from the CPMD and FM-MS-EVB simulations agree quite well with one another, 3.1 and 3.0, respectively. The coordination number calculated from the MS-EVB3 model also similarly agrees with the CPMD simulation.

The depletion between the first two peaks in the OH–OW RDF also has a contribution from waters that donate weak H-bonds to the hydronium being repelled too strongly. A signature for this effect is the presence of a peak/shoulder at separations just longer than 3 Å. This effect is much more pronounced, however, in the nonreactive hydronium model than the FM-MS-EVB model. In the OH–HW RDF, more evidence for water molecules donating H-bonds somewhat

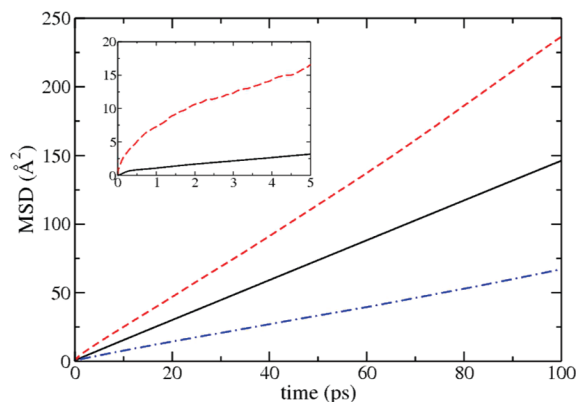


Figure 6. Mean squared displacements as a function of time for water (solid black line) and hydronium in the FM-MS-EVB (dashed red line) and nonreactive models (dashed-dotted blue line). The inset shows data from the longer CPMD trajectory for water (solid black line) and hydronium (dashed red line).

too weakly to the hydronium can be seen in the shoulder region near 2.0 Å just before the first major peak, as well as in the shift of the first peak toward larger separations. The FM-MS-EVB model recovers some of the density in this shoulder region which integrates to ~ 0.1 in the CPMD simulations. This shoulder is absent in simulations with the nonreactive hydronium model. The main contribution to the OH–HW RDF from waters that donate a H-bond to the hydronium accumulates at larger separations, ~ 3 Å, which gives rise to a slight shift of the first peak and first minimum toward larger separations. Again, the effect is more pronounced in the nonreactive hydronium model. The shifting of density in the OH–HW RDF compared to the CPMD simulations has a slightly larger effect on the integrated coordination numbers for the first peak. The coordination number calculated from the CPMD simulation was 9.4. The value calculated from the FM-MS-EVB and MS-EVB3 models, 11.2 and 11.4, respectively, are larger than the CPMD result, consistent with a shifting of the first peak maximum and minimum toward larger separations. The increased coordination number of the MS-EVB models as compared to CPMD is partly due to those waters donating H-bonds to hydronium and those waters that do not directly accept a H-bond from hydronium.

The peak positions and intensities for the other hydronium–water RDFs calculated from the FM-MS-EVB model similarly match well when compared to those calculated from the CPMD simulations. For comparisons between CPMD and the FM-MS-EVB model in the second solvation shell and beyond, it should be noted that the underlying water model from the CPMD simulation was replaced with the SPC/Fw model. This can make an exact comparison of the RDFs far from the hydronium cation difficult. Both the FM-MS-EVB and MS-EVB3 models use the SPC/Fw water model, and it can be seen in Figure 4 that all RDFs track one another very well in these two models beyond the second solvation shell, >5 Å.

The mean-squared displacements as a function of time for both water and hydronium are shown in Figure 6 for the force-matched nonreactive model, FM-MS-EVB model, and

the CPMD simulations (inset). For water, the calculated self-diffusion coefficient of 0.22 ± 0.01 Å²/ps from the FM-MS-EVB simulations agrees with the previously reported value for the SPC/Fw⁴⁵ model (0.23 ± 0.05 Å²/ps) and the experimental value, 0.23 Å²/ps.⁴⁹ The water self-diffusion coefficient calculated from the CPMD trajectories is 0.1 Å²/ps, in agreement with previously published results.^{36,50} The hydronium self-diffusion coefficient for the nonreactive model, which is solely vehicular in character (no Grotthuss shuttling), is 0.11 ± 0.01 Å²/ps, which is a factor of 2 smaller than the diffusion of the underlying SPC/Fw water model. The net excess proton diffusion constant (including the effect of Grotthuss shuttling) calculated from the FM-MS-EVB and the longer CPMD trajectory is 0.35 ± 0.06 and 0.31 Å²/ps, respectively. The same diffusion coefficient calculated from the MS-EVB3 model, 0.29 ± 0.03 Å²/ps,²⁴ is smaller than the value from both the FM-MS-EVB and CPMD simulations. The CPMD value reported here is larger than the previously reported value of 0.21 Å²/ps for the HCTH/120 density functional⁵⁰ that was calculated from a simulation containing 64 water molecules. All of the calculated excess proton diffusion coefficients are smaller than the experimental value of 0.94 Å²/ps,⁵¹ but it should be noted that the present simulations are classical and, thus, have no effects included from the quantization of the nuclear motion.²⁴

The proton hopping rate was calculated as the slope of the forward proton hop accumulation function.²⁴ The rate calculated from the FM-MS-EVB model was 0.12 ± 0.01 ps⁻¹. This value is somewhat smaller than that estimated from the CPMD simulations, 0.4 ± 0.2 ps⁻¹; however, the limited statistics from the CPMD trajectories makes it difficult to determine this value with sufficient accuracy. The proton hopping rate obtained from the MS-EVB3 model, 0.108 ± 0.009 ps⁻¹, is a little smaller than that calculated with the FM-MS-EVB model. The smaller proton hopping rate in the MS-EVB3 model is consistent with a smaller excess proton diffusion as compared with the FM-MS-EVB model.

The probability distribution and potential of mean force along the δ_{O1x} coordinate for the CPMD, MS-EVB3, and FM-MS-EVB simulations are shown in Figure 3. It can be seen that the curves generated from the FM-MS-EVB simulations agree quite well with the CPMD results over the whole range of sampled δ_{O1x} values. The barrier for proton transfer along this coordinate for the CPMD simulation is about 0.5 kcal/mol. The barrier for the FM-MS-EVB model is only slightly higher by 0.05 kcal/mol. The potential of mean force curve for the nonreactive model is also included for comparison. The equilibrium value for δ_{O1x} for the nonreactive model, the FM-MS-EVB model, and CPMD are all near 0.33 Å. The potential of mean force calculated from the MS-EVB3 model does not agree as well with that calculated from the CPMD simulations. The barrier for proton transfer is lower by 0.1 kcal/mol, and the minimum for the well is shifted toward shorter distances by 0.08 Å compared to the CPMD result. This result especially reflects the ability of the FM-MS-EVB methodology to improve the agreement with AIMD data over that obtained with the standard MS-EVB approach alone.

4. Conclusions

In this work, a reactive MS-MD force field was successfully parametrized using data from AIMD simulations as input to a force matching algorithm. The force matching algorithm also provided the freedom to only parametrize select portions of the model to the ab initio data, mainly the hydronium–water interactions and proton hopping portions of the model. The SPC/Fw water model, which agrees well with the experiment for a number of properties, was used in place of also parametrizing a water model based on the CPMD simulations that have been shown to be less accurate.⁴⁴ The resulting reactive FM-MS-EVB model then enabled the simulation of the excess hydrated proton in water for several nanoseconds, something not currently possible with AIMD simulation techniques, and it was found to reproduce a number of thermodynamic and dynamic properties calculated from the original CPMD trajectories. A comparison of the FM-MS-EVB model with the previously empirically parametrized MS-EVB3 model was also made where calculated properties were found to qualitatively agree with each other. The excess proton self-diffusion coefficient and proton hopping rate were found to be slightly larger in the FM-MS-EVB model compared to the values calculated from the MS-EVB3 model. The improved agreement of the FM-MS-EVB model, based on parametrization from condensed phase AIMD simulations, over the MS-EVB3 model suggests that the optimal parameters for the particular set of functions used in the MS-EVB3 model can be identified via the force matching approach. As more accurate AIMD methods become tractable for the simulation of condensed phase systems, the force matching algorithm presented in this work will facilitate the generation of empirical MS-MD models that can accurately reproduce the computationally expensive ab initio simulations, thus allowing for reactive MD simulations of much larger systems and for significantly longer time scales.

A key goal of future work will be to increase the flexibility of the force matching procedure discussed here and its application to complex reactions in the condensed phase. Effort toward this goal will be directed at the incorporation of more flexible functional forms for the interactions in the MS-MD models such as interpolated functions to describe the reactive portion of the model. The use of these interpolated functions would replace the use of functional forms that may be nonlinear with respect to the model parameters and, therefore, difficult to generalize to describe other chemical reactions. For example, in the work presented here, these interpolated functions would replace the empirical functions taken from the MS-EVB3 functional form, which constrain the range of accuracy of the resulting FM-MS-EVB model. The optimization of a nonlinear function of several variables can also be a difficult task due to the possible presence of many unphysical local minima and poor convergence properties. By constructing a model where the atomic forces are linear (or close too) with respect to all model parameters, one will be able to take advantage of linear least-squares algorithms^{30,31} to systematically and variationally determine those parameters that best represent the training set of data used to develop a reactive model.

Acknowledgment. This research was supported by the National Science Foundation (CHE-1036464). This work was supported in part by a grant of computer time from the DOD High Performance Computing Modernization Program at the Navy and Army Research Laboratory DOD Supercomputing Resource Centers. This research was also supported in part by the National Science Foundation Teragrid computing resources provided by the Texas Advanced Computing Center under Grant Number TG-MCA94P017. The authors thank Dr. Jessica Swanson for helpful discussions.

References

- (1) Grothuss, C. J. T. d. *Ann. Chim. (Paris)* **1806**, 58, 54.
- (2) Agmon, N. *Chem. Phys. Lett.* **1995**, 244, 456.
- (3) Tuckerman, M.; Laasonen, K.; Sprik, M.; Parrinello, M. *J. Chem. Phys.* **1995**, 103 (1), 150.
- (4) Marx, D.; Tuckerman, M. E.; Hütter, J.; Parrinello, M. *Nature* **1999**, 397, 601.
- (5) Marx, D. *ChemPhysChem* **2006**, 7 (9), 1848.
- (6) Pisani, C.; Maschio, L.; Casassa, S.; Halo, M.; Schütz, M.; Usvyat, D. *J. Comput. Chem.* **2008**, 29 (13), 2113.
- (7) Marsman, M.; Grüneis, A.; Paier, J.; Kresse, G. *J. Chem. Phys.* **2009**, 130 (18), 184103.
- (8) Duin, A. C. T. v.; Dasgupta, S.; Lorant, F.; Goddard, W. A., III *J. Phys. Chem. A* **2001**, 105 (41), 9396.
- (9) Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, 48 (7), 1198.
- (10) Schmitt, U. W.; Voth, G. A. *J. Phys. Chem. B* **1998**, 102 (29), 5547.
- (11) Schmitt, U. W.; Voth, G. A. *J. Chem. Phys.* **1999**, 111 (20), 9361.
- (12) Day, T. J. F.; Soudackov, A. V.; Čuma, M.; Schmitt, U. W.; Voth, G. A. *J. Chem. Phys.* **2002**, 117 (12), 5839.
- (13) Voth, G. A. *Acc. Chem. Res.* **2006**, 39 (2), 143.
- (14) Swanson, J. M. J.; Maupin, C. M.; Chen, H.; Petersen, M. K.; Xu, J.; Wu, Y.; Voth, G. A. *J. Phys. Chem. B* **2007**, 111 (17), 4300.
- (15) Marx, D.; Chandra, A.; Tuckerman, M. E. *Chem. Rev.* **2010**, 110, 2174.
- (16) Warshel, A.; Weiss, R. M. *J. Am. Chem. Soc.* **1980**, 102, 6218.
- (17) Åqvist, J.; Warshel, A. *Chem. Rev.* **1993**, 93, 2523.
- (18) Warshel, A. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, 32, 425.
- (19) Warshel, A. In *Computer Modeling of Chemical Reactions in Enzymes and Solutions*; John Wiley and Sons: New York, NY, 1991; pp 1–236.
- (20) Guo, Y.; Thompson, D. L. *J. Chem. Phys.* **2003**, 118 (4), 1673.
- (21) Nelson, K. V.; Benjamin, I. *J. Phys. Chem. C* **2010**, 114 (2), 1154.
- (22) Kim, Y.; Corchado, J. C.; Villà, J.; Xing, J.; Truhlar, D. G. *J. Chem. Phys.* **2000**, 112 (6), 2718.
- (23) Higashi, M.; Truhlar, D. G. *JCTC* **2009**, 5 (11), 2925.
- (24) Wu, Y.; Chen, H.; Wang, F.; Paesani, F.; Voth, G. A. *J. Phys. Chem. B* **2008**, 112, 7146.

- (25) Brewer, M. L.; Schmitt, U. W.; Voth, G. A. *Biophys. J.* **2001**, *80* (4), 1691.
- (26) Maupin, C. M.; Wong, K. F.; Soudackov, A. V.; Kim, S.; Voth, G. A. *J. Phys. Chem. A* **2006**, *110*, 631.
- (27) Maupin, C. M.; McKenna, R.; Silverman, D. N.; Voth, G. A. *J. Am. Chem. Soc.* **2009**, *131* (22), 7598.
- (28) Ercolessi, F.; Adams, J. B. *Eruophys. Lett.* **1994**, *26* (8), 583.
- (29) Tangney, P.; Scandolo, S. *J. Chem. Phys.* **2002**, *117* (19), 8898.
- (30) Izvekov, S.; Parrinello, M.; Burnham, C. J.; Voth, G. A. *J. Chem. Phys.* **2004**, *120* (23), 10896.
- (31) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109* (14), 6573.
- (32) Marx, D.; Hutter, J. Ab initio molecular dynamics: Theory and Implementation. In *Modern Methods and Algorithms of Quantum Chemistry*; Grotendorst, J., Ed.; John von Neumann Institute for Computing (NIC); Forschungszentrum Jülich: Jülich, Germany, 2000; Vol. 1, pp 301–449.
- (33) CPMD, <http://www.cpmc.org>, Copyright IBM Corp 1990-2008, Copyright MPI für Festkörperforschung Stuttgart 1997-2001 (accessed March 1, 2002).
- (34) Boese, A. D.; Doltsinis, N. L.; Handy, N. C.; Sprik, M. *J. Chem. Phys.* **2000**, *112* (4), 1670.
- (35) Izvekov, S.; Voth, G. A. *J. Chem. Phys.* **2005**, *123*, 044505.
- (36) VandeVondele, J.; Mohamed, F.; Krack, M.; Hutter, J.; Sprik, M.; Parrinello, M. *J. Chem. Phys.* **2005**, *112*, 014515.
- (37) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43* (3), 1993.
- (38) Grossman, J. C.; Schwegler, E.; Draeger, E. W.; Gygi, F.; Galli, G. *J. Chem. Phys.* **2004**, *1120*, 300.
- (39) Maupin, C. M.; Aradi, B.; Voth, G. A. *J. Phys. Chem. B* **2010**, *114* (20), 6922.
- (40) Smith, W.; Forester, T. R. *J. Mol. Graphics* **1996**, *14* (3), 136.
- (41) LAMMPS, <http://lammps.sandia.gov> (accessed February 1, 2010).
- (42) Plimpton, S. J. *J. Comput. Phys.* **1995**, *117*, 1.
- (43) Nosé, S. *J. Chem. Phys.* **1984**, *81* (1), 511.
- (44) Yoo, S.; Zeng, X. C.; Xantheas, S. S. *J. Chem. Phys.* **2009**, *130* (22), 221102.
- (45) Wu, Y.; Tepper, H. L.; Voth, G. A. *J. Chem. Phys.* **2006**, *124* (2), 024503.
- (46) Markovitch, O.; Chen, H.; Izvekov, S.; Paesani, F.; Voth, G. A.; Agmon, N. *J. Phys. Chem. B* **2008**, *112* (31), 9456.
- (47) Frenkel, D.; Smit, B. In *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed.; Academic Press: San Diego, CA, 2002; pp 291–316.
- (48) Nelder, J. A.; Mead, R. *Comput. J.* **1965**, *7* (4), 308.
- (49) Krynicki, K.; Green, C. D.; Sawyer, D. W. *Faraday Discuss. Chem. Soc.* **1978**, *66*, 199.
- (50) Izvekov, S.; Voth, G. A. *J. Chem. Phys.* **2005**, *123* (4), 044505.
- (51) Roberts, N. K.; Northey, H. L. *J. Chem. Soc., Faraday Trans.* **1974**, *1* (70), 253.

CT1004438

Effect of Water Polarizability on the Properties of Solutions of Polyvalent Ions: Simulations of Aqueous Sodium Sulfate with Different Force Fields

Erik Wernersson* and Pavel Jungwirth

Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, and Center for Biomolecules and Complex Molecular Systems, Flemingovo nám. 2, 16610 Prague 6, Czech Republic

Received August 18, 2010

Abstract: We show that aqueous sodium sulfate solutions exhibit an unrealistically large degree of ion pairing and clustering when modeled using nonpolarizable force fields, with clusters resembling precipitate readily forming in a 0.5 m solution at ambient conditions. This aggregation behavior was found to be persistent in nonpolarizable water for a range of parameters of the sulfate anion. In contrast, a polarizable potential performs satisfactorily, producing a well dissolved salt with a degree of association that is consistent with activity data for real solutions. Most of this improvement is due to polarization of water molecules in the vicinity of the divalent sulfate anion, which enhances its solvation.

1. Introduction

For computer simulations to be a useful aid for understanding the role of ions in complex chemical and biological systems, it is essential that realistic force fields are used. A necessary condition for an ionic force field to be considered realistic is that it gives a correct account of the properties of simple salt solutions. We have observed that a commonly used set of sulfate parameters¹ together with the Smith–Dang sodium parameters² in SPC/E water lead to formation of clusters, reminiscent of crystallites, of sodium sulfate well below the experimental solubility limit of 2 m at 25 °C.³ Similar aggregation has been reported from molecular dynamics (MD) simulations of ammonium sulfate in SPC/E, TIP3P Ewald, and TIP3P-F water⁴ with a different set of sulfate parameters.⁵ This suggests that the problem of spurious clustering is not unique to a particular sulfate parametrization or salt. A polarizable version⁶ of the sulfate model of ref 1, neutralized by Smith–Dang sodium ions² in polarizable POL3 water, has been applied to the study of the interfacial behavior of sulfate.⁷ In that work, excessive cluster formation was not observed.

The fact that similar ionic force fields predict dramatically different aggregation behavior in nonpolarizable vs polariz-

able simulations suggests that polarizability may be important for proper description of interactions between sodium and sulfate ions in water. To investigate this issue in detail, we made a systematic comparison of different variants of the sulfate model proposed in ref 1 in combination with different sodium and water models. In order to enable comparison with experiments, we analyzed the simulations in terms of the Kirkwood–Buff (KB) theory.⁸ Within this theoretical framework, the integrals of the radial distribution functions, which are readily obtainable from simulations, can be related to the concentration derivatives of thermodynamic quantities such as chemical potentials and partial molar volumes. The deviation of the salt chemical potential from ideality gives a measure of the overall degree of association in the solution. Comparison with this quantity is, therefore, a suitable way to ascertain whether the association behavior seen for a given set of parameters is realistic.

2. Simulation Details

We performed MD simulations of sodium sulfate in aqueous solution. The simulation box contained 12 sulfate and 24 sodium ions together with 1395 water molecules, which yielded a solution with a concentration of 0.48 m. After 0.5 ns equilibration, the trajectories were propagated in 10 ns increments until the radial distribution functions were well

* To whom correspondence should be addressed; E-mail: erik.wernersson@uochb.cas.cz.

Table 1. Force Field Parameters

model	atom	σ (Å)	ϵ (kcal/mol)	q
sulfate 1	S	3.55	0.250	2.4
	O	3.15	0.250	-1.1
sulfate 2 ^a	S	3.55	0.250	2.0
	O	3.15	0.200	-1.0
sulfate 3	S	3.55	0.250	2.8
	O	3.15	0.200	-1.2
sulfate 4	S	3.55	0.250	1.6
	O	3.15	0.200	-0.9
sulfate 5	S	3.55	0.250	2.0
	O	3.213	0.200	-1.0
sulfate 6	S	3.55	0.250	2.0
	O	3.087	0.200	-1.0
sodium 1 ^b	Na	2.35	0.130	1.0
sodium 2	Na	2.73	0.100	1.0

^a Used in the polarizable simulations with an oxygen polarizability of 1.0 \AA^3 and a sulfur polarizability of zero. ^b Used in the polarizable simulations with a polarizability of 0.24 \AA^3 .

converged or until persistent clusters had formed. The cutoff for short-range interactions was set to 12.0 \AA . Long-range electrostatic interactions were accounted for with the use of the particle mesh Ewald method.⁹ In all polarizable simulations, the induced dipoles of all atoms were self-consistently converged at each time step. The temperature in all simulations was kept at 300 K using the Berendsen weak coupling algorithm, and the pressure was held constant at 1 atm using an analogous algorithm.¹⁰ The AMBER 10 molecular dynamics package was used for all calculations.¹¹

In this study, we focused on the effects of the details of the force field parametrization on the properties of sodium sulfate solutions. In ref 1, two alternative sets of sulfate parameters were suggested. Sulfate model 1 (using the same numbering as in ref 1) has charges of $-1.1 e$ placed on each oxygen whereas sulfate model 2 has on each oxygen a partial charge of $-1.0 e$. As sulfate model 2 is the one that forms the basis for the polarizable model from ref 6, it is adopted here as the reference sulfate model. To investigate the influence of the sulfate partial charges with fixed Lennard-Jones parameters, we additionally considered two variants of sulfate model 2, denoted as sulfate models 3 and 4, with oxygen partial charges of -0.9 and $-1.2 e$, respectively. Also, variations in the Lennard-Jones σ parameter of oxygen were considered. Sulfate models 5 and 6 have σ reduced and increased, respectively, by 2% compared to sulfate model 2. The Lennard-Jones parameters and partial charges for all sulfate models considered are summarized in Table 1. To assess the importance of the cation parameters, we have used two different models of the sodium cation. Sodium model 1 is the one by Smith and Dang,² while sodium model 2 is taken from ref 12; see Table 1.

Next, we included polarizability into the force field. In ref 6, the polarizability of sulfate was determined, on the basis of ab initio MD simulations, to be almost isotropic with a value of about 7.1 \AA^3 . Because most of the electron density from the frontier orbitals was located on the oxygens, the authors suggested that the total polarizability should be evenly divided between the four oxygen atoms. The resulting model with a polarizability of 1.775 \AA^3 on each oxygen atom could not be used directly as the iterative procedure used to

Table 2. Summary of the Different Combinations of Sulfate, Sodium, and Water Models Considered

run	water	sulfate model	sodium model	time (ns)	clusters ^a
1	SPC/E	1	1	20	y
2	SPC/E	2	1	30	y
3	SPC/E	3	1	20	y
4	SPC/E	4	1	10	y
5	SPC/E	5	1	10	y
6	SPC/E	6	1	20	y
7	SPC/E	2	2	30	n
8	TIP4P/2005	2	1	20	n
9	SPC/E	2 ^b	1 ^b	10	y
10	POL3	2 ^c	1 ^c	20	n
11	POL3	2	1	20	n
12	Dang-Chang	2	1	30	n

^a Refers to persistent clusters; some degree transient clustering was seen in all solutions. ^b Polarizable ions. ^c Nonpolarizable ions.

calculate the induced dipole moment diverged due to the so-called polarization catastrophe.¹³ Therefore, we used a reduced value of 1.0 \AA^3 for the polarizability of each of the oxygens in all sulfate models. For sodium, the polarizability of 0.24 \AA^3 was used.²

Four water models were considered, two polarizable and two nonpolarizable (see Table 2). The polarizable ones were the POL3¹⁴ and Dang-Chang¹⁵ models and the nonpolarizable ones were SPC/E¹⁶ and TIP4P/2005.¹⁷ Polarizable water models were combined with polarizable ions and vice versa, unless otherwise stated. The POL3 and SPC/E models were chosen since they have been previously used with the polarizable and nonpolarizable sulfate models, respectively. The two other models, which are more recent and refined, were chosen for comparison. In the POL3 model, the polarizability is partitioned between the oxygen and hydrogen atoms. In the Dang-Chang model, the only polarizable site is the auxiliary site on the H-O-H bisector.^{14,15} The models thus differ, among other things, in the details of how the polarizability is handled. Comparison between these two models is, therefore, useful for discerning whether the polarizability itself, as opposed to other differences between them, is important for the qualitative behavior of the system. In order to assess separately the influence of the polarizability of the ions and water, we also performed simulations of polarizable ions in SPC/E water and of nonpolarizable ions in POL3 water.

Additionally, in order to analyze in detail polarization of water in the hydration shell of sulfate, we simulated a small system with one sulfate ion (model 2) and 512 water molecules for each of the polarizable water models above. For technical reasons related to the calculation of the induced dipole moment in AMBER 10, these simulations were done at constant volume. The volume was determined as the average volume from a 0.5 ns test run at constant pressure. The system was equilibrated for 0.5 ns, and data were collected during the following 1 ns. Similar calculations in POL3 water were carried out for sodium (model 1) as well as for a test monovalent variant of sulfate model 2 where all partial charges were halved.

Table 3. Values of Γ Calculated from Experimental Activity Coefficients

m (mol/kg)	Γ
0.15	0.45
0.25	0.48
0.35	0.51
0.45	0.52
0.55	0.54
0.65	0.55
0.75	0.54
0.85	0.58
0.95	0.56

3. Kirkwood-Buff Analysis

The Kirkwood-Buff integrals are defined as⁸

$$G_{ij} = 4\pi \int_0^\infty r^2 dr (g_{ij}(r) - 1) \quad (1)$$

where $g_{ij}(r)$ is the radial distribution function for species i and j . The physical interpretation of G_{ij} is best made in terms of the product $c_j G_{ij} = N_{ij}$ where c_j is the molar concentration of species j . N_{ij} can be interpreted as the excess number of particles of species j in the vicinity of a particle i .¹⁸

For electrolytes, special considerations are necessary in order to extract thermodynamic information from the KB integrals.¹⁹ Since the concentrations of charged species are subject to the electroneutrality condition, they cannot be varied independently. This interdependency must be taken into account for KB theory to give meaningful results. For this reason, the expressions relating the KB integrals to thermodynamic properties are different for systems containing charged particles compared to those for systems with only neutral species. The concentration derivative of the chemical potential, μ , of a binary electrolyte is given by¹⁹

$$\frac{1}{k_B T} \left(\frac{\partial \mu}{\partial c} \right)_{T,p} = \frac{1}{c\Gamma} \quad (2)$$

with

$$\Gamma = c(G_{+-} - G_{+s}) = c(G_{+-} - G_{-s}) = \frac{N_{\mp\pm}}{\nu_{\pm}} - \frac{cN_{\mp s}}{c_s} \quad (3)$$

where c is the molar concentration and ν_{\pm} is the stoichiometric coefficient of the salt, i.e., $c_i = \nu_i c$ for the ions. The subscripts $+$, $-$, and s stand for cation, anion, and solvent, respectively. Below, we refer to Γ as “the binding parameter”. (The definition of Γ is formally similar to that for the specific binding parameter in ternary solutions given as eq 9 in ref 20, but the thermodynamic significance is different.) Because of the electroneutrality condition for the KB integrals,¹⁹ the expression for Γ is not unique; it can be rewritten in terms of G_{++} or G_{--} or in terms of a linear combination of G_{++} , G_{--} , and G_{+-} .

Γ in sodium sulfate was calculated using experimental mean activity coefficients and densities from ref 3. (The densities were required to make the conversion from molal to molar concentration scales.) The derivative in eq 2 was evaluated as a centered finite difference ratio. The results are summarized in Table 3.

As the Kirkwood–Buff theory is strictly valid only for systems that are open with respect to exchange of particles with their environments, there are some subtleties inherent in the evaluation of eq 1 from standard NpT MD simulations. Briefly, the main requirement for the calculation of the Kirkwood–Buff integrals with acceptable accuracy from simulation of a closed system is that the system size is large enough. This is met if the subsystem implicitly defined by the choice of cutoff for the integral in eq 2 approximates an open system, while the remainder of the simulation box plays the role of the environment.

In practice, we do not calculate Γ from $g_{ij}(r)$ but from the cumulative numbers $n_{ij}(r)$ of species j around species i according to²⁰

$$\Gamma(r) = \frac{1}{\nu_{\pm}} \left(n_{\mp\pm}(r) - \frac{n_{\pm} - n_{\mp\pm}(r)}{n_s - n_{\mp s}(r)} n_{\mp s}(r) \right) \quad (4)$$

where n_i is the total number of particles of species i in the simulation box and $\Gamma(r)$ is the estimate of Γ for cutoff distance r . We approximate Γ by $\Gamma(r)$ for a sufficiently large value of r (14 Å). Comparison with test simulations for a larger system indicate that the error incurred by this procedure is less than ten percent. As the purpose of the comparison with experiments is not to find exact values of the optimal parameters for sulfate but rather to sort out models which are qualitatively unrealistic, this degree of accuracy is sufficient.

In systems where there is precipitation, Γ has no obvious thermodynamic meaning. The radial distribution functions obtained from simulation of such systems are not applicable to either the solid or solution bulk phase because the simulation contains the interfacial region between them, which is disproportionately important for small systems. Therefore, Γ could not be calculated for most of the systems where persistent clusters were formed. In run 3, precipitation did occur only after many nanoseconds and converged radial distribution functions unambiguously belonging to the solution phase could be obtained. Even though the association seen in run 8 (TIP4P/2005 water) may indicate either precipitation or association in the solution phase, see below, we have calculated Γ under the provisional assumption that the latter is the case.

4. Results

4.1. Ion Association in Sodium Sulfate Solutions. For sulfate models 1 and 2 together with sodium model 1 in SPC/E water (runs 1 and 2) extended clusters, comprising the majority of the ions in the simulation box, formed after several nanoseconds, see Figure 1. This behavior was observed also for the polarizable version of sulfate model 2 together with sodium model 1 in SPC/E water (run 9). For sulfate models 3 and 6 together with sodium model 1 (runs 3 and 6), clusters did not appear until after ~ 15 ns, whereas for sulfate models 4 and 5 with the same sodium model (runs 4 and 5), clusters formed almost immediately. The clusters, once formed, were persistent, and ions in their interiors were stripped of their solvation shells. Thus, the clusters have the appearance of an incipient solid phase. The tentative conclu-

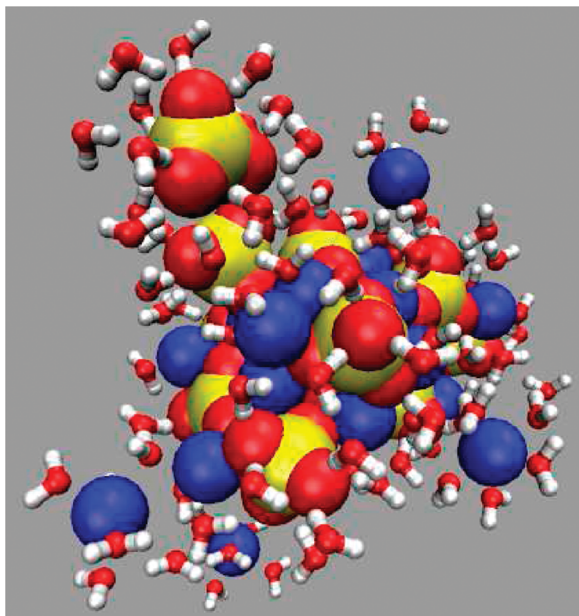


Figure 1. Snapshot from run 2 after 20 ns of simulation, showing the sodium and sulfate ions and all water molecules are within 3.0 Å of these.

sion is, therefore, that the cluster formation is due to precipitation, rather than association in the solution phase. We emphasize that further investigation is necessary to make a quantitative estimation of the solubility; the system size employed here is too small to draw quantitative conclusions about phase coexistence. Nevertheless, it is clear that almost complete association of the salt at 0.5 m concentration is not a realistic behavior for this electrolyte solution.

With the combination of sulfate model 2 and sodium model 2 (run 7), no persistent clusters were formed during the 30 ns simulation, although excessive ion pairing was still prevalent. For sulfate model 2 with sodium model 1 in TIP4P/2005 water (run 8), there was a large degree of

association, but the clusters that formed in this system were not persistent. The ions participating in the aggregates in this system remained solvated, and there was exchange of ions between the cluster and the bulk throughout the simulation. In this case, it is, therefore, much more difficult to judge whether the clusters represent incipient precipitation or aggregates in the solution phase. In nature, sodium sulfate can crystallize either as a decahydrate (mirabilite) or as an anhydride (thenardite), which are close to each other in free energy at room temperature (see ref 21 for a phase diagram). The fact that ions remain hydrated in the clusters is, therefore, not necessarily evidence against incipient precipitation but may imply that in TIP4P/2005 water a hydrated solid phase is favored. The issue could in principle be resolved using a methodology similar to that in ref 22, but this is beyond the scope of the current study.

In contrast to the above simulations, none of the systems with polarizable water models, including that with nonpolarizable ions in POL3 water (runs 10, 11, and 12), displayed any cluster formation resembling precipitation. The radial distribution functions for these runs are shown in Figure 2. As can be seen in this figure, the two polarizable water models show remarkably similar structures with respect to ion pairing. Even though the first peak in the radial distribution function, corresponding to contact ion pairs, is the highest, the cumulative numbers show that solvent separated ion pairs, corresponding to the second peak, are more abundant. Nonpolarizable ions in POL3 water (run 9) have an enhanced tendency to form contact ion pairs between sodium and sulfate compared to polarizable ions, but the ionic polarizability has little effect on the overall structure beyond the first peak. The sulfate–sulfate radial distribution functions display a solvent-separated peak at around 6 Å for both water models, with depletion for other separations. The sodium–sodium radial distribution function displays a very

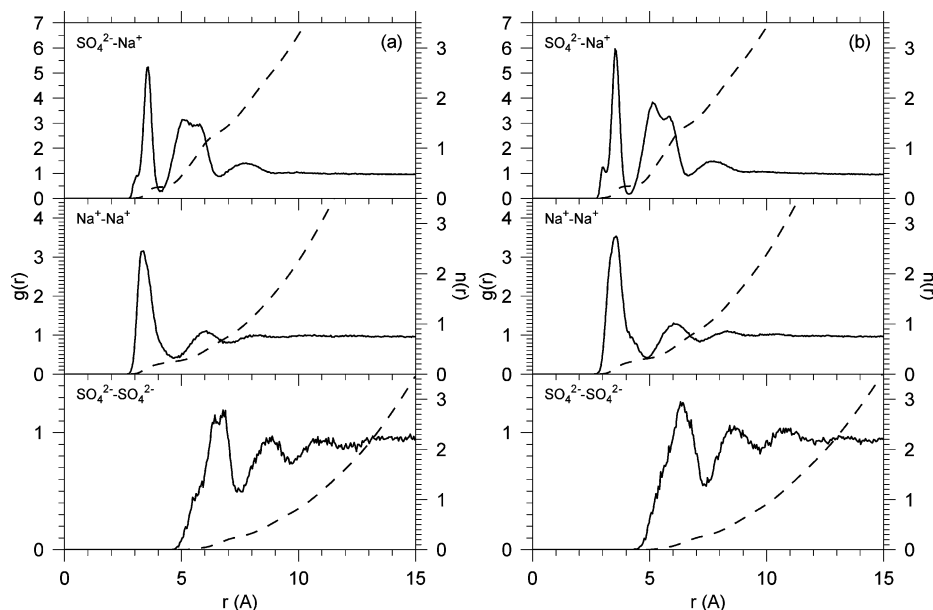


Figure 2. Radial distribution functions from runs 11 (subfigure a) and 12 (subfigure b), i.e., with POL3 and Dang–Chang water. The dashed lines are the cumulative numbers of ions (sodium ions for the $\text{SO}_4^{2-}\text{--Na}^+$ case).

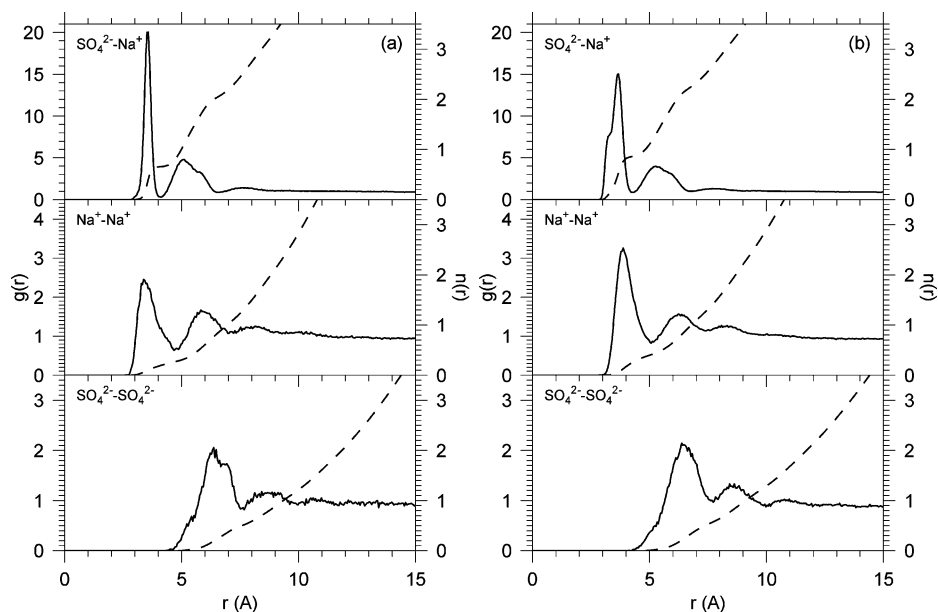


Figure 3. Radial distribution functions from the first 15 ns of run 3 (subfigure a) and run 7 (subfigure b), i.e., nonpolarizable runs with SPC/E water. The dashed lines are the cumulative numbers of ions (sodium ions for the $\text{SO}_4^{2-}-\text{Na}^+$ case).

pronounced peak at about 3 Å, which is due to simultaneous pairing of more than one sodium ion to a single sulfate ion.

The radial distribution functions for the two models with nonpolarizable water that remained unambiguously dissolved long enough for these to be calculated, i.e., sodium model 1 with sulfate model 3 (run 3) and sodium model 2 with sulfate model 2 (run 7), are shown in Figure 3. The sodium sulfate radial distribution functions for both of these models show a larger prevalence of both contact and solvent separated ion pairs than the models with polarizable water. Moreover, contact ion pairing is more strongly enhanced than solvent-separated pairing. The features of the sulfate–sulfate and sodium–sodium radial distribution functions are similar to those for the polarizable models, but for the nonpolarizable models, the values of the radial distribution functions are larger for all distances. This indicates a larger overall degree of association for the nonpolarizable models than for the polarizable ones. Although the overall structure is similar in the two nonpolarizable models, the shapes of the first peaks are somewhat different: for sulfate 3 with sodium 1, the peak is higher and narrower whereas for sulfate 2 with sodium 2 it is lower and has a pronounced shoulder toward smaller r .

The values of the binding parameter Γ for the different combinations of sodium, sulfate, and water parameters for which this quantity could be calculated are presented in Table 4. The experimental value of Γ for the concentration in the simulations of 0.48 m is 0.53. As can be seen from the table, all nonpolarizable models for which Γ could be evaluated, i.e., for which no persistent clusters formed, still predict values that are too large, by a factor two or more. The value for the TIP4P/2005 model of about 3 may reflect transient cluster formation and may, therefore, overestimate the association in the solution phase, as discussed above. Sulfate model 2 together with sodium model 1 in POL3 water agrees with the experimental value within the accuracy of the simulated values of Γ . Nonpolarizable ions in POL3 water and polarizable ions in Dang–Chang water give a bit higher

Table 4. Binding Parameter Γ Calculated According to Eq 4

water	sulfate model	sodium model	Γ
SPC/E	3	1	1.1
SPC/E	2	2	1.1
TIP4P/2005	2	1	3
POL3	2	1	0.54
POL3	2 ^a	1 ^a	0.71
Dang–Chang	2	1	0.72
experiment			0.53

^a Nonpolarizable ions.

value of Γ than the experiment and, thus, slightly overestimate the association in solution. The fact that Γ differs between POL3 and Dang–Chang water despite the observation that the radial distribution functions are similar, Figure 2, indicates that Γ is a very sensitive measure of the degree of association in the electrolyte.

4.2. Polarization of the Solvation Shell. The first solvation shell of sulfate in the simulations typically contained 11 to 12 water molecules, most commonly with three water molecules hydrogen bonding to each sulfate oxygen in a tetrahedral arrangement. A typical snapshot from the simulation of the solvent shell around a single sulfate ion in POL3 water is shown in Figure 4. The solvation structure did not depend significantly on the choice of water model.

To quantify the polarization in the vicinity of a sulfate ion, we calculated the average magnitude of the induced dipole moment of water molecules as a function of distance from the ion. In Figure 5, results are shown for sulfate model 2 in POL3 and Dang–Chang waters. In POL3 water, the average induced dipole moment in the water bulk is around 0.59 D, while in the first solvation shell of sulfate it reaches a value of 0.70 D. For the Dang–Chang water model, the average induced dipole moment in bulk is 0.88 D, reaching 0.99 D in the first solvation shell. The distance dependence of the deviation in induced dipole moment from the bulk value is almost identical for the two water models. The

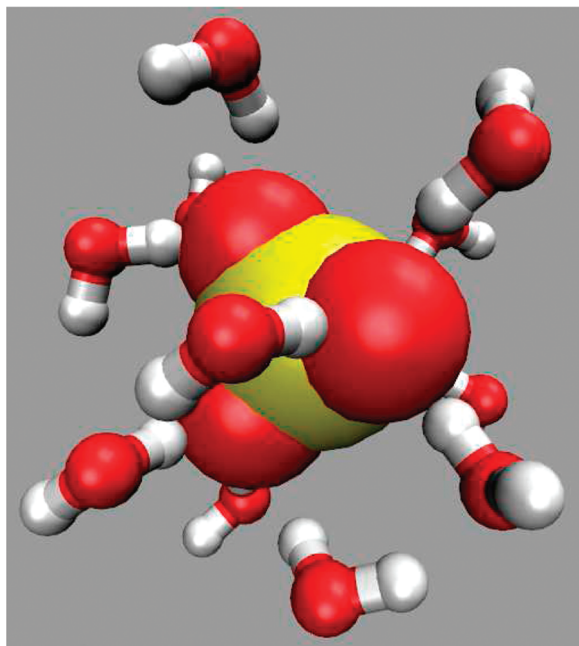


Figure 4. Snapshot from the simulation of one sulfate ion in water, showing the sulfate ion and all water molecules within 3.0 Å.

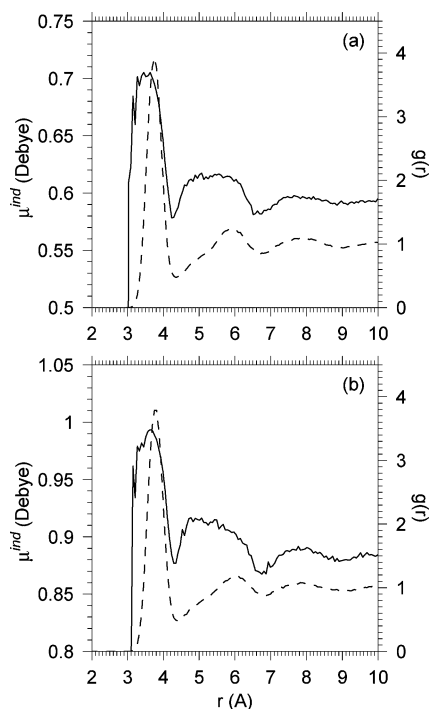


Figure 5. Average induced dipole moment of water molecules as a function of the center of mass distance from an ion (full curve). The ion–water center of mass radial distribution function is also shown (dashed curve) for reference. Panel (a) is for POL3 water and panel (b) is for Dang–Chang water.

structure of the induced dipole moment profile follows that of the radial distribution function with both the first and second solvation shells discernible. It is remarkable that the two water models agree to such an extent about the change in induced dipole moment caused by the presence of a sulfate ion, even though the bulk values differ by one-third.

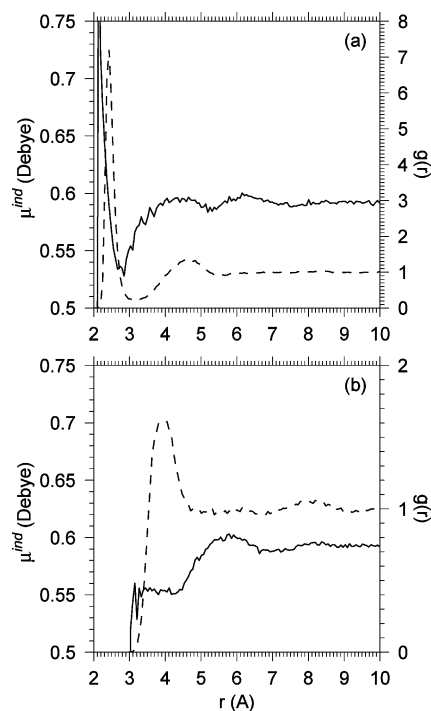


Figure 6. Same as Figure 5, but for sodium (panel a) and an ion model that is identical to sulfate model 2 except that all partial charges are halved (panel b).

For sodium, the polarization profile has a region close to the ion where the induced dipole moment is smaller than in bulk, but for very small sodium–water distances, the induced dipole moment becomes larger than in the bulk, Figure 6a. The polarization profile crosses the bulk value almost exactly at the position of the first maximum of the radial distribution function. The water molecules participating in the first solvation shell have, therefore, on average over the whole shell, an induced dipole moment similar to that of water molecules in bulk.

To separate the effect of ionic valency from geometric issues, we also carried out similar calculations for a monovalent version of sulfate model 2, where all partial charges are halved (Figure 6b). Thus, the electric field originating from the ion is halved, save for the contribution due to induced dipoles. For this model ion, the first peak in the induced dipole profile is absent. In fact, the first solvation shell is less polarized than bulk water.

5. Discussion

In the nonpolarizable simulations of sodium sulfate solutions, the values of the binding parameter Γ are significantly larger than in any of the polarizable ones. This indicates that polarizability strongly influences interactions in the solution phase as well as the apparent solubility. Since the binding parameter Γ is an integral quantity, the comparison of experimental and simulated values cannot constitute a definitive test of the quality of the description of the structural details of the electrolyte. The adequacy of a model with respect to finer structural details, in particular the relative abundances of contact and solvent-separated ion pairs, must be tested by other means. Raman spectroscopy does not give indication of a strong presence of contact ion pairs in aqueous

sodium sulfate solutions.²³ Similarly, dielectric relaxation spectra can be rationalized using the assumption that solvent separated and doubly solvent separated ion pairs are dominant over contact ion pairs.²⁴ Thus, also in structural aspects, the polarizable models, which predict predominance of solvent-separated ion pairs, are more realistic than the nonpolarizable ones, which yield extensive contact ion pairing and clustering.

Excessive association is found for all variations of the sulfate model examined in SPC/E water. An increase in the sodium ion σ can inhibit cluster formation on the time scale of the present simulations; however, the experimental binding parameter cannot be reached within realistic sizes for Na^+ . Moreover, the excessive association is not limited to sodium cations. Indeed, we have observed that polycationic peptides associate strongly with sulfate model 2 in SPC/E water, with association being reduced in POL3 water in the same way as for sodium sulfate.²⁵

From the differences between the simulations with polarizable and nonpolarizable water models and the similarity between the two polarizable models, it follows that partial suppression of contact ion pairing is primarily an effect of polarizability. A similar, albeit weaker, effect of polarizability was previously observed in suppressed contact ion pairing between cationic amino acid side chains and halide ions.²⁶ This behavior begs the explanation of the mechanism by which polarizability influences ion pairing.

If polarization of the first solvation shell of an ion is larger than the average polarization of bulk water, the polarization energy will give a stabilizing contribution to the free energy of solvation. It has been noted that inclusion of polarizability tends to make the free energy of solvation of both neutral and charged species more negative.²⁷ For charged solutes in water, however, this contribution accounts only for a small fraction of the total solvation free energy.

The free energy change associated with the replacement of a hydration water with a cation is sensitive to the marginal change in solvation energy due to solvent polarization since it is a result of a subtle balance between ion–ion and ion–water interactions. Replacement of a polarizable water molecule with a practically nonpolarizable sodium ion carries with itself a larger free energy cost than the corresponding replacement of a nonpolarizable water molecule. The free energy cost need not be large compared to the solvation free energy to have a significant effect on the overall behavior of the system; the difference between the radial distribution functions in Figures 2 and 3 corresponds to a difference in the minimum of the potential of mean force of about $1 k_{\text{B}}T$. As the hydration shell of sulfate is more polarized than the bulk water and therefore stabilized, the suppression of contact ion pairing in polarizable water models is consistent with this rationalization. Note that in other cases, where the solvation shell would be destabilized by polarization, contact ion pairing could be enhanced in polarizable water. In fact, the inclusion of polarizability in the force field has been found to cause an increase in contact ion pairing in model strontium chloride.²⁸ This illustrates that the effect of polarizability on interionic interactions is dependent on the details of the system.

It has recently been noted that certain combinations of commonly used models for the alkali metal cations and the halides display pathological behavior such as an unrealistically low solubility and excessive ion pairing, similar to what we have observed here.^{22,29–32} For the alkali halides, systematic parametrizations compatible with nonpolarizable water models have recently been successfully developed.^{22,33–35} Thus, in contrast to sulfate salt solutions, polarizability does not seem to play an essential role for the description of ion pairing in simple monovalent salt solutions. This difference can be reconciled by the fact that the strength of the field acting on, and thereby polarizing, the water molecules in the first solvation shell of divalent ions is stronger than that of monovalent ions.

6. Conclusions

We investigated different variants of nonpolarizable potential models of the sulfate dianion,¹ together with polarizable versions of that model⁶ and different models of water and the sodium cation. Nonpolarizable models consistently and significantly overestimate the degree of association in the system. For most of these models, there is precipitation of the salt already at relatively low concentrations and even when the salt is soluble there is excessive ion pairing. In contrast, for polarizable models, only a modest degree of association is observed, which is consistent with the activity coefficients of real sodium sulfate solutions. This effect is mainly due to polarization of water molecules; simulations with nonpolarizable ions in a polarizable water are more similar to the fully polarizable simulations than simulations with polarizable ions in a nonpolarizable water. While the degree of association is to some extent sensitive to the choice of Lennard–Jones parameters, we did not find a combination of parameters that would give a physically reasonable association behavior using a nonpolarizable water model. These results indicate that simulations of sulfate-containing systems using nonpolarizable water are liable to produce artifacts in the form of severely overestimated association.

Acknowledgment. Support from the Czech Ministry of Education (Grant LC 512), the Czech Science Foundation (Grant 203/08/0114), and the Academy of Sciences (Præmium Academie) is gratefully acknowledged.

References

- (1) Cannon, W. R.; Pettitt, B. M.; McCammon, J. A. *J. Phys. Chem.* **1994**, *98*, 6225–6230.
- (2) Smith, D. E.; Dang, L. X. *J. Chem. Phys.* **1994**, *100*, 3757–3766.
- (3) Lide, D. R., Ed. *Handbook of Chemistry and Physics (Internet Version)*, 90th ed.; CRC Press: Boca Raton, FL, 2010; Section 8, pp 112–117; Section 8, pp 52–77; Section 5, pp 79–80.
- (4) Cerutti, D. S.; Le Trong, I.; Stenkamp, R. E.; Lybrand, T. P. *Biochemistry* **2008**, *47*, 12065–12077.
- (5) Huige, C. J. M.; Altona, C. J. *Comput. Chem.* **1995**, *16*, 56–79.
- (6) Jungwirth, P.; Curtis, J. E.; Tobias, D. J. *Chem. Phys. Lett.* **2003**, *367*, 704–710.

- (7) Gopalakrishnan, S.; Jungwirth, P.; Tobias, D. J.; Allen, H. C. *J. Phys. Chem. B* **2005**, *109*, 8861–8872.
- (8) Kirkwood, J. G.; Buff, F. P. *J. Chem. Phys.* **1951**, *19*, 774–777.
- (9) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (10) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (11) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecker, T.; Gohlke, H.; Yang, L.; Tan, C.; Morgan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*; University of California: San Francisco, 2008.
- (12) Roselova, M.; Jungwirth, P.; Tobias, D. J.; Gerber, R. B. *J. Phys. Chem. B* **2003**, *107*, 12690–12699.
- (13) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341–350.
- (14) Caldwell, J. W.; Kollman, P. A. *J. Phys. Chem.* **1995**, *99*, 6208–6219.
- (15) Dang, L. X.; Chang, T.-M. *J. Chem. Phys.* **1997**, *106*, 8149–8159.
- (16) Berendsen, H. J. C.; Grigera, J. R.; Straasma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
- (17) Abascal, J. L. F.; Vega, C. *J. Chem. Phys.* **2005**, *123*, 234505.
- (18) Hall, D. G. *Trans. Faraday Soc.* **1971**, *67*, 2516–2524.
- (19) Kusalik, P. G.; Patey, G. N. *J. Chem. Phys.* **1987**, *86*, 5110–5116.
- (20) Pierce, V.; Kang, M.; Aburi, M.; Weerasinghe, S.; Smith, P. E. *Cell Biochem. Biophys.* **2008**, *50*, 1–22.
- (21) Flatt, R. J. *J. Cryst. Growth* **2002**, *242*, 435–454.
- (22) Joung, I. S.; Cheatham, T. E., III. *J. Phys. Chem. B* **2009**, *113*, 13279–13290.
- (23) Daly, F. P.; Brown, C. W.; Kester, D. R. *J. Phys. Chem.* **1972**, *76*, 3664–3667.
- (24) Buchner, R.; Capewell, S. G.; Hefter, G.; May, P. M. *J. Phys. Chem. B* **1999**, *103*, 1185–1192.
- (25) Wernersson, E.; Heyda, J.; Kubíčková, A.; Křifžek, T.; Coufal, P.; Jungwirth, P. *J. Phys. Chem. B* **2010**, *114*, 11934–11941.
- (26) Heyda, J.; Hrobárik, T.; Jungwirth, P. *J. Phys. Chem. A* **2009**, *113*, 1969–1975.
- (27) Geerke, D. P.; van Gunsteren, W. F. *J. Phys. Chem. B* **2007**, *111*, 6425–6436.
- (28) Smith, D. E.; Dang, L. X. *Chem. Phys. Lett.* **1994**, *230*, 209–214.
- (29) Auffinger, P.; Cheatham, T. E., III; Vaiana, A. C. *J. Chem. Theory Comput.* **2007**, *3*, 1851–1859.
- (30) Chen, A. A.; Pappu, R. V. *J. Phys. Chem. B* **2007**, *111*, 11884–11887.
- (31) Fennell, C. J.; Bizjak, A.; Vlachy, V.; Dill, K. A. *J. Phys. Chem. B* **2009**, *113*, 6782–6791.
- (32) Fennell, C. J.; Bizjak, A.; Vlachy, V.; Dill, K. A.; Sarupria, S.; Rajamani, S.; Garde, S. *J. Phys. Chem. B* **2009**, *113*, 14837–14838.
- (33) Horinek, D.; Mamatkulov, S. I.; Netz, R. R. *J. Chem. Phys.* **2009**, *130*, 124507.
- (34) Joung, I. S.; Cheatham, T. E., III. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.
- (35) Fyta, M.; Kalcher, I.; Dzubiella, J.; Vrbka, L.; Netz, R. R. *J. Chem. Phys.* **2010**, *132*, 024911.

CT100465G

Temperature Effects on Donor–Acceptor Couplings in Peptides. A Combined Quantum Mechanics and Molecular Dynamics Study

Frank H. Wallrapp,[†] Alexander A. Voityuk,^{*,‡,§} and Victor Guallar^{*,†,§}

*Barcelona Supercomputing Center, Nexus II Building, 08028 Barcelona, Spain,
Institute of Computational Chemistry, University of Girona, 17071 Girona, Spain, and
Institució Catalana de Recerca i Estudis Avançats, 08010 Barcelona, Spain*

Received March 17, 2010

Abstract: We report a quantum chemistry and molecular dynamics study on the temperature dependence of electronic coupling in two short model oligopeptides. Ten nanoseconds replica exchange molecular dynamics was performed on Trp–(Pro)3–Trp and Trp–(Pro)6–Trp peptides in the gas phase in combination with computation of the energy and electronic coupling for thermal hole transfer between Trp residues. The electron transfer parameters were estimated by using the semiempirical INDO/S method together with the charge fragment difference scheme. Conformational analysis of the derived trajectories revealed that the electronic coupling becomes temperature dependent when incorporating structural dynamics of the system. We demonstrate that Trp–(Pro)3–Trp, having only few degrees of freedom, results in relatively weak couplings at low and high temperature and a strong peak at 144 K, whereas the more flexible system Trp–(Pro)6–Trp shows monotonically decreased coupling. Only a few conformations with strong donor–acceptor couplings are shown to be crucial for the overall ET rates. Our results introduce the question whether the *T* dependence of ET coupling can also be found in large biological systems.

I. Introduction

Protein-mediated long-range electron transfer (ET) between separated donor (D) and acceptor (A) sites plays a major role in biochemistry.^{1–4} It is well established that direct electron transfer is roughly exponentially dependent on the donor–acceptor distance (d_{DA}),^{5,6} whereas there is still a lively discussion on different regimes for bridge-mediated electron transfer.^{7–11} In particular, structural fluctuations of the bridge have to be taken into account^{12–14} to properly describe conformational gating mechanisms.^{15–17} Here, the ET coupling of the system depends on the thermal population of few conformational states with high coupling values.

Several studies have been published on the temperature dependence of ET rates^{18–24} and electronic conductance through molecular wires.^{25–27} Eng et al. showed that the electronic coupling between bridge-mediated donor and acceptor depends on the temperature due to the different conformations available at different temperatures.^{23,24} If donor and acceptor have a strong coupling in the energetically most favorable conformations, an increase of temperature will result in a decrease of the average coupling as the system will explore conformations with lower coupling values. On the other hand, if in the energetically most favorable conformations the couplings are weak, a temperature increase will populate conformations of higher as well as lower coupling values. Such systems are predicted to be less sensitive to temperature. Oligo *p*-phenyleneethynylene and oligofluorene bridges exemplify the first and the second regimes, respectively.²⁴ There also exists a third regime, where the most favorable conformation has a weak coupling. In this case, the average coupling will increase with the

* Corresponding author e-mail: victor.guallar@bsc.es (V.G.), alexander.voityuk@icrea.es (A.A.V.).

[†] Barcelona Supercomputing Center.

[‡] University of Girona.

[§] Institució Catalana de Recerca i Estudis Avançats.

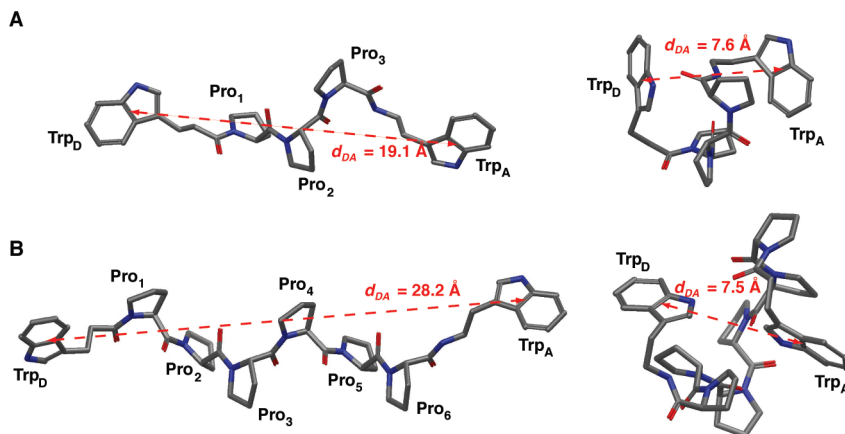


Figure 1. Oligopeptides Trp-(Pro)3-Trp and Trp-(Pro)6-Trp in panels A and B, respectively. Examples for extended and folded conformations are given for each system including their donor-acceptor distance.

temperature as demonstrated by fluorescence measurements on deoxytrinucleotides done by Jean et al.²² Additionally, there exist systems where the electronic coupling shows more complex temperature dependence, which is the case when the energy minimum does not correspond to a minimum (respectively maximum) of the electronic coupling. Examples therefore are theoretical electron transfer calculations on oligothiophene having maximal electronic coupling at about 25 K done by Eng and Albinsson²⁴ and experimental charge separation studies on 2-phenylenevinylenes with a maximal rate constant at about 210 K done by Davis et al.¹⁹ In general, we can say that electronic coupling becomes temperature dependent when structural dynamics is incorporated¹⁹ and the dependence regime strongly relies on the system itself.

In prior studies, the distance dependence of electronic coupling in oligopeptides has been considered at constant temperature.^{8,10,11} In the present work, we investigate the electronic coupling for thermal hole transfer in the positively charged Trp-(Pro)3-Trp and Trp-(Pro)6-Trp oligopeptides, at a wide range of temperatures. Within the hole transfer process, where a positive charge is transferred from the donor to the acceptor, the electronic coupling can be calculated applying the one-electron or Koopmans' theorem approximation.^{28,29} Here, the properties of the adiabatic states for a radical cation can be approximated through one-electron energies and occupied molecular orbitals of the corresponding neutral (close-shell) system.³⁰ There exist several studies comparing coupling values derived from semiempirical INDO/S to those derived from ab initio calculations (including CAS-PT2 as well as CASSCF conducted on DNA bases being analogues to tryptophans) showing that the INDO/S method provides good results for electronic couplings.^{6,10,31-33} Within the presented study, we estimate coupling values using semiempirical INDO/S calculations on a set of MD trajectories obtained at different temperatures. To overcome the effect of being trapped in local minima of the energy landscape at low temperature when applying classical MD, replica-exchange molecular dynamics (REMD) was employed.³⁴ This technique exchanges conformations from different temperatures with an acceptance criteria based on the energy difference. Hence, the system can faster explore more conformational space in comparison to classical MD, when describing ensembles of conformations. The derived results

show an unexpected temperature dependence of the electronic coupling with a maximum at 144 K for Trp-(Pro)3-Trp and a monotonic decrease of the coupling for Trp-(Pro)6-Trp.

II. Methods

The oligopeptides Trp-(Pro)3-Trp and Trp-(Pro)6-Trp, shown in Figure 1, were taken from our previous work.¹⁰ We used Impact³⁵ for the REMD calculations of 10 ns in the gas phase applying NVT with the OPLS2005 force field and a nonbonded cutoff of 12 Å. We chose a temperature range of 100–502 K with 16 temperature steps determined according to Patriksson et al.³⁶ and the exchange probability of 0.4 following Denschlag et al.³⁷ Snapshots were taken every single picosecond. We excluded the data from the first nanosecond of each trajectory due to equilibration reasons as well as trajectories with temperatures higher than 413 K included for deriving better REMD sampling.

To estimate the donor-acceptor electronic coupling for hole transfer in the polypeptides, we apply the Fragment Charge Difference method (FCD).^{30,38} Like the Generalized Mulliken-Hush (GMH) method,³⁹ FCD is based on the adiabatic-to-diabatic-state transformation. Within FCD, the adiabatic states are rotated to diabatic states to maximize the charge transferred between the fragments.^{30,38} The FCD method is general and can be applied to systems containing several redox centers.^{38,40} Recently, the scheme has been extended for treatment of electronic coupling for excitation energy transfer.⁴¹ The method provides accurate results when the adiabatic states are properly defined. Comparison of GMH and FCD methods shows that in most cases both methods give very similar results.^{30,38,42} To apply the Fragment Charge Difference method, one should explicitly define D and A sites involved in ET. Therefore, there is some limitation when the method is applied to a system where donor and acceptor sites are difficult to define (e.g., a π conjugated system consisting of several aromatic rings). In our case, however, D and A are defined straightforwardly (Trp residues are considered as redox sites of interest). On the other hand, GMH can also be applied for systems where more than two non-collinear sites are involved in ET.⁴²

Within the two-state model, the bridge-mediated electronic coupling can be calculated as:

$$V_{\text{DA}} = \frac{(E_2 - E_1)|\Delta q_{12}|}{\sqrt{(\Delta q_1 - \Delta q_2)^2 + 4\Delta q_{12}^2}} \quad (1)$$

Here, Δq_1 and Δq_2 are the donor–acceptor charges difference in the two adiabatic states with their respective energies E_1 and E_2 , and Δq_{12} is the corresponding off-diagonal term. For details on computation of Δq_1 , Δq_2 , and Δq_{12} , we refer to the original publication.³⁸ Using Koopmans' theorem,⁵ we estimate the adiabatic splitting $E_2 - E_1$ through the one-electron energies of the two highest occupied molecular orbitals (HOMO and HOMO–1) calculated for the closed-shell neutral system. According to our calculation, these MOs are almost completely localized on the donor and acceptor sites. As the measurement of the coupling within each trajectory, we use its root-mean-square (rms) V_{DA} ,

$$\text{rms } V_{\text{DA}} = \sqrt{\langle V_{\text{DA}}^2 \rangle} = \sqrt{\frac{1}{n} \sum_{i=1}^n V_{\text{DA},i}^2} \quad (2)$$

where $\langle \dots \rangle$ denotes the arithmetic mean and $V_{\text{DA},i}$ denotes the electronic coupling between donor and acceptor calculated on snapshot i of all n snapshots within a single trajectory. The use of rms V_{DA} rather than $\langle V_{\text{DA}} \rangle$ is in line with the expression for the nonadiabatic CT.^{1,5} In the following, we refer to direct coupling when computing V_{DA} for systems consisting of the donor and acceptor, and the bridge-mediated coupling is calculated for systems including the donor and acceptor sites and the bridging prolines (the whole system). The ET rate was estimated using the Marcus expression:¹

$$k_{\text{ET}} = \frac{2\pi}{\hbar} V_{\text{DA}}^2 \frac{1}{\sqrt{4\pi\lambda k_{\text{B}}T}} \exp\left(-\frac{(\lambda + \Delta G^\circ)^2}{4\lambda k_{\text{B}}T}\right) \quad (3)$$

where γ is Planck's constant, k_{B} is Boltzmann's constant, λ is the reorganization energy, T is the temperature, V_{DA} is electronic coupling, and ΔG° is the Gibbs free energy change of the electron transfer reaction.

In our previous study, we demonstrated that the averaged electronic couplings derived from semiempirical INDO/S⁴³ and more sophisticated Hartree–Fock calculations are almost identical.¹⁰ Hence, we believe that the derived electronic coupling is quite reliable, although we are aware of the approximate description of the hole transfer process in the systems.

III. Results and Discussion

System Flexibility. In a preliminary study, while applying classical MD at low temperatures we were facing the problem that the system kept trapping in a low-energy conformation depending on the starting structure. The mean couplings of those trajectories were not representative, as the system did not sample its conformational space sufficiently. As a consequence, we applied REMD, which can properly sample all conformations of the system. For illustration, we plotted the rmsd of Trp–(Pro)3–Trp for both trajectories, classical MD at 125 K (in black) and REMD at 128 K (in red), calculated against their respective average conformation,

shown in Figure 2. As seen, the system gets trapped in a single low energy conformation in the case of classical MD, while REMD generates more conformations. For low temperatures, the distribution of donor–acceptor distances (Figure 3) and the mean potential energy (shown in Figure S1) are converged after 5 ns of REMD. Nevertheless, our simulations were expanded to 10 ns.

We analyze here the REMD trajectory and coupling obtained for the neutral oligopeptide. Similar results are obtained when using an MD trajectory for the positively charged peptide (results shown in the Supporting Information).

The REMD trajectories of Trp–(Pro)3–Trp reveal a complex distribution of conformations over the investigated temperature range. We measured the donor–acceptor distance d_{DA} and plotted the distribution of each trajectory as well as the mean and the standard deviation in panel A of Figure 3. Furthermore, we computed the sum of the rmsd against a folded peptide and the rmsd against a fully extended peptide for each snapshot of the REMD. From this, we calculated the standard deviation of each trajectory to get a measure of the total fluctuation of the system at a given temperature. The plot is given in panel B of Figure 3.

Oligopeptide Trp–(Pro)3–Trp generally likes to be folded at low temperature, having a short mean d_{DA} . Also, it does hardly fluctuate as seen from the small standard deviation of d_{DA} . At $T = 128$ K, it begins to explore more conformational space, reaching conformations with shorter as well as longer d_{DA} . The fluctuations increase, reaching rmsd up to 3.4 Å at $T = 161$ K. Interestingly, at $T = 128$ K the peptide has a shorter average d_{DA} than at lower temperatures, reaching more folded conformations. Starting from $T = 249$ K, Trp–(Pro)3–Trp does not change its behavior anymore and keeps the mean d_{DA} of 14.5 Å with the standard deviation of 2.8 Å. The peptide has enough heat energy to cover the large conformational space and fluctuate strongly during the MD. The total fluctuation plot, given by panel B in Figure 3, closely follows the distributions of d_{DA} . We see very low fluctuations at low temperatures, a strong increase up to $T = 161$ K, stagnation within the temperature range of $T = 201$ – 306 K, followed by a significant increase of the fluctuations at very high temperatures.

Electronic Coupling. We computed the bridge mediated as well as direct couplings between the two tryptophans in Trp–(Pro)3–Trp. The data are given in the Supporting Information and shown in Figure 4.

In general, rms V_{DA} closely follows the change of d_{DA} except in the two regions of extreme temperature (see the following paragraphs). We have high V_{DA} values for low T and low V_{DA} for high T . It shows that the low T trajectories contain more high coupling conformations in comparison to the high temperature trajectories. This comes from the fact that low energy conformations of the oligopeptide are more folded and hence have shorter d_{DA} at low T . Furthermore, the underlying regime appears to be the direct DA electronic interaction as there is no increase in the coupling when including the orbitals of the bridging prolines.

At low temperature, the coupling values of Trp–(Pro)3–Trp show unexpected behavior: rms V_{DA} increases significantly

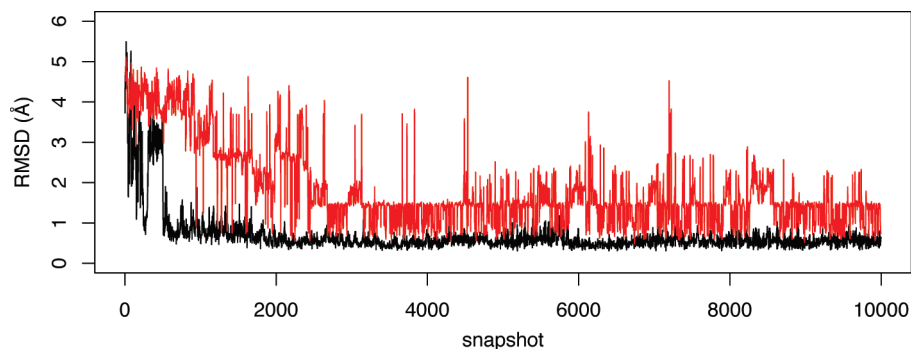


Figure 2. The rmsd of classical MD at 128 K (black) and REMD at 125 K (red) trajectories of Trp–(Pro)3–Trp.

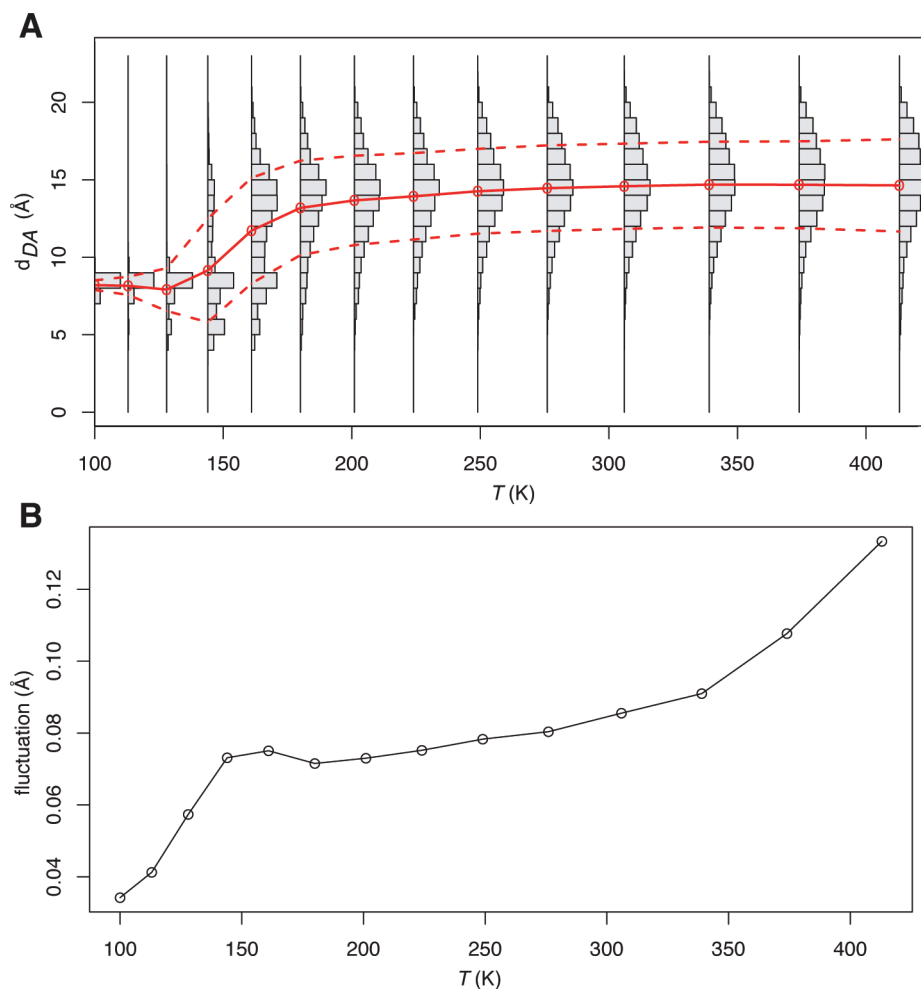


Figure 3. Panel A: Distribution of the donor–acceptor distance d_{DA} for REMD trajectories of Trp–(Pro)3–Trp. Solid and dashed red lines show the mean values and the standard deviations, respectively. Panel B: Total fluctuation of Trp–(Pro)3–Trp, given by the standard deviation within each trajectory.

from $T = 100$ to 144 K. We explain this by taking into account the fluctuation of d_{DA} at the different temperatures as shown in panel A of Figure 3. At very low temperatures (100–113 K), the peptide does not fluctuate much; the trajectories are composed of only a few conformations having very similar d_{DA} of about 8.2 Å. At slightly higher temperature (128 K), it begins to move within its limited conformational space and fluctuates only between structures of short d_{DA} . Thus, the resulting trajectories contain progressively more high-coupling conformations, increasing rms V_{DA} for $T = 128$ to 144 K. At $T = 144$ K, we see the peak of rms

V_{DA} . It matches perfectly with the distribution of d_{DA} , having higher mean d_{DA} values but, more importantly, containing also the highest fraction of short d_{DA} conformations. This discrepancy of rms V_{DA} expected from the mean conformation and actually measured rms V_{DA} can be described as conformational gating, where the electron transfer is triggered by a few strong-coupling conformations.⁴⁴

At $T > 161$ K, the peptide spends more time in extended conformations and has larger mean d_{DA} values. As a consequence, these trajectories have lower rms coupling values.

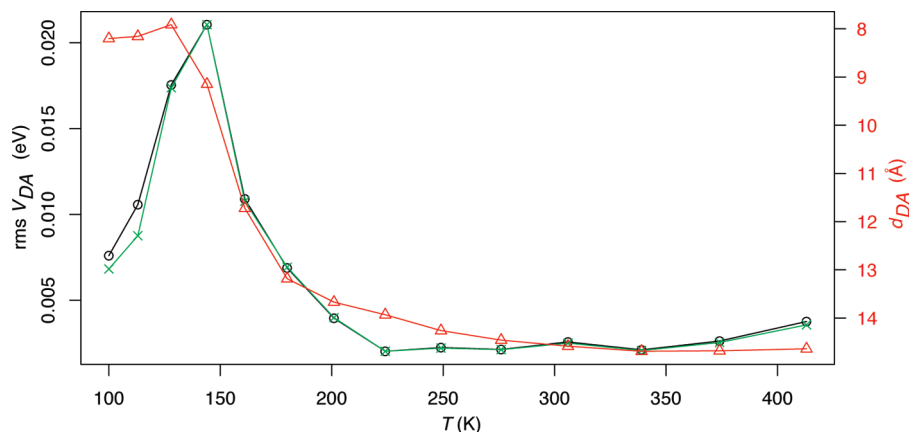


Figure 4. The rms V_{DA} plotted against temperature for Trp–(Pro)3–Trp. Color code: black, bridge-mediated coupling; green, direct coupling; and red, donor–acceptor distance d_{DA} .

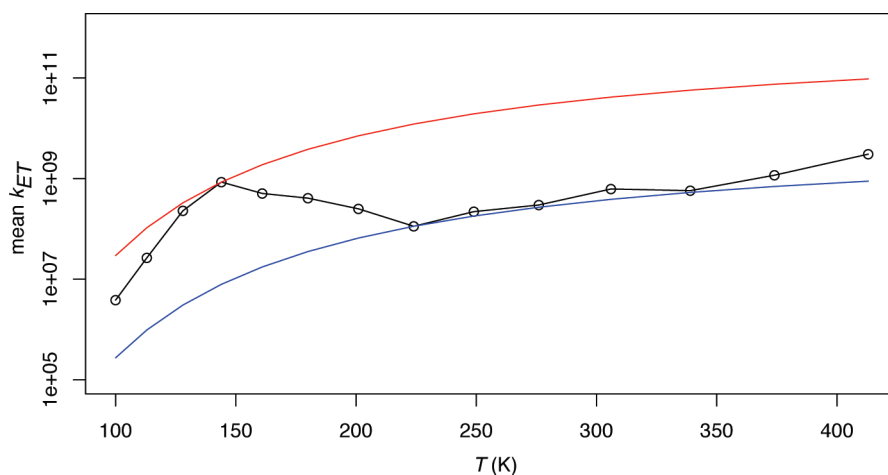


Figure 5. ET rates calculated for Trp–(Pro)3–Trp applying mean V_{DA} , $\lambda = 0.4$ eV, and $\Delta G^\circ = 0$ for each trajectory. Min k_{ET} (blue line) and max k_{ET} (red line) correspond to minimal and maximal mean values of V_{DA} .

Within the temperature range of 224–339 K, Trp–(Pro)3–Trp has enough energy to occupy nearly whole conformational space. Hence, the system contains significantly less low d_{DA} conformations when compared to the lower temperature trajectories. We see some increase of the coupling at high temperatures ranging from 376 to 413 K. This finding cannot be explained by the distance dependence of the coupling, as d_{DA} remains almost unchanged for these trajectories. For further insight, we point to the total fluctuation of the oligopeptide at these temperatures, shown in panel B of Figure 3. These high fluctuations overlay the pure d_{DA} dependence of the coupling and give the peptide the ability to acquire higher-coupling conformations, increasing the rms V_{DA} .

Rate Constants. We applied Marcus theory (eq 3) to compute the ET rate k_{ET} using the reorganization energy λ of 0.4 eV and ΔG° of zero due to identical donor and acceptor. The calculation of the reorganization energy was carried out at the B3LYP/6-31G* level. The internal reorganization energy λ_i in the gas phase was computed at the B3LYP/6-31G* level; for the Trp radical cation, the unrestricted Kohn–Sham method was applied. By definition, λ for an ET reaction is a sum of the reorganization energies of D and A. For charge transfer between identical donor and acceptor, $\lambda = 2\lambda$ (Trp). For molecule X ($X = \text{Trp}$), the

following terms were computed: (1) the energy of neutral X at the optimized geometry $E_0(X)$, (2) the energy $E_+(X^+)$ of the corresponding cation radical at the optimized geometry, (3) the energy $E_+(X)$ of neutral X calculated at the geometry of the anion radical X^+ , and (4) the energy $E_0(X^+)$ of the radical-cation state at the geometry of corresponding neutral molecule X. Thus, the reorganization energy $\lambda_i(X)$ becomes

$$\lambda = [E_+(X) - E_+(X^+) + E_0(X^+) - E_0(X)] \quad (4)$$

We estimated the mean k_{ET} by applying rms V_{DA} for each temperature trajectory of Trp–(Pro)3–Trp. Furthermore, we calculated the expected minimal and maximal rates when applying the lowest and highest coupling values, respectively. Figure 5 gives the logarithmic plot of the rates at different temperatures.

A strong increase of the rate from 100 to 144 K and then its slight decrease up to 224 K gives a peak at $T = 144$ K. At $T > 224$ K, as expected, $\log k_{ET}$ rises almost linearly with T .

Comparison between Short and Long Peptides. We also did the analysis of the temperature dependence of the coupling in the peptide having 6 proline residues instead of 3. Figure 6 shows the d_{DA} distribution, rms V_{DA} , and mean k_{ET} plots in panels A, B, and C, respectively. Trp–(Pro)6–Trp fluctuates significantly already at low temperatures (panel A). Furthermore,

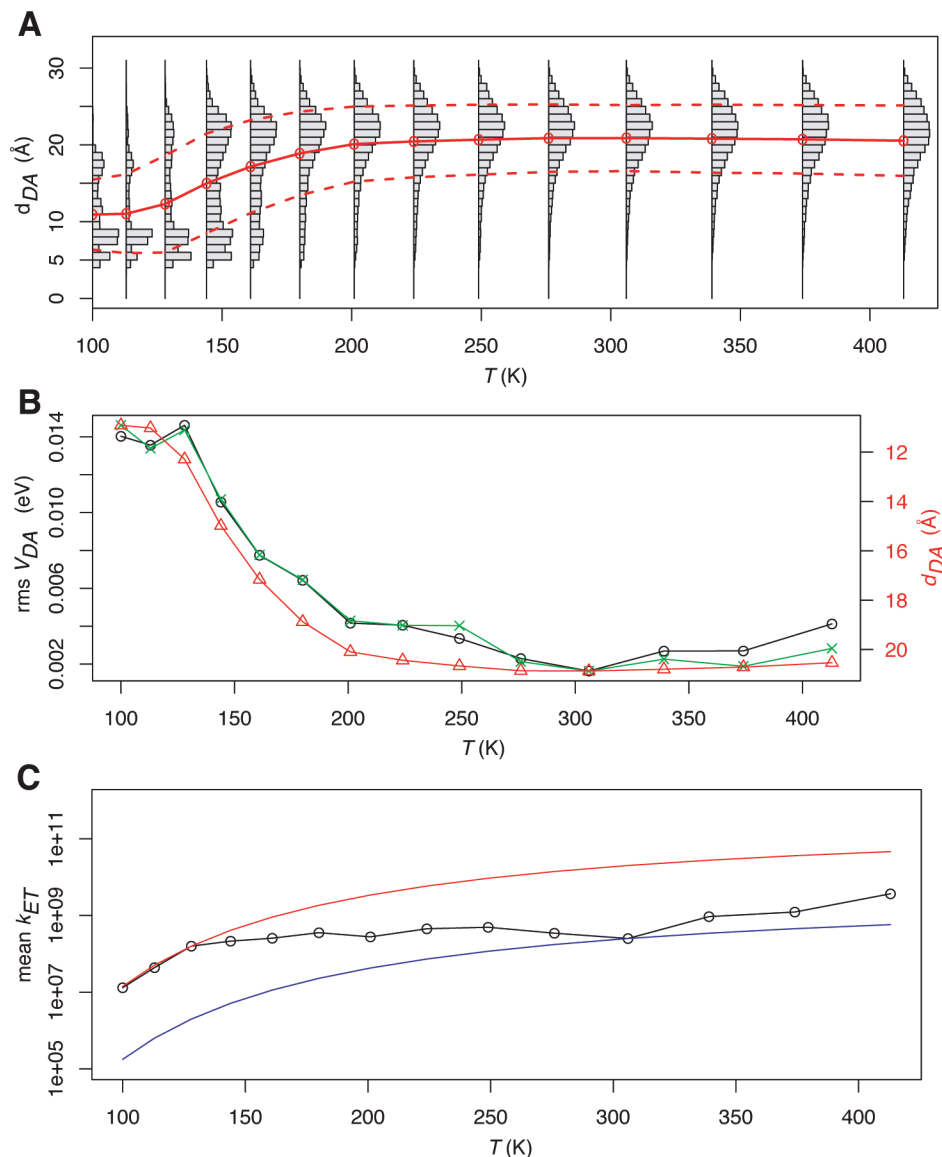


Figure 6. Results for Trp-(Pro)6-Trp. Panel A: Distribution of the donor-acceptor distance d_{DA} of each replica exchange trajectory. Solid and dashed red lines show the mean values as well as the standard deviations, respectively. Panel B: rms V_{DA} plotted against temperature. Color code: black, bridge-mediated coupling; green, direct coupling; and red, donor-acceptor distance d_{DA} . Panel C: Mean rate constant k_{ET} calculated with the rms V_{DA} , $\lambda = 0.4$ eV, and $\Delta G^\circ = 0$ for each trajectory. Min k_{ET} (blue line) and max k_{ET} (red line) correspond to minimal and maximal rms values of V_{DA} , respectively.

the system does spend most time in short d_{DA} conformations. The mean value of d_{DA} rises from about 11 Å at $T = 100$ K to about 20 Å at $T = 201$ K. At higher T , the system does not change its behavior anymore showing the normally distributed donor-acceptor distance. The rms V_{DA} as well as mean k_{ET} plots (panels B and C) show the usual temperature dependence of the coupling. We have an exponential decrease of the coupling with increasing temperature and relatively constant electron transfer rate. In trajectories higher than 300 K, we get some increase in the coupling and k_{ET} values.

We interpret the results by the following. Trp-(Pro)6-Trp has more degrees of freedom than does Trp-(Pro)3-Trp. At low temperature, each of these modes is restrained to a very little amplitude. For systems with only a few modes, these restraints prevent reaching conformations of very short d_{DA} , whereas more flexible systems are still capable of reaching low d_{DA} . As in our systems, mainly the donor-acceptor distance

gives rise to high coupling; the coupling follows closely the changes in d_{DA} , being weak for Trp-(Pro)3-Trp at very low temperature, strong around $T = 144$ K, and weak again at higher temperatures. In Trp-(Pro)6-Trp, a short d_{DA} is reached already at very low T and it increases with T , as shown in panel A of Figure 6. For completion, we also applied 10 ns REMD with a temperature range of 25–500 K on Trp-(Pro)6-Trp to study its d_{DA} at very low temperatures. The results show that Trp-(Pro)6-Trp has short d_{DA} already at 25 K. The plot of the d_{DA} distribution for $T = 25$ –418 K is given in Figure S3 of the Supporting Information.

IV. Conclusions

We have studied the temperature dependence of electronic coupling for thermal hole transfer in an oligopeptide with tryptophan-based donor and acceptor. The extensive replica

exchange molecular dynamics and conformational analysis have shown that a proper sampling at low T is very crucial in terms of electronic coupling and electron transfer rate calculations. We demonstrated that model systems with few degrees of freedom can show quite unexpected temperature-coupling dependence. For Trp-(Pro)3-Trp, there is a local peak of the rate constant at a temperature of about 144 K, whereas the more flexible peptide Trp-(Pro)6-Trp does not show such a feature. Furthermore, our results indicate that both considered oligopeptides undergo direct electron transfer as the inclusion of the bridging 3 respectively 6 prolines into the QM calculations does not increase the calculated coupling between the two tryptophans. Further characteristic therefore is the direct dependence of V_{DA} to d_{DA} of the applied oligopeptides as extensively discussed on our previous work.¹⁰

Previously, we showed for these model systems that solvent can influence the electronic coupling due to restriction of the thermally accessible conformational space.¹⁰ This may modulate the system behavior, reducing the effects discussed above. Nevertheless, our results raise the question of whether the temperature dependence similar to that we found for Trp-(Pro)3-Trp would also exist in biological systems where the protein matrix essentially affects the overall conformational dynamics.

Acknowledgment. This work was supported by a grant from the Spanish Ministry of Education and Science to V.G. through the project CTQ200762122 as well as to A.A.V. through the project CTQ2009-12346. Computational resources were provided by Barcelona Supercomputing Center and University of Girona.

Supporting Information Available: A plot of the average potential energy against the temperature for each trajectory of the REMD for Trp-(Pro)3-Trp, a plot of the d_{DA} distribution and the electronic coupling of the REMD on cationic Trp-(Pro)3-Trp, and the d_{DA} distribution for the REMD on Trp-(Pro)6-Trp with temperatures ranging from 25 to 500 K. Furthermore, two tables listing the mean d_{DA} , rms V_{DA} , and their fluctuations derived for both Trp-(Pro)3-Trp and Trp-(Pro)6-Trp. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Marcus, R. A.; Sutin, N. *Biochim. Biophys. Acta* **1985**, *811*, 265–322.
- Beratan, D. N.; Onuchic, J. N.; Winkler, J. R.; Gray, H. B. *Science* **1992**, *258*, 1740–1741.
- Gray, H. B.; Winkler, J. R. *Annu. Rev. Biochem.* **1996**, *65*, 537–561.
- Balzani, V.; Piotrowiak, P.; Rodgers, M. A. J.; Mattay, J.; Astruc, D.; Gray, H. B.; Fukuzumi, S.; Mallouk, T. E.; Haas, Y.; de Silva, A. P.; Gould, I. R. *Electron Transfer in Chemistry*; Wiley-VCH: Weinheim, Germany, 2001; Vols. I–V.
- Newton, M. D. *Chem. Rev.* **1991**, *91*, 767–792.
- Ungar, L. W.; Newton, M. D.; Voth, G. A. *J. Phys. Chem. B* **1999**, *103*, 7367–7382.
- Isied, S. S.; Ogawa, M. Y.; Wishart, J. F. *Chem. Rev.* **1992**, *92*, 381–394.
- Ogawa, M. Y.; Wishart, J. F.; Young, Z.; Miller, J. R.; Isied, S. S. *J. Phys. Chem.* **1993**, *97*, 11456–11463.
- Felts, A. K.; Pollard, W. T.; Friesner, R. A. *J. Phys. Chem.* **1995**, *99*, 2929–2940.
- Wallrapp, F.; Voityuk, A.; Guallar, V. *J. Chem. Theory Comput.* **2009**, *5*, 3312–3320.
- Malak, R. A.; Gao, Z.; Wishart, J. F.; Isied, S. S. *J. Am. Chem. Soc.* **2004**, *126*, 13888–13889.
- Kawatsu, T.; Kakitani, T.; Yamato, T. *J. Phys. Chem. B* **2002**, *106*, 11356–11366.
- Skourtis, S. S.; Balabin, I. A.; Kawatsu, T.; Beratan, D. N. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 3552–3557.
- Balabin, I. A.; Beratan, D. N.; Skourtis, S. S. *Phys. Rev. Lett.* **2008**, *101*, 158102–158104.
- Hoffman, B. M.; Ratner, M. A. *J. Am. Chem. Soc.* **1987**, *109*, 6237–6243.
- Balabin, I. A.; Onuchic, J. *Science* **2000**, *290*, 114–117.
- Hartings, M. R.; Kurnikov, I. V.; Dunn, A. R.; Winkler, J. R.; Gray, H. B.; Ratner, M. A. *Coord. Chem. Rev.* **2010**, *254*, 248–253.
- Read, I.; Napper, A.; Kaplan, R.; Zimmt, M. B.; Waldeck, D. H. *J. Am. Chem. Soc.* **1999**, *121*, 10976–10986.
- Davis, W. B.; Ratner, M. A.; Wasielewski, M. R. *J. Am. Chem. Soc.* **2001**, *123*, 7877–7886.
- Huppman, P.; Arlt, T.; Penzkofer, H.; Schmidt, S.; Bibikova, M.; Dohse, B.; Oesterhelt, D.; Wachtveit, J.; Zinth, W. *Biophys. J.* **2002**, *82*, 3186–3197.
- Napper, A. M.; Read, I.; Waldeck, D. H.; Kaplan, R. W.; Zimmt, M. B. *J. Phys. Chem. A* **2002**, *106*, 4784–4793.
- Jean, J. M.; Krueger, B. P. *J. Phys. Chem. B* **2006**, *110*, 2899–2909.
- Eng, M. P.; Martensson, J.; Albinsson, B. *Chem.-Eur. J.* **2008**, *14*, 2819–2826.
- Eng, M. P.; Albinsson, B. *Chem. Phys.* **2009**, *357*, 132–139.
- Selzer, Y.; Cabassi, M. A.; Mayer, T. S.; Allara, D. L. *J. Am. Chem. Soc.* **2004**, *126*, 4052–4053.
- Poot, M.; Osorio, E.; O'Neill, K.; Thijssen, J. M.; Vanmaekelbergh, D.; van Walree, C. A.; Jenneskens, L. W.; van der Zant, H. S. J. *Nano Lett.* **2006**, *6*, 1031–1035.
- Haiss, W.; Zalinge, H. v.; Bethell, D.; Ulstrup, J.; Schiffrin, D. J.; Nichols, R. J. *Faraday Discuss.* **2006**, *131*, 253–264.
- Liang, C.; Newton, M. D. *J. Phys. Chem.* **1992**, *96*, 2855–2866.
- Onuchic, J. N.; Beratan, D. N.; Hopfield, J. J. *J. Phys. Chem.* **1986**, *90*, 3707–3721.
- Rösch, N.; Voityuk, A. A. Quantum Chemical Calculation of Donor-Acceptor Coupling for Charge Transfer in DNA. *Long-Range Charge Transfer in DNA II*; 2004; pp 37–72.
- Prytkova, T. R.; Kurnikov, I. V.; Beratan, D. N. *J. Phys. Chem. B* **2005**, *109*, 1618–1625.
- Lambert, C.; Amthor, S.; Schelter, J. *J. Phys. Chem. A* **2004**, *108*, 6474–6486.
- Voityuk, A. *Chem. Phys. Lett.* **2006**, *427*, 177–180.

- (34) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- (35) *Impact*, 5.0 ed.; Schrödinger, LCC: New York, NY, 2008.
- (36) Patriksson, A.; Spoel, D. v. d. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2073–2077.
- (37) Denschlag, R.; Lingenheil, M.; Tavan, P. *Chem. Phys. Lett.* **2009**, *473*, 193–195.
- (38) Voityuk, A. A.; Rosch, N. *J. Chem. Phys.* **2002**, *117*, 5607–5616.
- (39) Cave, R. J.; Newton, M. D. *J. Chem. Phys.* **1997**, *106*, 9213–9226.
- (40) Voityuk, A. A. *J. Phys. Chem. B* **2005**, *109*, 17917–17921.
- (41) Hsu, C.-P.; You, Z.-Q.; Chen, H.-C. *J. Phys. Chem. C* **2008**, *112*, 1204–1212.
- (42) Subotnik, J. E.; Cave, R. J.; Steele, R. P.; Shenvi, N. *J. Chem. Phys.* **2009**, *130*, 234102–234102.
- (43) Ridley, J.; Zerner, M. *Theor. Chem. Acc.* **1973**, *32*, 111–134.
- (44) Berlin, Y. A.; Burin, A. L.; Siebbeles, L. D. A.; Ratner, M. A. *J. Phys. Chem. A* **2001**, *105*, 5666–5678.

CT100363E

Models To Approximate the Motions of Protein Loops

Aris Skliros, Robert L. Jernigan, and Andrzej Kloczkowski*

L. H. Baker Center for Bioinformatics and Biological Statistics, Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, Iowa 50011

Received March 17, 2010

Abstract: We approximate the loop motions of various proteins by using a coarse-grained model and the theory of rubberlike elasticity of polymer chains. The loops are considered as chains where only the first and the last residues thereof are tethered by their connections to the main structure, while, within the loop, the loop residues are connected only to their sequence neighbors. We applied these approximate models to five proteins. Our approximation shows that the loop motions can usually be computed locally which shows these motions are robust and not random. But most interestingly, the new method presented here can be used to compute the likely motions of loops that are missing in the structures.

Introduction

Coarse-grained elastic network models (ENM) have been extremely successful in predicting the large-scale motions of proteins, RNA, and other biological structures, even for the largest complexes such as the ribosome. The predicted fluctuations of the positions of amino acids in the coarse-grained representations usually give excellent agreement with the experimental *B*-factors reported by crystallographers,^{1–3} the ensembles reports by NMR scientists,⁴ and the variability in structures manifested in the known multiple structures of the same protein.^{4,5} The only information required in the ENMs is the structure of the protein, to furnish the coarse-grained coordinates of essential atoms, usually those of the of C^α atoms (but they could be other points representing the amino acids, such as centers of mass of side chains) for residue-level coarse graining. It has been shown that fluctuations of residues in proteins depend mostly on the protein's shape.⁶ Because of this, the ENMs give excellent results even for relatively low-resolution structural data, such as electron micrographs.

The problem of modeling the conformations of external protein loops is a really important problem. These are generally the most mobile parts of the protein localized on their surface and are the functional sites for many protein activities, particularly for encounter complexes and binding. Because of their relatively high mobilities they are often unresolved in the crystal structures, particularly for larger loops. Because of this, frequently the PDB coordinates for residues in loops are either missing or have alternative

positions. When obtaining an experimental structure, often the loops are the most uncertain parts of the structure.

Functions of biomolecules depend on their structures and are often exerted through functional motions. This makes understanding loop motions in proteins a particularly critical problem. Often interactions with other proteins or ligands can lead to apparent rearrangements of loops to accommodate functional ligands. For drug binding, the reconfiguration of target protein loops is relevant. Thus, the frequent involvement of external loops in function makes the prediction of their conformations essential for many structural applications in molecular biology, medicine, pharmacy, and drug design. The motions of loops computed by using ENMs have been intriguing, since often these move together with motions of the large-scale domains, either as rigid parts of domains or as separate parts moving in an anticorrelated way, but controlled by the domain motions. We have observed that the functionally meaningful loops most often appear to move under the control of the entire structure and its domain motions. Protein loops have been the focus of many previous studies. Modeling of loops has a significant role in making comparisons between protein structures,⁷ since these may be found in one structure and missing in the other. The field of structural genomics often requires building loop structures.⁸ Methods for the automated classification of the structures of protein loops have been developed.⁹ In principle, the study of loops can aid the understanding of protein evolution.¹⁰ Panchenko and Madej¹⁰ noted that protein loops are far from being random coils, regardless of their size.

Changes in the conformations of protein loops have also been a subject of some specific study.¹¹ The importance of protein loops for protein function has been widely acknowledged.¹² Conformations of loops play a large role in protein docking as has been pointed out in refs 13–16. The motion of protein loops, especially where they are flexible, is an important factor for understanding the various roles that proteins play. The Web site <https://simtk.org/home/looptk> provides a toolkit to model the kinematics of protein loops. In ref 17 a novel approach for loop prediction was presented and analyzed. Kolodny et al.¹⁸ describe an algorithm for generating conformations of candidate loops for a gap of a given size. A similar work also appeared in ref 19. Protein loops are also essential for protein folding.²⁰ Conformational evaluation of loops and their major role in protein design was discussed in ref 21. The importance of loop prediction was emphasized also in ref 22. In this paper, we devise a method for specifying how a protein loop can adopt different configurations. Our simple approximate model accounts only for the sequential connections within the loop and the loop's connections at the two ends, and this is a surprisingly successful model for generating loop forms. The issues of other interactions of the loop with the body of the protein are not explicitly taken into account in this work.

To overcome the difficulties with loop predictions, here we have applied the analytical theory of fluctuations in Gaussian Phantom Networks originally developed in the rubberlike elasticity theory of polymers by James and Guth^{23–26} and others.^{27–33} The theory assumes that polymer chains are phantomlike, i.e., they can pass freely through one another, so that excluded volume effects are to be completely neglected. It is also assumed that the distributions of the end-to-end vectors for polymer chains are Gaussian. This means that mechanically the network behaves as a collection of nodes (junctions) connected by simple Hookean springs and both chains and junctions fluctuate harmonically around their mean positions. Kloczkowski et al.²⁸ obtained analytical solutions for this model by assuming that all junctions in the network have the same connectivity ϕ (i.e., each junction is connected to ϕ other chains) and that a polymer network has the topology of an infinite tree. The theory provides analytical expressions for fluctuations of chains and junctions in such networks and for correlations of instantaneous fluctuations of two different points within the network.

The theory of phantom elastomeric networks was successfully adapted to treat protein motions originally as the Gaussian network model (GNM) by Bahar and Erman¹ and others.^{3,6,34–38} Their coarse-grained model was based on an earlier work of Tirion³⁹ who proposed that both nonbonded and covalently bonded atomic contacts in proteins could be modeled using a universal single spring constant in a harmonic analysis of protein dynamics. She assumed that two atoms are connected by a spring if they are separated by a distance smaller than a specified cutoff value. This defines a connectivity matrix for a system of nodes connected by springs. The coarse-grained GNM model enables computation of fluctuations of residues around their mean positions in protein structures directly from this connectivity matrix. Fluctuations of residues are simply expressed by the diagonal elements of the inverse of the connectivity

matrix. Theoretical predictions are usually in quite good agreement with crystallographic temperature factors (B -factors) that measure the extent of disorder in crystallographically determined positions of atoms resulting from thermal motions. Several variations of the elastic network approach to treat protein dynamics have been proposed recently^{4,35,36,40,41} that additionally improve agreement of theoretical results with B -factors.

In the present paper, we will apply analytical results from the theory of polymer networks obtained originally for tree-like networks to the external loops in proteins, and then we will compare these results with results from GNM computations (based on the known packing details within a protein structure). It is worthwhile mentioning that both the original theory of phantom polymer networks and the elastic network models of proteins are based on the assumption that excluded volume effects are completely negligible. The theory of phantom Gaussian networks, although developed for polymer networks with the topology of an ideal infinite tree, works well for real polymer networks that contain many loops. This means that the detailed topology of the network is really not so essential for studies of individual chains.

The theory of phantom Gaussian networks provides analytical expressions for fluctuations of chains and junctions in the polymer network having connectivity ϕ (where ϕ is the number of polymer chains connected at each junction), which is constant. However, since each end of the chain in an exterior loop of a protein can be connected to a different number of springs, the original theory²⁸ had to be modified to reflect having junctions at the two opposite ends of the loop with different functionalities ϕ_1 and ϕ_2 . We analytically compute the mean-square fluctuations and correlations of the instantaneous fluctuations for junctions and points along the polymer chains in such a treelike network,⁴² and here apply these analytical results to treat protein loops. The comparison of analytical predictions with the results of GNM computations for proteins with known crystallographic coordinates of loops overall show an excellent agreement. Our results demonstrate that it is possible to model theoretically the motions of protein loops using the Gaussian model from the polymer network without knowing the structural details of the loop itself.

The structure of the present paper is as follows. First, we describe briefly the Gaussian theory of random polymer networks and show how fluctuations of chains and junctions (cross-links) and covariances among them can be computed analytically for a network with an ideal treelike topology. In the next section we discuss the Gaussian network model (GNM) of proteins and its relationship to the earlier discussed theory of random polymer networks. Later, we compare the results from the GNM for several proteins with large external loops having known structures with the analytical results based on the theory of random polymer networks and the experimental B -factors. Other possible applications of our method for computing the structures of loops could be made to treat loops in nucleic acids, and other loops in large

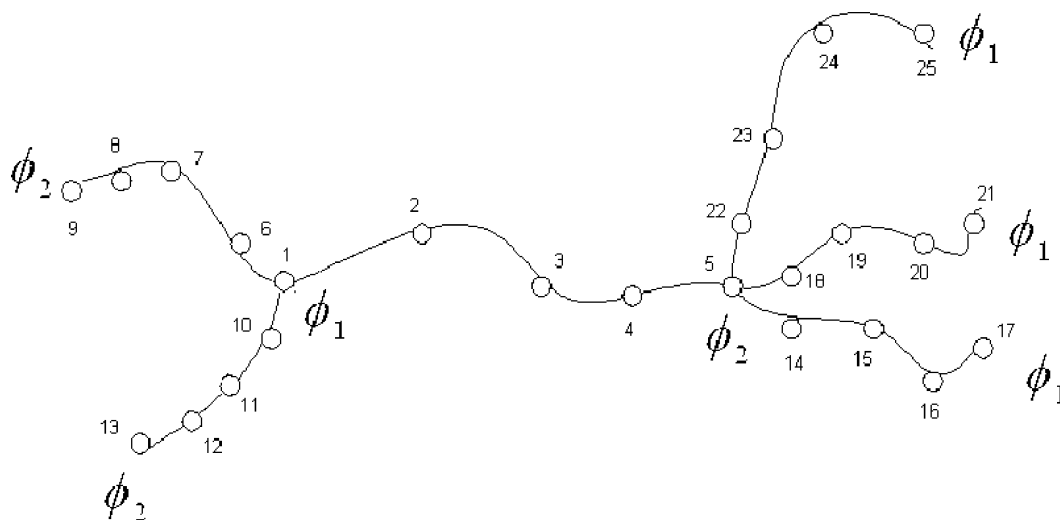


Figure 1. Treelike network with alternating functionalities separating each chain into four segments of equal length by three additional 2-functional junctions.

biological structures, such as the ribosome, for the prediction of protein function and for drug design.

Methods

Theory of Random Polymer Networks. The first theory of rubber elasticity was proposed by Kuhn in late 1930s.⁴³ The theory was further developed by Treloar.^{30,31} It was based on the assumptions that the rubber network consists on ν freely jointed Gaussian chains, which are cross-linked. It was also assumed that positions of the junctions (points of the chemical cross-links) deform affinely upon mechanical deformation of the rubber. The theory of phantom networks was developed in the 1940s by James and Guth.^{23–26,44} They also considered the network to be composed of cross-linked Gaussian chains. Additionally they assumed that there are two types of network junctions. Junctions which are at the surface of the rubber are fixed and deform affinely with the macroscopic strain, while the junctions inside the network are free to fluctuate around their mean positions. They assumed that the behavior of the network is determined only by the connectivities of network chains and neglected the effect of the excluded volume of the chains. The chains in their model are phantom-like; i.e., they may pass freely through one another.

Chain dimensions and fluctuations in random elastomeric networks were studied by Flory,²⁷ and by Kloczkowski, Mark, and Erman²⁸ who examined in detail the behavior of phantom Gaussian networks in the undeformed state. It can be shown that the mean-square fluctuations in position of a junction i $\langle\langle\Delta\mathbf{R}_i\rangle\rangle^2$ are related to the element Γ_{ii}^{-1} of the inverse of the connectivity matrix $\mathbf{\Gamma}$, and more generally the covariances in positions of points i and j $\langle\langle\Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j\rangle\rangle$ are related to Γ_{ij}^{-1}

$$\langle\langle\Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j\rangle\rangle = \frac{3\langle r^2 \rangle_0}{2} \Gamma_{ij}^{-1} \quad (1)$$

The elements of the inverse matrix have been calculated analytically for the network with the topology of an infinite tree, composed of chains of equal length (unimodal network), with equal mean square end-to-end distances $\langle r^2 \rangle_0$ in the

undeformed state. It is assumed that the network has functionality ϕ , i.e., that each free junction connects exactly ϕ chains.

Examples of unifunctional networks, recurrence relations between fluctuations of junctions in the neighboring tiers of the tree, recurrence relations between fluctuations of two junctions m and n separated by d other junctions along the path joining m and n , recurrence relations between fluctuations of points along the chains in the network and covariances of fluctuations among such points, recurrence relations between the elements of the inverse connectivity matrix $\mathbf{\Gamma}^{-1}$ are given by Kloczkowski et al. in ref 28 and presented briefly in Appendix A in the Supporting Information.

Because most models of real polymer networks use phantom network as a reference state for the construction of the real network models, these results are significant for rubber elasticity. The Gaussian network model has been also extended to proteins, as will be described later.

Theory of Random Polymer Networks with Alternating Functionality. The theory was developed in Skliros et al.^{42,45} and is presented briefly in Appendix B in the Supporting Information. To study fluctuations of points along the chain, we follow the method proposed in refs 28 and 46 and assume that all chains consist of n equal length segments and of $n - 1$ junctions of functionality 2, which connect these segments. Figure 1 illustrates this approach and the method of numbering all junctions for a treelike network with alternating multifunctional functionalities composed of two tiers.

Although we have obtained the most general solution of the problem when two points i and j can be separated by several multifunctional junctions,⁴² we show here only the results when these two points belong to the same chain; i.e., there are no multifunctional junctions between them. Additionally, since the network is assumed to be infinite, we will concentrate on the case for the central first tier shown in the center of Figure 1.

The positions of 2-functional junctions i and j can be expressed as the fraction of the chain between ϕ_1 -functional and ϕ_2 -functional junctions, counted from the closest ϕ_1 -functional junction on the left of points i or j in Figure 1; $\xi = (i - 1)/n$; $\theta = (j - 1)/n$.

The final result is

$$\left[\frac{\langle (\Delta R_i)^2 \rangle}{\langle \Delta \mathbf{R}_i \cdot \Delta \mathbf{R}_j \rangle} \quad \frac{\langle \Delta \mathbf{R}_i \cdot \Delta \mathbf{R}_j \rangle}{\langle (\Delta R_j)^2 \rangle} \right] = \frac{3n}{2\gamma_0} \times \left[\begin{array}{l} \frac{\phi_2(\phi_1 - 1)}{\phi_1(\phi_1\phi_2 - \phi_1 - \phi_2)} + \frac{\zeta(1 - \zeta)(\phi_1\phi_2 - \phi_1 - \phi_2) + \zeta(\phi_1 - \phi_2)}{\phi_1\phi_2} \quad \frac{\phi_2(\phi_1 - 1)}{\phi_1(\phi_1\phi_2 - \phi_1 - \phi_2)} + \frac{(\phi_1\phi_2 - \phi_1 - \phi_2)}{\phi_1\phi_2} [\min(\zeta, \theta) - \zeta\theta] + \frac{\min(\zeta, \theta) - \max(\zeta, \theta)}{\phi_2} \frac{\max(\zeta, \theta)}{\phi_1} \\ \frac{\phi_2(\phi_1 - 1)}{\phi_1(\phi_1\phi_2 - \phi_1 - \phi_2)} + \frac{(\phi_1\phi_2 - \phi_1 - \phi_2)}{\phi_1\phi_2} [\min(\zeta, \theta) - \zeta\theta] + \frac{\min(\zeta, \theta) - \max(\zeta, \theta)}{\phi_2} \frac{\max(\zeta, \theta)}{\phi_1} \quad \frac{\phi_2(\phi_1 - 1)}{\phi_1(\phi_1\phi_2 - \phi_1 - \phi_2)} + \frac{\theta(1 - \theta)(\phi_1\phi_2 - \phi_1 - \phi_2) + \theta(\phi_1 - \phi_2)}{\phi_1\phi_2} \end{array} \right] \quad (2)$$

Gaussian Network Model of Proteins. The Gaussian network model (GNM) was originally developed for the theory of rubberlike elasticity of random polymer networks^{27,28} to calculate fluctuations of junctions and chains inside the network. That physical situation is quite different from that prevailing in a protein because the polymer chains have random forms and the protein may be a more fixed form. The model has been adapted to coarse-grained proteins in 1997 by Bahar and Erman^{1,47} based on the earlier result of Tirion³⁹ with a single harmonic force parameter, which successfully described the large-scale motions in proteins.

The GNM is based on coarse-grained modeling of protein structure, with a single site per residue representing proteins. Positions of these sites are usually identified with the coordinates of C^α atoms in proteins, and it is assumed that both bonded and nonbonded contacts in protein structure are connected by uniform massless harmonic springs. Significantly, the atomic version gives only slightly better results than the coarse-grained model,³ indicating that the motions are mostly representative of the overall structure, and not so much of its details.

To define which sites are in contact, a uniform cutoff distance R_c is used.^{1,38,40,47} Residues separated by this distance or closer than R_c (including neighbors along the sequence) are assumed to be in contact and are connected with identical springs. This leads to the elastic network representation of a protein in the folded state that bears a resemblance to a random polymer network. While this model of a protein is closely similar to that of a rubbery network, the main difference is that in the rubber the coordinations are defined by covalent links whereas in the GNMs and ENMs the connections are primarily nonbonded contacts arising from close packing within the structure. While the GNM formally neglects the excluded volume, regions with a higher density of atoms are represented by higher density of springs, while less dense regions are represented by few springs.

The distance vector between the i th and j th sites is \mathbf{R}_{ij} , with $\Delta \mathbf{R}_{ij}$ being the instantaneous displacement of \mathbf{R}_{ij} from the mean value \mathbf{R}_{ij}^0 , and $\langle (\Delta \mathbf{R}_{ij})^2 \rangle$ is given by the scalar product $\langle \Delta \mathbf{R}_{ij}^T \cdot \Delta \mathbf{R}_{ij} \rangle$. The reference structure is usually the crystal structure taken from the Protein Data Bank (PDB), but could be a modeled structure or even the shape of a structure from an electron micrograph, which was filled with lattice points.⁴⁸ It can be shown^{27,28} that

$$\langle \Delta \mathbf{R}_i \cdot \Delta \mathbf{R}_j \rangle = \frac{3k_B T}{2\gamma} (\mathbf{\Gamma}^{-1})_{ij} \quad (3)$$

where $(\mathbf{\Gamma}^{-1})_{ij}$ is the ij th element of the inverse of the connectivity matrix $\mathbf{\Gamma}$.

It should be noted that the connectivity matrix $\mathbf{\Gamma}$ has been defined so that all elements in every row (or column) sum to zero. Because of this $\det \mathbf{\Gamma} = 0$, the matrix is singular, and only the pseudoinverse of $\mathbf{\Gamma}$ can be computed by the use of the singular value decomposition method. The pseudoinverse of $\mathbf{\Gamma}$ may be written as $\mathbf{\Gamma}^{-1} = \mathbf{U}(\mathbf{\Lambda}^{-1})\mathbf{U}^T$ where \mathbf{U} is the matrix composed of eigenvectors \mathbf{u}_i ($1 \leq i \leq N$) of $\mathbf{\Gamma}$, and $\mathbf{\Lambda}$ is the matrix having eigenvalues of $\mathbf{\Gamma}$ on the diagonal, and zeros off-diagonal. Additionally, it can be proven that all eigenvalues λ_i of $\mathbf{\Gamma}$ are nonnegative.

Mean-square fluctuations of each C^α computed from eq 3 can be compared with the Debye–Waller factors for the C^α atoms. These temperature factors are frequently measured by X-ray crystallography for all heavy atoms in the protein structure and are deposited in the Protein Data Bank as temperature B -factors. The computed B -factors for the i th residue are given by:

$$B_i = 8\pi^2 \langle (\Delta R_i)^2 \rangle / 3 \quad (4)$$

The B -factors computed by the GNM usually are in excellent agreement with experimental data,² although even better agreement is found when compared with the averages of internal distances from NMR ensembles.^{4,49}

The matrix $\mathbf{\Gamma}^{-1}$ can be written as the sum of contributions from individual modes⁵⁰

$$\mathbf{\Gamma}^{-1} = \sum_k \lambda_k^{-1} \mathbf{u}_k \mathbf{u}_k^T \quad (5)$$

where the zero eigenvalues (physically corresponding to motions of the center of mass of the system) are excluded from the sum. The i th component of the eigenvector \mathbf{u}_k (corresponding to the k th normal mode) specifies the magnitude of fluctuational motions of the i th residue in the protein exerted by the k th mode. If the eigenvalues are ordered according to an ascending order starting from zero, then the most meaningful contributions in eq 5 are given by the smallest nonzero eigenvalues λ_k , which correspond to the large-scale slow modes. The slowest modes play a dominant role in the fluctuational dynamics of structures, because their contributions to the mean-square fluctuations scale with λ_k^{-1} . It has been shown that the most essential motions of proteins^{51–53} or large biological structures such as the ribosome,^{54–57} which are associated with their biological function, are clearly identifiable within a few of the slowest modes of the GNM or ENM. The large-scale

changes of protein conformations between “open” and “closed” forms, or domain swapping in proteins, can be also explained well by these ENMs.^{58,59} The Gaussian network model is the simplest version of several different ENMs. It has been extended to treat anisotropic fluctuations with vector directions for the motions,⁴⁰ and hierarchical³⁵ or mixed³⁶ levels of coarse graining.

Results and Discussion

Prediction of Motions of Loops in Proteins. We have studied in detail the motions of external loops in five different proteins: tubulin (PDB code 1tub, tubulin α/β dimer), reverse transcriptase (1n5y), triose phosphate isomerase (1tph, human triose phosphate isomerase), protease (1j71, extracellular aspartic proteinase from *Candida tropicalis* yeast), and myoglobin (2v1k, ferrous deoxymyoglobin at pH 6.8). Tubulin, reverse transcriptase, and triose phosphate isomerase are composed of two monomers and have 867, 910, and 496 residues, respectively. Protease and myoglobin each contain single chains with 338 and 163 residues, respectively. We first locate loops of these proteins that are on the surface and compute the mean-square fluctuations of all the residues in these loops, and their cross correlations (covariances) between fluctuations of two different residues by using both the GNM (eq 4) and the analytical formula (eq 2) derived for a polymer network with alternating functionality. More specifically, we consider the loops as chains where the first and the last residue are junctions with functionalities equal to the actual connectivities for these residues with the remainder of the protein as given for a cut off distance (7 Å); the other residues of the loop are considered as junctions having functionality two. Now since the functionality of the two terminal loop residues may be different, in order to find the auto and cross covariances of the loop residues we use eq 2. For the case of the GNM model, we find the connectivity matrix (for the whole protein including PDB data for residues forming loops) and then we find the pseudoinverse thereof by using singular value decomposition. The fluctuations and the covariances of the residues of each loop residue are found based on eq 4.

We identify protein loops by first excluding helices and β -strands in the protein structure, leaving only coils. The criterion for a loop is the requirement that four or more consecutive coil residues are located on the protein surface. We illustrate these loops in protein structures for three of the studied proteins by coloring them blue in Figure 2.

We calculated covariances of instantaneous fluctuations of residues in loops both from eq 2 (polymer theory of rubberlike elasticity) and from eq 4 (GNM computations based on the complete protein structure). The results obtained are shown in Figures 3–7. Curves with squares show covariances calculated from eq 2 using only information on the connectivity of the terminal residues of protein loops (functionality of their junctions) and the length of a loop, while curves with dots display covariances calculated from eq 4 for GNM applied to the whole protein. The pattern for the residue–residue indexing is as follows: initially the index shows the covariance of the first residue in the loop with

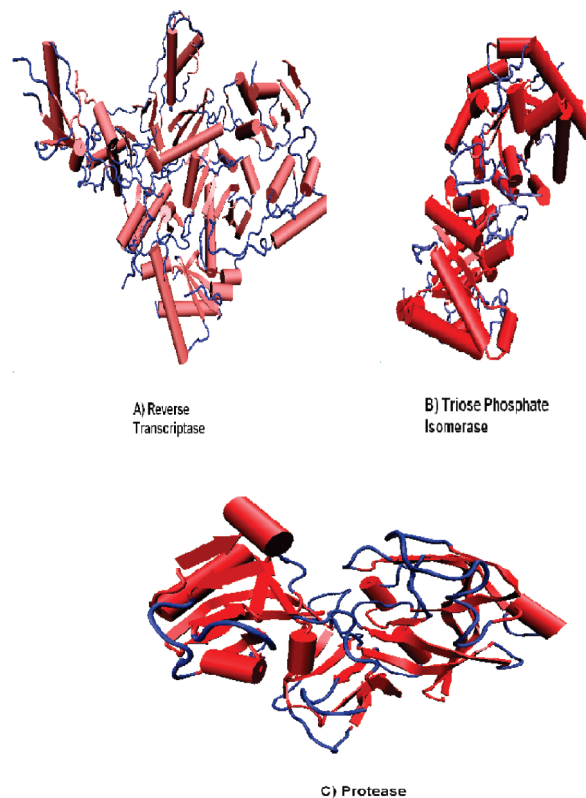


Figure 2. (A) Loops (colored in blue) of reverse transcriptase (A), triose phosphate isomerase (B), and protease (C).

itself and with all others, then of the second residue with itself and with all others (except the first one), etc. For a loop composed of n residues the residue–residues index changes from 1 to $n(n + 1)/2$.

Figure 3 shows the values of covariances calculated from polymer rubberlike elasticity theory (squares) and from GNM (dots) for the following loops (for indexing of all loops see Appendix C in the Supporting Information): (a) reverse transcriptase loop number 4 ($n = 7$), (b) tubulin loop number 4 ($n = 5$), (c) triose phosphate isomerase loop number 3 ($n = 5$), (d) protease loop number 7 ($n = 6$), and (e) myoglobin loop number 4 ($n = 5$). We see that for these cases the local approximation based on eq 2 provides an excellent result that very well approximates the whole structure result based on the GNM.

In addition to covariances of the instantaneous fluctuations, it is interesting to analyze correlations among them defined by

$$\text{corr} = \frac{\langle\langle \Delta \mathbf{R}_i \cdot \Delta \mathbf{R}_j \rangle\rangle}{\sqrt{\langle\langle (\Delta \mathbf{R}_i)^2 \rangle\rangle \langle\langle (\Delta \mathbf{R}_j)^2 \rangle\rangle}} \quad (6)$$

Figure 4 shows the correlations obtained by the GNM and the polymer elastic theory for the loops analyzed earlier in Figure 3.

Figure 5 shows the values of covariances computed for four loops of reverse transcriptase with asymmetry in the connectivities ϕ_1 , ϕ_2 of the terminal junctions for loops: (a) loop number 4 ($n = 7$, $\phi_1 = 4$, $\phi_2 = 11$), (b) number 5 ($n = 5$, $\phi_1 = 8$, $\phi_2 = 3$), (c) number 11 ($n = 6$, $\phi_1 = 5$, $\phi_2 = 10$), and (d) number 14 ($n = 4$, $\phi_1 = 5$, $\phi_2 = 13$). We see that

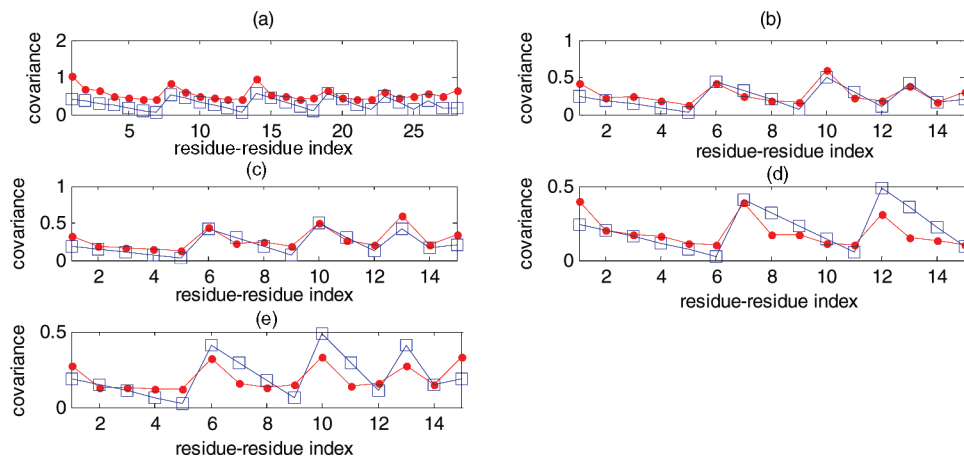


Figure 3. Covariances of instantaneous fluctuations calculated by using theory of rubberlike elasticity (eq 2) (squares) and GNM (eq 4) (dots) for the following individual loops: (a) reverse transcriptase loop no. 4, (b) tubulin loop no. 4, (c) triose phosphate isomerase loop no. 3, (d) protease loop no. 7, and (e) myoglobin loop no. 4. The abscissa shows the index for pairs of residues.

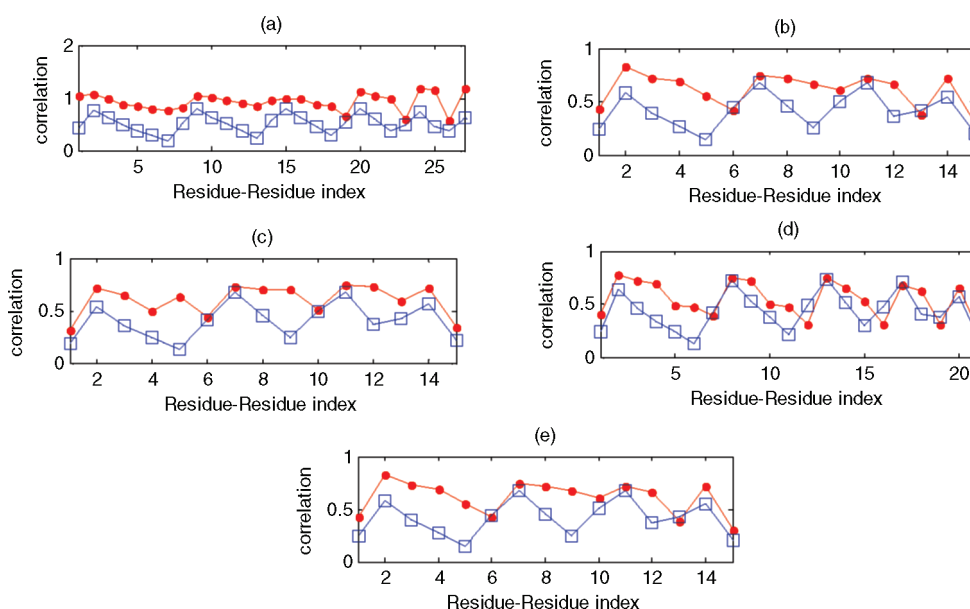


Figure 4. Correlations of instantaneous fluctuations computed by using the theory of rubberlike elasticity (eq 2) (squares) and the GNM (eq 4) (dots) for the following individual loops: (a) reverse transcriptase loop no. 4, (b) tubulin loop no. 4, (c) triose phosphate isomerase loop no. 3, (d) protease loop no. 7, and (e) myoglobin loop no. 4. The abscissa shows the index for pairs of residues.

the local approximations based on the polymer network model are more successful for longer loops than for short ones.

It is also relevant to examine the relationships between the fluctuations of the loop residues computed from polymer rubberlike elasticity model, and from GNM with experimental B -factors. Figure 6 shows plots of the mean square fluctuations of loop residues obtained from these three different sources for all 16 loops of the protease. For purposes of comparison, all three quantities are normalized. (For example $B_i^{\text{norm}} = (B_i - B_{\min}) / (B_{\max} - B_{\min})$, where i is the residue index for the particular loop. The same normalization is carried for the fluctuations computed from GNM and from polymer phantom network model.) We see that for some of these cases the local approximation based on our model approximates the GNM results very well. The main reasons

for the good or bad approximation by the local method to the GNM results are the connectivities of the residues of the loops.

Another crucial issue is whether the covariances of instantaneous fluctuations decay similarly with respect to the sequence distance between the residues. To address this problem, we have plotted in Figure 7 the covariances of the first residue in each loop with respect to the other residues of the same loop as a function of the sequence distance between these two residues.

Figures 3–7 indicate that the covariances of instantaneous fluctuations obtained both by considering the loops as individual entities (theoretical model) and as a whole structure (GNM) are closely similar. We computed the correlations of these covariances for all loops for all the proteins studied here. Our computations indicate that the

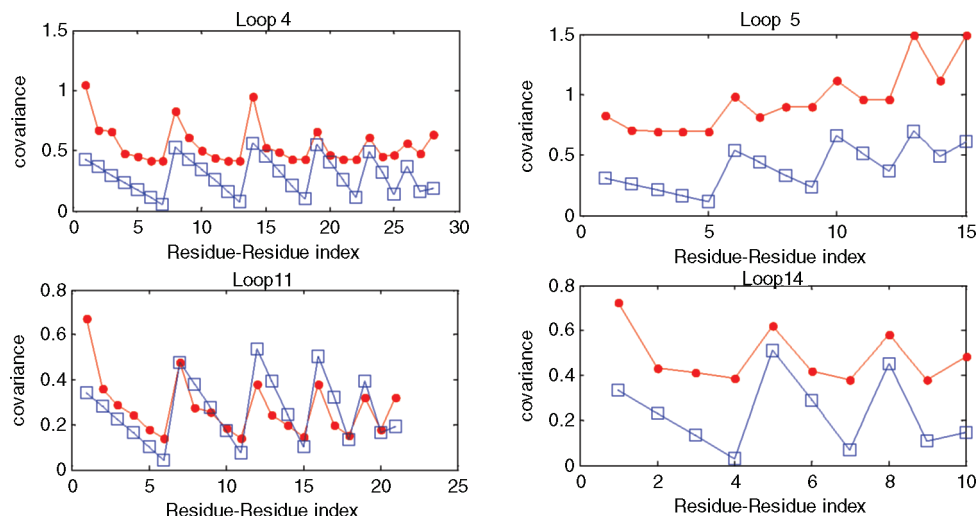


Figure 5. Covariances of the instantaneous fluctuations calculated by using the theory of rubber elasticity (eq 2) (squares) and GNM (eq 4) (dots) for loops no. 4, 5, 11, and 14 of reverse transcriptase. The abscissa shows the index for pairs of residues, with indexing described in the text.

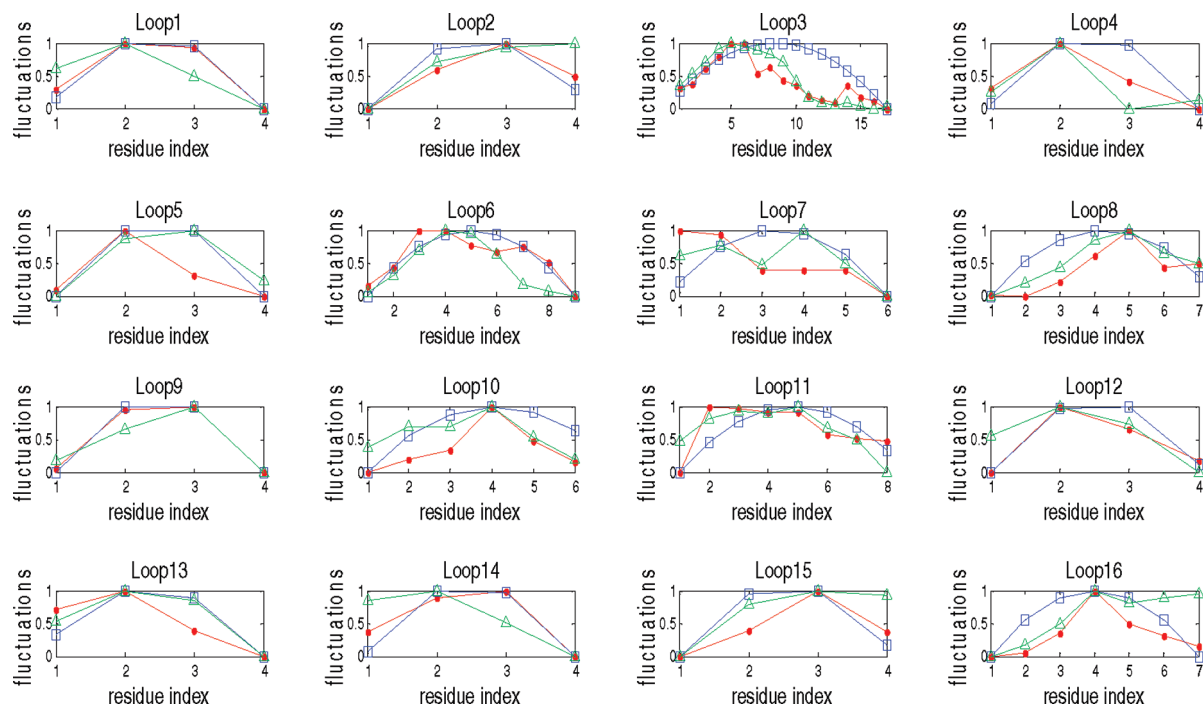


Figure 6. Fluctuations of the residues for all loops of protease. Fluctuations computed from polymer theory of rubberlike elasticity (squares) and from GNM (dots) are compared with *B*-factors (triangles). All fluctuations have been normalized. The abscissa is the residue index in the loop.

average correlation of covariances (averaged over all loops in a given protein) is the largest (0.70) for triose phosphate isomerase, 0.66 for reverse transcriptase, and the smallest (0.38) for myoglobin, which has only four very short loops. All profiles shown in these figures show a close resemblance between the behavior of our new loop modeling and the whole protein modeling with the Gaussian network model. This is at first surprising, since previous results with the Gaussian network model indicated that the whole structure was needed in order to compute the motions of any part. What we are seeing here is that the individual loops and their simplified representations are generally sufficient to compute the relative mobilities for the individual parts of the loop. Information contained in Figures 3–7 for additional

loops of the proteins are provided in Appendix D in the Supporting Information. Appendix E in the Supporting Information gives the correlations of the covariances for every loop of each protein we have studied as a function of the sum of the functionalities of the two terminal junctions in each loop. We have not noticed any apparent relationship between these two quantities. Appendix F in the Supporting Information shows results of computations of covariances of instantaneous fluctuations of residues belonging to helices for the proteins of study, similarly as was done earlier for loops. Our computations show that polymer network approximation does not work as well for helices as for the loops. Appendix G in the Supporting Information shows the computed correlation coefficients between the predicted

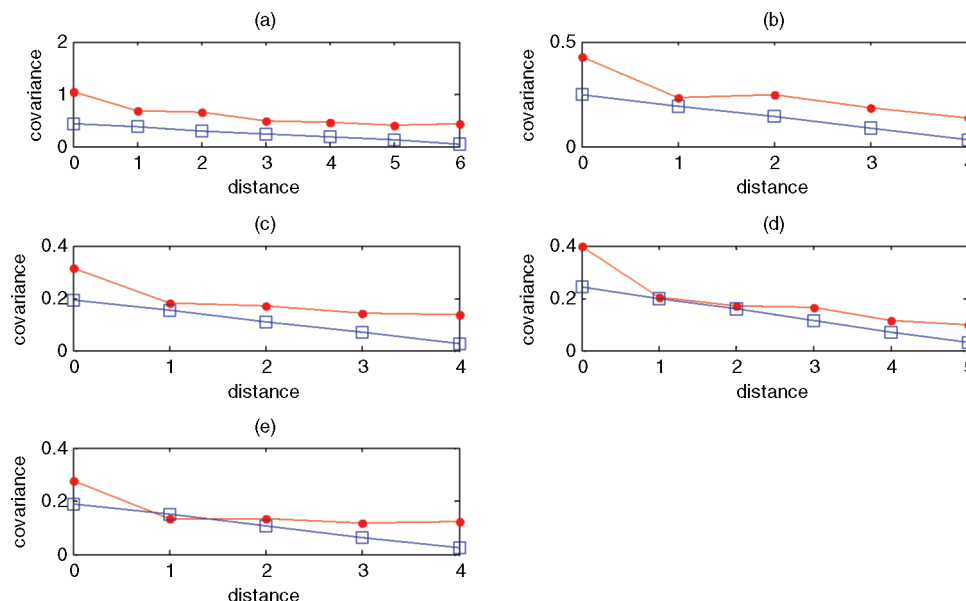


Figure 7. Covariances of instantaneous fluctuations of the first residue of the loop with other residues in the same loop as a function of the sequence distance between them for (a) reverse transcriptase loop no. 4, (b) tubulin loop no. 4, (c) triose phosphate isomerase loop no. 3, (d) protease loop no. 7, and (e) myoglobin loop no. 4. Results computed from polymer theory of rubberlike elasticity (squares) and from GNM (dots) are compared.

amplitudes of fluctuations of loop residues (using both the present analytical model and GNM theory) and the experimental B -factors.

Discussion. We have presented a comparison of the loop motions using both the GNM and a new approach based on the theory of rubberlike elasticity of polymer networks. For the latter, the loop is modeled by assuming that the two ends have functionalities ϕ_1 and ϕ_2 connecting them with the remainder of the protein structure, whereas the intermediate residues of the loop uniformly have the functionality of two, connecting them only to their sequence neighbors. We then calculated the mean square fluctuations (variances) and covariances for the loop residues by using formulas derived analytically by us for a treelike polymer network with alternating functionalities. We then applied this new approach to all external loops in five different proteins (reverse transcriptase, triose phosphate isomerase, tubulin, protease, and myoglobin) and compared analytical results for the mean square fluctuations and the covariances with GNM computations based on the coordinates for all residues of the loops. For each loop we have plotted the covariances between instantaneous fluctuations of pairs of residues of the loop, and the mean square fluctuations of individual loop residues. We have also compared the covariance of instantaneous fluctuations between two residues of the loop as a function of the distance between them. The comparisons between these two models show that the local approximation of the loops that describes the motion of the loop residues based on polymer treelike network topology independently from the rest of the protein structure closely approximates the results obtained from GNM where the whole protein structure is taken into account.

Loop mobilities have long been considered to be important for function. Loops on the surfaces of proteins are often the most uncertain parts of structures. It is quite likely that many loops are artificially immobilized and compressed against

the body of the protein in the crystal environment. What has been done in the present work is to develop and compare two simple models for loop mobilities. These two quite different approaches yield similar results. Our finding, that *computed loop motions are similar whether treated as independent or in the context of the whole protein structure*, implies that the loop motions are occurring along relatively well-defined pathways, which likely could be mapped out in future computations. This approach may be used not only for loops of proteins but also for the large loops frequently occurring in nucleic acid stem-loop structures, as well as the much larger loops originating in double-stranded DNA, when multiple subunit proteins bind at widely separated positions. This critical problem for transcription regulation requires coarse graining of the structure since the loops can be thousands of base pairs in length. The more difficult problem may be how to introduce additional interactions within these DNA loops when supercoiling is introduced.

Utility of This Approach for Describing the Motions of Loops with Unknown Structures. Often in protein structures loops are missing in the crystal structure. The approach described in this paper could be used to predict probably structures for these missing loops since the only requirement is knowing the connectivity numbers for the two ends and the number of residues in the loop.

In the present work, we have considered only the relative magnitudes of these motions, which compare favorably. In the future, it will be essential to develop a vector version of the present loop modeling to define actual pathways, to compare with the ANM and the anisotropic temperature factors. These pathways will be presented in future work, where we also investigate the effects of interactions between loop residues and the body of the protein. The results from the present work indicate that our new approach based on

polymer elasticity theory provides a close approximation to the GNM model that includes the effects of the whole structure.

The polymer rubberlike elasticity model only predicts a relatively simple pattern of fluctuations of residues across a loop with a convex shape, and with the residues close the center of the loops having a higher amplitude. However, according to this model the central residue of the loop does not always have the maximum amplitude of fluctuations. We have also an effect from the functionality of junctions on both ends of the loop. The maximum is shifted toward the junction which having lower functionality. This effect is also observed for experimental *B*-factors for loop residues, although sometimes there are exceptions to this rule, due to significant interactions of loop residues with the remainder of the protein. Our predictions are limited somewhat by the simplicity of the theoretical model, and by the assumption that the motions of loops are unobstructed by the remaining part of the protein, but nevertheless they do enable predicting the basic features of these motions.

Acknowledgment. We are pleased to acknowledge the financial support provided by the National Institutes of Health through grants R01GM081680, R01GM072014, and R01GM073095.

Supporting Information Available: In Appendix A we provide a synopsis of the theory of random polymer networks. In the Appendix B we present a summary of the theory of random polymer networks with alternating functionality. In Appendix C we show tables listing the loops and the loop residues for all the proteins of this study. In Appendix D we provide information (supplementary to that contained in Figures 3–7) for more loops for all the proteins of study. In Appendix E we show the calculated correlation of covariances of instantaneous fluctuations for every loop of each protein as a function of the sum of the functionalities of their terminal junctions. In Appendix F we have calculated the covariances of instantaneous fluctuations of residues belonging to helices for all proteins studied, as we did earlier for the loops. Appendix G lists the computed correlation coefficients between the predicted amplitudes of fluctuations (using both the present analytical model and GNM theory) and the experimental *B*-factors for protein loops.

This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Bahar, I.; Atilgan, A. R.; Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding Des.* **1997**, *2* (3), 173–181.
- (2) Kundu, S.; Melton, J. S.; Sorensen, D. C.; Phillips, G. N. Dynamics of proteins in crystals: Comparison of experiment with simple models. *Biophys. J.* **2002**, *83* (2), 723–732.
- (3) Sen, T. Z.; Feng, Y. P.; Garcia, J. V.; Kloczkowski, A.; Jernigan, R. L. The extent of cooperativity of protein motions observed with elastic network models is similar for atomic and coarser-grained models. *J. Chem. Theory Comput.* **2006**, *2* (3), 696–704.
- (4) Yang, L.; Song, G.; Carriquiry, A.; Jernigan, R. L. Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure* **2008**, *16* (2), 321–330.
- (5) Yang, L.; Song, G.; Jernigan, R. L. How well can we understand large-scale protein motions using normal modes of elastic network models. *Biophys. J.* **2007**, *93* (3), 920–929.
- (6) Lu, M. Y.; Ma, J. P. The role of shape in determining molecular motions. *Biophys. J.* **2005**, *89* (4), 2395–2401.
- (7) Fiser, A.; Do, R. K. G.; Sali, A. Modeling of loops in protein structures. *Protein Sci.* **2000**, *9* (9), 1753–1773.
- (8) Espadaler, J.; Fernandez-Fuentes, N.; Hermoso, A.; Querol, E.; Aviles, F. X.; Sternberg, M. J. E.; Oliva, B. ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic Acids Res.* **2004**, *32*, D185–D188.
- (9) Oliva, B.; Bates, P. A.; Querol, E.; Aviles, F. X.; Sternberg, M. J. E. An automated classification of the structure of protein loops. *J. Mol. Biol.* **1997**, *266* (4), 814–830.
- (10) Panchenko, A. R.; Madej, T. Structural similarity of loops in protein families: toward the understanding of protein evolution. *BMC Evolut. Biol.* **2005**, *5*, Art. No. 10.
- (11) Groban, E. S.; Narayanan, A.; Jacobson, M. P. Conformational changes in protein loops and helices induced by post-translational phosphorylation. *PLoS Comput. Biol.* **2006**, *2* (4), 238–250.
- (12) Hu, X. Z.; Wang, H. C.; Ke, H. M.; Kuhlman, B. High-resolution design of a protein loop. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (45), 17668–17673.
- (13) Bos, C.; Lorenzen, D.; Braun, V. Specific in vivo labeling of cell surface-exposed protein loops: Reactive cysteines in the predicted gating loop mark a ferrichrome binding site and a ligand-induced conformational change of the Escherichia coli FhuA protein. *J. Bacteriol.* **1998**, *180* (3), 605–613.
- (14) Li, C.; Banfield, M. J.; Dennison, C. Engineering copper sites in proteins: Loops confer native structures and properties to chimeric cupredoxins. *J. Am. Chem. Soc.* **2007**, *129*, 709–718.
- (15) Smith, J. W.; Tachias, K.; Madison, E. L. Protein loop grafting to construct a variant of tissue-type plasminogen activator that binds platelet integrin α (IIb) β (3). *J. Biol. Chem.* **1995**, *270* (51), 30486–30490.
- (16) Sudarsanam, S.; Dubose, R. F.; March, C. J.; Srinivasan, S. Modeling Protein Loops Using A Phi-I+1, Psi-I Dimer Database. *Protein Sci.* **1995**, *4* (7), 1412–1420.
- (17) vanVlijmen, H. W. T.; Karplus, M. PDB-based protein loop prediction: Parameters for selection and methods for optimization. *J. Mol. Biol.* **1997**, *267* (4), 975–1001.
- (18) Kolodny, R.; Guibas, L.; Levitt, M.; Koehl, P. Inverse kinematics in biology: The protein loop closure problem. *Int. J. Robotics Res.* **2005**, *24* (2–3), 151–163.
- (19) Gerstein, M.; Chothia, C. Analysis of Protein Loop Closure - 2 Types of Hinges Produce One Motion in Lactate-Dehydrogenase. *J. Mol. Biol.* **1991**, *220* (1), 133–149.
- (20) Krieger, F.; Fierz, B.; Axthelm, F.; Joder, K.; Meyer, D.; Kiefhaber, T. Intrachain diffusion in a protein loop fragment from carp parvalbumin. *Chem. Phys.* **2004**, *307* (2–3), 209–215.
- (21) Li, W. Z.; Liu, Z. J.; Lai, L. H. Protein loops on structurally similar scaffolds: Database and conformational analysis. *Biopolymers* **1999**, *49* (6), 481–495.
- (22) Burke, D. F.; Deane, C. M. Improved protein loop prediction from sequence alone. *Protein Eng.* **2001**, *14* (7), 473–478.

- (23) James, H. M.; Guth, E. Theory of the Increase in Rigidity of Rubber During Cure. *J. Chem. Phys.* **1947**, *15* (9), 669–683.
- (24) James, H. M. Statistical Properties of Networks of Flexible Chains. *J. Chem. Phys.* **1947**, *15* (9), 651–668.
- (25) James, H. M.; Guth, E. Simple Presentation of Network Theory of Rubber, with A Discussion of Other Theories. *J. Polym. Sci.* **1949**, *4* (2), 153–182.
- (26) James, H. M.; Guth, E. Statistical Thermodynamics of Rubber Elasticity. *J. Chem. Phys.* **1953**, *21* (6), 1039–1049.
- (27) Flory, P. J. Statistical Thermodynamics of Random Networks. *Proc. R. Soc. London Ser. A—Math. Phys. Eng. Sci.* **1976**, *351* (1666), 351–380.
- (28) Kloczkowski, A.; Mark, J. E.; Erman, B. Chain Dimensions and Fluctuations in Random Elastomeric Networks 0.1. Phantom Gaussian Networks in the Undeformed State. *Macromolecules* **1989**, *22*, 1423–1432.
- (29) Kloczkowski, A.; Mark, J. E.; Frisch, H. L. The relaxation spectrum for Gaussian Networks. *Macromolecules* **1990**, *23*, 3481–3490.
- (30) Treloar, L. R. G. The Elasticity of A Network of Long-Chain Molecules 0.3. *Trans. Faraday Soc.* **1946**, *42* (1–2), 83–94.
- (31) Treloar, L. R. G. The Statistical Length of Long-Chain Molecules. *Trans. Faraday Soc.* **1946**, *42* (1–2), 77–82.
- (32) Kloczkowski, A.; Mark, J. E.; Erman, B. Fluctuations, Correlations, and Small-Angle Neutron-Scattering from End-Linked Gaussian Chains in Regular Bimodal Networks. *Macromolecules* **1991**, *24*, 3266–3275.
- (33) Kloczkowski, A.; Mark, J. E.; Erman, B. A Diffused-Constraint Theory for the Elasticity of Amorphous Polymer Networks 0.1. Fundamentals and Stress-Strain Isotherms in Elongation. *Macromolecules* **1995**, *28*, 5089–5096.
- (34) Bahar, I.; Rader, A. J. Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.* **2005**, *15* (5), 586–592.
- (35) Doruker, P.; Jernigan, R. L.; Bahar, I. Dynamics of large proteins through hierarchical levels of coarse-grained structures. *J. Comput. Chem.* **2002**, *23* (1), 119–127.
- (36) Kurcuoglu, O.; Jernigan, R. L.; Doruker, P. Collective dynamics of large proteins from mixed coarse-grained elastic network model. *QSAR Comb. Sci.* **2005**, *24* (4), 443–448.
- (37) Tama, F.; Gadea, F. X.; Marques, O.; Sanejouand, Y. H. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins: Struct., Funct., Genet.* **2000**, *41* (1), 1–7.
- (38) Tama, F.; Sanejouand, Y. H. Conformational change of proteins arising from normal mode calculations. *Protein Eng.* **2001**, *14* (1), 1–6.
- (39) Tirion, M. M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* **1996**, *77* (9), 1905–1908.
- (40) 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* (1), 505–515.
- (41) Song, G.; Jernigan, R. L. vGNM: A better model for understanding the dynamics of proteins in crystals. *J. Mol. Biol.* **2007**, *369* (3), 880–893.
- (42) Skliros, A.; Mark, J. E.; Kloczkowski, A. Chain Dimensions and Fluctuations in Elastomeric Networks in which the Junctions Alternate Regularly in their Functionality. *J. Chem. Phys.* **2009**, *130*, 064905.
- (43) Kuhn, W. Relationship between molecular size, static molecular shape and elastic properties of high polymer materials. *Kolloid-Z.* **1936**, *76*, 258.
- (44) Jensen, J. H.; Gordon, M. S. An approximate formula for the intermolecular Pauli repulsion between closed shell molecules. II. Application to the effective fragment potential method. *J. Chem. Phys.* **1998**, *108* (12), 4772–4782.
- (45) Skliros, A.; Mark, J. E.; Kloczkowski, A. Small-Angle Neutron Scattering from Elastomeric Networks in which the Junctions Alternate Regularly in their Functionality. *Macromol. Theory Simul.* **2009**, *18* (9), 537–544.
- (46) Erman, B.; Kloczkowski, A.; Mark, J. E. Chain Dimensions and Fluctuations in Random Elastomeric Networks 0.2. Dependence of Chain Dimensions and Fluctuations on Macroscopic Strain. *Macromolecules* **1989**, *22*, 1432–1437.
- (47) Haliloglu, T.; Bahar, I.; Erman, B. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.* **1997**, *79* (16), 3090–3093.
- (48) Doruker, P.; Jernigan, R. L. Functional motions can be extracted from on-lattice construction of protein structures. *Proteins: Struct., Funct., Genet.* **2003**, *53* (2), 174–181.
- (49) Yang, L. W.; Eyal, E.; Chennubhotla, C.; Jee, J.; Gronenborn, A. M.; Bahar, I. Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions. *Structure* **2007**, *15* (6), 741–749.
- (50) Haliloglu, T.; Bahar, I.; Erman, B. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.* **1997**, *79* (16), 3090–3093.
- (51) Keskin, O.; Durell, S. R.; Bahar, I.; Jernigan, R. L.; Covell, D. G. Relating molecular flexibility to function: A case study of tubulin. *Biophys. J.* **2002**, *83* (2), 663–680.
- (52) Keskin, O.; Bahar, I.; Flatow, D.; Covell, D. G.; Jernigan, R. L. Molecular mechanisms of chaperonin GroEL-GroES function. *Biochemistry* **2002**, *41*, 491–501.
- (53) Navizet, I.; Lavery, R.; Jernigan, R. L. Myosin flexibility: Structural domains and collective vibrations. *Proteins: Struct., Funct., Genet.* **2004**, *54* (3), 384–393.
- (54) Wang, Y. M.; Rader, A. J.; Bahar, I.; Jernigan, R. L. Global ribosome motions revealed with elastic network model. *J. Struct. Biol.* **2004**, *147* (3), 302–314.
- (55) Wang, Y. M.; Jernigan, R. L. Comparison of tRNA motions in the free and ribosomal bound structures. *Biophys. J.* **2005**, *89* (5), 3399–3409.
- (56) Yan, A. M.; Wang, Y. M.; Kloczkowski, A.; Jernigan, R. L. Effects of Protein Subunits Removal on the Computed Motions of Partial 30S Structures of the Ribosome. *J. Chem. Theory Comput.* **2008**, *4* (10), 1757–1767.
- (57) Kurcuoglu, O.; Doruker, P.; Sen, T. Z.; Kloczkowski, A.; Jernigan, R. L. The ribosome structure controls and directs mRNA entry, translocation and exit dynamics. *Phys. Biol.* **2008**, *5* (4), XXX.
- (58) Kundu, S.; Jernigan, R. L. Molecular mechanism of domain swapping in proteins: An analysis of slower motions. *Biophys. J.* **2004**, *86* (6), 3846–3854.
- (59) Feng, Y. P.; Yang, L.; Kloczkowski, A.; Jernigan, R. L. The energy profiles of atomic conformational transition intermediates of adenylate kinase. *Proteins: Struct., Funct., Bioinf.* **2009**, *77* (3), 551–558.

JCTC

Journal of Chemical Theory and Computation

Fast Proton Titration Scheme for Multiscale Modeling of Protein Solutions

Andre Azevedo Reis Teixeira,[†] Mikael Lund,[‡] and Fernando Luís Barroso da Silva^{*,†,‡}

Department of Physics and Chemistry, 14040-903 Av. do café, s/no., FCFRP—USP, Ribeirão Preto, SP, Brazil, Department of Theoretical Chemistry Chemical Center, Lund University, P.O. Box 124-S-221 00, Lund, Sweden, and Department of Physics and Chemistry, 14040-903 Av. do café, s/no., FCFRP—USP, Ribeirão Preto, SP, Brazil

Received June 7, 2010

Abstract: Proton exchange between titratable amino acid residues and the surrounding solution gives rise to exciting electric processes in proteins. We present a proton titration scheme for studying acid–base equilibria in Metropolis Monte Carlo simulations where salt is treated at the Debye–Hückel level. The method, rooted in the Kirkwood model of impenetrable spheres, is applied on the three milk proteins α -lactalbumin, β -lactoglobulin, and lactoferrin, for which we investigate the net-charge, molecular dipole moment, and charge capacitance. Over a wide range of pH and salt conditions, excellent agreement is found with more elaborate simulations where salt is explicitly included. The implicit salt scheme is orders of magnitude faster than the explicit analog and allows for transparent interpretation of physical mechanisms. It is shown how the method can be expanded to multiscale modeling of aqueous salt solutions of many biomolecules with nonstatic charge distributions. Important examples are protein–protein aggregation, protein–polyelectrolyte complexation, and protein–membrane association.

1. Introduction

Electrostatic interactions are crucial for biological^{1–3} and technological phenomena involving proteins. Examples are protein foams in the food industry,⁴ food and pharmaceutical formulations,^{5,6} enzyme immobilization,⁷ protein separation, and other bioprocesses.^{8–11} These interactions are directly related to a multitude of ionized amino acid residues: aspartic acid, glutamic acid, cysteine not involved in SS bonds, tyrosine, C-terminal, arginine, histidine, lysine, and N-terminal. Consequently, salt and pH have a significant effect on protein solution behavior, and a wealth of studies of electrostatic interactions in and between biomolecules is available.^{1–3,12–14}

The tendency of amino acids to change their ionization states due to an external electrostatic perturbation is directly related to the charge fluctuation of the protein—the so-called

protein charge capacitance, C .¹² This fluctuation, quantified by C , is responsible for an important purely electrostatic attractive interaction observed in several systems.^{13–21} While the *charge regulation* mechanism has been known for a long time,^{22,23} only recently has it become possible to accurately extract C from the three-dimensional protein structure.

In many practical applications the molar concentration of salt is relatively high—100 mM or more—which invariably increases the computation time of any continuum electrostatic model that incorporates salt particles *explicitly*. The situation becomes even worse in the case of low protein concentrations. In such studies, often the main focus is not on the salt particles themselves but rather on their effect on the interactions between the proteins or biopolymers. Since many practical experimental conditions are not far from the isoelectric point (pI), the electrostatic coupling²⁴ tends to be small, and salt can be treated at the Debye–Hückel level.^{25–27}

Monte Carlo (MC) titration schemes using explicit ions have been successfully applied to describe ionization equilibria and protein complexation.^{12,28–31} Titrating MC simulations have also been used to study the impact of the charge

* To whom correspondence should be addressed. Tel.: +55 (16)3602 42 19. Fax: +55 (16)3602 48 80. E-mail: fernando@fcfrp.usp.br.

[†] Department of Physics and Chemistry, FCFRP—USP.

[‡] University of Lund.

Table 1. Details of the Three Studied Proteins^a

protein	PDB	residues	radius (Å)	mass (kDa)
α-LA	1F6S (A)	122	18	14
β-LG	1BEB (A+B)	312	24	35
LF	1BLF	685	30	76

^a The radius is that of a sphere with the same volume as the three dimensional structure of the protein.

regulation mechanism mentioned above.^{12–14} In these approaches the explicit inclusion of ions requires excessive computational resources at conditions where the salt concentration is high when at the same time the protein concentration is low.

The Poisson–Boltzmann (PB) approximation has been widely used for protein electrostatics, ranging from the early Tanford–Kirkwood model^{32,33} to advanced finite element methods where boundary conditions are numerically matched on the anisotropic protein surface.³⁴ For studying *many* proteins, a fairly simple approach is required, while maintaining the essential physics. In this work, we present a fast algorithm for proton titration in Metropolis Monte Carlo simulations where salt is treated implicitly. This drastically reduces the computation time while capturing the main effects of pH and salt.

Milk proteins have diverse physicochemical properties and extensive applications within formulation technology in the pharmaceutical and food industries.^{20,35,36} They are further used as model systems for biophysical studies,^{37,38} and their electrostatic properties are hence of considerable interest. In this study, we use α-lactalbumin (α-LA), β-lactoglobulin (β-LG), and lactoferrin (LF) as test cases for our theoretical models. The properties and structures of all three proteins are listed and shown in Table 1 and Figure 1, respectively, and we now briefly introduce each protein:

The globular 14 kDa protein α-LA is part of the C-type lysozyme family and corresponds to approximately 20% of the proteins present in bovine milk serum. α-LA acts as an enzyme for the biosynthesis of lactose,³⁵ and the tertiary structure is available at 2.2 Å resolution.³⁹

β-LG is the main protein constituent of bovine milk serum, corresponding to roughly 50%. In mammalian animals, it has two main functions: binding of retinol (vitamin A) and stimulating lipolysis. Under normal milk conditions, β-LG is found in a dimer/monomer equilibrium which is shifted toward the dimer as the pH approaches pI. We here study the dimeric form since its high molecular dipole moment³⁵ provides a more challenging case for the theoretical models. The used tertiary structure (a mixture of two genetic variants) was obtained by X-ray diffraction with a 1.8 Å resolution.⁴⁰

With its molecular mass of 76 kDa, LF is the largest protein studied. LF is a bacteriostatic agent that sequesters iron ions. Commercial applications are many, including nutraceuticals production, antioxidant agents, and oral care products. The tertiary structure of LF is available at 2.8 Å resolution, and contrary to the previous proteins that have pI values between 4.2 to 5.4, the pI ≈ 9 of LF is on the basic side.³⁵

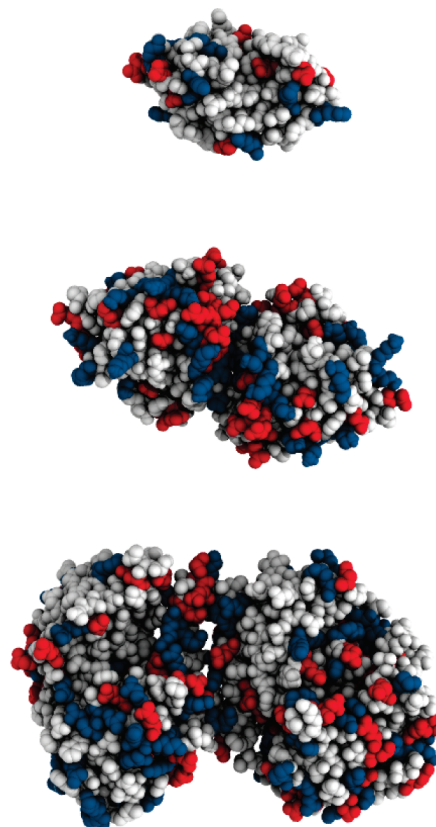


Figure 1. Structures of the three studied milk proteins. Acidic and basic residues are shown in red and blue, respectively. From top to bottom: α-LA (PDB ID: 1F6S), β-LG (PDB ID: 1BEB), and LF (PDB ID: 1BLF).

2. Models and Theory

2.1. Titration Behavior at Ideal Conditions. In this section, we describe general equilibrium properties of weak acids which shall serve as a foundation for the two MC titration schemes presented later. We start with the dissociation process for a monoprotic acid



with the corresponding thermodynamic equilibrium constant

$$K_a = \frac{\gamma_{\text{H}^+} \gamma_{\text{A}^-}}{\gamma_{\text{HA}}} \times \frac{c_{\text{H}^+} c_{\text{A}^-}}{c_{\text{HA}}} \quad (1)$$

where γ and c are activity coefficients and concentrations, respectively. It then follows that the free energy difference between the protonated and deprotonated form is

$$\beta \Delta A_{\text{HA} \rightarrow \text{A}^-} = -\ln \frac{c_{\text{A}^-}}{c_{\text{HA}}} = -\ln \frac{\gamma_{\text{A}^-}}{\gamma_{\text{HA}}} - (\text{pH} - \text{p}K_a) \ln 10 \quad (2)$$

where $\beta = 1/k_B T$ is the inverse thermal energy and $\text{pH} = -\log(c_{\text{H}^+} \gamma_{\text{H}^+})$. We now choose a reference state so that at infinite dilution, $\gamma \rightarrow 1$ and the degree of dissociation can hence be written as

$$\alpha(\text{pH}) = \frac{c_{\text{A}^-}}{c_{\text{HA}} + c_{\text{A}^-}} = \frac{10^{\text{pH} - \text{p}K_a}}{1 + 10^{\text{pH} - \text{p}K_a}} \quad (3)$$

If for a protein with many titratable groups, we (briefly!) assume that each residue is unaffected by the rest, the average net-charge number of the protein, Z_p , can be estimated by summing over all N_{acid} acidic and N_{basic} basic residues, requiring only the primary structure

$$Z_p(\text{pH}) = \sum_{i=1}^{N_{\text{basic}}} [1 - \alpha_i(\text{pH})] - \sum_{i=1}^{N_{\text{acid}}} \alpha_i(\text{pH}) \quad (4)$$

Due to proton exchange with the surrounding solution, the charge state of ionizable groups will *fluctuate* over time. This fluctuation leads to the charge regulation mechanism^{23,41,42} and can be defined as a capacitance¹²

$$C = \langle Z_p^2 \rangle - \langle Z_p \rangle^2 = -\frac{1}{\ln 10} \times \frac{\partial Z_p}{\partial \text{pH}} \quad (5)$$

where the brackets denote statistical mechanical ensemble averages. Hence, C can be obtained simply by differentiating eq 4 with respect to pH.

Naturally, two proteins with the same amino acid composition have identical charge and capacitance at all pH values if ideal behavior is assumed. That is, neglecting intra- and intermolecular interactions ($\gamma \rightarrow 1$) precludes any distinction between different conformations of the same molecule. Nevertheless, the ideal contribution to the net charge and capacitance is a useful first approximation, and by comparing with full titration and capacitance data, we can deduce the effects of the tertiary structure and surrounding salt. Further, at high salt concentrations, electrostatic interactions are screened, and experimental data approach the ideal case.⁴³

2.2. Proton Titration with Explicit Salt. The Monte Carlo (MC) Metropolis algorithm^{44–46} is a stochastic method based on random walks in coordinate space where each configuration has a statistical weight defined by the Boltzmann distribution.^{25,45,46} Within the primitive model of electrolytes, macromolecular titration behavior has previously been studied using MC simulations,^{31,47,48} and we here rely on the following procedure: A single, rigid protein is placed at the center of a spherical simulation cell that also encapsulates mobile ions (counterions and added salt), see Figure 2. The protein is described in mesoscopic detail so that the full, experimentally determined structure is replaced by a coarse grained representation where each amino acid is represented by a sphere.⁴⁹ Except where otherwise stated, proteins are used in their monomeric form (first structure if more present in PDB). Mobile ions and amino acid residues are described by hard spheres of diameter σ_i and charge $q_i = z_i e$, where e is the elementary charge and z_i is the valency. Any two particles, i and j , in the system interact via a solvent-mediated Coulomb potential, augmented by a repulsive hard sphere term, i.e., the primitive model of electrolytes²⁵

$$u^{\text{el}}(r_{ij}) = \begin{cases} \infty, & r_{ij} \leq (\sigma_i + \sigma_j)/2 \\ \frac{e^2 z_i z_j}{4\pi\epsilon_0\epsilon_r r_{ij}}, & \text{otherwise} \end{cases} \quad (6)$$

where ϵ_0 is the permittivity of vacuum ($\epsilon_0 = 8.854 \times 10^{-12}$ C²/Nm²), $\epsilon_r = 78.7$ is the relative dielectric constant of water,

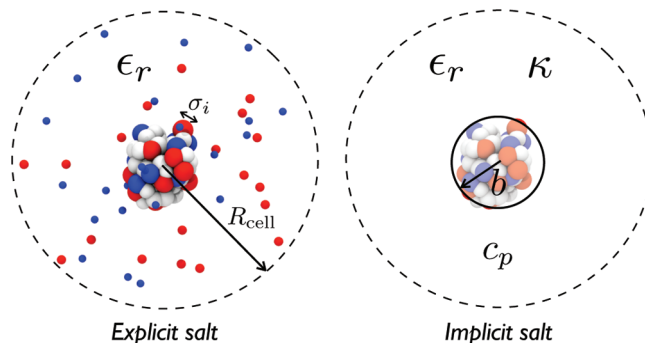


Figure 2. Schematic representation of the Monte Carlo simulation schemes. Left: Amino acids, salt, and counterions are described by (charged) hard spheres of diameter σ_i . Concentrations of salt, protein, and counterions are determined by the radius of the simulation cell, R_{cell} . Right: Salt is implicitly accounted for by the inverse Debye screening length, κ , which also varies with the counterion concentration and hence the protein concentration, c_p . The volume of the salt-free region defined by the sphere b matches that of the protein.

Table 2. Intrinsic pK_a Values for Titratable Amino Acid Groups⁵⁰

	Asp	Glu	Cys ^a	Tyr	Ctr	Arg	His	Lys	Ntr
pK_a	4.0	4.4	9.5	9.6	3.8	12.0	6.3	10.4	7.5

^a Only cysteins not engaged in sulfide bridges can titrate.

and $T = 298.15$ K is the temperature. During the MC simulation, mobile ions ($\sigma_{\pm} = 4$ Å) are allowed to translate in any direction while the charges on the protein fluctuate according to the following semicanonical MC move:

1. A titratable site on the protein is chosen by random.
2. If deprotonated, move the charge from the bulk solution to the site (protonation process). If protonated, move the proton to the bulk solution (deprotonation process). Note that no new particle is introduced into the dense protein environment. The proton binding site is already there, given by the experimental structure, and we merely increase (or decrease) its charge by $+1$.
3. The move is accepted with the probability,

$$\min(1, e^{-\beta\Delta U_{\text{el}} \pm (\text{pH} - \text{pK}_a) \ln 10}) \quad (7)$$

where ΔU_{el} is the change in total electrostatic energy, calculated by summing over all particles before and after the move according to eq 6. The second term in the exponential accounts for the free energy change of the (de)protonation process for a single amino acid, not affected by the presence of the rest of the protein nor by salt (see eq 2). The magnitude of this term is determined by pH and the intrinsic pK_a value. The latter is a measured property which *implicitly* includes nonelectrostatic interactions (dispersion, solvation effects, polarization, etc.) that are not explicitly accounted for in the MC Hamiltonian. In this work, we employ pK_a values given by Nozaki and Tanford,⁵⁰ see Table 2.

In summary, the MC simulation enables us to calculate statistical mechanical ensemble averages, integrated over all salt positions ($\mathbf{r}_i^{N_s}$) and protonation states (λ^{N_p})

$$\langle x \rangle = \frac{\int \int x(\mathbf{r}^{N_s}, \lambda^{N_p}) e^{-\beta U(\mathbf{r}^{N_s}, \lambda^{N_p})} d\mathbf{r}^{N_s} d\lambda^{N_p}}{\int \int e^{-\beta U(\mathbf{r}^{N_s}, \lambda^{N_p})} d\mathbf{r}^{N_s} d\lambda^{N_p}} \quad (8)$$

Important examples are the protein net charge, $\langle Z_p \rangle$; the molecular dipole moment, $\langle \mu \rangle$; and the capacitance, C .

An intense debate about the protein dielectric constant can be seen in the literature.^{51–56} In the presented scheme we assume a uniform dielectric constant for the entire system, including the protein. For a range of globular proteins, it has been shown that this approximation is indeed valid, mainly because most titratable sites are located at the surface and are thus well hydrated. In addition, a major part of any polarization contribution is implicitly included in the intrinsic pK_a values used. That is, we need only account for the *difference* in free energy between the residue in the protein and the reference model peptide for which the pK_a value was originally obtained. Unless the site is deeply buried, this difference will be small. For a detailed description of—perhaps surprisingly—*why* the uniform dielectric model works, the reader should consult ref 51.

2.3. Proton Titration with Implicit Salt. In the scheme described in the previous section, salt particles were explicitly included in the MC simulation. Here, we present a scheme for implicit salt titration based on Kirkwood's theory of impenetrable spheres.^{32,33} Consider a set of charges ez_1, \dots, ez_n distributed in a sphere of diameter σ_p immersed in a salt solution. The distance of closest approach between salt and the spherical macromolecule is defined by $b = (\sigma_p + \sigma_{\text{salt}})/2 \approx \sigma_p/2$. Assuming a uniform dielectric constant throughout the medium⁵¹ and neglecting higher-order multipolar terms for the salt–protein interactions, the free energy can be approximated by the leading terms in the Tanford–Kirkwood model,^{33,53}

$$w_{\text{tk}} \approx \frac{e^2}{4\pi\epsilon_0\epsilon_r} \left(\sum_{i>j}^n \frac{z_i z_j}{r_{ij}} - \frac{Z_p^2 \kappa}{2(1 + \kappa b)} \right) \quad (9)$$

where $Z_p = \sum_i^n z_i$, n is the number of protein ionizable sites, and κ is the inverse Debye screening length.^{57–59} Note that solvation free energies are implicitly accounted for by the intrinsic pK_a values, assuming that the residue is not deeply buried. The first term in eq 9 involves site–site interactions, while the second term accounts for all interactions with the surrounding salt.

For low ionic concentrations, it has been suggested^{60–62} that the concentration of counterions should be included in κ . We therefore employ the following definition of the screening length that includes both salt and counterions:

$$D = \kappa^{-1} = \sqrt{\frac{\epsilon_0 \epsilon_s k_B T}{2N_a e^2 I}} \quad (10)$$

where N_a is Avogadro's number, k_B is Boltzmann's constant ($k_B = 1.3807 \times 10^{-23} \text{ J} \cdot \text{K}^{-1}$), and T is the temperature. The ionic strength, I , is defined as,

$$I = \frac{1}{2} \sum_{i=1}^{N_{\text{ionic}}} c_i z_i^2 \quad (11)$$

where c_i is the molar concentration of species i and N_{ionic} is the total number of the ionic species (added salt particles and the counterions).

The MC protocol for implicit titration is similar to the one used for explicit salt in the previous section. The differences are that (1) salt particles and their MC moves are omitted and (2) the acceptance probability is

$$\min(1, e^{-\beta \Delta w_{\text{tk}} \pm (\text{pH} - \text{p}K_a) \ln 10}) \quad (12)$$

remembering that κ depends on the protonation state due to the inclusion of counterions in the ionic strength, c.f., eq 11. The new MC scheme is implemented in the Faunus package⁶³—a free, open-source framework for biomolecular simulation.

All Monte Carlo simulations—explicit and implicit salt alike—were performed with a protein concentration of 117.5 μM obtained through a cell radius of 150 Å. The salt concentration was varied between 5 and 150 mM. The protein structures were modeled in mesoscopic detail—i.e., each amino acid was treated as a single sphere.⁴⁹ Although this simplification greatly decreases the simulation time in explicit salt simulations, it has no effect with implicit salt—at least not for single protein systems. For the explicit and implicit salt models, typical simulation runs were carried out during the equilibration and production phases with 2×10^4 and 10^6 MC steps, respectively.

Finally, let us discuss a subtle difference between the two titration methods. In the case of explicit salt, charges literally move in and out of the bulk, and as such, ΔU_{el} in eq 7 accounts for the excess chemical potential of the protons. To satisfy eq 2, pH must be defined on the *concentration* scale, i.e., $\text{pH}^* = -\log c_{\text{H}^+}$. This double counting is absent from the implicit salt titration scheme where pH is correctly defined on the activity scale. It is however trivial to convert between the two scales via

$$\text{pH} = -\log(c_{\text{H}^+} \gamma_{\text{H}^+}) = \text{pH}^* - \log \gamma_{\text{H}^+} \quad (13)$$

In the presented explicit salt MC results, we estimate γ_{H^+} using the well-known Debye–Hückel expression²⁵ so as to transform to the *activity*-based pH scale. The maximum correction, needed at 150 mM salt, is less than 0.2 pH units.

3. Results and Discussion

In this section, we analyze and compare results obtained from the three different models: ideal (the analytical solution based on the amino acid composition), MC with explicit salt ions (the protocol with tertiary structures largely used in this field), and MC with the implicit salt description (the new scheme suggested here). Data are presented for the three whey proteins (α -LA, β -LG, and LF) at different pH and salt concentrations.

Protein Charge. We first investigate the effect of pH and salt on the protein net-charge number, $Z_p = \langle Z_p \rangle$. This is shown for α -LA, β -LG, and LF in Figure 3. In all cases, increasing the ionic strength tends to make the protein more charged, independent of the salt description. This is due to the reduced internal repulsion caused by salt screening, which pushes the titration curve closer to the ideal curve. This can

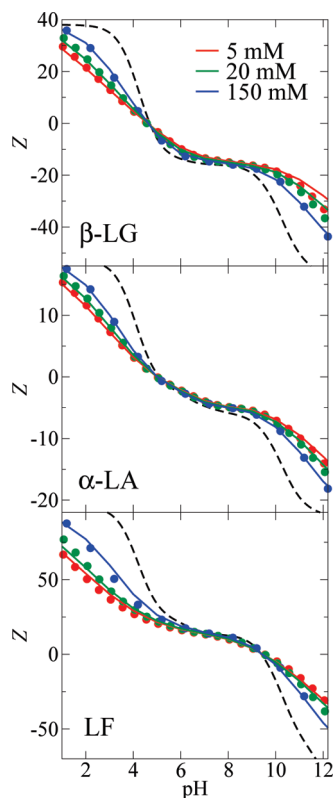


Figure 3. Net charges, Z_p , of the three proteins at different pH and salt concentrations. Calculated using explicit (symbols) and implicit (lines) salt MC simulations. The dashed lines represent the ideal titration curves—i.e., when no salt or intramolecular interactions are taken into account.

be understood from eq 9 by noting that when the salt concentration and hence κ become large, the second term approaches $Z_p^2/2b$, which is nothing but the self-energy of spreading Z_p on b . Neglecting multipolar terms, this self-energy approximately cancels out the first term whereby $\lim_{\kappa \rightarrow \infty} W_{tk} \approx 0$. This is in perfect agreement with experimental observations as demonstrated by the classical titration data for ribonuclease in KCl.⁴³ The agreement between the explicit and implicit schemes is very good. Even as the salt concentration is increased to 150 mM, one finds excellent agreement between the two MC procedures.

A weakness of the implicit salt model is that the isoelectric point, pI, is unaffected by the salt concentration since anisotropic protein–salt interactions are neglected in the truncated Kirkwood approach (eq 9). Nevertheless, explicit salt simulations—which account for *all* protein–salt interactions—do not reveal noticeable salt-dependent pI shifts for the three studied proteins. The calculated iso-electric points for the three proteins are 4.8, 5.4, and 9.7 for α -LA, β -LG, and (apo) LF, respectively. The corresponding experimental values are 4.2–4.5, 5.1, and 8.8.^{64,65} Despite its high affinity for iron, LF releases the iron ions from its binding sites below pH 6.0.⁶⁶ Differences between theoretical and experimental values for LF's pI have been observed before. Our MC data are in agreement with the theoretical value (9.4) reported by Steijns and van Hooijdonk.⁶⁷

Note also that in our model the protein structure is oblivious to pH changes. While this may be problematic at pH extremes, it is not a serious limitation at intermediate

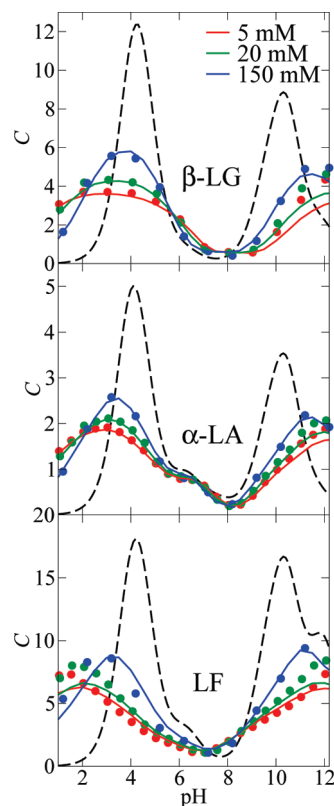


Figure 4. Electric capacitance, C , of the three proteins at different pH and salt concentrations, calculated using explicit (symbols) and implicit (lines) salt MC simulations. The dashed lines represent the ideal capacitance curves—i.e., when no salt or intramolecular interactions are taken into account.

pH values where they proteins are not subject to denaturation.⁶⁸ For mere comparison between explicit and implicit salt simulations, this is of course unimportant.

Protein Capacitance. Next, let us examine the protein charge capacitance, C —see eq 5. This property is important for macromolecular complexation as it quantifies charge regulation mechanisms.^{12,14,20,23} As C is a derivative, it is a more sensitive measure than the protein net charge. As can be seen in Figure 4, good agreement between the explicit and implicit salt descriptions is found—in particular for intermediate pH values where the net charge is relatively low. The peaks at high and low pH values are due to the large number of basic and acidic residues that titrate in these regions since proton fluctuations are maximized for $\text{pH} \approx \text{p}K_a$. Electrostatic interactions between titratable sites tend to lower and widen the capacitance peaks. This effect is well captured by both models, and when salt is added, we approach the ideal curve due to reasons already discussed.

Protein Dipole Moment. The molecular dipole moment, $\mu = |\sum_i^N r_i z_i|$, embodies the first step toward an arbitrary, anisotropic charge distribution. Since in the implicit titration model we treat protein–salt interactions as *isotropic*, we expect the dipole moment to be the hardest property to accurately reproduce. Yet, as can be seen in Figure 5, good agreement between the explicit and implicit model is found for α -LA and β -LG. LF, with dipole moments of up to 1600 D (interestingly at physiological pH), shows the largest discrepancies between the two models—especially at high

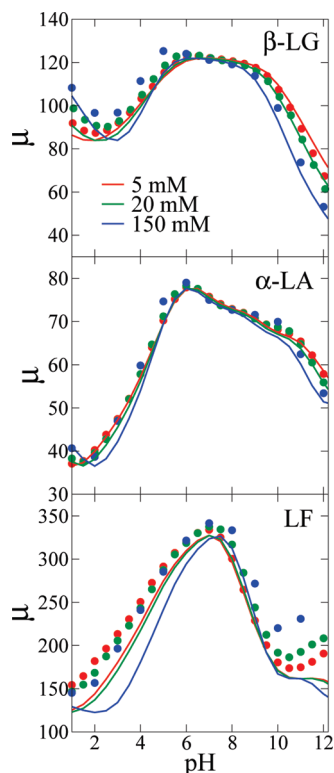


Figure 5. Molecular dipole moments, μ , of the three proteins at different pH and salt concentrations, calculated using explicit (symbols) and implicit (lines) salt MC simulations. The unit is electron $e \text{ \AA}$, which corresponds to 4.8 Debye (D).

and low pH values. LF deviates most from spherical symmetry, and it is expected that anisotropic protein–salt interactions play a larger role. As true also for the charge and capacitance, when $\text{pH} = \text{pI}$, the salt dependency disappears for the implicit titration scheme.

Many-Body Interactions. So far we have focused on protonation properties of single protein molecules in salt solutions. However, in the presence of other charged macromolecules,^{12,23} the protonation state may change. This perturbation can be captured by, for each proton move, evaluating the electrostatic interaction of the site with all *extra-molecular* charges in the system. The acceptance criteria then becomes

$$\min(1, e^{-\beta\Delta w_{ik} \pm (\text{pH} - \text{p}K_a) \ln 10 - \beta\Delta\sum\phi_i e z_i}) \quad (14)$$

where ϕ_i is the potential on i due to all external charges. Assuming that salt exclusion from the protein has only a minor effect on intermolecular interactions, we can approximate ϕ_i with a plain Debye–Hückel potential. Following this scheme, it is possible to introduce proton equilibria in many-body protein simulations at a minimal computational cost.

Performance. Lastly, we briefly discuss the computational performance of implicit versus explicit salt simulations. The computation time of an N -body simulation scales as N^2 , and explicit salt simulations hence decelerate for large salt concentrations when at the same time the protein concentration is low. This is illustrated in Figure 6, where we have plotted the number of salt particles per protein as a function

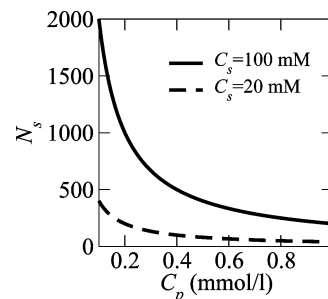


Figure 6. Number of salt particles, N_s , per protein plotted as a function of the protein concentration, C_p , for typical salt concentrations, C_s .

of the protein concentration. From this picture, it is clear that an implicit salt description leads to speed-ups of several orders of magnitude and that the computational time is independent of the salt concentration.

4. Conclusions

We present a Monte Carlo titration scheme within the Debye–Hückel implicit salt description. The model is tested against reference data obtained using explicit salt simulations for three different whey proteins in the concentration range 5–150 mM. The model very well reproduces electric properties such as net charge, dipole moment, and capacitance, even at relatively high ionic concentrations and in a wide span of pH values.

Elongated proteins show the largest deviations from the explicit salt results. This is explained by the fact that the implicit salt model neglects anisotropic salt–protein interactions and further assumes that salt exclusion from the protein is spherically symmetric. This becomes more evident at higher salt concentrations where the protein net charges tend to be larger.

Treating the salt implicitly drastically reduces the computation time and makes it independent of the ionic strength. Conversely, for explicit salt, the CPU cost scales with the square of the number of salt particles and can be orders of magnitude slower than the model presented here. We expect the presented model to be valuable for including proton equilibria in multiscale simulations with many ionizable macromolecules.

Acknowledgment. This work was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and the Swedish Research Council through a Linnaeus grant. We thank CENAPAD-SP for computational resources and, for useful comments, Jan Forsman, Bo Jönsson, and Björn Persson.

References

- (1) Perutz, M. F. Electrostatic Effects in Proteins. *Science* **1978**, *201*, 1187–1191.
- (2) Linse, S.; Johansson, C.; Brodin, P.; Grundström, T.; Drakenberg, T.; Forsén, S. Electrostatic contributions to the binding of Ca^{2+} in calbindin D_{9k} . *Biochemistry* **1991**, *30*, 154–162.

- (3) Woodward, C. E.; Svensson, B. Potentials of mean force in charged systems: Application to Superoxide Dismutase. *J. Phys. Chem.* **1991**, *95*, 7471–7477.
- (4) Foegeding, E. A.; Luck, P.; Davis, J. Factors determining the physical properties of protein foams. *Food Hydrocolloids* **2006**, *20*, 284–292.
- (5) Proctor, V. A.; Cunningham, F. E. The Chemistry of Lysozyme and its use as a Food Preservative and a Pharmaceutical. *CRC Crit. Rev. Food Nutrition* **1988**, *26*, 359–3958.
- (6) Dickinson, E. Interfacial structure and stability of food emulsions as affected by protein-polysaccharide interactions. *Soft Matter* **2008**, *4*, 932–942.
- (7) Biró, E.; Németh, A. S.; Sisak, C.; Feczko, T.; Gyenis, J. Preparation of chitosan particles suitable for enzyme immobilization. *J. Biochem. Biophys. Methods* **2008**, *70*, 1240–1246.
- (8) Saksena, S.; Zydney, A. L. Effect of solution pH and ionic strength on the separation of albumin from immunoglobulins (IgG) by selective filtration. *Biotechnol. Bioeng.* **1994**, *43*, 960–968.
- (9) Wang, Y.-F.; Gao, J. Y.; Dubin, P. L. Protein Separation via Polyelectrolyte Coacervation: Selectivity and Efficiency. *Biotechnol. Prog.* **1996**, *12*, 356–362.
- (10) Wang, Y.; Dubin, P. Protein binding on polyelectrolyte-treated glass. Effect of structure of adsorbed polyelectrolyte. *J. Chromatogr. A* **1998**, *808*, 61–70.
- (11) van Eijndhoven, R. H. C. M.; Saksena, S.; Zydney, A. L. Protein fractionation using electrostatic interactions in membrane filtration. *Biotechnol. Bioeng.* **1995**, *48*, 406–414.
- (12) Lund, M.; Jönsson, B. On the charge regulation of proteins. *Biochemistry* **2005**, *44*, 5722–5727.
- (13) Jönsson, B.; Lund, M.; da Silva, F. L. B. Electrostatics in Macromolecular Solution. In *Food Colloids: Self-Assembly and Material Science*; Dickinson, E., Leser, M. E., Eds.; Royal Society of Chemistry: London, 2007; Vol. 9.
- (14) Da Silva, F. L. B.; Jönsson, B. Polyelectrolyte-protein complexation driven by charge regulation. *Soft Matter* **2009**, *5*, 2862–2868.
- (15) Timasheff, S. N.; Dintzis, H. M.; Kirkwood, J. G.; Coleman, B. D. Studies of molecular interaction in isoionic protein solutions by light-scattering. *Proc. Natl. Acad. Sci. U.S.A.* **1955**, *41*, 710–714.
- (16) Timasheff, S. N.; Dintzis, H. M.; Kirkwood, J. G.; Coleman, B. D. Light Scattering Investigation of Charge Fluctuations in Isoionic Serum Albumin Solutions. *J. Am. Chem. Soc.* **1957**, *79*, 782–791.
- (17) Timasheff, S. N. Specific interactions in proteins due to proton fluctuations. *Biopolymers* **1966**, *4*, 107–120.
- (18) Ståhlberg, J.; Jönsson, B. Influence of Charge Regulation in Electrostatic Interaction Chromatography of Proteins. *Anal. Chem.* **1996**, *68*, 1536–1544.
- (19) Menon, M. K.; Zydney, A. L. Determination of effective protein charge by capillary electrophoresis: effects of charge regulation in the analysis of charge ladders. *Anal. Chem.* **2000**, *72*, 5714–5717.
- (20) da Silva, F. L. B.; Lund, M.; Jönsson, B.; Åkesson, T. On the Complexation of Proteins and Polyelectrolytes. *J. Phys. Chem. B* **2006**, *110*, 4459–4464.
- (21) de Vos, W. M.; Leermakers, F. A. M.; de Keizer, A.; Cohen Stuart, M. A.; Kleijn, J. M. Field Theoretical Analysis of Driving Forces for the Uptake of Proteins by Like-Charged Polyelectrolyte Brushes: Effects of Charge Regulation and Patchiness. *Langmuir* **2010**, *26*, 249–259.
- (22) Linderstrøm-Lang, K. Om proteinstoffernes ionization Compt. *Rend. Trav. Lab. Carlsberg* **1924**, *15*, 1–29.
- (23) Kirkwood, J. G.; Shumaker, J. B. Forces Between Protein Molecules in Solution Arising from Fluctuations in Proton Charge and Configuration. *Proc. Natl. Acad. Sci. U.S.A.* **1952**, *38*, 863–871.
- (24) Netz, R. R. Electrostatics of counter-ions in and between planar charged walls: From Poisson-Boltzmann to the strong-coupling theory. *Eur. Phys. J. E.* **2001**, *5*, 557–574.
- (25) Hill, T. L. *An Introduction to Statistical Thermodynamics*; Dover Publications: Mineola, NY, 1987.
- (26) Evans, D. F.; Wennerström, H. *The Colloidal Domain: Where Physics, Chemistry, Biology, and Technology Meet*; John Wiley & Sons, Ltd.: New York, 1999.
- (27) Holmberg, K.; Jönsson, B.; Kronberg, B.; Lindman, B. *Surfactants and Polymers in Aqueous Solution*, 2nd ed.; John Wiley & Sons, Ltd.: New York, 2002.
- (28) Jönsson, B.; Ullner, M.; Peterson, C.; Sommelius, O.; Söderberg, B. Titrating Polyelectrolytes - Variational Calculations and Monte Carlo Simulations. *J. Phys. Chem.* **1996**, *100*, 409–417.
- (29) Kesvatera, T.; Jönsson, B.; Thulin, E.; Linse, S. Ionization Behavior of Acidic Residues in Calbindin D_{9k}. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 106–115.
- (30) André, I.; Kesvatera, T.; Jönsson, B.; Åerfeldt, K. S.; Linse, S. The Role of Electrostatic Interactions in Calmodulin-Peptide Complex Formation. *Biophys. J.* **2004**, *87*, 1929–1938.
- (31) Labbez, C.; Jönsson, B.; Skarba, M.; Borkovec, M. Ion-Ion Correlation and Charge Reversal at Titrating Solid Interfaces. *Langmuir* **2009**, *25*, 7209–7213.
- (32) Kirkwood, J. G. Theory of Solutions of Molecules Containing Widely Separated Charges with Special Application to Zwitterions. *J. Chem. Phys.* **1934**, *2*, 351–361.
- (33) Tanford, C.; Kirkwood, J. G. Theory of Protein Titration Curves I. General Equations for Impenetrable spheres. *J. Am. Chem. Soc.* **1957**, *79*, 5333–5339.
- (34) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037–10041.
- (35) Thompson, A.; Boland, M.; Singh, H. *Milk Proteins: From Expression to Food*; Academic Press: New York, 2008.
- (36) Dickinson, E.; Leser, M. E. *Food Colloids: Self-Assembly and Material Science*; Royal Society of Chemistry: London, 2007.
- (37) Gottschalk, M.; Nilsson, H.; Roos, H.; Halle, B. Protein self-association in solution: The bovine β -lactoglobulin dimer and octamer. *Protein Sci.* **2003**, *12*, 2404–2411.
- (38) Fast, J.; Mossberg, A.-K.; Svanborg, C.; Linse, S. Stability of HAMLET - A kinetically trapped α -lactalbumin oleic acid complex. *Protein Sci.* **2005**, *14*, 329–340.
- (39) Chrysina, E.; Brew, K.; Acharya, K. Crystal structures of apo- and holo-bovine alpha-lactalbumin at 2.2-Å resolution reveal an effect of calcium on inter-lobe interactions. *J. Biol. Chem.* **200**, *275*, 37021–37029.

- (40) Brownlow, S.; Morais Cabral, J.; Cooper, R.; Flower, D.; Yewdall, S.; Polikarpov, I.; North, A.; Sawyer, L. Bovine beta-lactoglobulin at 1.8 Å resolution—still an enigmatic lipocalin. *Structure* **1997**, *5*, 481–495.
- (41) Phillis, G. D. J. Excess chemical potential of dilute solutions of spherical polyelectrolytes. *J. Chem. Phys.* **1974**, *60*, 2721–2731.
- (42) Grant, M. L. Nonuniform Charge Effects in Protein-Protein Interactions. *J. Phys. Chem. B* **2001**, *105*, 2858–2863.
- (43) Tanford, C.; Hauenstein, J. D. Hydrogen Ion Equilibria of Ribonuclease. *J. Am. Chem. Soc.* **1956**, *78*, 5287–5291.
- (44) Metropolis, N.; Ulam, S. The Monte Carlo method. *J. Am. Stat. Assoc.* **1949**, *44*, 335–341.
- (45) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: New York, 1989.
- (46) Frenkel, D.; Smit, B. *Understanding Molecular Simulation*; Academic Press: New York, 2001.
- (47) Ullner, M.; Jönsson, B. A Monte Carlo Study of Titrating Polyelectrolytes in the Presence of Salt. *Macromolecules* **1996**, *29*, 6645–6655.
- (48) Kesvatera, T.; Jönsson, B.; Thulin, E.; Linse, S. Measurement and Modelling of Sequence-specific pK_a Values of Calbindin D_{9k}. *J. Mol. Biol.* **1996**, *259*, 828.
- (49) Persson, B. A.; Lund, M. Association and electrostatic steering of alpha-lactalbumin-lysozyme heterodimers. *Phys. Chem. Chem. Phys.* **2009**, *11*, 8879–8885.
- (50) Nozaki, Y.; Tanford, C. Examination of titration behavior. *Methods Enzymol.* **1967**, *11*, 715–734.
- (51) Lund, M.; Jönsson, B.; Woodward, C. E. Implications of a high dielectric constant in proteins. *J. Chem. Phys.* **2007**, *126*, 225103–225110.
- (52) Jönsson, B.; Svensson, B. *Monte Carlo simulation of ion-protein binding. In Computer Simulation of Biomolecular Systems*, Vol. 2; van Gunsteren, W. F.; Weiner, P. K.; Wilkinson, A., Eds.; ESCOM: Leiden, 1993.
- (53) da Silva, F. L. B.; Jönsson, B.; Penfold, R. A critical investigation of the Tanford-Kirkwood scheme by means of Monte Carlo simulations. *Protein Sci.* **2001**, *10*, 1415–1425.
- (54) de Carvalho, S. J.; Fenley, M. O.; da Silva, F. L. B. Protein-Ion Binding Process on Finite Macromolecular Concentration. A Poisson-Boltzmann and Monte Carlo Study. *J. Phys. Chem. B* **2008**, *112*, 16766–16776.
- (55) Penfold, R.; Warwicker, J.; Jönsson, B. Electrostatic Models for Calcium Binding Proteins. *J. Phys. Chem. B* **1998**, *102*, 8599–8610.
- (56) Warwicker, J. Simplified methods for pK_a and acid pH-dependent stability estimation in proteins: Removing dielectric and counterion boundaries. *Protein Sci.* **1999**, *8*, 418–425.
- (57) Lin, S.-C.; Lee, W. I.; Shurr, J. M. Brownian Motion of Highly Charged Poly(L-lysine). Effects of salt and polyion concentration. *Biopolymers* **1978**, *17*, 1041–1064.
- (58) Schmitz, K. S. *Macro-ion Characterization: From Dilute Solutions to Complex Fluids*; American Chemistry Society: Washington, DC, 1994.
- (59) Kjellander, R.; Ulander, J. Effective ionic charges, permittivity and screening length: dressed ion theory applied to 1:2 electrolyte solutions. *Mol. Phys.* **1996**, *95*, 495–505.
- (60) Beresford-Smith, B.; Chan, D. Y. C. Electrical double-layer interactions in concentrated colloidal systems. *Faraday Discuss. Chem. Soc.* **1983**, *76*, 65–75.
- (61) da Silva, F. L. B.; Linse, S.; Jönsson, B. Binding of Charged Ligands to Macromolecules. Anomalous Salt Dependence. *J. Phys. Chem. B* **2005**, *109*, 2007–2013.
- (62) Dobnikar, J.; Castañeda Priego, R.; Grünberg, H. H. V.; Trizac, E. Testing the relevance of effective interaction potentials between highly-charged colloids in suspension. *New J. Phys.* **2006**, *8*, 277+.
- (63) Lund, M.; Trulsson, M.; Persson, B. Faunus: An object oriented framework for molecular simulation. *Source Code Biol. Med.* **2008**, *3*:1 [doi:10.1186/1751-0473-3-1].
- (64) Jimenez-Flores, R.; Bleck, G. T.; Brown, E. M.; Butler, J. E.; Creamer, L. K.; Hicks, C. L.; Hollar, C. M.; Ng-Kwai-Hang, K. F.; Swaisgood, H. E. Nomenclature of the Proteins of Cows- Milk - Sixth Revision. *J. Dairy Sci.* **2004**, *87*, 1641–1674.
- (65) Shimazaki, K.-I.; Kawaguchi, A.; Sato, T.; Ueda, Y.; Tomimura, T.; Shimamura, S. Analysis of human and bovine milk lactoferrins by rotofor and chromatofocusing. *Int. J. Biochem.* **1993**, *25*, 1653–1658.
- (66) Bezwoda, W. R.; Mansoor, N. Lactoferrin from human breast milk and from neutrophil granulocytes. Comparative studies of isolation, quantitation, characterization and iron binding properties. *Biomed. Chromatogr.* **1989**, *3*, 121–126.
- (67) Steijns, J. M.; van Hooijdonk, A. C. M. Occurrence, structure, biochemical properties and technological characteristics of lactoferrin. *Br. J. Nutr.* **2000**, *84*, 11–17.
- (68) Baker, E.; Baker, H. A structural framework for understanding the multifunctional character of lactoferrin. *Biochimie* **2009**, *91*, 3–10.

Simulating POPC and POPC/POPG Bilayers: Conserved Packing and Altered Surface Reactivity

Lorant Janosi and Alemayehu A. Gorfe*

Department of Integrative Biology and Pharmacology, University of Texas Health Science Center at Houston, 6431 Fannin Street, MSB 4.108, Houston, Texas 77030

Received July 8, 2010

Abstract: Molecular dynamics (MD) simulation is a popular technique to study bilayer structural properties, but it has not been widely used in mixed bilayers of neutral and charged lipids. Here, we present results from constant temperature and pressure MD simulations of a 2-oleoyl-1-palmitoyl-*sn*-glycero-3-phosphocholine (POPC) bilayer containing 23% 2-oleoyl-1-palmitoyl-*sn*-glycero-3-glycerol (POPG). The simulations were performed using the recently updated CHARMM force field and involved two bilayers of 104 and 416 lipids. A control simulation of a pure POPC bilayer of 128 lipids yielded equilibrium structural properties that compare very well with experimental data. The average equilibrium properties of the mixed bilayer systems were very similar to those of the pure POPC. However, nearly one-half of all the POPG lipids were found to be involved in hydrogen bonding with POPC lipids. Furthermore, the hydration of the mixed bilayer is different from that of the pure POPC, with the former inducing ordering of water molecules at longer distances. Thus, a phospholipid bilayer with ~23% negative charge content in the liquid crystalline phase differs from its neutral counterpart only at the headgroup.

1. Introduction

Many cytosolic proteins that have a cluster of basic residues or a polycationic domain, such as the signaling mediator K-ras, bind to the negatively charged inner leaflet of the plasma membrane.¹ The mechanism by which these proteins target the host bilayer continues to be a subject of intense experimental scrutiny.² Computational approaches, such as molecular dynamics simulation (MD), play key roles in the study of pure bilayers with and without bound proteins and peptides.^{3,4} They complement the experimental efforts by characterizing the atomic interactions responsible for the protein–membrane recognition.^{3,5} However, such an MD study requires constructing a well-equilibrated bilayer system containing the right balance of neutral and charged phospholipids. On the basis of data from fully equilibrated MD simulations, this work describes the structural and dynamic behaviors of a mixture of 2-oleoyl-1-palmitoyl-*sn*-glycero-3-phosphocholine (POPC) and 2-oleoyl-1-palmitoyl-*sn*-glycero-3-glycerol (POPG) lipid bilayer (Figure 1).

A number of MD studies targeted the POPC bilayer using several different force fields (FFs), including AMBER,⁶ CHARMM,^{7,8} GROMOS,⁹ and others.¹⁰ The ensemble-averaged bilayer properties derived from these simulations were in overall agreement with each other and with experimental data. However, there were discrepancies too, particularly in the characterization of the lateral organization of the bilayers by the area per lipid (A_L), which varied between 63.8 and 69.3 Å² in the simulations.¹¹ Interestingly, the A_L from experiments also varied between 63.0 and 68.3 Å².^{12–15} Both the simulations (298–303 K) and the experiments (297–310 K) were done at temperatures well above the gel–liquid crystalline (L_α) phase transition temperature of 268 K.¹⁶ Such ambiguity in the experimental A_L presented a particular challenge for the fully atomistic CHARMM27 (C27) FF,¹⁷ which best preserves bilayer properties for simulations under constant area or surface tension condition.^{18,19} A number of improvements^{20–22} have been made to mitigate the dramatic lateral contraction that occurs during an isothermal isobaric, that is, constant temperature and pressure (NPT) simulation with C27. The improvements ranged from updating only the acyl chain torsions,^{20,21} to reparametrizing

* Corresponding author phone: (713) 500-7538; fax: (713) 500-7444; e-mail: alemayehu.g.abebe@uth.tmc.edu.

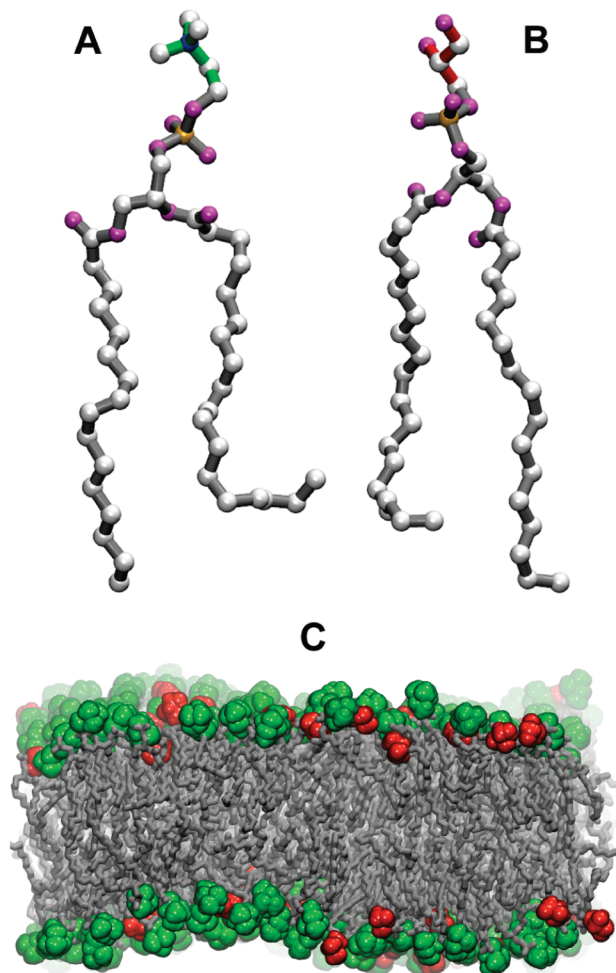


Figure 1. The structure of the POPC (A) and POPG (B) lipids represented by ball-and-stick models, with bonds at the headgroup colored green (choline) and red (glycerol), whereas carbon is in white, oxygen is in purple, nitrogen is in blue, and phosphorus is in tan. (C) A snapshot from the large simulation (SIIb, see Table 1) of the mixed bilayer containing 320 POPC (green) and 96 POPG (red) lipids. Water molecules, ions, and hydrogen atoms were omitted for clarity.

charges on the headgroup and upper chain atoms,²² to modification of selected torsional, Lennard-Jones, and partial atomic charge parameters.²³ The latter has been incorporated into the new CHARMM36 (C36) parameter set and was tested with six lipids (including POPC). In all cases, it yielded bilayer structural properties that are in very good agreement with experiments.²³

In the current work, the C36 FF was used to perform tensionless NPT simulations of pure POPC and a POPC/POPG mixture. The POPC/POPG bilayer was simulated in two system sizes: a small system of 104 lipids and a large one of 416 lipids. In both cases, we found that $\sim 75\%$ of all the POPG molecules are engaged in intra- or intermolecular hydrogen bonding, in agreement with several previous reports on the strong propensity of POPG lipids to form hydrogen bonds.^{6,24,25} The pure POPC simulation also yielded results that are consistent with experimental data. The simulations predict that bilayers of POPC without and with 23% POPG have nearly identical equilibrium structural properties,

including area per lipid, bilayer thickness, and area compressibility modulus.

2. Methods

Three MD simulations, a pure POPC bilayer (simulation SI), a small POPC/POPG bilayer (SIIa), and a large POPC/POPG bilayer (SIIb), were carried out in the NPT ensemble at $T = 310$ K and $P = 1.01325$ atm (Table 1). The simulations used the C36 FF and were run with the NAMD program.²⁶ Visualization and some of the analyses were carried out with VMD.²⁷

The initial model for the pure POPC bilayer (SI) was built using the CHARMM GUI bilayer builder^{28,29} and consisted of 128 lipids (64 per leaflet) solvated by 6052 water molecules in a box of $67.1 \times 67.1 \times 76.1 \text{ \AA}^3$. The starting point for the small POPC/POPG bilayer (SIIa) was originally constructed as pure POPC bilayer of 104 lipids. After a short minimization and equilibration (see below), 24 POPC lipids were randomly selected and replaced by POPG lipids, resulting in a system of 52 lipids per leaflet (40 POPC and 12 POPG) solvated by 6224 water molecules in a box of $58.7 \times 58.7 \times 96.0 \text{ \AA}^3$. Twenty-four sodium ions were added to neutralize the total charge of the system. After SIIa was run for 70 ns, a snapshot was taken and multiplied into four copies, which were then assembled into a large bilayer consisting of 320 POPC and 96 POPG lipids (a total of 130 096 atoms).

Each system was minimized for 2000 steps with all non-hydrogen atoms fixed and for another 2000 steps with only the phosphorus atoms harmonically restrained ($k = 4 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$). This was followed by 200 ps equilibration, and by another 300 ps equilibration with k scaled by 0.75, 0.5, and 0.25 every 100 ps. A single production simulation for each system was then run for the durations listed in Table 1. The integration time step, δt , was 1 fs during the equilibration runs. For the production phase, $\delta t = 2$ fs was used in conjunction with constraints applied to all bonds involving hydrogens using the SHAKE algorithm.³⁰ The NAMD multi-timestepping procedure was used with the bonded and nonbonded forces computed at every δt and the particle mesh Ewald (PME) calculations at every $2\delta t$; the step cycle was 10. The cutoffs for nonbonded interactions and for pair-list updates were 12 and 14 \AA , respectively, with the switch function turned on at 10 \AA .

3. Results and Discussion

In the following sections, we first examine the performance of the C36 FF and then describe the structure and dynamics of the POPC bilayer in the presence and absence of POPG lipids.

3.1. Performance of C36 Relative to C27. Membrane simulation with C27 FF works best when the area is fixed (see Introduction), which can be done for pure bilayers of the common lipids whose A_L is well-documented (e.g., DMPC and DPPC). Constant area simulation becomes problematic when unambiguous A_L is not available or when the lateral dimensions of the simulation box must change during the simulation. The former is typically true for

Table 1. Simulations Performed and Bilayer Structural Properties (Average \pm S.D.)^a

sim.	composition	no. of lipids (atoms)	length (ns)	A_L (\AA^2) exp./prev. sim. ^b	K_A (mN/m) exp./prev. sim. ^c	D_{P-P} (\AA) exp./prev. sim. ^d	D_{C2-C2} (\AA)
SI	POPC	128 (35 340)	75	64.7 \pm 1.3 54, ¹³ 63, ¹⁵ 68.3, ¹² 66 ¹⁴ /65.5, ⁷ 63.8, ¹¹ 66.8, ¹⁰ 63.5 ⁶	272 180–330 ³⁴ / 404 ¹¹	39.1 \pm 0.6 37 ¹² / 34.6, ¹¹ 35.5 ⁶	28.2 \pm 0.6
SIa	POPC/POPG	104 (32 524)	70	63.6 \pm 1.2	346	39.7 \pm 0.6	28.7 \pm 0.5
SIb	POPC/POPG	416 (130 096)	120	64.4 \pm 0.8	295 \pm 74 (238)	39.1 \pm 0.4	28.2 \pm 0.3

^a The first 20 ns of the data from each trajectory was excluded from the analyses. D_{P-P} represents bilayer thickness, measured as the average distance along the membrane normal between the centers of mass of the phosphorus atoms of the two leaflets; A_L is the area per lipid, measured as the average area of the simulation box (obtained from its lateral dimensions) normalized by the number of lipids per leaflet; K_A is the area compressibility modulus at constant temperature, calculated as in eq 1 over the last 50 ns (and over the last 100 ns in parentheses for SIb); and D_{C2-C2} represents the hydrophobic thickness, measured as the average distance along the membrane normal between the centers of mass of the first methyl carbon atoms of the two leaflets. ^b Experiments at 275 K,¹³ 297 K,¹⁵ 303 K,¹² and 310 K;¹⁴ simulations at 300 K,⁷ 303 K,¹¹ 303 K,¹⁰ and 310 K.⁶ ^c Simulation at 303 K.¹¹ ^d Experiment at 303 K,¹² simulations at 303 K¹¹ and 310 K.⁶ The current simulations were carried out at 310 K.

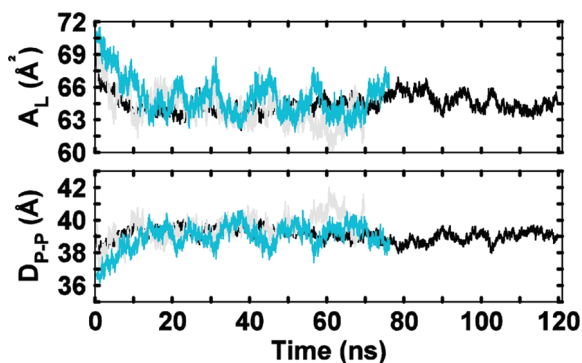


Figure 2. Time evolution of the area per lipid (A_L) and bilayer thickness (D_{P-P}) in the pure POPC (SI, cyan) simulation, and small (SIa, light gray) and large (SIb, black) POPC/POPG bilayer simulations.

multicomponent bilayers. The latter occurs during insertion of a molecule of non-negligible size into small bilayer patches. In the current work, we used the C36 FF that allowed us to run NPT simulations in which the system is able to adjust its area to minimize unfavorable atomic contacts. For comparison, we performed a 60 ns NPT simulation with the C27 FF using a copy of the system in SIa. Figure S1 plots the time evolution of the bilayer thickness (D_{P-P}) from this simulation, which can be compared to the corresponding plot in Figure 2 obtained from simulation SIa (i.e., performed with the C36 FF). The continuous compaction (or thickening) of the bilayer in Figure S1 relative to the equilibration achieved in Figure 2 clearly shows that NPT simulation with C27 yields unacceptable results. In contrast, the simulations with C36 produced results that are in very good agreement with available experimental data (Table 1).

3.2. Bilayer Structural Properties. **3.2.1. Pure POPC.** The time evolution of the bilayer thickness, D_{P-P} , and area per lipid, A_L , in the pure POPC simulation (Figure 2, cyan) shows that the bilayer has equilibrated after approximately 20 ns. The symmetric number density distributions along the bilayer normal, z , calculated for various components of POPC (Figure 3) are also consistent with a well-equilibrated bilayer system. The distributions further indicate that water penetrates the bilayer up to the glycerol-ester region, while the methyl groups of the lipid tails are fully dehydrated, in agreement with experiments.^{12,31,32} Furthermore, the ensemble-averaged D_{P-P} , A_L , and D_{C2-C2} (hydrophobic thickness) of

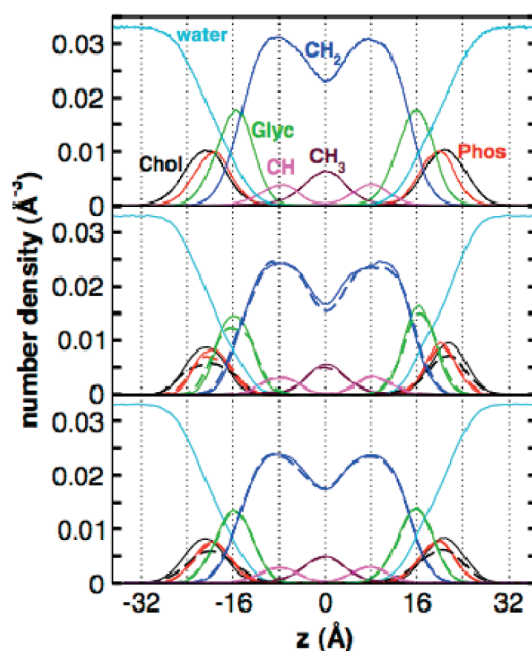


Figure 3. Number density distributions for the various components of POPC (solid lines) and POPG (dashed lines) lipid derived from simulation SI (top panel), SIa (middle panel), and SIb (bottom panel). In this and subsequent figures, only the equilibrated portions of the trajectories were used, that is, excluding the first 20 ns. All the groups (labeled in the top panel) contain their heavy atoms only. The dashed black line in the two bottom panels corresponds to the glycerol headgroup of the POPG lipid. The number densities for POPG were rescaled by a factor that matches the CH_2 profiles to those of POPC.

the POPC bilayer agree well with results from previous simulations and experiments (Table 1). For example, Rog et al.⁶ and Poger et al.¹¹ reported A_L values of 63.5 and 63.8 \AA^2 from simulations with the AMBER94³³ FF and an improved version of the GROMOS96 FF, respectively. Our A_L of 64.6 \AA^2 is close to these values, as well as to the experimental data reported by Smaby et al. (63 \AA^2)¹⁵ and Hyslop et al. (66 \AA^2)¹⁴. However, the current A_L is somewhat smaller than the 68.3 \AA^2 obtained by Kucerka et al.¹² The difficulty associated with the lack of unambiguous experimental data, particularly for A_L , has been discussed in a recent work that systematically compared various structural parameters from MD and experiments.¹¹

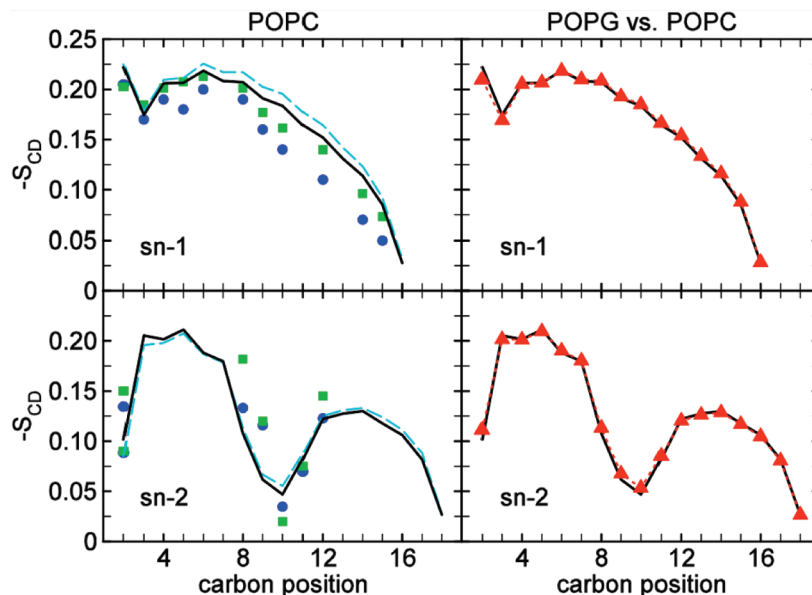


Figure 4. Deuterium order parameter ($-S_{CD}$) profiles. Left column shows a comparison between POPC order parameters (sn-1, top and sn-2, bottom) obtained from the pure POPC system (dashed line), the large POPC/POPG (black solid line), and experimental data at 300 K (■) and 315 K (●) from Seelig and colleagues.^{16,41} The column on the right compares profiles for POPG sn-1 and sn-2 tails (symbols) with those for POPC (lines) extracted from SIIb.

Fluctuation in A_L at constant temperature is related to an important bilayer property, the isothermal area compressibility modulus, K_A , given by

$$K_A = \frac{2k_B \langle T \rangle \langle A_L \rangle}{N_L \sigma^2} \quad (1)$$

where k_B is the Boltzmann constant, σ^2 is the variance associated with A_L , N_L is the number of lipids, and the angle brackets denote time and ensemble averages. The large fluctuations in the area (Figure 2) resulted in large variance, σ^2 , leading to $K_A = 272 \text{ mN m}^{-1}$ (Table 1), which falls within the experimental range of 180–330 mN m^{-1} .³⁴

The behavior of the hydrocarbon tails was examined using the deuterium order parameter, S_{CD} , calculated from the trajectory as

$$S_{CD} = \frac{1}{2} \langle 3 \cos^2 \theta - 1 \rangle \quad (2)$$

where θ is the angle between a C–H bond of the methylene/methyl group in a given acyl chain and the bilayer normal. Overall, the calculated S_{CD} values are in very good agreement with the experimental data (Figure 4), particularly at the often poorly predicted chain termini. The agreement with experiment is best for the sn-2 chain, whereas the S_{CD} 's associated with carbons in the middle of the sn-1 chain are slightly overestimated. These results are consistent with those reported by Klauda et al. in their evaluation of the C36 force field.²³ In so far as S_{CD} partly measures chain ordering, and given the connection between chain order and membrane thickness, it is safe to assume that the somewhat high D_{P-P} may be due to the enhanced ordering of the saturated sn-1 tail. This could be an area for additional improvements in future updates of the FF.

3.2.2. POPC/POPG Mixed Bilayer. Global Structure. The structural properties of the small POPC/POPG bilayer derived from simulation SIIa are surprisingly close to those of the pure POPC bilayer of comparable size (Table 1, Figures 2 and 3). Both the instantaneous (Figure 2) and the ensemble-averaged (Table 1) values of A_L , D_{P-P} , and D_{C2-C2} are nearly identical in the two simulations. The K_A calculated from SIIa is also fairly close to that from the pure POPC bilayer simulation (Table 1). The number density distributions for POPC and POPG in the binary mixture are also very similar to each other and to those of the pure POPC bilayer (Figure 3, top and middle panels). Minor differences are present only between the distributions of the glycerol-ester oxygen atoms of the two leaflets in SIIa (Figure 3). However, the position of the peaks did not change when the distributions were computed using the last 10, 20, 30, and 40 ns data (not shown), indicating that the equilibrium structural properties discussed above remain valid.

To ensure that these results are not artifacts due to the small size of the binary mixture, we simulated a larger system for a longer duration under identical simulation conditions (see Methods). Comparison of the results from this simulation (SIIb) with those of the pure POPC or the smaller POPC/POPG (SIIa) trajectories (Table 1, Figures 2 and 3) clearly shows that the calculated averages from each of the three trajectories are within error of each other. The only noticeable differences are in the symmetry of the distributions for the glycerol-ester oxygens (Figure 3, lower panel) and in the amplitude of the fluctuations in A_L and D_{P-P} (Figure 2, see also the SD in Table 1). The large simulation yielded perfectly symmetric distributions and smaller fluctuations due perhaps to the improved statistics afforded by the increased number of lipids and longer simulation time. The change in fluctuation has a direct impact on K_A (see eq 1), so that the obtained K_A from SIIb became closer to that from the pure

POPC bilayer (Table 1). Thus, in the following sections, we will use the larger system (SIIB) to characterize the POPC/POPG bilayer structure and dynamics in more detail.

In the past, several NPT MD simulation studies have been done on the negatively charged POPG in the pure form^{24,25} or mixed with POPE,⁶ but to our knowledge none in its mixture with POPC. Comparing the current results with those studies is nevertheless instructive. By analyzing POPC and POPE/POPG trajectories, Rog et al.⁶ obtained very similar area and thickness for POPC ($A_L = 63.5 \text{ \AA}^2$ and $D_{P-P} = 35.6 \text{ \AA}$) and POPG ($A_L = 62.8 \text{ \AA}^2$ and $D_{P-P} = 35.5 \text{ \AA}$) lipids. On the other hand, Elmore²⁵ calculated $A_L(\text{POPG}) = 56.1 \text{ \AA}^2$ and $A_L(\text{POPC}) = 68.4 \text{ \AA}^2$ based on 50 and 10 ns trajectories of pure POPG and POPC bilayers, respectively. In the case of the POPC bilayer, the discrepancy between our and Elmore's results may be explained by the different lengths of the simulations, because we were able to reproduce their data from the first 10 ns of our trajectory (not shown). We would like to note that parameters used by Elmore for POPG were taken from the PRODRG server,^{35,36} which does not give a rigorous, consistent set of parameters. Despite the lack of experimental data to validate a PG force field, the CHARMM parameter set for PGs is consistent with the general CHARMM force field for biomolecules (in terms of bonded terms, van der Waals terms, and charges). Therefore, it would not be surprising if Elmore's parameters showed some discrepancy. However, a more recent study²⁴ based on much longer simulations of pure POPG and POPC reported $A_L(\text{POPG}) = 53.0 \text{ \AA}^2$ and $A_L(\text{POPC}) = 65.8 \text{ \AA}^2$. Although differences in length, force field, and composition among these simulations preclude a direct comparison, the large variations in these results suggest that more work is required to determine the structural properties of POPG in the pure form and in its mixture with other lipids. Our results suggest that the global structural properties of a POPC bilayer containing 23% POPG are very similar to those of the pure POPC bilayer.

Chain Order and Dynamics. Because POPC and POPG lipids differ only at the headgroup, where choline ($-\text{CH}_2-\text{CH}_2-\text{N}^+(\text{CH}_3)_3$) of the former is replaced by glycerol ($-\text{CH}_2-[(\text{CH})(\text{OH})]-\text{CH}_2-\text{OH}$) in the latter, we wanted to know if the nearly identical average A_L , D_{P-P} , and $D_{C_2-C_2}$ discussed above are a consequence of similar packing and dynamics of the identical hydrophobic tails of the two lipid types. Thus, we compared the POPC and POPG lipid tails using the deuterium order parameter, S_{CD} , and the rotational autocorrelation function. The S_{CD} profiles for the POPC tails in the binary mixture show ordering similar to that of the pure bilayer and are almost identical to the POPG ones (Figure 4). The lateral diffusion coefficients, D , calculated from the slope of the mean square displacement of individual lipids (see Figure S2) and averaged over the number of the POPC and POPG lipids are also within error of one another. In the pure phase, $D_{\text{POPC}} \approx 8.9 \pm 0.7 \times 10^{-8} \text{ cm}^2/\text{s}$ (error estimated by block averaging, see Supporting Information), which compares very well with the experimental data for the fully hydrated POPC bilayer at 313 K, $\sim 14 \times 10^{-8} \text{ cm}^2/\text{s}$,³⁷ and at 322 K, $\sim 19 \times 10^{-8} \text{ cm}^2/\text{s}$,³⁸ as well as with results from previous simulations (e.g., $6.5 \times 10^{-8} \text{ cm}^2/\text{s}$).³⁹

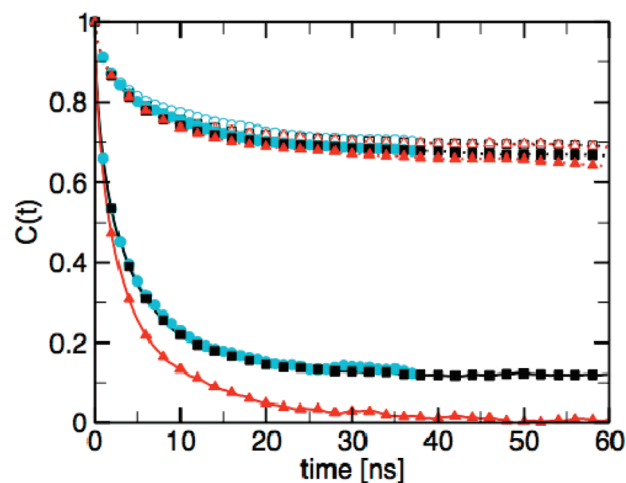


Figure 5. First rank rotational autocorrelation functions calculated for the headgroups (solid lines) and hydrophobic tails (dotted lines, sn-1, empty symbols; sn-2, filled symbols) for POPC in the pure (circles) and mixed (squares) bilayer, and for POPG in the mixed bilayer (triangles). The autocorrelation functions are computed for the headgroups using vectors P→N (C_{12} for POPG – central carbon of the glycerol headgroup) and the tails using the vector between the first ethyl and the terminal methylene carbon atoms.

In the binary mixture, we obtained $D_{\text{POPC}} \approx 9.1 \pm 0.6 \times 10^{-8} \text{ cm}^2/\text{s}$ and $D_{\text{POPG}} \approx 9.2 \pm 0.5 \times 10^{-8} \text{ cm}^2/\text{s}$.

The average rotational autocorrelation functions, calculated for each tail and for the headgroups, are shown in Figure 5. In contrast to the clear differences in the fast-relaxing headgroups, where the two lipids are chemically different, the POPC and POPG tails exhibit nearly identical rotational behavior. The rotational autocorrelation functions for the sn-1 tails overlap, while those of the sn-2 tails of POPG decorrelate slightly faster than for POPC. We conclude that the lipid tails of POPC and POPG in a mixture behave in the same manner as the pure POPC tails. As a consequence, the bilayer thickness, area per lipid, and other structural quantities that are predominantly dictated by lipid packing at the L_α phase remain unaffected. The effect of the POPG lipids is thus limited to the lipid–water interfacial region.

Lipid–Lipid Interaction. Previous studies of POPG and POPE/POPG bilayers have found that POPG molecules have a high propensity to form intra-POPG or interlipid hydrogen bonds.^{24,25} Thus, we monitored intra- and intermolecular hydrogen bonds using donor–acceptor distance and angle cutoffs of 3.1 \AA and 150° . The distributions of the mean number of hydrogen bonds formed per molecule of POPG and POPC are shown in Figure 6A. The histograms show that about 30% of the POPG lipids are engaged in POPG–POPG hydrogen bonding. This interaction is almost exclusively intramolecular, where the hydroxyl headgroup donates a hydrogen to a phosphate (and to some extent the glycerol ester) oxygen atom. Another 45% of the POPG molecules are involved in hydrogen bonding with POPC lipids. Therefore, 75% of all the POPG lipids are engaged in hydrogen bonding. In contrast, an insignificant number of POPC–POPC hydrogen bonds were detected. An important conclusion from these data is the following. Negatively

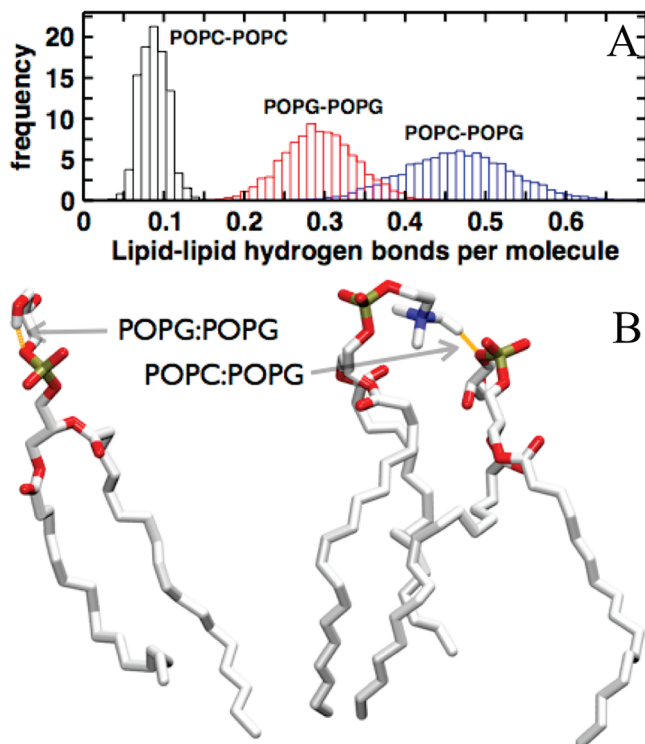


Figure 6. Lipid–lipid hydrogen bonds in the large POPC/POPG bilayer. (A) Histograms showing the frequency distributions for the fraction of POPC molecules involved in hydrogen bonding with other POPC molecules (black), as well as the fraction of POPG molecules involved in intramolecular (red), and in intermolecular hydrogen bonds with POPC (blue). (B) Examples of the intramolecular POPG and interlipid POPC–POPG hydrogen bonds. Hydrogen bonds were defined by a donor–acceptor distance cutoff of 3.1 Å and donor–hydrogen–acceptor angle cutoff of 150°.

charged lipids with the same fatty acid chains as the predominant lipid in the host bilayer preserve the overall structure of the bilayer while inducing a drastically different surface reactivity. These conditions are potentially important for polycationic proteins that preferentially target membrane patches of regular thickness and mechanical property, unlike proteins that target ordered membrane domains or lipid rafts.

Hydration. It is well-known that the phosphate oxygens of lipids polarize the surrounding water molecules in part through the formation of hydrogen bonds.⁴⁰ As a result, the average water dipole moment is directed toward the center of the membrane, resulting in a preferred orientation, or ordering, of the interfacial water molecules. To check if the hydration of the mixed bilayer differs from that of the pure POPC, we computed the orientational order parameter of the water molecules in both systems as the average cosine of the angle between the O–H bond and the bilayer normal. The results are plotted in Figure 7 as a function of the water molecules' distance from the membrane center. Water is most ordered around 20 Å, which is the average position of the phosphate group for both the pure POPC and the POPC/POPG bilayers (see Figure 3 and Table 1). As the distance to the bilayer surface increases, the ordering of the water molecules disappears quickly (~ 26 Å) for the pure POPC membrane. However, the average ordering of the water around the binary mixture is decreasing very slowly and

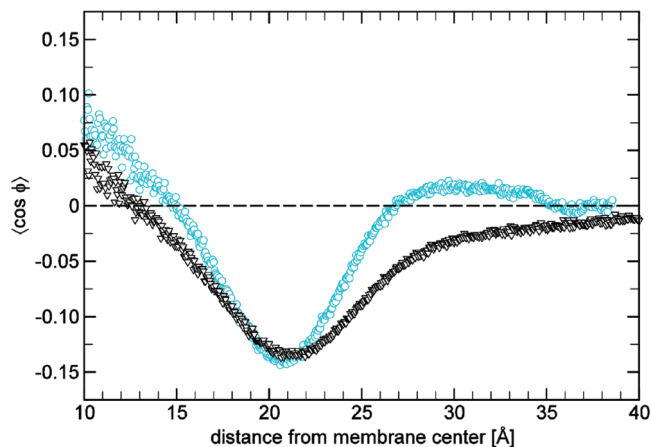


Figure 7. Rank one order parameter of water molecules as a function of distance from the center of the membrane in the pure POPC (○) and large POPC/POPG (▽) bilayers.

vanishes at about 40 Å. This is in qualitative agreement with previous simulations that have shown that pure POPG bilayers induce long-range ordering of water molecules around the bilayer due both to their net negative charge and to the polar hydroxyl oxygens of their glycerol headgroup.²⁴ Consistent with the conservation of the bilayer structure, the number of hydration water molecules per lipid is similar in the pure POPC bilayer and the binary mixture (see Supporting Information). These results support the conclusions of the previous paragraph that the charged POPG increases the surface reactivity, leading to interlipid interactions and enhanced polarization of the surrounding solvent without affecting the overall structure of the bilayer.

4. Conclusion

In this Article, we presented results from MD simulations of bilayers made up of pure POPC and a binary mixture of POPC lipids containing 23% anionic POPG lipids. The simulations were carried out in the NPT ensemble using the C36 force field. The obtained results for the pure POPC are in very good agreement with experimental data, validating the new C36 force field. The average structural properties of the binary mixture, such as area per lipid, bilayer thickness, and isothermal area compressibility modulus, remain nearly the same as for the pure POPC bilayer. Furthermore, deuterium order parameters for the tails of both POPC and POPG lipids of the binary mixture maintain their values from the pure POPC bilayer. Nonetheless, considerable differences were observed in the behavior of the headgroups, including the strong hydrogen-bonding potential, both intramolecular and interlipid, of the POPG lipids and the enhanced long-range ordering of water molecules at the hydrophobic–hydrophilic and headgroup–water interfaces of the mixed bilayer.

Supporting Information Available: Additional figures and hydration waters per lipid. This material is available free of charge via the Internet at <http://pubs.acs.org>.

Acknowledgment. We gratefully acknowledge Dr. J. Klauda (University of Maryland) for making the updated CHARMM36 force available to us before its publication,

the Texas Advanced Computing Center, and the National Center for Supercomputing Applications for computational resources.

References

- (1) McLaughlin, S.; Murray, D. *Nature* **2005**, *438*, 605–611.
- (2) Yeung, T.; Gilbert, G. E.; Shi, J.; Silvius, J.; Kapus, A.; Grinstein, S. *Science* **2008**, *319*, 210–213.
- (3) Lindahl, E.; Sansom, M. S. *Curr. Opin. Struct. Biol.* **2008**, *18*, 425–431.
- (4) Gorfe, A. A. *Curr. Med. Chem.* **2010**, *17*, 1–9.
- (5) Biggin, P. C.; Bond, P. J. *Methods Mol. Biol.* **2008**, *443*, 147–160.
- (6) Rog, T.; Murzyn, K.; Pasenkiewicz-Gierula, M. *Acta Biochim. Pol.* **2003**, *50*, 789–798.
- (7) Heller, H.; Schaefer, M.; Schulten, K. *J. Phys. Chem.* **1993**, *97*, 8343–8360.
- (8) Li, Z.; Venable, R. M.; Rogers, L. A.; Murray, D.; Pastor, R. W. *Biophys. J.* **2009**, *97*, 155–163.
- (9) Poger, D.; Van Gunsteren, W. F.; Mark, A. E. *J. Comput. Chem.* **2010**, *31*, 1117–1125.
- (10) Jojart, B.; Martinek, T. A. *J. Comput. Chem.* **2007**, *28*, 2051–2058.
- (11) Poger, D.; Mark, A. E. *J. Chem. Theory Comput.* **2010**, *6*, 325–336.
- (12) Kucerka, N.; Tristram-Nagle, S.; Nagle, J. F. *J. Membr. Biol.* **2005**, *208*, 193–202.
- (13) Pabst, G.; Rappolt, M.; Amenitsch, H.; Laggner, P. *Phys. Rev. E* **2000**, *62*, 4000–4009.
- (14) Hyslop, P. A.; Morel, B.; Sauerheber, R. D. *Biochemistry* **1990**, *29*, 1025–1038.
- (15) Smaby, J. M.; Momsen, M. M.; Brockman, H. L.; Brown, R. E. *Biophys. J.* **1997**, *73*, 1492–1505.
- (16) Seelig, J.; Waespe-Sarcevic, N. *Biochemistry* **1978**, *17*, 3310–3315.
- (17) Feller, S. E.; MacKerell, A. D. *J. Phys. Chem. B* **2000**, *104*, 7510–7515.
- (18) Jensen, M. O.; Mouritsen, O. G.; Peters, G. H. *Biophys. J.* **2004**, *86*, 3556–3575.
- (19) Feller, S. E.; Pastor, R. W. *J. Chem. Phys.* **1999**, *111*, 1281–1287.
- (20) Klauda, J. B.; Brooks, B. R.; MacKerell, A. D.; Venable, R. M.; Pastor, R. W. *J. Phys. Chem. B* **2005**, *109*, 5300–5311.
- (21) Klauda, J. B.; Pastor, R. W.; Brooks, B. R. *J. Phys. Chem. B* **2005**, *109*, 15684–15686.
- (22) Sonne, J.; Jensen, M. O.; Hansen, F. Y.; Hemmingsen, L.; Peters, G. H. *Biophys. J.* **2007**, *92*, 4157–4167.
- (23) Klauda, J. B.; Venable, R. M.; Freites, J. A.; O'Connor, J. W.; Tobias, D. J.; Mondragon-Ramirez, C.; Vorobyov, I.; Mackerell, A. D.; Pastor, R. W. *J. Phys. Chem. B* **2010**, *114*, 7830–7843.
- (24) Zhao, W.; Rog, T.; Gurtovenko, A. A.; Vattulainen, I.; Karttunen, M. *Biophys. J.* **2007**, *92*, 1114–1124.
- (25) Elmore, D. E. *FEBS Lett.* **2006**, *580*, 144–148.
- (26) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (27) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.
- (28) Jo, S.; Kim, T.; Im, W. *PLoS One* **2007**, *2*, e880.
- (29) Jo, S.; Lim, J. B.; Klauda, J. B.; Im, W. *Biophys. J.* **2009**, *97*, 50–58.
- (30) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (31) Kucerka, N.; Liu, Y. F.; Chu, N. J.; Petrache, H. I.; Tristram-Nagle, S. T.; Nagle, J. F. *Biophys. J.* **2005**, *88*, 2626–2637.
- (32) Gawrisch, K.; Gaede, H. C.; Mihailescu, M.; White, S. H. *Eur. Biophys. J.* **2007**, *36*, 281.
- (33) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (34) Binder, H.; Gawrisch, K. *J. Phys. Chem. B* **2001**, *105*, 12378–12390.
- (35) Schuttelkopf, A. W.; van Aalten, D. M. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 1355–1363.
- (36) van Aalten, D. M.; Bywater, R.; Findlay, J. B.; Hendlich, M.; Hoof, R. W.; Vriend, G. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 255–262.
- (37) Filippov, A.; Oradd, G.; Lindblom, G. *Biophys. J.* **2003**, *84*, 3079–3086.
- (38) Gaede, H. C.; Gawrisch, K. *Biophys. J.* **2003**, *85*, 1734–1740.
- (39) Bockmann, R. A.; Hac, A.; Heimburg, T.; Grubmüller, H. *Biophys. J.* **2003**, *85*, 1647–55.
- (40) Aman, K.; Lindahl, E.; Edholm, O.; Hakansson, P.; Westlund, P. O. *Biophys. J.* **2003**, *84*, 102–115.
- (41) Seelig, A.; Seelig, J. *Biochemistry* **1977**, *16*, 45–50.

CT100381G

Activating the Prolactin Receptor: Effect of the Ligand on the Conformation of the Extracellular Domain

Flora S. Groothuizen,^{†,§} David Poger,[†] and Alan E. Mark^{*,†,‡}

School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD 4072, Australia, and Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia

Received July 16, 2010

Abstract: The prolactin receptor resides on the surface of the cell as a preformed dimer. This suggests that cell signaling is triggered by conformational changes within the extracellular domain of the receptors. Here, by using atomistic molecular dynamics simulations, we show that the removal of the ligand placental lactogen from the dimeric form of the prolactin receptor results in a relative reorientation of the two extracellular domains by 20–30°, which corresponds to a clockwise rotation of the domains with respect to each other. Such a mechanism of activation for the prolactin receptor is similar to that proposed previously in the case of the growth hormone receptor. In addition to the effect of the removal of the ligand, the mechanical coupling between the extracellular and transmembrane domains within a model membrane was also examined.

1. Introduction

The prolactin (PRL) receptor (PRLR) regulates the production of milk in mammals and is involved in physiological functions, ranging from fetal growth to the regulation of hormonal balance and development of the reproductive system.¹ It is a class I cell surface cytokine receptor, a family that also includes the growth hormone receptor (GHR), the erythropoietin receptor (EpoR), and several interleukin receptors.² Three ligands that are similar in sequence and in their four- α -helix bundle structure are known to bind to PRLR: prolactin, growth hormone (GH), and placental lactogen (PL). The three PRLR-binding ligands have several overlapping functions in the body, and their metabolic states are interdependent and tightly regulated in mammals during pregnancy.³

The full-length 68.6 kDa PRLR consists of 592 amino acids, of which 210 form the N-terminal extracellular domain (ECD) that is connected to a 358-residue intracellular domain

(ICD) by a single-pass transmembrane (TM) helix of 24 amino acids. Alternative splicing of the PRLR gene results in multiple isoforms,⁴ of which the full-length receptor is generally termed the long form.⁵ The long form of PRLR mainly exerts its activity via the JAK2/STAT5 pathway, starting a cascade of tyrosine phosphorylations that eventually leads to the activation of target genes. The functions of the other isoforms remain uncertain, although one isoform termed the short form, containing a shorter ICD (57 residues), has been reported to have an inhibitory effect on the activity of the long form.^{6,7}

The structure of the ECD of the long form of the PRLR has been determined as a single receptor bound to GH⁸ or to a PRL antagonist.⁹ However, consistent with other class I cytokine receptors, the active form of the PRLR is assumed to be a homodimer that binds its ligand in a 2:1 ratio. A crystal structure of the homodimeric rat prolactin receptor (rPRLR) ECD in complex with ovine placental lactogen (oPL) has been published.¹⁰ Note that a cross-species PRLR-complex was used as the proteins in this complex can have higher affinities for each other than in the same-species complexes, which only exist transiently.¹¹ Despite the ECDs of GHR and PRLR being only ~30% identical in sequence, the tertiary structures of the PRLR ECD and GHR ECD are very similar.¹² Both consist of two fibronectin type III (FN-III) domains [an N-terminal domain (named D_N, residues

* Corresponding author phone: +61 7 3365 4180; fax: +61 7 3365 3872; e-mail: a.e.mark@uq.edu.au.

[†] School of Chemistry and Molecular Biosciences.

[‡] Institute for Molecular Bioscience.

[§] Present address: Division of Biochemistry and Center for Biomedical Genetics, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands.

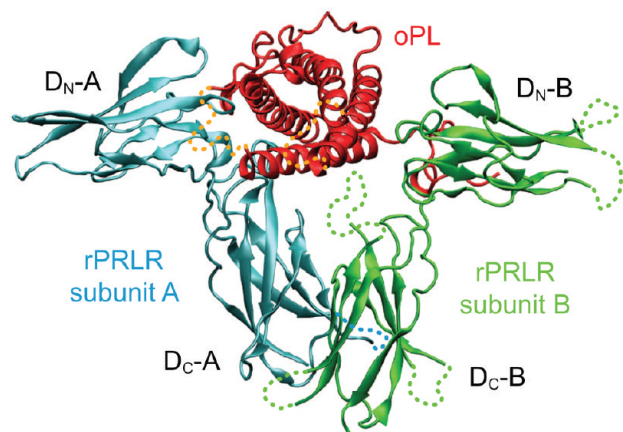


Figure 1. Crystal structure (Protein Data Bank entry 1F6F) of the extracellular domain of the homodimeric prolactin receptor (cyan and green) bound to placental lactogen (red) shown as a cartoon.¹⁰ Each receptor subunit consists of two FN-III domains (D_N and D_C) connected by a short hinge. Missing loops are shown as dotted lines.

Gln1–Asp96) and a C-terminal domain (D_C , residues Val101–Asp210] connected by a four-residue hinge (Val97–Ile100). The dimer (consisting of subunits A and B) is stabilized through interactions between the two C-terminal domains of each receptor subunit (termed D_{C-A} and D_{C-B}). The ligand that binds asymmetrically is primarily associated with the two N-terminal domains (termed D_{N-A} and D_{N-B}), as shown in Figure 1.

The key event in the activation of PRLR had been hypothesized to be a ligand-induced homodimerization that assumes that PRLR is predominantly monomeric on the surface of the cell¹³ (Figure 2a). However, recent experiments such as yeast two-hybrid studies and immunoprecipitation studies suggest that, on the surface of cells, PRLR is present as a constitutive dimer even in the absence of a ligand.⁶ More generally, the mechanism of activation of class I cytokine receptors remains unclear. If the receptor resides on the surface of cells as a preformed dimer, then the collection or cross-linking of individual receptor molecules by a ligand is unlikely to be the primary mechanism of activation. Activation could instead involve ligand-induced conformational changes within the ECD that would, in turn, induce the transmission of a mechanical signal through the plasma membrane, via the TM helices. Different motions that could give rise to activation are illustrated in Figure 2b–d. These include a translational motion (Figure 2b), a rotational motion (Figure 2c), and/or a scissor-like motion (Figure 2d) that would change the angle or the distance between the TM helices and, as a consequence, trigger changes in the orientation of the associated intracellular kinases.

In the case of GHR, a rotation of the individual ECDs within the GHR dimer has been proposed as the primary mechanism of activation.¹⁴ This result has also been supported in recent simulation studies.¹⁵ Specifically, atomistic molecular dynamics simulations of the GHR ECD were performed in the presence and absence of GH. Removal of GH from the crystal structure of the GH-bound GHR ECD dimer resulted in a rotation of the receptor subunits relative to each other by an angle of 45° on average, in close

agreement with experimental results in both direction and magnitude. In this study, atomistic molecular dynamics simulations have been used to investigate the mechanism of activation of PRLR, which is structurally similar to GHR. Starting from the crystal structure of the activated (hormone-bound) form of the receptor (PL-PRLR₂),¹⁰ a series of simulations have been conducted with and without PL. The effect of counterions and the inclusion of regions of the protein not observed in the initial crystal structure were also examined. Analysis of the structural changes within the receptor dimer consecutive to the removal of the ligand suggests a modification in the relative orientation of the extracellular domains consistent with a scissor-like mechanism. In addition, the interaction of the ECDs with a model phospholipid bilayer and the coupling of the ECDs with the TM domains were also studied.

2. Methods

2.1. Simulation of the Soluble Receptor Extracellular Domain. To examine the effect of ligand binding on the conformation of the ECD of PRLR, the ECD was simulated in water in the presence and absence of the ligand oPL. The coordinates of the 2:1 ternary complex of the rPRLR bound to oPL were taken from the structure of ref 10 [Protein Data Bank (PDB) entry 1F6F]. To assess the extent to which the loop regions that are disordered, thus not observed, in the crystal structure affect the structural stability of the domains in the receptor molecules, two systems were constructed. In the first system, only those residues present in the crystal structure were included in the model. This results in gaps in the sequence. To minimize the effects of these gaps, the amino acids preceding or following a missing residue were capped with a neutral C-amide or N-acetyl group. The residues missing included Gly48–Lys56, Ser107–Ser111, Glu198, and Thr199 of oPL; Gln1–Pro4, Gln115–Asp118, and Asn204–Asp210 in subunit A of the PRLR dimer; and Gln1–Gly5, Asp29–Pro33, Asn83–Ser87, Gln115–Lys119, Thr131–Phe140, Pro150–Glu154, and Asp205–Asp210 in subunit B of the PRLR dimer. In the second system, the missing residues were modeled using SwissPDB Viewer version 4.0.¹⁶ Specifically, possible loop geometries were identified on the basis of loops in the PDB with similar sequences and a geometry chosen that had a low conformational energy and did not overlap with other atoms in the protein.

To obtain the oPL-free form of the PRLR dimer, we removed oPL from the ternary complex in both systems. In addition to the effect of modeling the missing residues, the effect of salt concentration was also examined. Each of the four systems was simulated in pure water and in the presence of a physiological salt concentration via replacement of some water molecules in the hydrated systems with Na⁺ and Cl[−] ions, according to the most favorable electrostatic potential, to give a final salt concentration of 150 mM NaCl.

For the sake of simplicity, the systems that contain the bound oPL will be named with the letter B (bound) whereas the systems from which oPL has been removed will be named with the letter U (unbound). The inclusion of the

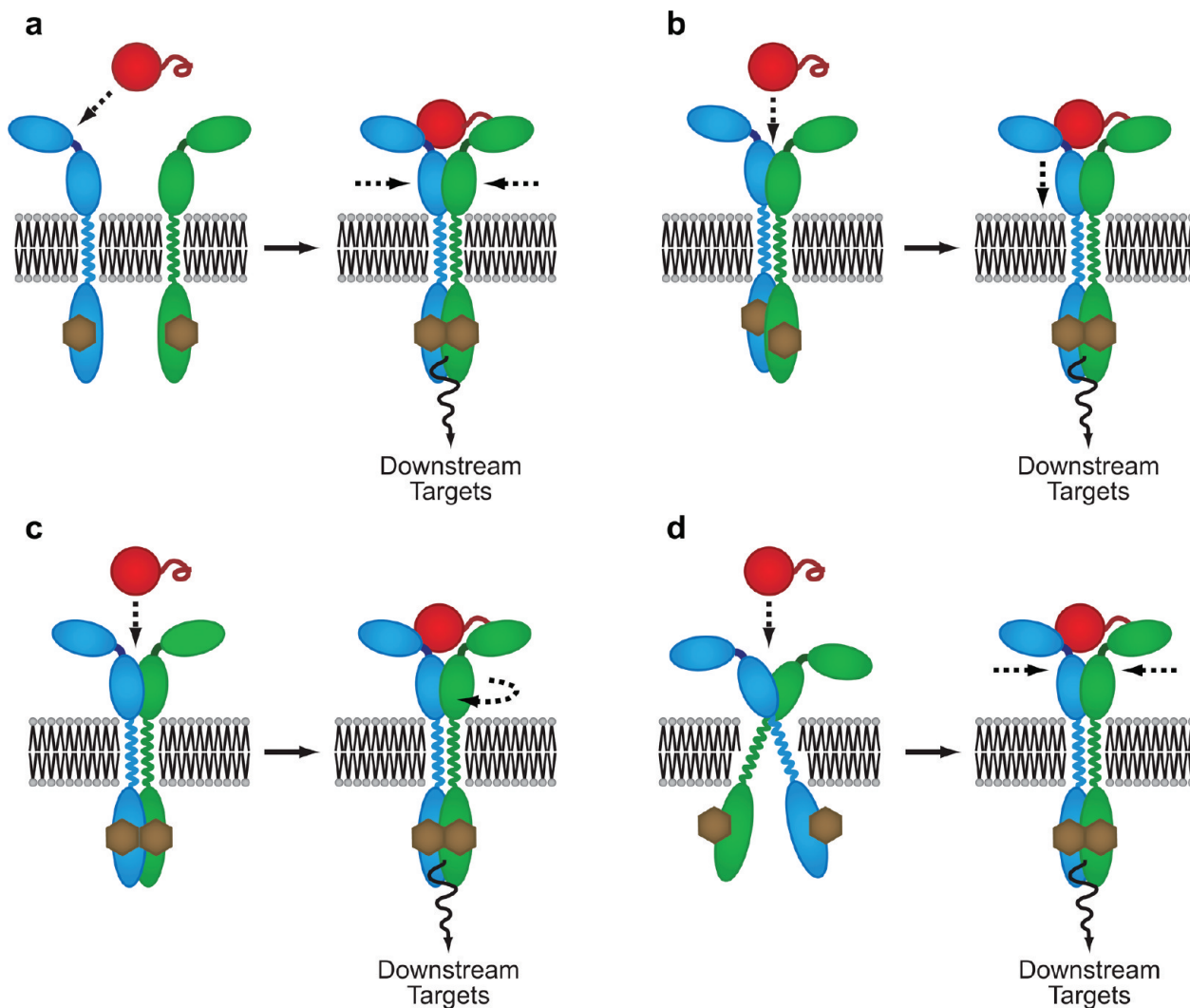


Figure 2. Schematic views of possible motions that may be involved in the activation of the prolactin receptor. Receptor subunits are colored blue and green, and the ligand is colored red. The brown hexagons represent kinases associated with the ICDs of the receptor subunits. (a) Ligand-induced homodimerization. A ligand binds first to one receptor molecule after which a second receptor molecule binds to form the ternary complex. (b–d) Activation of the preformed receptor dimer induced by ligand binding through (b) a translational motion, (c) a rotational motion, and (d) a scissor-like motion.

Table 1. Overview of the PRLR Systems That Were Simulated

system	description	[NaCl] (mM)	loops reconstructed ^a	lipid bilayer	linker structure ^b	TMD ^c	no. of simulations	simulation time (ns)
B	oPL–rPRLR ₂	0	no	no	no	no	2	37
B _I	oPL–rPRLR ₂	150	no	no	no	no	3	37
B _L	oPL–rPRLR ₂	0	yes	no	no	no	1	25
B _{L_I}	oPL–rPRLR ₂	150	yes	no	no	no	1	25
U	rPRLR ₂	0	no	no	no	no	6	25–58
U _I	rPRLR ₂	150	no	no	no	no	6	18–57
U _L	rPRLR ₂	0	yes	no	no	no	1	49
U _{L_I}	rPRLR ₂	150	yes	no	no	no	1	25
M _R	oPL–rPRLR ₂	0	yes	yes	random coil	yes	1	12
M _H	oPL–rPRLR ₂	0	yes	yes	α-helix	yes	1	10
M _{RΔ}	rPRLR ^{ΔD_N}	0	yes	yes	random coil	yes	1	5.5

^a Residues missing in the crystal structure: Gly48–Lys56, Ser107–Ser111, Glu198, and Thr199 of oPL; Gln1–Pro4, Gln115–Asp118, and Asn204–Asp210 in subunit A of the rPRLR dimer; and Gln1–Gly5, Asp29–Pro33, Asn83–Ser87, Gln115–Lys119, Thr131–Phe140, Pro150–Glu154, and Asp205–Asp210 in subunit B of the rPRLR dimer. ^b Linker region between the extracellular and transmembrane domains (residues Asp205–Asp210). ^c Transmembrane domain (residues Thr211–Met240). ^d Truncated form of the extracellular domain of PRLR from which N-terminal domain D_N (residues Gln1–Asp96) has been deleted.

missing loops in the model and the presence of a physiological salt concentration will be indicated by the subscripts L and I, respectively (see Table 1).

2.2. Simulation of the Membrane-Bound Receptor. To study the coupling between the extracellular and transmembrane domains within the oPL-bound PRLR dimer, the two

domains were simulated in the presence of a model of the cell membrane. As the addition of the missing loops was shown to have a minor effect on the structure of the ECD of the PRL receptor and because some of the loops found to be disordered in the crystal may be involved in interactions with the plasma membrane, all missing loops were included in the model as described above. Furthermore, the C-terminal end of the ECD of each receptor subunit was extended with residues Thr211–Met240; the segment of Thr211–Leu234 has been predicted to be membrane-spanning.¹⁷ Whereas the TM domain can be assumed to be helical, little is known about the structure of the six-residue linker (Asp205–Asp210) that connects the ECD of the PRLR to the TM region. The structure of this linker is, however, critical when positioning the receptor with respect to the membrane. Therefore, the linker was constructed in two alternative ways. (1) The linker was modeled as a random coil, and the ECDs were placed above the membrane (system M_R). (2) The linker was modeled as a continuation of the transmembrane α -helix (system M_H). The objective when modeling the linker region as a random coil was to allow the system to fold spontaneously to an appropriate configuration. As this might not be possible on the time scale accessible when simulating the complete ECD dimer, a single ECD consisting of only the D_C domain (rPRLR ^{Δ D_N}, residues Val101–Asp210) connected to a single TM region via an unstructured linker was also simulated (system M_{RA}). In this case, the N-terminus was capped with a neutral acetyl group. In all cases, the C-terminal end of the TM domain was capped with a neutral amide group. The receptor was inserted into a fully equilibrated lipid bilayer consisting of 512 POPC (2-oleoyl-1-palmitoyl-*sn*-glycero-3-phosphocholine) molecules,¹⁸ and the system was hydrated. The PRLR was oriented such that the long axis of the TM helices laid close to the bilayer normal, the *z*-axis in our coordinate system. An overview of all the PRLR systems simulated is given in Table 1.

2.3. Simulation Parameters. All simulations were performed using Gromacs version 3.2.1¹⁹ in conjunction with the Gromos 53a6 united-atom force field.²⁰ The parameters for POPC were taken from the revised Gromos 53a6 parameter set for lipids.²¹ Each system was subjected to periodic boundary conditions, using a truncated octahedral box for the soluble protein complexes and a rectangular box for the protein–membrane systems. The simple point charge (SPC) model²² was used to describe the water. Protonation states of ionizable groups were chosen so that they were appropriate for pH 7.0.

Water, lipids, and protein were coupled separately to an external temperature bath at 298 K by using a Berendsen thermostat²³ with a coupling constant τ_T of 0.1 ps. For the systems consisting of the ECD in solution (systems B, B_I, B_L, B_{LI}, U, U_I, U_L, and U_{LI}), the pressure was maintained at 1 bar by isotropically coupling the system to an external bath again using the method of Berendsen²³ with an isothermal compressibility of $4.6 \times 10^{-5} \text{ bar}^{-1}$ and a coupling constant τ_P of 1 ps. For those systems in which the receptor was embedded within a membrane (systems M_R, M_H, and M_{RA}), semi-isotropic pressure coupling was used. The same parameter values were used for the directions normal and

parallel to the plane of the POPC bilayer as for the simulations in water.

A twin-range cutoff scheme was used for the evaluation of nonbonded interactions: interactions falling within the short-range cutoff of 0.8 nm were calculated every step, whereas interactions falling within the long-range cutoff of 1.4 nm were updated every three steps. A reaction-field correction²⁴ was applied to account for the truncation of the electrostatic interactions beyond the long-range cutoff with an ϵ_{RF} of 78.

During the simulations, bond lengths within the solute molecules were constrained using the LINCS algorithm.²⁵ To extend the time scale that could be simulated, hydrogen atoms in the proteins were replaced with dummy interaction sites, the positions of which were constructed at each step from the coordinates of the heavy atoms to which they were attached. This allowed a 4 fs time step to be used without affecting the thermodynamic properties of the systems significantly.²⁶

To remove possible bad contacts between atoms stemming from the original crystal structure or introduced during the modeling of the loops, the systems were first energy-minimized in vacuo and then in solution using a steepest-descent algorithm. The systems were then equilibrated by gradually increasing the temperature from 50 to 298 K in 50 K steps over a time span of 200 ps before the simulations commenced. Starting velocities were randomly assigned from a Maxwellian distribution with different random seeds for each simulation. In total, the isolated ECD of the PRLR was simulated independently seven times in the presence of oPL and 14 times after oPL was removed. Each simulation in solution was at least 18 ns in length, and configurations were saved every 10 ps for analysis.

2.4. Analysis. Prior to analysis, the initial structure of the receptor dimer was reoriented such that the longest axis of the D_C domain of subunit A (D_C-A) was aligned along the *z*-axis. The longest axis in the D_C domains was chosen as the vector connecting the C α atoms of His188 and Pro203 in each receptor subunit. Furthermore, the axis defined by the *x*–*y* coordinates of the center of mass of the D_C-A and D_C-B domains was aligned along the *x*-axis. Then, the structures from the trajectories of the simulations were superimposed with respect to the backbone of D_C-A. The relative angle of rotation θ between the D_C domains was defined as the angle between the long axis of the two D_C domains when projected onto the *y*–*z* plane, as depicted in Figure 3.

The root-mean-square deviation (rmsd) of the coordinates of the backbone atoms (N, C α , C, and O) was calculated after a least-squares fit on the backbone atoms of the initial structure of every domain, subunit, or complex had been performed separately. The rmsd values for the region of interest were calculated with respect to the initial X-ray structure. For these calculations, the (N-terminal) D_N domain and the (C-terminal) D_C domain of a given PRLR subunit were defined by including only those residues present in the crystal structure. The D_N domain comprised Gln1–Asp96 and the D_C domain Val101–Asn204.

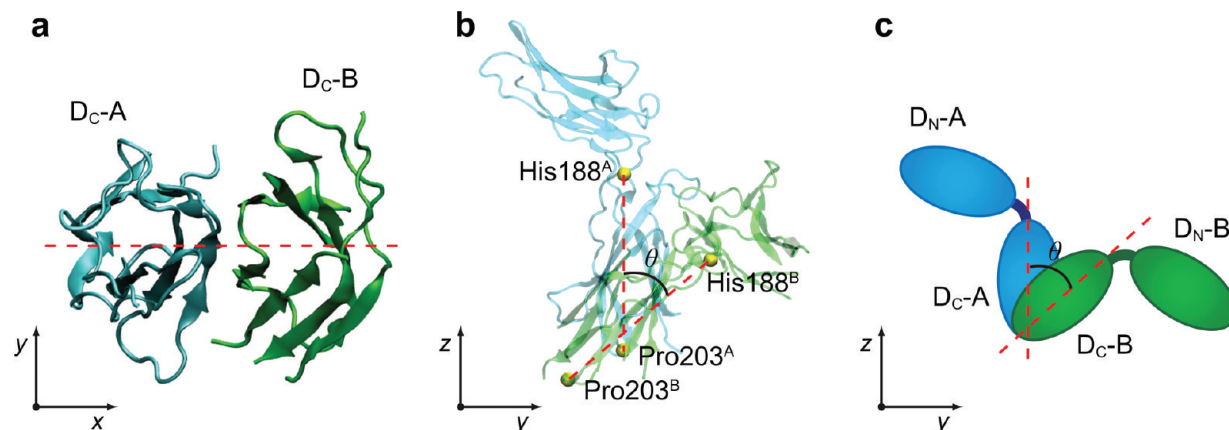


Figure 3. Method for measuring the angle θ between the two D_C domains. (a) The two domains are first aligned along the x -axis. (b) The long axis of each D_C domain was taken as the vector connecting the C_α atoms of His188 and Pro203 (His188^A and Pro203^A in subunit A and His188^B and Pro203^B in subunit B). This axis in subunit A is then aligned along the z -axis. The angle θ is defined as the angle between the long axis of subunits A and B, projected onto the y - z plane. (c) Schematic view of the angle θ between the two D_C domains in the receptor dimer.

3. Results

3.1. Dynamic Properties of the Isolated Receptor Complex.

We assessed the dynamics of the ternary oPL-rPRLR₂ complex in solution by conducting seven independent simulations of the receptor ECD with its ligand in a box of water. In five of these simulations, the model included only those residues for which coordinates were given in the X-ray crystal structure. In the two other simulations, the model was augmented to include the effect of modeling the missing loop residues. The effect of the physiological concentration of salt was also investigated. Of the five simulations that involved only those residues in the crystal structure, two were performed in pure water (simulations B-1 and B-2), while the three other were performed in the presence of 150 mM NaCl (simulations B_I-1–B_I-3). Of the two simulations including the modeled loops, one was performed in pure water (simulation B_L-1) and one in 150 mM NaCl (simulation B_L-1).

The dynamics of the homodimeric rPRLR₂ complex from which the ligand oPL had been removed was studied via 14 independent simulations. Of these, 12 comprised only those residues observed in the crystal structure, while in two, the missing loops were included in the model. Half of the simulations were performed in pure water (simulations U-1–U-6 and U_L-1) and half in the presence of 150 mM NaCl (simulations U_I-1–U_I-6 and U_L-1).

3.1.1. Structural Stability. As shown in Figure 1, the ECD of the PRLR consists of two fibronectin type III (FN-III) domains, the N-terminal domain (D_N) and the C-terminal domain (D_C), connected via a flexible hinge region. To examine whether the force field could maintain the structure of the protein and to examine the effect on the structural stability of the protein caused by the introduction of residues missing in the crystal structure and/or the inclusion of 150 mM NaCl, the rmsd of the positions of the backbone atoms of each domain with respect to the starting crystal structure was determined. The structure of the individual FN-III domains was stable under all conditions. The average rmsd over the last 5 ns of each simulation varied between 0.13

and 0.30 nm. The inclusion of missing amino acids in the model and/or 150 mM NaCl had no obvious effect on the stability of the individual domains (Figure 4a,b). For example, in the case of the unliganded dimer in which residues missing in the crystal structure were not included in the model, resulting in a discontinuous backbone with multiple breaks (simulations U-1–U-6), the average backbone rmsd with respect to the crystal structure was only 0.21 ± 0.03 nm for both the D_N and D_C domains. Essentially identical rmsd values (0.19 ± 0.04 and 0.20 ± 0.05 nm in domains D_N and D_C , respectively) were obtained in the presence of 150 mM NaCl (simulations U_I-1–U_I-6).

3.1.2. Effect of the Removal of oPL from the rPRLR₂ Complex. Although the structure of the individual FN-III domains was stable, significant motions between the domains were observed in all cases. The average rmsd over the last 5 ns of all the simulations of the ternary oPL-PRLR₂ complex was 0.45 ± 0.09 nm. This was largely because of changes in the position of the oPL ligand as the rmsd of the PRLR₂ dimer alone was 0.37 ± 0.12 nm. As the rmsd is a highly nonlinear measure, such values are easily obtained because of slight changes in the relative positions of the domains. Again, the effect of the inclusion of the missing residues or 150 mM NaCl was minor.

The removal of oPL from the ternary complex led in contrast to marked fluctuations in the overall rmsd. The large variations in rmsd with values reaching as high as 1.4–1.6 nm were a result of rigid-body hinge-bending motions within the individual subunits of the receptor dimer, as well as changes in the orientation of the two subunits with respect to each other. This type of interdomain motion is exemplified in Figure 4c, which shows the time evolution of the backbone rmsd with respect to the starting crystal structure for simulation U_I-3. Figure 4d shows various snapshots from the trajectory illustrating changes in the relative position of the D_N and D_C domains. As is evident from Figure 4d, the relative motion of the domains is reversible with the subunit flexing around the hinge region, opening and then closing the ligand-binding site. Again, similar results were obtained

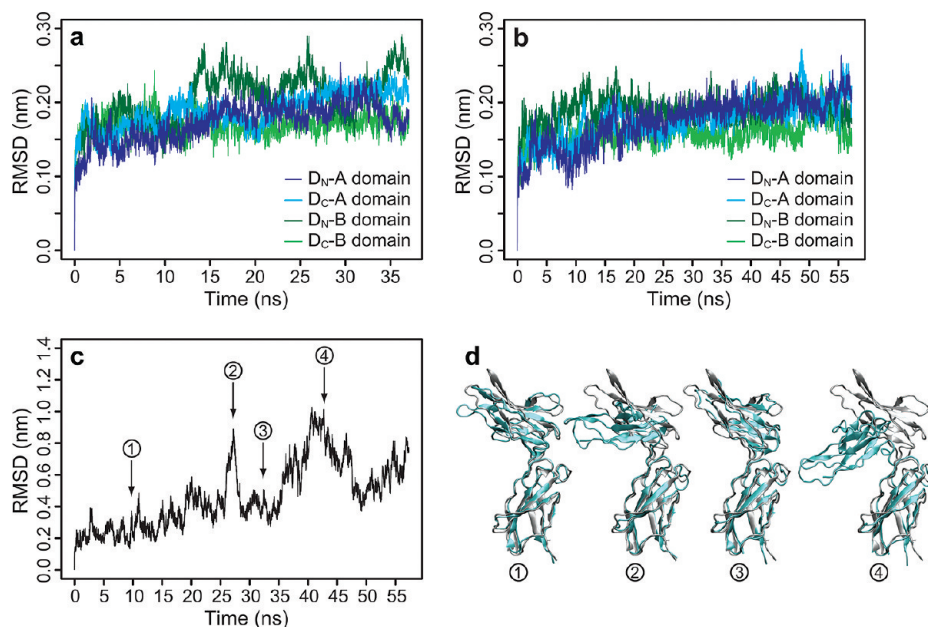


Figure 4. Conformational flexibility within the subunits of the prolactin receptor dimer. (a) Time evolution of the backbone rmsd calculated for the two fibronectin type III domains in subunits A and B of the oPL-bound receptor dimer in simulation B₁-1 with respect to the crystal structure. (b) Time evolution of the backbone rmsd calculated for the two fibronectin type III domains in subunits A and B of the unliganded receptor dimer in simulation U₁-3 with respect to the crystal structure. (c) Time evolution of the backbone rmsd of subunit B in simulation U₁-3 with respect to the crystal structure. (d) Snapshots from simulation U₁-3 at times corresponding to the arrows in panel c.

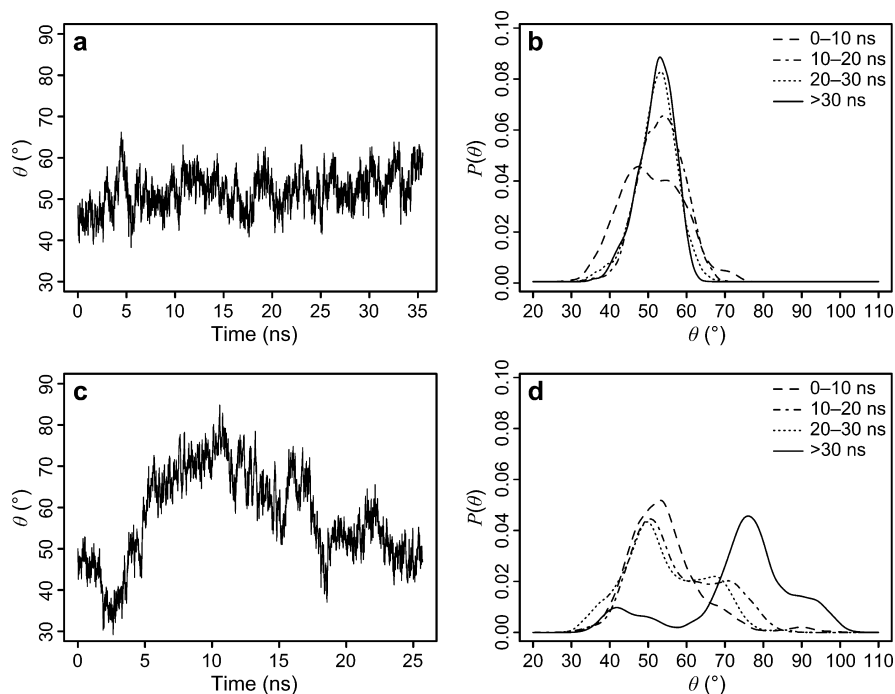


Figure 5. Time evolutions (a and c) and probability distributions (b and d) of the angle of rotation θ between the long axes of the two D_C domains of the subunits over the course of the simulations of the oPL-bound (a and b) and oPL-free (c and d) PRLR dimer. (a) Time evolution of θ in simulation B₁-2. (b) Probability distribution of θ calculated over all seven ligand-bound PRLR simulations and over four time ranges. (c) Time evolution of θ in simulation U-6. (d) Probability distribution of θ calculated over all 14 PRLR simulations after removal of oPL and over four time ranges.

in the presence or absence of 150 mM NaCl and regardless of whether the missing residues were included in the model.

3.2. Changes in the Relative Orientation of the Receptor Subunits. In addition to fluctuations in the relative positions of the D_N and D_C domains, the removal of the oPL

ligand was associated with changes in the relative orientation of the receptor subunits. Figure 5 shows the time evolution and distribution of the angle θ formed by the long axis of the two D_C domains in simulations in the presence (panels a and b) and absence (panels c and d) of oPL. In the X-ray

structure, the long axes of the two D_C domains make an angle of 49° with respect to each other. In the simulations of the oPL–PRLR₂ complex, the angle θ is essentially unchanged, fluctuating around an average value of 53° . This is illustrated in Figure 5a, which shows the variation of θ as a function of the simulation time for one of the simulations of the oPL–PRLR₂ complex in the presence of 150 mM NaCl (B1-2). As one can see, θ fluctuates between 40° and 65° . Similar results were obtained in all simulations of the ternary complex. The probability distribution of the value of θ calculated over all seven liganded simulations shows a single peak centered at 53° that becomes progressively narrower with time. In contrast, the removal of oPL from the complex resulted in large variations in the angle during the simulations. This is illustrated by the time evolution of θ during simulation U-6 in Figure 5c. In this case, it can be seen that θ increases to approximately 80° before falling again to around 50° . An increase in θ is associated with a clockwise rotation of one domain relative to the other.

As is evident from Figure 5c, the change in the relative orientation of the D_C domains after the removal of the oPL ligand is a stochastic process, and little can be inferred from a single simulation. However, upon combination of the results of all 14 simulations performed in the absence of oPL (Figure 5d), a bimodal behavior clearly appears, with the receptor dimer having two preferred angles: one centered between 45° and 55° and the other centered between 70° and 80° . Furthermore, there was a distinct shift as a function of time toward higher values of θ .

As noted above, the change in the relative orientation of the D_C domains is also associated with a clockwise rotation of the domains with respect to each other that, in turn, leads to an increase in the distance between their D_N domains as illustrated in Figure 6. Figure 6 shows the initial and final configurations from simulation U-6. The reorientation of the D_C domains results in an opening of the ligand-binding site and the exposure of the binding surfaces on the D_N domains. Nonetheless, the primary hydrogen bonds and hydrophobic interactions between residues positioned at the interface of the D_C domains (subunit A, mainly Phe167–Asp171; subunit B, mainly Phe160–Gln164) were maintained (Figure 6a,b). At least on the time scale of the simulations, there was no evidence to suggest that the removal of the ligand would cause the complex to dissociate.

3.3. Interaction of the Extracellular Domain with a POPC Membrane. A major unanswered question with regard to the mechanism of action of all class I cytokine receptors is how the extracellular domains are mechanically coupled to the transmembrane domains. If the transmission of a signal is the result of a structural change in the ECDs as has been suggested here and in previous studies,^{14,15,27} the structure of the linker that connects the ECD to its TM helix is critical. Specifically, if the linker is to transmit a mechanical signal, then it is expected that the linker would be rigid and tightly coupled to both the ECD and the TMD. Such a coupling could occur if the linker adopted a helical or β -sheet structure or folded onto the ECD to lock the TMD to the receptor ECD at the surface of the membrane. As no information about the structure of the linker is available, a

series of simulations of the complete receptor complex embedded within a POPC bilayer were performed. The aim of these studies was two-fold: to attempt to shed light on the ECD–TMD coupling via the linker region and to understand the nature of the interaction of the ECD with the lipid bilayer. Several different approaches were used to model the linker region in the context of the ligand-bound receptor dimer. In the first approach (simulation M_R), the linker between the ECD and the TM helix was modeled in an extended, random-coil conformation to produce an unbiased starting configuration (Figure 7a). In this case, the linker rapidly associated (within 3 ns of simulation) with residues of the base of the ECD, forming an array of hydrogen bonds and salt bridges. This resulted in the ECD interacting directly with the membrane. In particular, charged residues within the loops that were disordered in the crystal structure interacted directly with the lipid headgroups (Figure 7b). However, despite the ECD lying in the proximity of the TMD, the linker region itself did not adopt a clearly defined structure.

In the second approach (simulation M_H), the linker region was modeled as a helical elongation of the transmembrane helix as depicted in Figure 7c. Such a helical linker would provide a direct way to couple changes in the relative orientation of the D_C domains via the TMDs to the ICDs. Nonetheless, in this case, the helical linker immediately

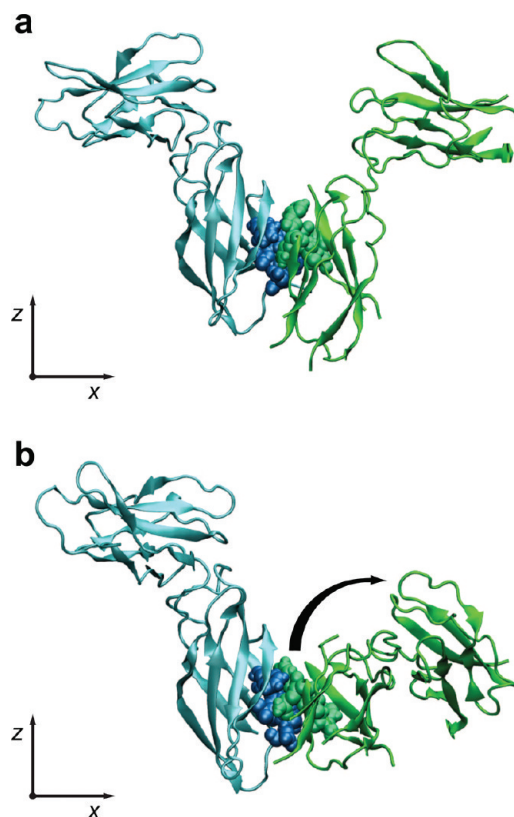


Figure 6. Rotation between the subunits of the rPRLR within the dimer when simulated after removal of oPL. (a) Initial structure (subunits A and B colored cyan and green, respectively) with residues that keep interacting when the subunits move with respect to each other depicted as spheres. (b) Structure after simulation for 17.5 ns with subunit B tilted backward (simulation U-1).

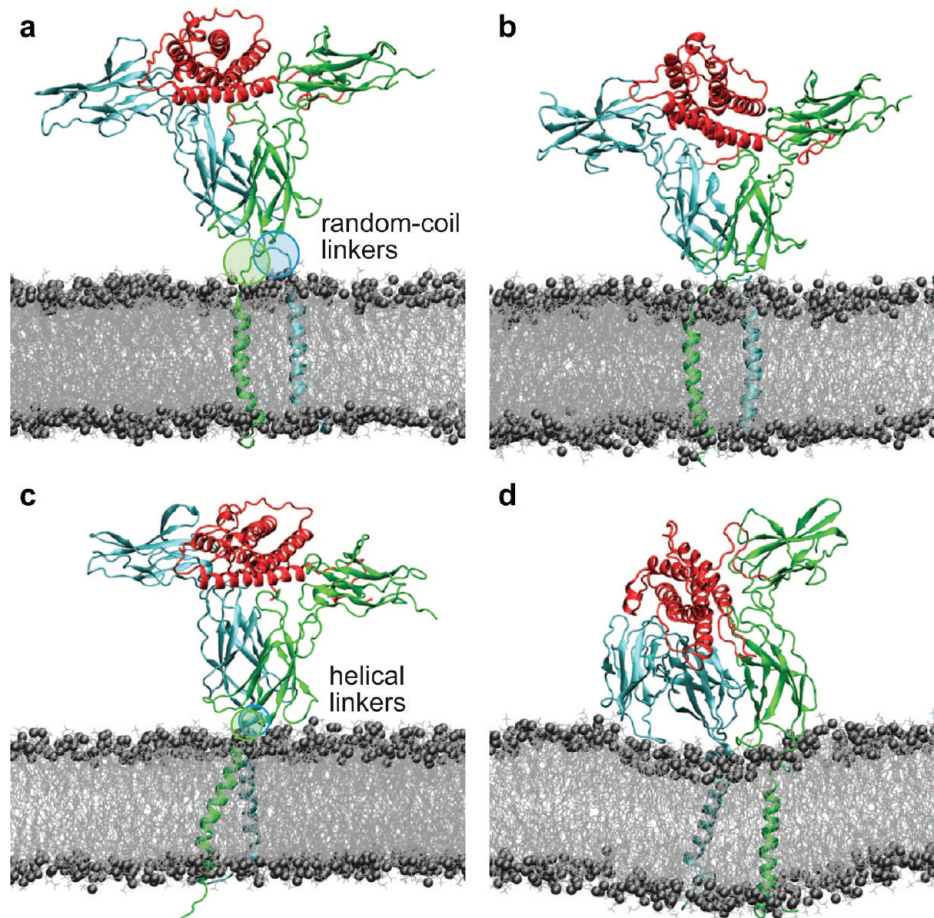


Figure 7. Initial and representative final configurations of simulations of the ECD and TMD of the rPRLR₂ dimer in a complex with oPL and embedded within a POPC bilayer performed using two different initial configurations of the linker region. (a and b) The linker region was modeled in an extended structure: (a) initial structure and (b) structure after simulation for 3 ns. (c and d) The linker region was helical: (c) initial structure and (d) structure after simulation for 9 ns. The ECDs of the rPRLR subunits are colored cyan and green, and oPL is colored red. The lipids are shown as light gray lines, and phosphorus atoms shown as dark gray spheres. The linkers are circled in cyan and green.

unfolded, and again, the base of the ECD interacted strongly with the membrane (Figure 7d). Although this does not mean the linker cannot be helical, it does imply that the structure of the receptor dimer determined in a nonphysiological environment and the arrangement in which the TMDs lie in an approximately parallel configuration are incompatible with the linker being a helical extension of the TMD.

A third system in which the D_C domain of a single subunit (the D_N domain of the ECD was removed) was also coupled to the TMD via a linker in an extended conformation was simulated (simulation M_{RA}). Again, the base of the D_C domain embedded in the membrane before the linker region could not fold into a specific conformation.

4. Discussion

Despite many years of investigation, the precise mechanism by which the binding of a class I cytokine to the extracellular domain of its corresponding receptor transmits a signal through the cell membrane is unknown. While it is possible that activation may involve the collection of two receptor subunits by a ligand, there is growing evidence that class I cytokine receptors such as the growth hormone receptor (GHR), the erythropoietin receptor (EpoR) and the prolactin

receptor (PRLR) reside on the membrane as preformed dimers.^{7,28} In this case, the activation of the receptor is most likely linked to the relative position or orientation of the receptor subunits, such as the relative rotation and/or translation of the subunits or a scissor-like movement of the subunits (illustrated in Figure 2b–d). In this study, we have attempted to determine which of these underlying mechanisms could lead to activation of the PRLR by examining the effect of the removal of ligand from the crystal structure of the activated complex and by examining the mechanical coupling of the ECD to the TM domain.

Our results suggest that in the absence of ligand, the receptor dimer shows a high degree of flexibility and that it can adopt two distinct conformations that differ in the relative subunit orientation of the membrane-proximal (D_C) domains. Furthermore, the simulations suggest that the binding of a ligand, in this case placental lactogen, stabilizes one of these two preexisting conformations, rigidifying the complex. In this respect, ligand binding can be viewed as introducing a clockwise rigid-body rotation of one subunit with respect to the other by ~20–30°. Importantly, this was associated with only minor changes in the nature of the interactions at the interface between the two subunits. These findings indicate

that the activation mechanism for PRLR may involve changes in the relative orientation of the ECD similar to those proposed in the case of the activation of GHR in both experiment and simulation.^{14,15} As the monomeric GHR and PRLR and their corresponding dimers are structurally very similar, it is not surprising that the simulations showed related, but slightly different, motions for the two receptors.

It has been found experimentally that the deletion of the N-terminal tail of PL greatly reduces the activity of the ligand.²⁹ From the crystal structure (Figure 1), one can see that oPL is mainly associated with the D_N-A domain and only the N-terminal end of placental lactogen binds to the D_N-B domain. Thus, it is likely that it is the binding of the N-terminal tail to the D_N-B domain that stabilizes the rotated (activated) form of the complex.

The mechanism that we would propose is one in which in the absence of ligand, both D_N domains are solvent-exposed and the ECD is predominantly in an open, inactive form. The binding of a ligand stabilizes the rotated form, leading to a change in the relative orientation of the TMDs and, thus, the relative positions of the intracellular domains.

A critical aspect of the model is the degree of mechanical coupling between the ECD and the TMD. The simulations that were performed in an attempt to address this issue suggest that the base of the D_C domains of the PRLR interact significantly with the membrane as opposed to the ECD standing proud of the membrane, and that the structure of this region in the available crystal structure may not be representative of that in a physiologically relevant environment. In addition, inasmuch as the linker between the ECD and the TMD was relatively flexible and did not fold into a specific conformation in any of the simulations, the ECD and the TMD may interact directly to transmit the mechanical signal through the membrane.

In summary, we have proposed a model for activation of the prolactin receptor, taking into account the necessity for the receptor to transfer a signal through the cell membrane that involves changes in the relative orientation of the extracellular domains upon binding of a ligand. This suggests that a relative rotation of the extracellular domains may represent a general model for the activation of class I cytokine receptors. Furthermore, we have shown that the extracellular domains interact strongly with the membrane and that to understand the mechanical coupling between the extracellular and transmembrane domains, an appropriate representation of the environment of the membrane–water interface will be critical.

Abbreviations

D_C, C-terminal domain of the extracellular domain of the prolactin receptor; D_N, N-terminal domain of the extracellular domain of the prolactin receptor; ECD, extracellular domain; FN-III, fibronectin type III; GH, growth hormone; GHR, growth hormone receptor; ICD, intracellular domain; oPL, ovine placental lactogen; PL, placental lactogen; POPC, 2-oleoyl-1-palmitoyl-*sn*-glycero-3-phosphocholine; PRL, prolactin; PRLR, prolactin receptor; rmsd, root-mean-square deviation; rPRLR, rat prolactin receptor; TM, transmembrane; TMD, transmembrane domain.

Acknowledgment. This work was funded by the Australian Research Council (ARC). A.E.M. is an ARC Federation Fellow. All the calculations were performed using high-performance computing resources of The University of Queensland and of the National Computational Infrastructure (NCI) National Facility at the Australian National University under the Merit Allocation Scheme through the Queensland Cyber Infrastructure Foundation (QCIF) partner share scheme.

References

- (1) Bole-Feysot, C.; Goffin, V.; Edery, M.; Binart, N.; Kelly, P. A. Prolactin (PRL) and its receptor: Actions, signal transduction pathways and phenotypes observed in PRL receptor knockout mice. *Endocr. Rev.* **1998**, *19*, 225–268.
- (2) Bazan, J. F. Structural design and molecular evolution of a cytokine receptor superfamily. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 6934–6938.
- (3) Grattan, D. R.; Steyn, F. J.; Kokay, I. C.; Anderson, G. M.; Bunn, S. J. Pregnancy-induced adaptation in the neuroendocrine control of prolactin secretion. *J. Neuroendocrinol.* **2008**, *20*, 497–507.
- (4) Hu, Z. Z.; Zhuang, L.; Dufau, M. L. Prolactin receptor gene diversity: Structure and regulation. *Trends Endocrinol. Metab.* **1998**, *9*, 94–102.
- (5) Trott, J. F.; Hovey, R. C.; Koduri, S.; Vonderhaar, B. K. Alternative splicing to exon 11 of human prolactin receptor gene results in multiple isoforms including a secreted prolactin-binding protein. *J. Mol. Endocrinol.* **2003**, *30*, 31–47.
- (6) Gadd, S. L.; Clevenger, C. V. Ligand-independent dimerization of the human prolactin receptor isoforms: Functional implications. *Mol. Endocrinol.* **2006**, *20*, 2734–2746.
- (7) Qazi, A. M.; Tsai-Morris, C.-H.; Dufau, M. L. Ligand-independent homo- and heterodimerization of human prolactin receptor variants: Inhibitory action of the short forms by heterodimerization. *Mol. Endocrinol.* **2006**, *20*, 1912–1923.
- (8) Somers, W.; Ultsch, M.; de Vos, A. M.; Kossiakoff, A. A. The X-ray structure of a growth hormone-prolactin receptor complex. *Nature* **1994**, *372*, 478–481.
- (9) Svensson, L. A.; Bondensgaard, K.; Nørskov-Lauritsen, L.; Christensen, L.; Becker, P.; Andersen, M. D.; Maltesen, M. J.; Rand, K. D.; Breinholt, J. Crystal structure of a prolactin receptor antagonist bound to the extracellular domain of the prolactin receptor. *J. Biol. Chem.* **2008**, *283*, 19085–19094.
- (10) Elkins, P. A.; Christinger, H. W.; Sandowski, Y.; Sakal, E.; Gertler, A.; de Vos, A. M.; Kossiakoff, A. A. Ternary complex between placental lactogen and the extracellular domain of the prolactin receptor. *Nat. Struct. Biol.* **2000**, *7*, 808–815.
- (11) Gertler, A.; Grosclaude, J.; Strasburger, C. J.; Nir, S.; Djiane, J. Real-time kinetic measurements of the interactions between lactogenic hormones and prolactin-receptor extracellular domains from several species support the model of hormone-induced transient receptor dimerization. *J. Biol. Chem.* **1996**, *271*, 24482–24491.
- (12) de Vos, A.; Ultsch, M.; Kossiakoff, A. A. Human growth hormone and extracellular domain of its receptor: Crystal structure of the complex. *Science* **1992**, *255*, 306–312.
- (13) Sakal, E.; Elberg, G.; Gertler, A. Direct evidence that lactogenic hormones induce homodimerization of membrane-anchored prolactin receptor in intact Nb2-11C rat lymphoma cells. *FEBS Lett.* **1997**, *410*, 289–292.

- (14) Brown, R. J.; Adams, J. J.; Pelekanos, R. A.; Wan, Y.; McKinstry, W. J.; Palethorpe, K.; Seeber, R. M.; Monks, T. A.; Eidne, K. A.; Parker, M. W.; Waters, M. J. Model for growth hormone receptor activation based on subunit rotation within a receptor dimer. *Nat. Struct. Mol. Biol.* **2005**, *12*, 814–821.
- (15) Poger, D.; Mark, A. E. Turning the growth hormone receptor on: Evidence that hormone binding induces subunit rotation. *Proteins* **2010**, *78*, 1163–1174.
- (16) Guex, N.; Peitsch, M. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **1997**, *18*, 2714–2723.
- (17) Boutin, J.-M.; Jolicoeur, C.; Okamura, H.; Gagnon, J.; Edery, M.; Shirota, M.; Banville, D.; Dusanter-Fourt, I.; Djiane, J.; Kelly, P. A. Cloning and expression of the rat prolactin receptor, a member of the growth hormone/prolactin receptor gene family. *Cell* **1988**, *53*, 69–77.
- (18) Poger, D.; Mark, A. E. On the validation of molecular dynamics simulations of saturated and *cis*-monounsaturated phosphatidylcholine lipid bilayers: A comparison with experiment. *J. Chem. Theory Comput.* **2010**, *6*, 325–336.
- (19) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. Gromacs: Fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
- (20) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: The Gromos force-field parameter sets 53a5 and 53a6. *J. Comput. Chem.* **2004**, *25*, 1656–1676.
- (21) Poger, D.; van Gunsteren, W. F.; Mark, A. E. A new force field for simulating phosphatidylcholine bilayers. *J. Comput. Chem.* **2010**, *30*, 117–1125.
- (22) 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*; Reidel: Dordrecht, The Netherlands, 1981; pp 331–342.
- (23) 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.
- (24) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. A generalized reaction field method for molecular dynamics simulations. *J. Chem. Phys.* **1995**, *102*, 5451–5459.
- (25) 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.
- (26) Feenstra, K.; Hess, B.; Berendsen, H. J. C. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.* **1999**, *20*, 786–798.
- (27) Seubert, N.; Royer, Y.; Staerk, J.; Kubatzky, K. F.; Moucadel, V.; Krishnakumar, S.; Smith, S. O.; Constantinescu, S. N. Active and inactive orientations of the transmembrane and cytosolic domains of the erythropoietin receptor dimer. *Mol. Cell* **2003**, *12*, 1239–1250.
- (28) Gent, J.; van Kerkhof, P.; Roza, M.; Bu, G.; Strous, G. J. Ligand-independent growth hormone receptor dimerization occurs in the endoplasmic reticulum and is required for ubiquitin system-dependent endocytosis. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 9858–9863.
- (29) Gertler, A.; Hauser, S. D.; Sakal, E.; Vashdi, D.; Staten, N.; Freeman, J. J.; Krivi, G. G. Preparation, purification, and determination of the biological activities of 12 N-terminus-truncated recombinant analogues of bovine placental lactogen. *J. Biol. Chem.* **1992**, *267*, 12655–12659.

CT1003934