JCTC Journal of Chemical Theory and Computation

First-Principle Molecular Dynamics Study of Selected Schiff and Mannich Bases: Application of Two-Dimensional Potential of Mean Force to Systems with Strong Intramolecular Hydrogen Bonds

Aneta Jezierska*,[†,§] and Jarosław J. Panek[‡]

University of Wrocław, Faculty of Chemistry, F. Joliot-Curie 14,
50-383 Wrocław, Poland, and National Institute of Chemistry, Hajdrihova 19,
1001 Ljubljana, Slovenia

Received October 9, 2007

Abstract: Car−Parrinello Molecular Dynamics simulations were performed for selected anharmonic systems, i.e., Schiff and Mannich base-type compounds, to investigate the vibrational properties associated with O−H stretching. All calculations were performed in the gas phase to compare them with available experimental data. First the vibrational properties of the two compounds were analyzed on the basis of well-established approaches: Fourier transformation of the autocorrelation function of both the atomic velocities and dipole moments. Then path integral molecular dynamics simulations were performed to demonstrate the influence of quantum effects on the proton's position in the hydrogen bridge. In addition, quantum effects were incorporated a posteriori into calculations of O−H stretching envelopes for the Schiff and Mannich bases. Proton potential snapshots were extracted from the ab initio molecular dynamics trajectory. Vibrational Schrödinger equations (one- and two-dimensional) were solved numerically for the snapshots, and the O−H stretching envelopes were calculated as a superposition of the 0→1 transitions. Subsequently, one- and two-dimensional potentials of mean force (1D and 2D pmf) were calculated for the proton stretching mode from the proton vibrational eigenfunctions and eigenvalues incorporating statistical sampling and nuclear quantum effects. The results show that the applied methodologies are in good agreement with experimental infrared spectra. Additionally, it is demonstrated that the 2D pmf method could be applied in systems with strong anharmonicity to describe the properties of the O−H stretching mode more accurately. Future applications of the 2D pmf technique include, in principle, large biomolecular systems treated within the QM/MM framework.

I. Introduction

Intramolecular hydrogen-bonding plays a crucial role in biologically relevant systems and materials science, two important areas of contemporary research.[1−5] Its advantage over intermolecular interactions stems from the fact that the molecular scaffolding provides a rigid, durable framework for the weaker, modifiable hydrogen bridge of interest. The formation of an intramolecular hydrogen bond is entropically favorable over intermolecular contacts because of the formation of a pseudoring. Two examples of groups of compounds possessing strong intramolecular hydrogen bonds are aromatic Schiff and Mannich bases. The introduction of substituents either in the phenyl ring or at the acceptor

* Corresponding author phone: (+48) 71-3757-308; fax: (+48) 71-3282-348; e-mail: anetka@elrond.chem.uni.wroc.pl.

[†] University of Wrocław.

[‡] National Institute of Chemistry.

[§] Current address: The International School for Advanced Studies (SISSA), INFM DEMOCRITOS, Italian Institute of Technology- SISSA Unit, Via Beirut 2-4, 34014 Trieste, Italy.

10.1021/ct7002644 CCC: $40.75 © 2008 American Chemical Society
Published on Web 02/12/2008

nitrogen atom leads to fine-tuning of the bridge properties[6−9] with direct influence on the macroscopic characteristics valuable for the practical application of the studied compounds. The main difference between the molecular skeletons of Schiff and Mannich bases is the presence of a double bond in the imine group, a hydrogen bond acceptor in the former class of compounds. The double bond enables coupling between the hydrogen bridge and the aromatic π-electron system of the phenyl ring. This in turn leads to a shortening of the bridge, which is classified as a low-barrier hydrogen bond (LBHB).[10] The flattening of the potential energy surface (PES) of the proton motion results in observable proton-transfer phenomena. The substitution in the phenyl ring or the imine group can influence the hydrogen bridge either by induction or steric effects.[11−13] Induction is a classical mechanism of electron withdrawal or electron donation, which is dependent on the electronic character of the substituents. On the other hand, a steric influence is provided by bulky substituents especially in the *ortho* position of the aromatic ring or in the imine group of both classes of these compounds. The microscopic results of the above effects are visible in the large variations of the hydrogen bridge's geometrical and spectroscopic parameters. The most significant are shifts observed in the vibrational, electronic, and NMR spectra.[14,15] The microscopically observed effects are related to and further responsible for the exhibited molecular properties important from the biological and industrial points of view. Schiff bases were found to be involved in the biological processes of vision and photoconversion.[16−21] Their potential practical applications are due to photochromic, thermochromic, magnetic, and conducting properties.[22−26] Mannich bases exhibit cytotoxic properties which lead to diverse biological activity found experimentally[27−31] and commercially exploited in, for example, derivatives of Norfloxacin (an antibiotic used to treat certain infections caused by bacteria—such as gonorrhea—and prostate and urinary tract infections).[32−34] They have found industrial application as lubricating oil additives[35] and epoxy resin hardeners.[36]

The current study discusses in detail the hydrogen bridge properties of the two model systems chosen from the Schiff and Mannich base families (Figure 1).[37,9] The selected Schiff base is N-methyl-2-hydroxybenzylidene amine (HBZA) and the representative Mannich base is *ortho*-dimethylaminomethylphenol (DMAP). Their structural similarity allows us to compare directly the molecular properties of the two classes of compounds with emphasis on the proton dynamics in the hydrogen bridge. The size of the studied molecules is small enough to allow a clear understanding of the proton dynamics and the exclusion of unwanted interactions. On the other hand, the models are sufficiently large to require statistical description via ab initio molecular dynamics.[38] Molecular dynamics provides a bridge between the microscopic and macroscopic levels of describing the studied systems. The potential energy surface obtained from first-principle methods (in our case, density functional theory, DFT,[39,40] propagated in time using the Car−Parrinello[38] scheme) allows us to study chemical reaction pathways such as the proton dynamics in the hydrogen bridge. Time evolution of the electronic and structural features gives us a more detailed description of the hydrogen bridge's properties. In particular, a set of frozen-nuclei proton potential functions can be converted into an ensemble-averaged free energy profile defined as the potential of mean force[41] for the proton motion. Using free energy profiles instead of static proton potential functions might be advantageous for describing processes involving large systems (e.g., enzymes), where the huge number of degrees of freedom prevents the use of only one representative total-energy profile. In the current study we extend the scheme for estimating the potential of mean force of the proton motion from the one-dimensional (1D pmf) to the two-dimensional (2D pmf) case. This helps us visualize the qualitative and quantitative differences in the molecular properties provided by the molecular frameworks of the Schiff and Mannich base. Our calculated results are verified by comparison with available experimental infrared (IR) spectra and previous theoretical investigations.[9,37] Summarizing, the main goal of our study is a description of the proton dynamics in two closely related molecular skeletons exhibiting short, strong, low-barrier (Schiff base) and medium-strong (Mannich base) intramolecular hydrogen bonds. The outline of the article is as follows: the theory and methods applied in the study are presented in section II, the results and discussion are given in section III, and concluding remarks are presented in section IV.

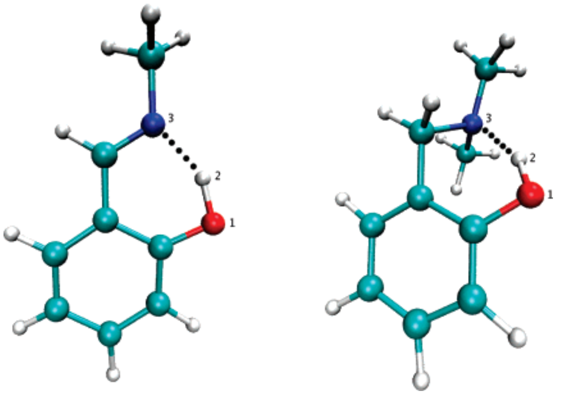


Figure 1. Computational models of the investigated Schiff (N-methyl-2-hydroxybenzylidene amine, HBZA) (left side) and Mannich (*ortho*-dimethylaminomethylphenol, DMAP) (right side) bases with the atom numbering scheme. Only selected atoms, important in the results discussion, are marked. Color coding of atoms is as follows: cyan − carbon; dark blue − nitrogen (atom 3); red − oxygen (atom 1); white − hydrogen.

II. Theory and Methods

A. Car−Parrinello Molecular Dynamics (CPMD) in Vacuo. Car−Parrinello Molecular Dynamics[38] (CPMD) on the basis of Density Functional Theory[39,40] (DFT) was applied to investigate vibrational features of the selected Schiff (N-methyl-2-hydroxybenzylidene amine, HBZA) and Mannich (*ortho*-dimethylaminomethylphenol, DMAP) bases (see Figure 1) in vacuo. The initial structure optimizations were carried out using the Schlegel Hessian matrix at the starting point.[42] The cubic cell dimension for both compounds was $a = 15$ Å. The size of the cell was dictated by the need to avoid artifacts at the cell boundary. The Hockney periodic image removal scheme was applied. The functional proposed by J. P. Perdew, K. Burke, and M. Ernzerhof (PBE) in conjunction with the plane-wave basis set was used for the study.[43] The pseudopotentials proposed by Troullier and Martins[44] were used to replace the core electrons of the atoms in the system studied. A kinetic energy cutoff of 70 Ry was applied for the plane-wave basis set. The initial period of the molecular dynamics (ca. 10 000 steps) was used to equilibrate both studied systems and was not further analyzed or discussed. During the molecular dynamics (MD) simulations, the time step was consistently set to 3 au (0.0725 fs), and a fictitious electron mass parameter (EMASS) of 400 au was used to reproduce the orbital dynamics. The simulations were performed at room temperature ($T = 300$ K) for the Schiff base, and $T = 390$ K was used for the Mannich base. The data collection was 10.5 ps for the Schiff base and 12.5 ps for the Mannich base. Vibrational features of the two compounds were analyzed on the basis of the power spectra of the atomic velocities. In addition, the dipole moment values were collected during the MD run and were subsequently used to generate the predicted IR spectra. A program for Fourier transform autocorrelation function calculations was used for this purpose. The time evolution of the interatomic distances related to the intramolecular hydrogen bond was analyzed using programs developed in our laboratory.

B. Path Integral Molecular Dynamics (PIMD). Path Integral Molecular Dynamics[45−47] (PIMD) simulations were then carried out for the studied compounds in vacuo. PIMD was performed using a setup similar to the one applied for the CPMD simulations. The calculations were also performed at $T = 300$ K (for Schiff base) and $T = 390$ K (for Mannich base), controlled by a Nosé-Hoover thermostat chain.[48−51] Eight Trotter replicas ($P = 8$) were used for imaginary time path integration. The staging representation of the path integral propagator was applied.[52,47] The data were collected for 9 ps after an initial equilibration and further used to obtain a histogram presenting the hydrogen (H2) position in the intramolecular hydrogen bond.

C. A Posteriori Inclusion of Quantum Effects of Nuclear Motion into Calculation of the Vibrational Features of the Intramolecular Hydrogen Bond. The inclusion of quantum effects of the O−H stretching in the studied Schiff and Mannich bases was performed using an envelope methodology.[5] The method consists of a posteriori quantum corrections obtained by solving the vibrational Schrödinger equation[53] for a set of proton potential functions. These functions are obtained from the Car−Parrinello Molecular Dynamics trajectory of the investigated molecules sampled at regular intervals (0.2 ps). Such an interval, being ca. ten times larger than heavy atom vibrations, was chosen to minimize correlation between snapshots while providing sufficient number of structures. Regular, unbiased sampling of the CPMD trajectory, corresponding to the NVT ensemble by the virtue of Nosé-Hoover thermostat, ensures that the snapshots also closely follow the canonical statistical ensemble. Subsequently, the resulting snapshots are processed in the following way: the selected proton of the hydrogen bridge is displaced along a circular arc defined uniquely by the positions of the donor, proton, and acceptor atoms.[54] A total of 40 evenly spaced points are generated on this arc, and after rejecting those giving too close contact (less than 0.7 Å) with either donor or acceptor atom, 18 to 24 positions for a bridge proton are obtained for each snapshot. Then, the total energy is calculated for each frozen structure with the proton placed in the subsequent positions on the arc. Thus, for each snapshot, an instantaneous proton potential function is obtained which serves in solving the quantum vibrational problem with two-step methodology and software described in ref 53. First, the proton potential function is approximated by a ninth degree polynomial; the fit accuracy was usually better than 0.05 kcal/mol at each point. Second, the vibrational Schrödinger equation is solved using Fourier grid technique with 300 grid points evenly spaced in the range from 0.7 Å to 2.0 Å as a real-space basis set. This procedure yields vibrational energy levels and wavefunctions for each snapshot taken from the CPMD trajectory. The set of quantum-corrected anharmonic vibrational frequencies then serves finally to construct an envelope of the O−H stretching mode by summing a set of Gaussian functions centered at each of the calculated frequencies. This procedure has been previously successfully applied to a model structure (a Mannich base-type compound) with an intramolecular hydrogen bond.[55,56]

The dynamics simulations concerning parts A, B, and C above were performed using the CPMD v.3.9.2 program[57] compiled with parallel support to maximize the efficiency of the time-consuming calculations.

D. One- and Two-Dimensional Potentials of Mean Force (1D and 2D pmf). Finally, the one and two-dimensional potentials of mean force (1D and 2D pmf) were calculated for the O−H stretching mode on the basis of eigenfunctions obtained by solving the vibrational Schrödinger equation. The potential of mean force is a *free energy* profile along the postulated reaction coordinate.[41] Combined use of statistical sampling by Car−Parrinello molecular dynamics and quantization of the nuclear motions provides the following scheme of computing the pmf applied in this study. For each selected molecular dynamics snapshot, the 1D vibrational Schrödinger equation was solved, and the resulting eigenfunctions are stored (see previous subsection C). The squared wavefunction represents the probability density $\rho(x)$, where x is the chosen coordinate, in this case the O−H distance. In more detail, the probability density reads

$$\rho(x) = \frac{\sum_{i=0}^{\infty} \Psi_i^{\,2}(x) e^{-E_i/kT}}{\sum_{i=0}^{\infty} e^{-E_i/kT}}$$

where i runs over the vibrational eigenfunctions and eigenvalues.

The vibrational energy scale is assumed to be shifted so that the ground-state energy E_0 is 0 in the above formula. The main contribution to $\rho(x)$ comes from the ground vibrational state; the excited states contribute much less. However, in this study we chose to include contributions from the two lowest-lying excited states (i.e., $i = 0, 1, 2$) as well, since we observed large anharmonicity and correspondingly flat potential energy surfaces in one of the studied systems (see below). Subsequently, the $\rho(x)$ is averaged over the molecular dynamics trajectory, and the pmf is calculated directly from the expression:

$$\mathrm{pmf}(x) = -k_{\mathrm{B}}T\ln <\rho(x)>$$

The probability density $\rho(x)$ is in this case a Boltzmann-averaged sum of the squared wavefunctions, and the angle brackets denote averaging over the molecular dynamics snapshots. The applied method for the pmf calculations directly incorporates the quantum nature of the proton motion.[55] Additionally, this procedure is based on the potential energy surface calculated using nonempirical electronic structure methods only. This fact provides some advantage in systems for which force-field parametrization would be difficult, such as metals, organometallics, etc. It should be noted that many elegant and successful methods of calculating the potential of mean force are based on the force-field approach, including studies of intramolecular hydrogen bonds[58] parametrized to the DFT potential energy surface. Detailed discussion of the pmf calculation method employed here—its applicability range and relation to other techniques—is given at the end of the Results and Discussion.

The extension of the one-dimensional pmf technique to the two-dimensional case was carried out by choosing two reaction coordinates significant for proton dynamics. The extraction of snapshots from the CPMD trajectory suggests the application of the clamped nuclei model, in which the donor−acceptor distance is fixed at the value present in a given trajectory frame. The two coordinates applied in our study are therefore the O−H distance (as in the 1D case) and the O−H...N angle. For each extracted frame, the bridge proton position was scanned along these two coordinates. The angle was set to a value ranging from 110° to 180° in 5° increments, and at each fixed value of the angle the proton was displaced along the arc defined analogously to the one-dimensional case. This approach assumes a semicylindrical symmetry of the proton potential surface with respect to the donor − acceptor axis. The assumption is reasonable in our small-molecule, gas-phase models by the lack of bulky groups or intermolecular contacts and provides significant reduction of computational effort with respect to the full 3D scan of proton potential function. The set of the DFT total energy values for the generated coordinates forms a grid on which the 2D vibrational Schrödinger equation is solved, using the computational approach described above in subsection C, but extended to two dimensions. The resulting eigenfunctions are then stored for final ensemble averaging, which yields the 2D pmf according to the formula

$$\mathrm{pmf}(x, y) = -k_{\mathrm{B}}T\ln\left\langle\frac{\sum_{i=0}^{\infty}\Psi_i^2(x, y)e^{-E_i/kT}}{\sum_{i=0}^{\infty}e^{-E_i/kT}}\right\rangle$$

where x and y are the chosen internal coordinates.

Test calculations indicated that because of the smaller gaps between the eigenvalues in comparison with the 1D case, it is obligatory to include the eigenfunctions of the ground state and two subsequent excited states from the solution of the 2D vibrational problem during the pmf calculation. The one- and two-dimensional potentials of mean force (1D and 2D pmf) were calculated using programs written especially for the purpose of this project. The graphical representation of the obtained results was prepared using the VMD[59] and Gnuplot[60] programs.

Table 1. Average Values and Standard Deviations of the Distance Parameters of the Hydrogen Bridge in the Studied Molecules[a]

	HBZA		DMAP	
interatomic distance	average	SD	average	SD
O1−H2	1.053	0.085	1.018	0.031
O1...N3	2.601	0.108	2.769	0.178
H2...N3	1.641	0.162	1.857	0.223

[a] Results of the CPMD simulation. All values are in Å.

III. Results and Discussion

The models of the investigated Schiff (N-methyl-2-hydroxybenzylidene amine, HBZA) and Mannich (*ortho*-dimethylaminomethylphenol, DMAP) bases are presented in Figure 1. Car−Parrinello molecular dynamics simulations were performed in vacuo using the conditions applied during the experimental measurements of infrared (IR) spectra. Therefore, the correctness of the applied theoretical protocols was verified by comparison with the experimental data reported previously in the literature.[37,9] The features of the vibrational spectra are sensitive measures of the dynamic processes in the studied systems, and our attention will be concentrated on the properties of the intramolecular hydrogen bridge. Previous calculations for the compounds HBZA and DMAP performed on the basis of Density Functional Theory (DFT) and Møller-Plesset (MP2) perturbation theory[37,9] described the molecular properties based on static models; therefore, information on the intramolecular hydrogen bridge dynamics was not investigated. Our current study will discuss the results obtained on the basis of ab initio (CPMD) and path integral (PIMD) molecular dynamics and postprocessing analysis of the obtained trajectories.

The vibrational properties are direct derivatives of the time evolution of the structural parameters of the studied molecules. Therefore, the interatomic distances of the atoms involved in the intramolecular hydrogen bond were analyzed at the beginning of our study. The average values of the distance parameters of the hydrogen bridge are presented in Table 1 together with their standard deviations (SD). Increased SD might indicate delocalization events during the CPMD run. This might be suspected for the gas-phase simulation of HBZA, where not only the O−H bond length is longer than in DMAP but also the corresponding SD is almost three times larger. Accordingly, the donor−acceptor distance is shorter in HBZA and exhibits a smaller SD. This means that the intramolecular hydrogen bond of HBZA is stronger than that of DMAP. More details of the phenomena occurring in the analyzed bridges will be revealed in the time-domain analysis. The graphical representation of the obtained distances as a function of simulation time is presented in Figure 2. The main difference between the graphs for the investigated Schiff and Mannich bases is the presence of the bridged proton-transfer event after 3.5 ps of simulation time.

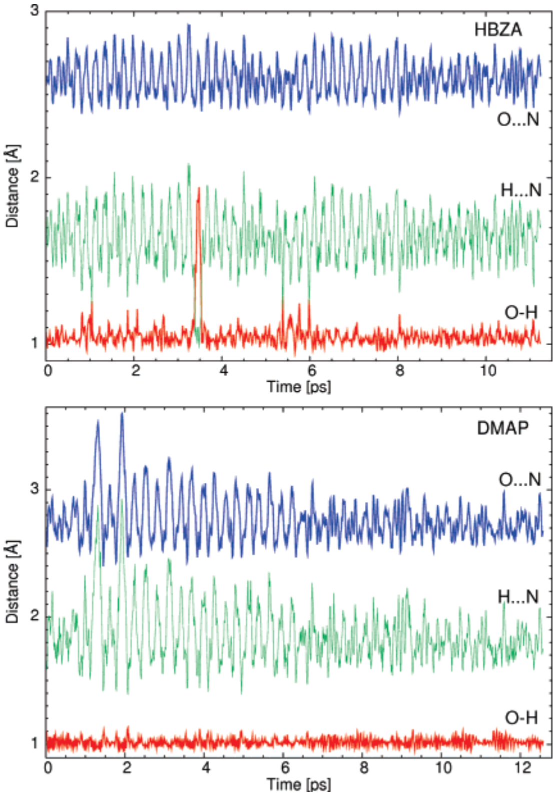


Figure 2. Time evolution of interatomic distances of atoms involved in the hydrogen bridge as a result of ab initio molecular dynamics.

Additionally, a few instances of equal donor−proton and acceptor−proton bond lengths are also visible which do not develop into full proton transfer. We have to stress that the transfer is not a permanent event but lasts only for ca. 0.2 ps, after which the proton returns to the donor site. In the case of the Mannich base, proton-transfer phenomena are not observed. The bridged proton remains totally at the oxygen atom, and the donor−proton distance is always significantly shorter than the proton−acceptor separation. This reflects the fact that the proton potential barrier is higher in DMAP than in HBZA, and it quite possibly has only one minimum at the donor site.

The atomic velocity power spectra generated from the CPMD trajectory of the investigated compounds are presented in Figure 3. In agreement with the discussion presented above, the spectra indicate various extents of proton delocalization in both molecules. The Schiff base HBZA is characterized experimentally by a very intense and broad absorption ascribed to the O−H stretching mode strongly coupled with other vibrational modes of the system.[37] The experimental data were collected in the gas phase at 300 K;[37] therefore, this is an internal property of the HBZA molecule, i.e., not induced or modified by environmental effects. Our calculations reflect this feature: the atomic velocity power spectrum for the bridged proton is almost continuous, with the high-frequency 2000−3000 cm^{-1} range corresponding to the O−H stretching mode. However, the gap between the low end of this band and the high end of the low-frequency proton modes is only 400 cm^{-1}. The experimentally available low-frequency part of the vibrational spectrum indeed shows bands at 1405, 1457, 1494, and 1639−1648 cm^{-1} which might correspond to our computational power spectra and which belong to modes coupled with bridge motions. The greater localization of the proton position in the dynamics of DMAP is related to the narrower proton absorption range (2550−3300 cm^{-1}) and the much larger separation from the low-frequency motions, which are present up to 1600 cm^{-1}. The experimental range attributed to the O−H stretching mode of DMAP's hydrogen bridge[9] is 2600−3450 cm^{-1}, centered at 3030 cm^{-1}. Our observations are strengthened by the dipole moment power spectra (Figure 3), which provide correct absorption intensities and are thus able to show the spectral effects of the hydrogen bond formation. In the case of HBZA, the calculated IR spectrum exhibits a broad region of absorption in the 2000−3000 cm^{-1} range. This feature does not appear in DMAP, a Mannich base with a markedly weaker intramolecular hydrogen bond. Concluding, we observed that our computational spectra are red-shifted with respect to the experimental data. For example, the maximum peak of the low-frequency region for HBZA is found experimentally[37] at 1639−1648 cm^{-1}, while our calculated value is 1580−1590 cm^{-1}. This red shift is a combined result of the use of the PBE functional within the framework of the DFT theory and application of Car−Parrinello dynamics, in which the fictitious mass used to propagate the electronic degrees of freedom introduces a delaying effect.

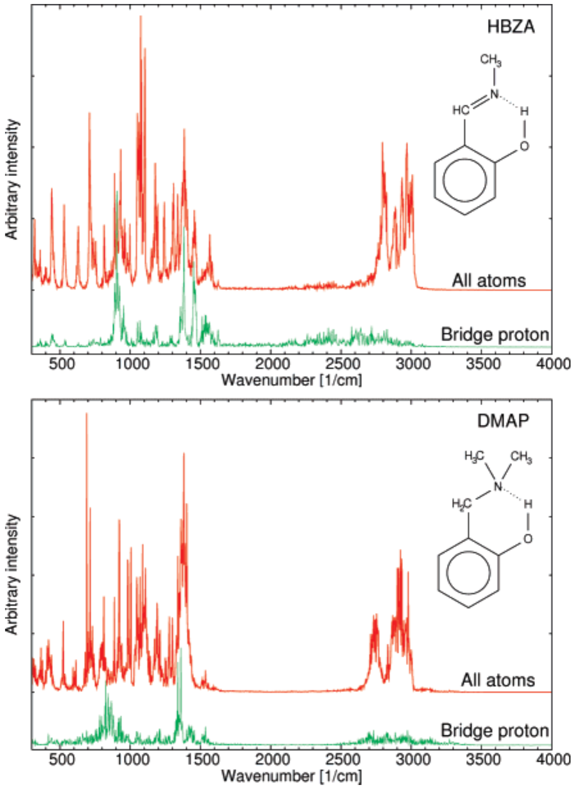


Figure 3. Power spectra of atomic velocities of all atoms and the hydrogen bridge proton. The intensities are in arbitrary units, while the wave numbers correspond to the actual vibrational features of the system.

Path Integral Molecular Dynamics (PIMD) was applied to investigate the influence of quantum effects on proton delocalization. High computational overhead of the PIMD scheme restricted the number of Trotter replicas in our study

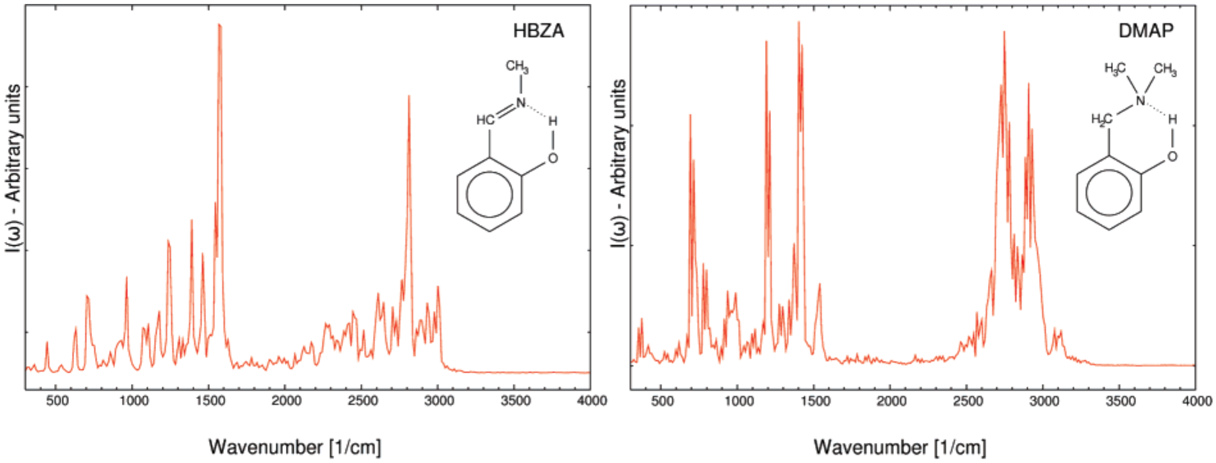


Figure 4. Predicted infrared spectra of studied Schiff and Mannich bases as results of ab initio molecular dynamics.

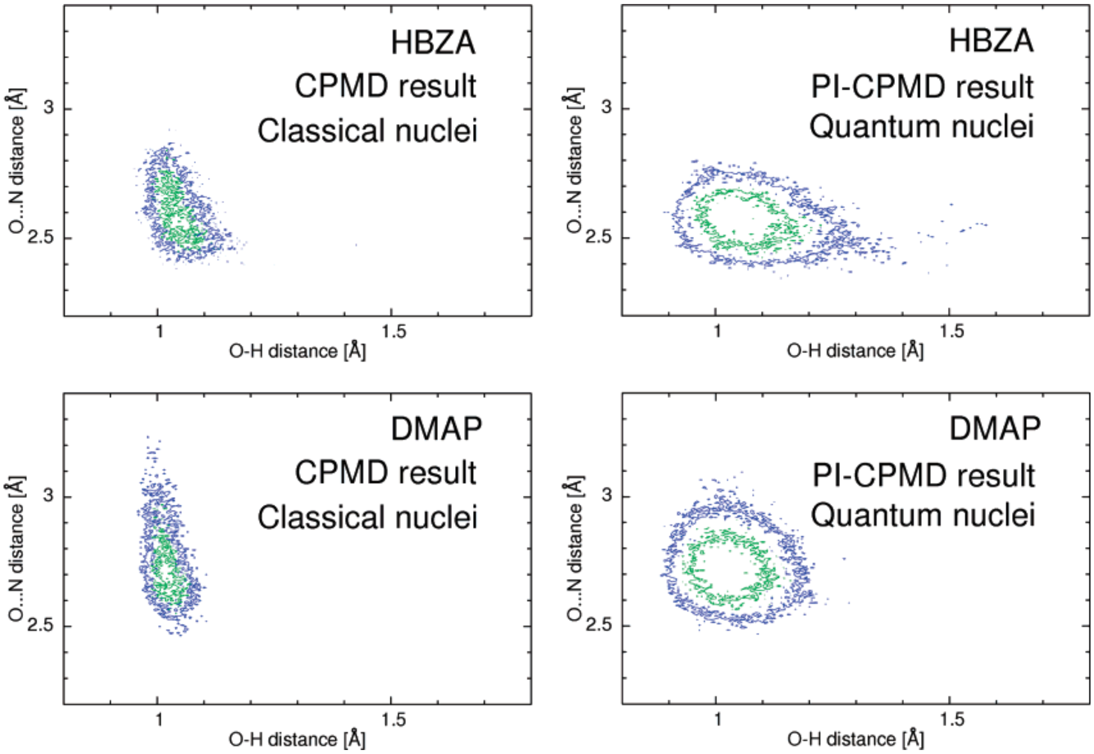


Figure 5. Comparison of the proton position in the hydrogen bridge reproduced by classical CPMD simulations (left side) and with quantum inclusion using the path integral (PI) method (right side) of studied Schiff (top) and Mannich (bottom) bases. The probability density isolines are 5 Å^{-2} for the blue line and 15 Å^{-2} for the green line.

to $P = 8$. Values of P as high as 16 were used to obtain converged potentials of mean force for model proton-transfer systems; the best example is a PIMD study on malonaldehyde.[61] However, the calculation of the pmf (described later in the text) proceeds in our case not from PIMD results but from a posteriori quantum corrections to the CPMD trajectory for the bridge proton. In view of the possible limitations of the convergence of the PIMD-derived pmf, we choose to discuss the effect of the quantization on the geometrical parameters only. Values of primitive[62] and virial[63] energy estimators provide an additional test of the PIMD convergence in terms of both simulation length and number of replicas. The PIMD run for HBZA yields the value of 0.06874 ± 0.01127 au for the primitive estimator and 0.06873 ± 0.00490 au for the virial expression. Corresponding run-averaged values and their standard deviations for the DMAP simulation are 0.09885 ± 0.01811 au (primitive estimator) and 0.09796 ± 0.00530 au (virial estimator), respectively. The agreement of average values of both estimators within a particular simulation suggests that the adopted calculation protocol provides reasonably converged results. Additionally, the standard deviations for the virial energy estimator are, correctly, smaller than for the primitive formula.[64] The tests described above enabled us to proceed with further analysis of the PIMD simulation. The two-dimensional (2D) histograms obtained for HBZA and DMAP are presented in Figure 5. A comparison of the results obtained on the basis of CPMD simulation (classical description of the nuclei) with PIMD is given in this chart. The two chosen coordinates, i.e., donor−acceptor distance and

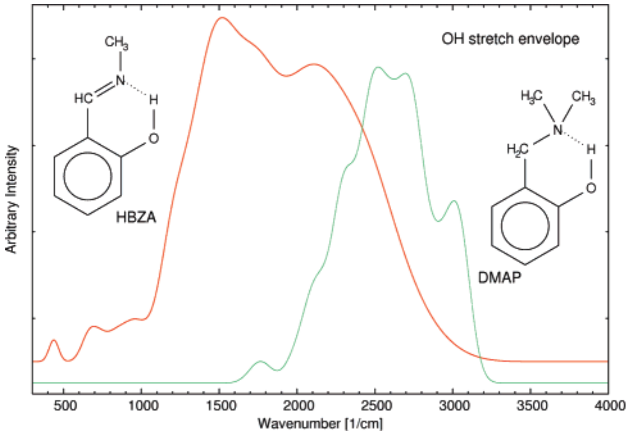


Figure 6. Simulated hydrogen bridge O−H stretching envelope in gas phase (arbitrary intensities on the Y axis) of the studied Schiff and Mannich bases.

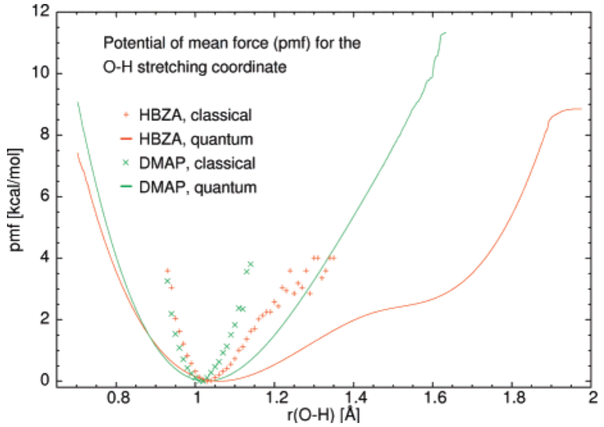


Figure 7. One-dimensional potential of mean force (1D pmf) for the O−H stretching mode calculated on the basis of CPMD simulation of the studied Schiff and Mannich bases. Markers denote the pmf resulting from classical nuclear probability density, while the solid lines are pmf curves augmented with quantization of nuclear motions.

O−H bond length, provide the most important characteristics of the investigated intramolecular hydrogen bond. Quantum effects seem to have a significantly stronger influence on the proton dynamics in HBZA compared with the DMAP molecule. The provided data analysis indicates the qualitative change in the proton’s position (stronger delocalization) in the Schiff base. On the other hand, the classical CPMD description of the Mannich base does not exhibit significant differences from the PIMD histogram. The probability of proton transfer is higher at shorter donor−acceptor distances. It is worth mentioning that even the additional kinetic energy introduced by the higher simulation temperature (in the case of the Mannich base) was not able to promote proton transfer within the time frame of the simulation. Detailed analysis of Figure 5 reveals that the inclusion of quantum effects in the nuclear dynamics affects not only the proton’s position but also the bridge as a whole. For HBZA, the histogram for donor−acceptor distances covers the range of 2.40−2.85 Å in the CPMD simulation, which shortens slightly to 2.40−2.80 Å when the PIMD technique is used. A similarly small shortening of the bridge due to quantum effects is also visible in DMAP, where the corresponding ranges are 2.50−3.05 Å for CPMD and 2.50−3.00 Å for PIMD. The respective values of the O−H distances indicate increased proton delocalization in PIMD, as mentioned above. In HBZA, the O−H bond lengths range from 0.98−1.15 Å for classical nuclei to 0.92−1.35 Å in the PIMD simulation. In the latter case there are also isolated instances of much larger O−H separations, corresponding to instantaneous proton-transfer events. DMAP does not display such a behavior, but the accessible O−H distance range increases from 0.98−1.10 Å in CPMD to 0.90−1.20 Å for quantized nuclear degrees of freedom. In summary, inclusion of a quantum description of the nuclei seems to affect the Mannich base to a smaller degree than the Schiff base and does not change the qualitative description of the hydrogen bridge in DMAP.

A posteriori inclusion of quantum effects of the proton motion provides us with the possibility of computing the spectral features corresponding to the proton motion in the bridge (Figure 6) and the one-dimensional potential of mean force (1D pmf) for the O−H coordinate (Figure 7). Figure 6 shows that for DMAP the calculated spectral feature corresponds to the O−H stretching region, while for the stronger intramolecular hydrogen bond the coupling between various modes broadens the feature toward the low-frequency part of the spectrum. This leads to the formation of an almost continuous, very broad band. Comparison with Figure 3, where classical power spectrum is presented, shows that the broadening is a result of quantum corrections allowing the proton to probe larger range of positions, where anharmonicity is pronounced much stronger. This is more important for the HBZA with its stronger hydrogen bond, than for DMAP. The corresponding 1D pmf for the O−H coordinate (Figure 7) reflects this difference in the proton dynamics. Most significantly, the classical pmf calculated directly from the O−H probability distribution of the CPMD run is more localized than the result of a posteriori quantum corrections. The difference between HBZA and DMAP proton dynamics is visible even in the classical pmf. However, it suffers from inadequate statistical sampling above 3.0 kcal/mol (oscillations in Figure 7), and further we discuss only the quantum-corrected results. While for DMAP the potential is anharmonic but similar to the Morse potential, the HBZA molecule has a very asymmetric 1D pmf with a distinct shoulder related to possible instantaneous proton-transfer events. The minimum of the 1D pmf for DMAP (containing quantum effects) corresponds to the average O−H distance from the classical-nuclei CPMD simulation (Table 1), confirming the PIMD result stating that the quantization of proton motion does not have a strong impact on the ensemble averages in the case of the Mannich base. The 1D pmf minimum is shifted to a larger O−H bond length value (ca. 1.07 Å) for HBZA, but the difference from the classical simulation is visible rather in the flattening of the potential.

A more detailed overview of the proton dynamics in the hydrogen bridge is provided by the graphs of the two-dimensional potential of mean force (2D pmf) for both compounds (Figure 8). The choice of the analyzed coordinates, the O−H distance and O−H...N angle, was dictated by their structural relevance to proton mobility. We decided not to discuss the motion of the heavy atoms involved in

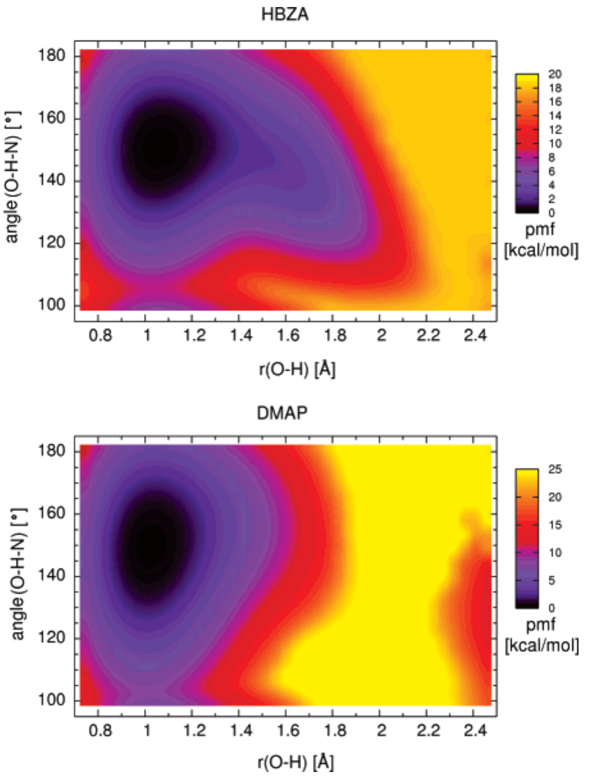


Figure 8. Two-dimensional potential of mean force (2D pmf) for the O−H stretching mode calculated on the basis of CPMD simulation augmented by quantization of nuclear motions. Upper graph: results for the Schiff base, lower graph: for the Mannich base.

the hydrogen bridge, i.e., the O...N distance is not a coordinate chosen for the graph. Instead, the graph provides a statistically averaged potential for the proton motion “as seen by the proton” and defined by the coordinates directly involving this nucleus. The 2D pmf map for HBZA exhibits a potential well centered at r(O−H) equal to 1.05 Å and an O−H...N angle equal to 150°. The same parameters are found for the Mannich base. However, a second, very shallow minimum of the pmf is present in HBZA but not in DMAP. The coordinates of this feature are $r(\mathrm{O-H}) = 1.60$ Å and an O−H...N angle of 140°. This secondary minimum corresponds to instantaneous proton-transfer phenomena, which are more probable within a quantized proton-motion framework. The picture obtained by investigation of the 2D free energy profile is totally consistent with the PIMD results for both compounds.

The methodology of calculation of the pmf by a posteriori quantum corrections to the CPMD trajectory snapshots, used previously in 1D case[55] and extended here to a 2D problem, is an addition to the large set of schemes for free energy computation. Methods such as replica exchange molecular dynamics[65] can provide rapid access to the complicated conformational space of macromolecules, describing, e.g., the process of reversible protein folding.[66] Constrained molecular dynamics[67] also provides a convenient route for obtaining free energy profiles directly from the MD simulation. Additionally, in the low-dimensional case there exist numerous enhanced sampling methods, a few of which are discussed below. Umbrella sampling, in its native[68,69] or semiautomated, adaptative version,[70] modifies the potential energy surface (PES) of the system to overcome locality of standard Boltzmann sampling. Metadynamics[71] modifies the PES with history-dependent potential terms which fill up the PES minima allowing for further recovery of rare-event statistics, which is especially useful in the context of first-principle molecular dynamics.[72] Finally, adiabatic free energy dynamics[73] separates reaction coordinate subspace from the rest of the phase space and enhances probability of rare events by applying large temperature to the reaction subspace only. The methodologies described above are general and applicable to any reaction coordinate. Moreover, they can be implemented within classical MD, first-principle MD, or path integral nuclear quantization schemes. The pmf calculation method of the current paper is, in principle, restricted to the description of localized vibrational coordinates. It is intended as a relatively inexpensive way of further analysis of a classical-nuclei CPMD trajectory, providing at the same time pmf profile and spectral signature of a selected coordinate. Therefore, this method can be advantageous in studies emphasizing important local interactions, especially intra- or intermolecular hydrogen bonds.

IV. Conclusions

Our computational investigations showed that the applied methodologies were able to describe faithfully the molecular properties of the studied Schiff and Mannich bases. Car−Parrinello molecular dynamics augmented by a posteriori quantum corrections is particularly valuable in studies of hydrogen bridge dynamics. The application of path integral molecular dynamics showed that in the case of the bridged proton of the studied Schiff base, the quantum effects improve the description of the proton’s position in the intramolecular hydrogen bond and enable the possibility of proton transfer. The bridged proton position is localized on the donor side in the Mannich base. The vibrational properties were also analyzed, and they are closely related to the available experimental measures of proton delocalization. The one- and two-dimensional free energy profiles were obtained from the eigenfunctions of the vibrational states of the studied molecules. The resulting potential of mean force for the proton motion describes the proton-transfer pathway, which includes quantum corrections to the classical picture. In the case of the Schiff base, proton transfer can occur with a quite large probability, while in the Mannich base the proton is mostly localized on the donor side. The computational strategy employed in the study has shown its potential for describing the influence of structural modifications of the molecular skeleton on the properties of the O−H...N hydrogen bridge. We are currently extending the use of the methodology to other strongly anharmonic systems. The proposed 2D pmf technique could be applied to study the free energy profiles of large systems with biological relevance, especially when coupled with the QM/MM simulation framework.

Acknowledgment. The authors would like to thank Dr. Sonia M. Melikova (University of St. Petersburg), Dr. Riccardo Mazzarello (SISSA, Trieste), and Prof. Aleksander

Koll (University of Wrocław) for fruitful discussions and critical reading of the manuscript and Dr. Harald Forbert (Ruhr-Universität Bochum) for the program for dipole moment Fourier transforms. In addition, the authors gratefully acknowledge the Wrocław Center for Networking and Supercomputing (WCSS), the Academic Computer Center CYFRONET-KRAKÓW (grants KBN/SGI/UWrocl/029/1998 and KBN/SGI/UWrocl/078/2001), and the Poznań Supercomputing and Networking Center for providing computer time and facilities.

References

(1) Warshel, A. *Computer Modelling of Chemical Reactions in Enzymes and Solutions*, 1st ed.; Wiley: New York, NY, 1997; pp 136−152.

(2) Jeffrey, G. A. *An Introduction to Hydrogen Bonding*, 1st ed.; Oxford University Press: New York, NY, 1997; pp 184−212.

(3) Cleland, W. W.; Kreevoy, M. M. *Science* **1994**, *264*, 1887−1890.

(4) Gao, J.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467−505.

(5) Denisov, G. S.; Mavri, J.; Sobczyk, L. Potential energy shape for the proton motion in hydrogen bonds reflected in infrared and NMR spectra. In *Hydrogen bonding - new insights, (Challenges and advances in computational chemistry and physics*, *3*), 1st ed.; Grabowski, S. J., Ed.; Springer: Dordrecht, The Netherlands, 2006; pp 377−416.

(6) Zundel, G. Hydrogen-bonded systems with large proton polarizability due to collective proton motion as pathways of protons in biological systems. In Müller, A.; Ratajczak, H.; Junge, W.; Diemann, E. *Electron and Proton Transfer in Chemistry and Biology*; Elsevier: Amsterdam, The Netherlands, 1992; pp 313−327.

(7) Król-Starzomska, I.; Filarowski, A.; Rospenk, M.; Koll, A.; Melikova, S. *J. Phys. Chem. A* **2004**, *108*, 2131−2138.

(8) Filarowski, A. *J. Phys. Org. Chem.* **2005**, *18*, 686−698.

(9) Koll, A.; Melikova, S. M.; Karpfen, A.; Wolschann, P. *J. Mol. Struct.* **2001**, *559*, 127−145.

(10) Filarowski, A.; Koll, A.; Głowiak, T. *J. Chem. Soc. Perkin Trans.* **2002**, *2*, 835−842.

(11) Filarowski, A.; Głowiak, T.; Koll, A. *J. Mol. Struct.* **1999**, *484*, 75−89.

(12) Rospenk, M.; Król-Starzomska, I.; Filarowski, A.; Koll, A. *Chem. Phys.* **2003**, *287*, 113−124.

(13) Filarowski, A.; Koll, A.; Głowiak, T. *J. Mol. Struct.* **2002**, *615*, 97−108.

(14) Filarowski, A.; Koll, A. *Vib. Spectrosc.* **1996**, *12*, 15−24.

(15) Schilf, W.; Kamiński, B.; Kołodziej, B.; Grech, E.; Rozwadowski, Z.; Dziembowska, T. *J. Mol. Struct.* **2002**, *615*, 141−146.

(16) Washington, I.; Brooks, C.; Turro, N. J.; Nakanishi, K. *J. Am. Chem. Soc.* **2004**, *126*, 9892−9893.

(17) Shimono, K.; Furutani, Y.; Kamo, N.; Kandori, H. *Biochemistry* **2003**, *42*, 7801−7806.

(18) Maeda, A.; Gennis, R. B.; Balashov, S. P.; Ebrey, T. G. *Biochemistry* **2005**, *44*, 5960−5968.

(19) Furutani, Y.; Shichida, Y.; Kandori, H. *Biochemistry* **2003**, *42*, 9619−9625.

(20) Tanimoto, T.; Furutani, Y.; Kandori, H. *Biochemistry* **2003**, *42*, 2300−2306.

(21) Lee, Y-S.; Krauss, M. *J. Am. Chem. Soc.* **2004**, *126*, 2225−2230.

(22) Hadjoudis, E.; Vittorakis, M.; Moustakali-Mavridis, I. *Tetrahedron* **1987**, *43*, 1345−1360.

(23) Harada, J.; Uekusa, H.; Ohashi, Y. *J. Am. Chem. Soc.* **1999**, *121*, 5809−5810.

(24) Evans, O. R.; Lin, W. *Acc. Chem. Res.* **2002**, *35*, 511−522.

(25) Toftlund, H. *Coord. Chem. Rev.* **1989**, *94*, 67−108.

(26) León-Clemente, M.; Coronado, E.; Delhaes, P.; Galán Mascarós, J. R.: Gómez-García, C. J.; Mingotaud, C. Hybrid materials formed by two molecular networks. Magnetic conductors, magnetic multilayers and magnetic films. In *Supramolecular Engineering of Synthetic Metallic Materials: Conductors and Magnets*, *NATO ASI Series*; Veciana, J., Rovira, C., Amabilino, D. B., Eds.; Kluwer: Dordrecht, The Netherlands, 1998; Vol. C-518, pp 291−312.

(27) Vashishtha, S. C.; Zello, G. A.; Nienaber, K. H.; Balzarini, J.; De Clercq, E.; Stables, J. P.; Dimmock, J. R. *Eur. J. Med. Chem.* **2004**, *39*, 27−35.

(28) Holla, B. S.; Veerendra, B.; Shivananda, M. K.; Poojary, B. *Eur. J. Med. Chem.* **2003**, *38*, 759−767.

(29) Malinka, W.; Karczmarzyk, Z.; Sieklucka-Dziuba, M.; Sadowski, M.; Kleinrok, Z. *Il Farmaco* **2001**, *56*, 905−918.

(30) Sridhar, S. K.; Saravanan, M.; Ramesh, A. *Eur. J. Med. Chem.* **2001**, *36*, 615−625.

(31) Sriram, D.; Bal, T. R.; Yogeeswari, P. *Med. Chem. Res.* **2005**, *14*, 211−228.

(32) Goldstein, E. J. *Am. J. Med.* **1987**, *82*, 3−17.

(33) Pandeya, S. N.; Sriram, D.; Nath, G.; De Clercq, E. *Eur. J. Med. Chem.* **2000**, *35*, 249−255.

(34) Pandeya, S. N.; Sriram, D.; Yogeeswari, P.; Ananthan, S. *Chemotherapy* **2001**, *47*, 266−269.

(35) Kleist, R. A.; Gutierrez, A.; Lundberg, R. D.; Song, W. R. U.S. Patent No. 5,433,874 (July 18, 1995).

(36) Air Products and Chemicals, Inc., U.S. Patent No. 5,854,-312 (December 29, 1998).

(37) Filarowski, A.; Koll, A.; Karpfen, A.; Wolschann, P. *Chem. Phys.* **2004**, *297*, 323−332.

(38) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471−2474.

(39) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864−B871.

(40) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133−A1138.

(41) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300−313.

(42) Schlegel, H. B. *Theor. Chim. Acta* **1984**, *66*, 333−340.

(43) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.

(44) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993−2006.

(45) Marx, D.; Parrinello, M. *Science* **1996**, *271*, 179−181.

(46) Marx, D.; Parrinello, M. *J. Chem. Phys.* **1996**, *104*, 4077−4082.

(47) Tuckerman, M. E.; Marx, D.; Klein, M. L.; Parrinello, M. *J. Chem. Phys.* **1996**, *104*, 5579−5588.

(48) Nosé, S. *Mol. Phys.* **1984**, *52*, 255−268.

(49) Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511−519.

(50) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695−1697.

(51) Martyna, G. J.; Klein, M. L.; Tuckerman, M. E. *J. Chem. Phys.* **1992**, *97*, 2635−2643.

(52) Tuckerman, M. E.; Berne, B. J.; Martyna, G. J.; Klein, M. L. *J. Chem. Phys.* **1993**, *99*, 2796−2808.

(53) Stare, J.; Mavri, J. *Comput. Phys. Commun.* **2002**, *143*, 222−240.

(54) Panek, J.; Stare, J.; Hadži, D. *J. Phys. Chem. A* **2004**, *108*, 7417−7423.

(55) Jezierska, A.; Panek, J. J.; Koll, A.; Mavri, J. *J. Chem. Phys.* **2007**, *126*, 205101.

(56) Jezierska, A.; Panek, J. J.; Borštnik, U.; Mavri, J.; Janežič, D. *J. Phys. Chem. B* **2007**, *111*, 5243−5248.

(57) CPMD, Copyright IBM Corp. 1990−2004, Copyright MPI fuer Festkoerperforschung Stuttgart, 1997−2001.

(58) Ventura, K. M.; Greene, S. N.; Halkides, C. J.; Messina, M. *Struct. Chem.* **2001**, *12*, 23−31.

(59) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33−38.

(60) Williams, T.; Colin, K. Copyright 1986−1993, 1998, 2004.

(61) Tuckerman, M. E.; Marx, D. *Phys. Rev. Lett.* **2001**, *86*, 4946−4949.

(62) Barker, J. A. *J. Chem. Phys.* **1979**, *70*, 2914−2918.

(63) Herman, M. F.; Bruskin, E. J.; Berne, B. J. *J. Chem. Phys.* **1982**, *76*, 5150−5155.

(64) Parrinello, M.; Rahman, *J. Chem. Phys.* **1984**, *80*, 860−867.

(65) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolym. Pept. Sci.* **2001**, *60*, 96−123.

(66) Rao, F.; Caflisch, A. *J. Chem. Phys.* **2003**, *119*, 4035−4042.

(67) Sprik, M.; Ciccotti, G. *J. Chem. Phys.* **1998**, *109*, 7737−7744.

(68) Patey, G. N.; Valleau, J. *J. Chem. Phys.* **1975**, *63*, 2334−2339.

(69) Torrie, G.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187−199.

(70) Mezei, M. *J. Comput. Phys.* **1987**, *68*, 237−240.

(71) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562−12566.

(72) Asciutto, E.; Sagui, C. *J. Phys. Chem. A* **2005**, *109*, 7682−7687.

(73) Rosso, L.; Tuckerman, M. E. *Mol. Simul.* **2002**, *28*, 91−112.

CT7002644

JCTC Journal of Chemical Theory and Computation

Solutions of the Optimized Closure Integral Equation Theory: Heteronuclear Polyatomic Fluids

M. Marucho,[†] C. T. Kelley,[‡] and B. Montgomery Pettitt*,[†]

Chemistry Department, University of Houston, Houston, Texas 77204-5003, and Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205

Received August 13, 2007

Abstract: Recently, we developed a thermodynamically optimized integral equation method which has been successfully tested on both simple and homonuclear diatomic Lennard-Jones fluids [*J. Chem. Phys*. **2007***, 126*, 124107]. The systematic evaluation of correlation functions required by the optimization of the chemical potential has shown a clear need for more efficient algorithms to solve these integral equations. In the present paper we introduce a high-performance algorithm which is found to be faster and more efficient than the direct Picard iteration. Here we have utilized this to solve the aforementioned optimized theory for molecules more complex than those considered previously. We analyzed representative models for heteronuclear diatomic and triatomic polar molecular fluids. We include results for several modified SPC-like models for water, obtaining site−site correlation functions in good agreement with simulation data.

I. Introduction

Solvents play an important role in many physical, biological, and chemical processes. Indeed, they govern conformational stability, binding, and functions of biomolecules as well as the details of chemical reactions.[1] Including such solvent effects via computer simulation techniques can be costly since the resulting systems involve a large number of particles with long-range interactions and demand substantial computing time and memory storage.

Alternatively, a variety of computational schemes have been developed to capture the dominant solvation effects on large solutes using approximate representations of solvents. The simplest proposals, namely the continuum models, neglect many essential correlations in solvents and are thus seriously limited.[2] Integral equation theories (IET) offer a compromise between accuracy and computational expense.[3] They provide a powerful theoretical tool for computing approximate pair distribution functions, $g(r)$, from which it is possible to get structural and thermodynamic properties for fluids.[1,4,5] These theories are usually formulated by providing two set of equations, namely the integral equation and a closure relationship. Beyond obtaining a good qualitative description of such systems, closure approximations borrowed from atomic fluids like Percus−Yevick(PY)[6,7] and hypernetted chain (HNC)[8,9] have not produced uniformly good quantitative results for molecular fluids in general. Whether this is due to the propagating integral equation or the closure is not always clear.[10]

As with methods arising from interaction site model representations of molecular liquids,[10] these approximations generate thermodynamic inconsistency as well as quantitative limitations in the description of the short- and long-range structure.[11−21] Indeed, to find an approximate theory describing these properties as accurately as possible and at low computational cost requires the development of more sophisticated theories. In fact, during the last quarter of the century, a significant number of different approaches have been proposed to overcome those difficulties, including new integral equations and/or new closure approximations.[22−29]

Recently, we have developed a thermodynamically consistent integral equation theory which has been successfully tested on both simple and homonuclear diatomic Lennard-

* Corresponding author e-mail: pettitt@uh.edu.

[†] University of Houston.

[‡] North Carolina State University.

10.1021/ct700202h CCC: $40.75
Published on Web 02/21/2008

Jones fluids.[30] A new closure approximation is obtained by using the Percus' functional expansion to first order in the density as a target functional generator.[6,31,32] It depends on parameters providing a smooth transition interpolating between PY and HNC closures which can be variationally optimized. The chemical potential functional is minimized then to yield the values of the parameters. This optimized closure approximation can be coupled with a variant of the diagrammatically proper integral equation introduced by this laboratory,[29] having a density matrix with nondiagonal elements containing screened densities. The simplicity of the expressions involved in the resulting theory have allowed us to also obtain an approximate analytic expression for the molecular excess chemical potential which is minimized to estimate the numerical value of the free parameters that defines the closure. Indeed, the success of this approach is largely due to the fact that for molecular fluids in general there are regions of the phase diagram where PY and HNC usually bracket the simulated pressure and densities. In such a case, one achieves thermodynamic self-consistency at the free energy optimum.

Given that approach, the purpose of this paper centers on implementing a more efficient numerical approach to solve the aforementioned optimized closure approximation for molecules more complex than the homonuclear diatomic fluids considered previously and in particular water. The paper is organized as follows. In section II, we introduce the approximate theory including the expression for the closure approximation, the integral equation, and the analytic approximation for the molecular excess chemical potential μ_{ex}. In section III, we describe the computational scheme used to solve these equations numerically. We also compare this algorithm with direct Picard iteration and other solvers to compare with the literature.[34−38] In section IV, we present the numerical solution obtained for the pair site−site correlation function of two representative models of heteronuclear Lennard-Jones fluids as well as of that modified SPC models for water. Further, we compare our predictions with the corresponding results of MD simulation. Finally, in section V, we summarize the central finding of the article, leaving the details of our computational scheme for the Appendix.

II. Theory

In our recently introduced optimized integral equation theory for homonuclear fluids[30] we found that the optimal parametrization is not universal. As anticipated, it depends on the thermodynamic state, the internal structure of the molecules, and consequently on the different species conforming the molecules. In this way, the extension of the closure approximation for a diagrammatically proper interaction site model (PISM) representation of heteronuclear molecular fluids[24,26,27,39,40] is easily obtained by assigning species labels to the parametrization performed for one-component fluids. Specifically, it reads

$$c^o_{\alpha\gamma}(r) = -\,a^o_{\alpha\gamma}e^{-\beta u_{\alpha\gamma}(r)} + (1 + a^o_{\alpha\gamma})e^{[-\beta u_{\alpha\gamma}(r)+t_{\alpha\gamma}o(r)/(1+a_{\alpha\gamma}o)]} - \\ 1 - t^o_{\alpha\gamma}(r)$$

$$c^r_{\alpha\gamma}(r) = \frac{(1 + a^o_{\alpha\gamma})}{(1 + a^r_{\alpha\gamma})}t^r_{\alpha\gamma}(r)e^{[-\beta u_{\alpha\gamma}(r)+t^o_{\alpha\gamma}(r)/(1+a^o_{\alpha\gamma})]} - t^r_{\alpha\gamma}(r)$$

$$c^l_{\alpha\gamma}(r) = \frac{(1 + a^o_{\alpha\gamma})}{(1 + a^r_{\gamma\alpha})}t^l_{\alpha\gamma}(r)e^{[-\beta u_{\alpha\gamma}(r)+t^o_{\alpha\gamma}(r)/(1+a^o_{\alpha\gamma})]} - t^l_{\alpha\gamma}(r)$$

$$c^b_{\alpha\gamma}(r) = (1 + a^o_{\alpha\gamma})\left[\frac{t^b_{\alpha\gamma}(r)}{(1 + a_\alpha\gamma^b)} + \frac{t^r_{\alpha\gamma}(r)t^l_{\alpha\gamma}(r)}{(1 + a^r_{\alpha\gamma})(1 + a^r_{\gamma\alpha})}\right] \times \\ \mathrm{e}^{[-\beta u_{\alpha\gamma}(r)+t^o_{\alpha\gamma}(r)/(1+a^o_{\alpha\gamma})]} - t^b_{\alpha\gamma}(r) \quad (1)$$

where $t^i_{\alpha\gamma}(r)$ and $c^i_{\alpha\gamma}(r)$ represent the contribution associated with the *i*th group of terms to the indirect and direct site−site correlation function between sites α and γ, respectively. Here the superscript represents the proper subclasses ($i = o$, l, r, or b) for none, left, right, and both sets of integrals in standard notation.[22,23] The set of parameters, $a^i_{\alpha\gamma}$, is composed of unknowns to be determined in the optimization. The site−site correlation functions are given by the sum of the four components, for instance for the indirect part

$$t_{\alpha\gamma}(r) = t^o_{\alpha\gamma}(r) + t^r_{\alpha\gamma}(r) + t^l_{\alpha\gamma}(r) + t^b_{\alpha\gamma}(r)$$

and so on for h and c. In the latter expression $h_{\alpha\gamma}(r) = g_{\alpha\gamma}(r) - 1$ represents the total site−site correlation function. This closure becomes PISM-HNC and PISM-PY approximations[24] for $a^o_{\alpha\gamma} = a^r_{\alpha\gamma} = a^b_{\alpha\gamma} = a_{\alpha\gamma}$ and $a_{\alpha\gamma} \to 0$ and $a_{\alpha\gamma} \to \infty$, respectively. Expressions 1 can be seen to be a simple generalization of our approximation for heteronuclear proper site−site molecular fluids. Moreover, in the absence of intramolecular correlations, only the o elements are nonzero, and the closure for $c^o_{\alpha\gamma}(r)$ becomes the result previously obtained for multicomponent atomic fluids. We refer the reader to ref 30 for a detailed description of the previous approach.

This approximate theory is completed by coupling the closure approximation 1 with a recently introduced integral equation[29] which, in Fourier space, reads

$$\hat{\mathbf{H}}(k) = \hat{\mathbf{C}}(k) + [\hat{\mathbf{C}}(k) + \hat{\mathbf{S}}(k)]\hat{\bar{\boldsymbol{\rho}}}[\hat{\mathbf{H}}(k) + \hat{\mathbf{S}}(k)] \quad (2)$$

where $\hat{\mathbf{H}}(k)$ and $\hat{\mathbf{C}}(k)$ are the Fourier transform of $\mathbf{H}(r)$ and $\mathbf{C}(r)$, respectively. Each of the correlation functions appearing above is a symmetric matrix in the form

$$\mathbf{Q}(k) = \begin{pmatrix} \mathbf{Q}_{11}(k) & \mathbf{Q}_{12}(k) \\ \mathbf{Q}_{21}(k) & \mathbf{Q}_{22}(k) \end{pmatrix}, \quad \mathbf{Q}_{\alpha\gamma}(k) = \begin{pmatrix} q^o_{\alpha\gamma}(k) & q^r_{\alpha\gamma}(k) \\ q^l_{\alpha\gamma}(k) & q^b_{\alpha\gamma}(k) \end{pmatrix}$$

where $\mathbf{Q}$ represents $\mathbf{H}$ or $\mathbf{C}$. The screened density and the renormalized intramolecular matrices $\bar{\boldsymbol{\rho}}$ and $\hat{\mathbf{S}}(k)$ for triatomic heteronuclear molecules are given by the following expressions

$$\bar{\boldsymbol{\rho}} = \begin{pmatrix} \rho & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \rho & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \rho \end{pmatrix}, \quad \rho = \begin{pmatrix} \rho & \eta \\ \eta & 0 \end{pmatrix}, \quad \hat{\mathbf{S}}(k) = \begin{pmatrix} 0 & \hat{\mathbf{s}}_{12}(k) & \hat{\mathbf{s}}_{13}(k) \\ \hat{\mathbf{s}}_{21}(k) & 0 & \hat{\mathbf{s}}_{23}(k) \\ \hat{\mathbf{s}}_{31}(k) & \hat{\mathbf{s}}_{32}(k) & 0 \end{pmatrix},$$

$$\hat{\mathbf{s}}_{ij}(k) = \begin{pmatrix} 0 & 0 \\ 0 & (\sin(kL_{ij})/\eta kL_{ij}) \end{pmatrix} \quad (3)$$

with η being the screened density whose approximation to first order[29] can be numerically computed by the following expression

$$\eta = \rho/[1 + \rho \int d\mathbf{r}_2\{f_{21}(r_{12})f_{11}(r_2) + f_{23}(r_{12})f_{31}(r_2) + f_{22}(r_{12})f_{21}(r_2)\}|_{r_1=L_{12}} + \rho\int d\mathbf{r}_2\{f_{33}(r_{12})f_{31}(r_2) + f_{32}(r_{12})f_{21}(r_2) + f_{31}(r_{12})f_{11}(r_2)\}|_{r_1=L_{13}}] \quad (4)$$

In the latter formula, ρ represents the molecular density, $L_{\alpha\gamma}$ is the bond length between the two sites α and γ, and $f_{\alpha\gamma}$ is the corresponding Mayer function.

The simplicity of the previous expressions has allowed us to also obtain an approximate analytic expression for the molecular excess chemical potential μ_{ex}. This is convenient; much as with density functional theories this is the quantity which is minimized to estimate the value of the free parameters that define the closure approximation 1. In fact, it is obtained by a straightforward extension of the calculation carried out in section IV of ref 30. The resulting expression reads

$$\beta\mu_{ex}(\{a^i_{ab}\}) \simeq - \rho \sum_{\alpha\gamma} \int d\mathbf{r}\Bigg\{h_{\alpha\gamma}(r) - \mathscr{I}[t^o_{\alpha\gamma}(r),h^o_{\alpha\gamma}(r),a^o_{\alpha\gamma}] \Bigg[t^o_{\alpha\gamma}(r) + (1 + a^o_{\alpha\gamma})\Bigg[\frac{t^r_{\alpha\gamma}(r)}{(1 + a^r_{\alpha\gamma})} + \frac{t^l_{\alpha\gamma}(r)}{(1 + a^r_{\gamma\alpha})}\Bigg] + \frac{(1 + a^o_{\alpha\gamma})t^b_{\alpha\gamma}(r)}{(1 + a^b_{\alpha\gamma})}\Bigg] - \frac{h^r_{\alpha\gamma}(r)t^l_{\alpha\gamma}(r)}{(1 + a^r_{\gamma\alpha})} - \frac{[h^r_{\alpha\gamma}(r) + h^l_{\alpha\gamma}(r) + h^b_{\alpha\gamma}(r)]t^o_{\alpha\gamma}(r)}{2(1 + a^o_{\alpha\gamma})} \quad (5)$$

where $\mathscr{I}[u, v, w]$ is given by the following formula (eq 12 in ref 30)

$$\mathscr{I}[u, v, w] = \frac{\Big\{[v + 1]\ln\Big[\frac{y}{w}\Big] + \ln[w] + \frac{v(w + 1)}{u}\mathrm{Re}\Big[Li_2\Big(\frac{y}{w} + 1\Big) - Li_2\Big(\frac{w + 1}{w}\Big)\Big]\Big\}}{u} \quad (6)$$

with $Li_2(x)$ being the well-known dilogarithm function[41,42] and $y = - w + (w + 1) \exp[u/(w + 1)]$.

The aforementioned optimization procedure certainly requires a fast and efficient computational scheme to evaluate the correlation functions appearing in expression 5. This topic is analyzed in the next section, where a high-performance numerical algorithm is presented.

III. Numerical Analysis

As shown initially by Gillan for atomic fluids,[43] the multiresolution representation of correlation functions into a "coarse'' part, which is expanded in a set of bases functions with coefficients determined by the Newton-Rapshon (NR) scheme, and the "fine'' part, which is evaluated numerically by direct Picard iteration, leads to a significant reduction in the computational complexity and cost to solve liquid state integral equation theories. Subsequently modified and adapted for more complex fluids, this method has provided excellent results in general.[44−46] In fact, it is well-known that a fairly rapidly convergence to a solution is achieved when these iterative solvers are started from a sufficiently good initial guess. Further, the choice of the basis functions employed to expand the correlation functions and the method utilized to determine the corresponding weight of their projections play a fundamental role in the rate of convergence. These methods, however, require the analytical expression for the Jacobian matrix to be computationally most efficient.

Based on these considerations, the numerical method proposed in the present study consists of expanding the indirect correlation function on a sine function basis set, a natural expansion within a scheme that already involves a systematic evaluation of sine Fourier transforms. Indeed, unidimensional transforms appear in the integral eq 2 using the angular average of the spherically symmetric correlation functions defined in the three-dimensional Fourier transforms. As shown in ref 45, the so named "unidimensional'' correlation functions

$$\mathscr{T}_{\alpha\gamma}(r) \equiv rt_{\alpha\gamma}(r), \quad \mathscr{C}_{\alpha\gamma}(r) \equiv rc_{\alpha\gamma}(r), \quad \alpha, \gamma = 1, .., 2m \quad (7)$$

satisfy the following properties

$$\mathscr{A}_{\alpha\gamma}(r) = \frac{1}{2\pi^2}\int_0^\infty dk \sin(kr)\hat{\mathscr{A}}_{\alpha\gamma}(k) \quad (8)$$

$$\hat{\mathscr{A}}_{\alpha\gamma}(k) = 4\pi\int_0^\infty dk \sin(kr)\mathscr{A}_{\alpha\gamma}(r) \quad (9)$$

where $\mathscr{A}$ represents $\mathscr{C}$ or $\mathscr{T}$. Consequently, eqs 1 and 2 become

$$\mathscr{C}^o_{\alpha\gamma}(r) = -ra^o_{\alpha\gamma}e^{-\beta u_{\alpha\gamma}(r)} + r(1 + a^o_{\alpha\gamma})e^{[-\beta u_{\alpha\gamma}(r)+\mathscr{T}_{\alpha\gamma}o(r)/[r(1+a_{\alpha\gamma}o)]]} - r - \mathscr{T}^o_{\alpha\gamma}(r)$$

$$\mathscr{C}^r_{\alpha\gamma}(r) = \frac{(1 + a^o_{\alpha\gamma})}{(1 + a^r_{\alpha\gamma})}\mathscr{T}^r_{\alpha\gamma}(r)e^{[-\beta u_{\alpha\gamma}(r)+\mathscr{T}_{\alpha\gamma}{}^o(r)/[r(1+a_{\alpha\gamma}o)]]} - \mathscr{T}^r_{\alpha\gamma}(r)$$

$$\mathscr{C}^l_{\alpha\gamma}(r) = \frac{(1 + a^o_{\alpha\gamma})}{(1 + a^r_{\gamma\alpha})}\mathscr{T}^l_{\alpha\gamma}(r)e^{[-\beta u_{\alpha\gamma}(r)+\mathscr{T}_{\alpha\gamma}o(r)/[r(1+a_{\alpha\gamma}o)]]} - \mathscr{T}^l_{\alpha\gamma}(r)$$

$$\mathscr{C}^b_{\alpha\gamma}(r) = (1 + a^o_{\alpha\gamma})\Bigg[\frac{\mathscr{T}^b_{\alpha\gamma}(r)}{r(1 + a^b_{\alpha\gamma})} + \frac{\mathscr{T}^r_{\alpha\gamma}(r)\mathscr{T}^l_{\alpha g}(r)}{r^2(1 + a^r_{\alpha\gamma})(1 + a^r_{\gamma\alpha})}\Bigg] \times e^{[-\beta u_{\alpha\gamma}(r)+\mathscr{T}^o_{\alpha\gamma}(r)/[r(1+a^o_{\alpha\gamma})]]} - \mathscr{T}^b_{\alpha\gamma}(r) \quad (10)$$

and

$$\hat{\mathscr{T}}(k) = k[k\bar{\boldsymbol{\rho}}^{-1}[k\bar{\boldsymbol{\rho}}^{-1} - \hat{\mathscr{C}}(k) - k\hat{\mathbf{S}}(k)]^{-1}\,\bar{\boldsymbol{\rho}}^{-1} - \bar{\boldsymbol{\rho}}^{-1}] - k\hat{\mathbf{S}}(k) - \hat{\mathscr{C}}(k) \equiv IE(\hat{\mathscr{C}}(k)) \quad (11)$$

respectively. Clearly, the discretization of eq 8 yields the following definition of the coarse and fine part of the indirect correlation function

$$\mathscr{T}_{\alpha\gamma,j} = \frac{\Delta k}{2\pi^2}\sum_{n=1}^{N-1}\sin\left(\frac{jn\pi}{N}\right)\hat{\mathscr{T}}_{\alpha\gamma,n} = \frac{\Delta k}{2\pi^2}\sum_{n=1}^{M}\sin\left(\frac{jn\pi}{N}\right)\hat{\mathscr{T}}_{\alpha\gamma,n} + \frac{\Delta k}{2\pi^2}\sum_{n=M+1}^{N-1}\sin\left(\frac{jn\pi}{N}\right)\hat{\mathscr{T}}_{\alpha\gamma,n} \equiv \text{coarse part} + \text{fine part} \qquad (12)$$

The latter equation is valid for $j = 1, .., N - 1$ and $\alpha, \gamma = 1, .., 2m$, being $\mathscr{A}_{\alpha\gamma}(r_j) \equiv \mathscr{A}_{\alpha\gamma,j}$, Δr and $\Delta k = \pi/(\Delta r N)$ the corresponding mesh size in distance and reciprocal space, respectively, N is the number of points on the grid, and M is an integer to be fixed later. Obviously, the coefficients for these expansions are directly the sine Fourier components, which can be rapidly evaluated, when required, via Fast Fourier Transform (FFT) techniques. In fact, the discretization of eq 9 yields

$$\hat{\mathscr{A}}_{\alpha\gamma,n} \equiv 4\pi\Delta r\sum_{j=1}^{N-1}\sin\left(\frac{jn\pi}{N}\right)\mathscr{A}_{\alpha\gamma,j}, \quad \alpha, \gamma = 1, .., 2m, \quad n = 1..N-1 \qquad (13)$$

In this way, the numerical scheme deriving from the decomposition 12 has two components. First, a fast and robust Newton-GMRES algorithm,[47] which approximates the solution for the weight of the projections in each iteration, is initially utilized to solve the set of nonlinear eqs 10 and 11 for the first M sine Fourier components $\{\hat{\mathscr{T}}_{\alpha\gamma,n}\}$ (primary contribution), keeping the remaining components fixed. This is followed by one direct Picard iteration to refine the higher sine Fourier components, namely for $n = M + 1,.., N - 1$. This sequence is repeated to get convergence. Other variants of GMRES have been used before in this area.[33−38]

The first phase involves the solution of the system of equations

$$R_{\alpha\gamma,n} \equiv \hat{\mathscr{T}}_{\alpha\gamma,n} - [IE(\hat{\mathscr{C}})]_{\alpha\gamma,n} = 0, \quad \alpha, \gamma = 1,..,2m \qquad (14)$$

with respect to the first M sine Fourier components $\hat{\mathscr{T}}_{\alpha\gamma,n}$ ($n = 1,..M$). This requires the evaluation of the Jacobian matrix

$$J_{\alpha\gamma,n;\beta\zeta,p} = \frac{\partial R_{\alpha\gamma,n}}{\partial\hat{\mathscr{T}}_{\beta\zeta,p}} = \delta_{\alpha\beta}\delta_{\gamma\zeta}\delta_{np} - \frac{\partial[IE(\hat{\mathscr{C}})]_{\alpha\gamma,n}}{\partial\hat{\mathscr{T}}_{\beta\zeta,p}}, \quad n, p = 1,.., M, \quad \alpha, \gamma = 1, .., 2m \qquad (15)$$

in which the derivative appearing on the right side of the latter equation is computed by the chain rule. Indeed, each sine Fourier component n of the element α, γ of the direct correlation function, $\hat{\mathscr{C}}_{\alpha\gamma,n}$, depends on the corresponding direct correlation function defined over the entire grid as given by eq 13, and each one of those components, namely $\mathscr{C}_{\alpha\gamma,j}$, depends on elements of the indirect correlation function via the closure relationships 10. Note that the latter relationships can be conveniently written in terms of correlation function matrices of dimensions $[2m \times 2m]$ like those appearing in the integral eq 11. Using the following identities between the different matrix notations: $\mathscr{A}^{o}_{\alpha\gamma}(r) = \mathscr{A}_{2\alpha-1,2\gamma-1}(r)$, $\mathscr{A}^{r}_{\alpha\gamma}(r) = \mathscr{A}_{2\alpha-1,2\gamma}(r)$, $\mathscr{A}^{l}_{\alpha\gamma}(r) = \mathscr{A}_{2\alpha,2\gamma-1}(r)$, and $\mathscr{A}_{2\alpha,2\gamma}(r) = \mathscr{A}_{2\alpha,2\gamma}(r)$ for $\alpha, \gamma = 1, .., m$, we can formally express the four discretized closure approximations 10 by the following general relationship

$$\mathscr{C}_{2\alpha-i,2\gamma-j,p} = CL^{ij}\{\mathscr{T}_{2\alpha-1,2\gamma-1,p}, \mathscr{T}_{2\alpha-1,2\gamma,p}, \mathscr{T}_{2\alpha,2\gamma-1,p}, \mathscr{T}_{2\alpha,2\gamma,p}\}, \quad \alpha, \gamma = 1, .., m, \quad i, j = 0, 1 \qquad (16)$$

the set of pairs $(i,j) = (0,0), (0,1), (1,0)$, and $(1,1)$ representing the subclasses o, r, l, and b, respectively.

Eq 16 clearly shows that the analytic expression for $\partial\mathscr{C}_{\beta\zeta,j}/\partial\mathscr{T}_{\alpha\gamma,j}$ contains new nonzero elements and new contributions for those found in other IETs requiring only one of these relationships (subclass o for instance) to close the theory and having the correspondence between each element of the matrices $\mathscr{C}$ and $\mathscr{T}$. For instance, the derivative $\partial\mathscr{C}_{\beta\zeta,j}/\partial\mathscr{T}_{11,j}$ is nonzero for the following pair of elements $(\beta, \zeta) = (1,1), (1,2), (2,1)$, and $(2,2)$. Further, each is different from the other. Finally, each one of the elements $\mathscr{T}_{\alpha\gamma,j}$ appearing in eq 16 depends on the sine Fourier components of the indirect correlation function defined over the entire grid in the reciprocal space as shown by eq 12. As a result of these complicated relationships, the chain rule yields the following analytical expression for the Jacobian matrix

$$J_{\alpha\gamma,n;\beta\zeta,p} = \delta_{\alpha\beta}\delta_{\gamma}\zeta_{\delta np} - \sum_{\mu,\nu=1}^{2m}\{-\delta_{\alpha\mu}\delta_{\gamma\nu} + [\mathbf{I} + [\hat{\mathbf{H}} + \hat{\mathbf{S}}]\bar{\boldsymbol{\rho}}]_{\alpha\mu,n}[\mathbf{I} + \bar{\boldsymbol{\rho}}[\hat{\mathbf{H}} + \hat{\mathbf{S}}]]_{\nu\gamma,n}\}\Phi_{\mu\nu,n;\beta\zeta,p} \qquad (17)$$

in which

$$\Phi_{\mu\nu,n;\beta\zeta,p} \equiv \frac{2}{N}\sum_{j=1}^{N-1}\sin\left(\frac{jn\pi}{N}\right)\sin\left(\frac{jp\pi}{N}\right)\frac{\partial\mathscr{C}_{\mu\nu,j}}{\partial\mathscr{T}_{\beta\zeta,j}} \qquad (18)$$

Note that the expression for $\partial\mathscr{C}_{\mu\nu,j}/\partial\mathscr{T}_{\beta\zeta,j}$ is easily obtained from eqs 16 and 10 and that the coefficients 18 are numerically computed using FFTs. The computational cost of computing the full Jacobian and the comparison to the cost of a matrix-free approach are left for a future discussion.

Finally, the analytical expressions for the Jacobian and residual are utilized in the aforementioned nonlinear solve as follows. We provide an initial guess for $\mathscr{T}^{guess}_{\alpha\gamma,j}$ for $j = 1, .., N - 1$, and $\alpha, \gamma = 1, .., 2m$, from where we evaluate the first M sine Fourier components $\hat{\mathscr{T}}^{guess}_{\alpha\gamma,k}$ using FFT (eq 13). We also evaluate the elements $\mathscr{C}^{guess}_{\alpha\gamma,j}$ defined by the closure relationship 16 and subsequently the entire set of sine Fourier components for $\hat{\mathscr{C}}^{guess}_{\alpha\gamma,n}$ via FFT (eq 13). These elements are required to evaluate the expressions given for the residual (14) and the Jacobian (17) matrices. Then, the new estimates of the first M components $\hat{\mathscr{T}}^{new}_{\alpha\gamma,k}$ ($k = 1, .., M$) are obtained by solving the set of linear equations

$$\sum_{\mu,\nu=1}^{2m}\sum_{p=1}^{M}J^{guess}_{\alpha\gamma,n;\mu\nu,p}\Delta\hat{\mathscr{T}}_{\mu\nu,p} = R^{guess}_{\alpha\gamma,n}, \quad \alpha, \gamma = 1, .., 2m, \quad n = 1, .., M \qquad (19)$$

for $\Delta\hat{\mathscr{T}}_{\alpha\gamma,k} \equiv \hat{\mathscr{T}}^{new}_{\alpha\gamma,k} - \hat{\mathscr{T}}^{guess}_{\alpha\gamma,k}$. To achieve this, we use GMRES. We repeat the previous calculation iteratively to get the desired convergence. The first phase is accomplished by choosing a positive value for the parameter δ such that

$$\sqrt{\sum_{\mu,\nu=1}^{2m}\sum_{p=1}^{M}[\Delta\hat{\mathscr{T}}_{\mu\nu,p}]^2} < \sqrt{M}\delta \qquad (20)$$

The second phase is subsequently implemented to minimize the norm of the matrix 14 with respect to the remaining elements $\hat{\mathscr{T}}_{\alpha\gamma,k}$ ($k = 1 + M$, .., $N - 1$). The new estimates for the higher components are easily obtained by direct evaluation of eqs 10, 13, and 11 as performed by one Picard iteration. The input for this iteration is given by the indirect correlation function coming from eq 12 in which the first M components are those obtained in the first phase and the remaining are those initially kept fixed.

Last, we check the convergence for the entire cycle on a distancelike norm. We use the new sine Fourier components previously obtained in the first and second phases to calculate new estimates for the indirect correlation function, $\mathscr{T}^{new}_{\mu\nu,p}$ ($p = 1$, .., $N - 1$) as given by eq 12. We will obtain the solution at the required precision η when the following condition

$$\sqrt{\sum_{\mu,\nu=1}^{2m}\sum_{p=1}^{N-1}[\mathscr{T}^{new}_{\mu\nu,p} - \mathscr{T}^{guess}_{\mu\nu,p}]^2} < 2m\sqrt{(N-1)}\eta \qquad (21)$$

is fulfilled. Otherwise, we redefine $\mathscr{T}^{guess}_{\mu\nu,p} = \mathscr{T}^{new}_{\mu\nu,p}$ and go back to the first phase.

Table 1. Nonpolar Hydrogen Chlorine Fluid: Computing Times (Expressed in Seconds) Obtained on a Common PC Using the Nested Picard Iteration NPI ($M = 0$,$j_{\max} = 6$), Our Proposal as Described in Section III OP ($M \approx 50$, $j_{\max} = 6$), Plain Picard Iteration PPI ($M = 0$, $j_{\max} = 0$), and Our Proposal without the Nested Iteration OP ($M \approx 50$, $j_{\max} = 0$)[a]

	HCL[b]							
	$j_{max} = 6$[c]				$j_{max} = 0$ (directly on the target grid)[c]			
	NPI[d]			OP[d]	PPI[d]			OP2[d]
α	0.75	0.65	0.45		0.75	0.65	0.45	
OPT	19.0	20.9	29.0	4.1	66.2	46.5	65.18	10.0
PY	14.84	18.3	22.6	4.2	32.6	28.4	40.4	10.5
HNC	29.7	5.8	11.1	4.4	55.4	13.7	55.6	7.6
CL1	5.2	5.8	8.4	4.55	34.6	16.6	18.3	13.3
CL2	14.3	16.6	23.8	4.3	25.3	30.9	39.9	7.5
CL3	14.7	17.0	24.4	3.9	27.9	28.1	44.1	6.9
CL4	14.9	17.2	25.1	4.7	26.9	27.8	43.2	10.6
CL5	15.0	17.2	24.8	5.5	24.7	28.0	50.2	9.2
CL6	16.5	14.8	21.6	4.7	35.3	44.9	73.4	9.2
CL7	15.0	17.4	25.3	4.5	31.3	34.5	134.8	10.4
CL8	32.1	21.9	28.8	6.2	63.9	45.7	56.2	9.8
CL9	17.0	19.5	28.2	4.0	30.5	36.5	52.3	9.9

[a] α represents the mixing parameter, being the minimum relaxation for α→1. The set of parameters $\{a^i_{\mu\nu}\}$ defines the closure approximation. They are written in the same order as the one presented in section IV. We named OPT = $\{a^i_{\mu\nu}$ = optimized parameters$\}$, PY = $\{a^i_{\mu\nu} = 100\}$, HNC = $\{a^i_{\mu\nu} = 0\}$, CL1 = $\{a^i_{\mu\nu} = 1\}$, CL2 = $\{a^i_{\mu\nu} = 5\}$, CL3 = $\{a^i_{\mu\nu} = 10\}$, CL4 = $\{a^i_{\mu\nu} = 25\}$, CL5 = $\{a^i_{\mu\nu} = 50\}$, CL6 = $\{2,2,1,4,2,2,3,3,1,4\}$, CL7 = $\{4,3,3,3,1,2,2,4,3,9\}$, CL8 = $\{13,11,11,9,9,9,6,6,13,9\}$, and CL9 = $\{23,21,31,33,20,26,24,24,23,29\}$. [b] Molecule. [c] Nested levels. [d] Algorithm.

In considering the global convergence of this proposal, it is well-known that numerical solvers based on an NR scheme have no guarantees of convergence. Rapid local convergence is reachable when a guess sufficiently close to the solution is provided. To achieve this, a nested algorithm is implemented, which approximately solves the nonlinear equation systems 10 and 11 using the above approach on a sequence of meshes, ending with a solution at the target or finest mesh.[47]

We obtain a solution initially at low resolution for a large mesh size $\Delta\xi_o$ and with few points $N_{\xi o}$ such that $\Delta\xi_o(N_{\xi o} - 1) = L$. The convergence is fast and efficient since the complexity and dimensionality of the problem has been significantly reduced and the results are not required to be as accurate as that for the finest mesh. At subsequent levels of iterations, such complexity is gradually increased without affecting the rate of convergence considerably. The reason is the fact that a very good initial guess is obtained at each level (except for the first nested iteration) by using an interpolating linear polynomial splines on the nodes of the solution generated in the previous level. As the number of points N increases, the error in using the aforementioned interpolating polynomial as an approximation to the desired solution tends to zero like N^{-2}. Thus, higher levels in the nested iteration provides closer initial guesses for the iterative solver described above. This guarantees a rapid local convergence without requiring a larger set of basis functions. Consequently, it provides a fast and efficient scheme by which a sufficiently good initial guess to the solution required for the target grid is reached at very low computational cost. For convenience, we used in the present paper a nested grid defined by $\Delta\xi_j = 2^{-j}\Delta\xi_o$ and $N_j - 1 = L/\Delta\xi_j$ with $j = 0$, .., $j_{\max}$, in such a way that the mesh size is reduced by a factor of 2 at each level. Accordingly, the guess for the jth nested level ($j > 0$) is transferred from the solution obtained in the previous level as follows: $\mathscr{T}^{guess}_{\mu\nu,2p} = \mathscr{T}^{solution}_{\mu\nu,p}$ and $\mathscr{T}^{guess}_{\mu\nu,2p+1} = 1/2(\mathscr{T}^{solution}_{\mu\nu,p+1} + \mathscr{T}^{solution}_{\mu\nu,p})$ for $p = 0$, .., $(N_{j-1} - 1)$ and μ, $\nu = 1$, .., $2m$, where $\mathscr{T}_{\mu\nu,0} \equiv 0$ and we use $\mathscr{T}^{guess}_{\mu\nu,2N-1} \cong \mathscr{T}^{solution}_{\mu\nu,N-1}$.

Just for the purpose of comparing relative computing times between several numerical solvers (including the purely nested Picard iteration ($M = 0$) and the present proposal ($M \neq 0$)) we present in Tables 1 and 2 the numerical results obtained for different closure approximations and thermodynamic states for HCl and water (see the next section for details about the feature of these fluids). We have tabulated in columns named NPI, OP, PPI, and OP2 the results obtained on a common PC using the nested Picard iteration ($M = 0$, $j_{\max} > 0$), our proposal as described above ($M > 0$, $j_{\max} > 0$), plain Picard iteration ($M = 0$, $j_{\max} = 0$), and our proposal without the nested iteration ($M > 0$, $j_{\max} = 0$), respectively. The results were obtained in the interval [0, $L = 35.84$ Å] on a target grid of mesh size $\Delta r = 0.004375$ Å and number of points $N = 8193$, with a precision parameter η set to 10^{-10}. We initially obtained the approximate solution of the nonlinear equation systems on the coarsest grid of mesh size $\Delta\xi_o = 0.28$ Å and number of points $N_o = 129$, with a precision required of only one significant digit ($\eta_o =$

Table 2. Nonpolar Waterlike Model[a]

	H_2O[b]							
	$j_{max} = 6$[c]				$j_{max} = 0$ (directly on the target grid)[c]			
	NPI[d]			OP[d]	PP[d]			OP2[d]
α	0.75	0.65	0.45		0.75	0.65	0.45	
OPT	130.4	150.3	217.8	13.8	198.4	228.4	324.2	25.7
PY	219.5	253.1	364.5	13.9	328.8	378.7	534.0	23.1
HNC	102.0	118.8	171.0	13.2	157.3	180.5	255.6	22.9
CL1	162.1	186.5	269.9	14.7	210.2	242.6	343.5	24.5
CL2	215.6	249.2	361.3	14.0	278.8	319.0	452.7	24.0
CL3	232.9	269.9	390.8	14.0	301.9	345.1	488.9	24.0
CL4	247.5	286.6	413.0	13.9	371.7	365.5	519.5	23.3
CL5	252.5	292.1	422.9	13.9	325.3	373.8	529.6	23.3
CL6	185.1	213.2	309.3	13.9	240.0	275.8	389.2	23.6
CL7	199.4	229.1	333.2	16.9	257.5	296.5	420.2	25.8
CL8	229.2	275.2	384.6	15.4	301.6	346.8	491.2	23.7
CL9	244.9	283.0	411.1	14.8	316.1	363.1	512.5	23.5

[a] The notation is the same as that in Table 1. [b] Molecule. [c] Nested levels. [d] Algorithm.

10^{-1}). At intermediate levels, this precision was increased gradually. This means that just a few levels ($j_{\max} = 6$) were required in the nested iteration to obtain a very good initial guess for the target grid.

The resulting convergence in the nested Picard iteration was found to depend notably on the value of the parameters defining the closure approximation and on the relaxation parameter. In contrast with our proposal, we had to restart Picard several times on the coarsest grid, adjusting the relaxation parameter to get convergence for many closure approximations. Specifically, we shifted slightly the value of the relaxation parameter and restarted the iteration every time that the rate between residuals of consecutive iterations increases by 30 times or more. In other cases this was not necessary, but the rate of convergence for different closure approximations obtained from the same value of the relaxation parameter was very slow. These facts show a clear restriction in using a purely Picard iteration in optimization programs which demand the fast evaluation of correlation functions at different closure approximations.

In contrast, our algorithm was shown to be faster and more efficient than direct Picard iteration. The reason presumably lies in the fact that only the first $M \sim 0.01N$ sine Fourier components were sufficient to get a good representation of the indirect correlation functions. Clearly, a very high value for M would make inefficient this approach since it gives an enormous system of linear equations. On the other hand, a very small value for M would not dramatically improve the computing times obtained by a direct Picard iteration. We found particularly useful a fixed value for $M \cong 50$ to get the optimal rate of convergence of the present computational scheme for the models analyzed in this article. The parameter δ was updated after each global iteration (second phase) in such a way that $\delta_i = a_i/10$, being that a_i is the error generated for the indirect correlation function after performing i global iterations as given by the left term in expression 21. This algorithm is summarized in the Appendix.

Table 3. Performance of the Nested Iteration Obtained in Our Proposal for Nonpolar Fluids[a]

			molecule			
			HCL		H_2O	
nested levels	GRID	DP	GI	RC	GI	RC
$j = 0$	129	1	4	0.25	5	0.2
$j = 1$	257	2	7	0.33	8	0.33
$j = 2$	513	3	10	0.33	11	0.33
$j = 3$	1025	4	13	0.33	14	0.33
$j = 4$	2049	5	16	0.33	17	0.33
$j = 5$	4097	7	19	0.67	20	0.67
$j = 6$	8193	10	23	0.75	24	0.75

[a] We presented the digits of precision (DP) obtained for the residual at each level of the nested iteration as well as the number of global iterations (GI) performed in the corresponding levels. We also presented the rate of convergence (RC = ΔDP/ΔGI) achieved between consecutive levels. These results correspond to the optimized closure approximation.

On the other hand, performing matrix operations and computing Fourier transforms directly on the target grid increased significantly the computing times in both algorithms as shown in columns PPI and OP2. These algorithms also became more unstable and dependent on the initial guess. By comparing the columns NPI and PPI for a Picard iteration and the columns OP and OP2 for our proposal, we conclude that the nested iteration sped up the rate of convergence in both approaches for all the cases analyzed in this work. It is worth mentioning that the results tabulated in the OP2 columns provide an estimation of the computing times that are obtained by numerical solvers based on Gillian/GMRES algorithms.[33−38]

To illustrate the importance of the nested iteration in the performance of our proposal we tabulated in Table 3 the number of global iterations (GI) and the corresponding number of digits of precision (DP) achieved for the residual at each level j of the nested iteration. We also tabulated the rate of convergence (RC = ΔDP/ΔGI) between consecutive levels. Since we have obtained similar results for all the closures previously analyzed in Tables 1 and 2, we only presented in Table 3 those corresponding to the optimized closure approximation. As expected, the results tabulated in column RC indicate that the increase in the dimensionality and complexity of the set of nonlinear equations does not reduce the rate of convergence. Thus, we reached a sufficiently good initial guess for the target grid at a very low computational cost from where the set of nonlinear equations was solved by performing only 4 global iterations.

In the next section, representative models of heteronuclear molecular fluids are analyzed using this computational scheme. To determine the accuracy and efficiency of the theory in predicting the structure of such systems, pair site−site correlation functions are quantitatively compared against MD simulation. The results presented below were obtained with a finest grid of 4097 points and mesh size of 0.00875 Å.

IV. Results and Discussions

A. Diatomics. As a preliminary test, we have numerically solved the equations for two quite different heteronuclear

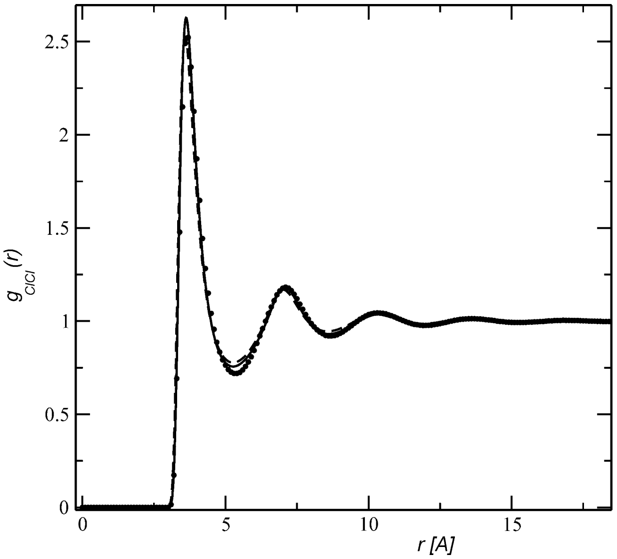


Figure 1. Nonpolar hydrogen chloride: Site−site pair correlation function $g_{ClCl}(r)$ at density $\rho = 0.018$ Å^{-3}, screened density $\eta = \rho/1.361$, and temperature $T = 210$ K. Dashed and solid lines correspond to the PISM-HNC prediction[24] and this work, respectively. Circles correspond to data simulation.

molecular fluids. The first system corresponds to a Lennard-Jones fluid corresponding to a nonpolar hydrogen-chloride liquid (*HCl*).[27] This is a severe test due to the chemical fact that the hydrogen is completely enveloped in the van der Waals sphere of the chlorine. The predicted site−site pair correlation functions, $g_{ClCl}(r)$, $g_{HCl}(r)$, and $g_{HH}(r)$, at density $\rho = 0.018$ Å^{-3}, screened density $\eta = \rho/1.361$, temperature $T = 210$ K, and bond length $L_{HCl} = 1.3$ Å are plotted in Figures 1−3, respectively. The solid and dashed lines correspond to our prediction and the PISM-HNC approximate theory, respectively, and circle symbols correspond to MD simulation. The parameters which describe the Lennard-Jones potential are $\sigma_{HH} = 0.4$ Å, $\epsilon_{HH}/k_B = 20$ K, $\sigma_{ClCl} = 3.353$ Å, and $\epsilon_{ClCl}/k_B = 259$ K, with k_B being the Boltzmann constant. The cross-interaction terms are given by the Lorentz−Berthelot rules.[5]

Since the smaller interaction site (H) is fully enclosed within the larger interaction site (Cl), there is a strong physical screening of the pair correlation between hydrogen interaction sites. As shown in Figure 3, the resulting behavior for $g_{HH}(r)$ (circle symbols) is similar to that of uncorrelated sites. This characteristic property of this model system is correctly reproduced by our solution. In this sense, it successfully eliminates the unphysical behavior exhibited by PISM-HNC approximation, in particular the negative region appearing in $g_{HH}(r)$ and $g_{HCL}(r)$ pair probability distribution functions and a peak predicted at about 0.4 Å in $g_{HH}(r)$. The reason probably lies in the fact that the different contributions to our correlation functions over those regions are appropriately cancelled when the optimal parametrization is reached.

In contrast, Lue and Blankschtein[27] showed that the aforementioned cancellation is not accomplished by adding a certain set of bridge diagrams to the PISM-HNC theory. In fact, this problem is also found in other known approximate theories based on diagrammatically proper integral equations. The chlorine−chlorine pair distribution function predicted by PISM-HNC is otherwise fairly similar to our solution, both being in good agreement with MD simulation. It is a consequence of the particular geometry exhibited by this fluid which makes the correlation between chlorine interaction sites quite affected by the internal structure of the molecules. As a result, the pair correlation function looks like the one obtained from atomic Lennard-Jones fluids where the thermodynamics and structure are rather insensitive to the specific approximation for the closure relationship.

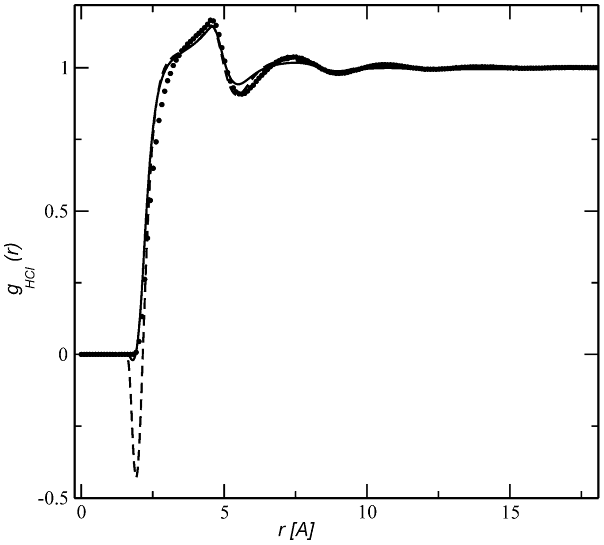


Figure 2. Nonpolar hydrogen chloride: Site−site pair correlation function $g_{HCl}(r)$ at density $\rho = 0.018$ Å^{-3}, screened density $\eta = \rho/1.361$, and temperature $T = 210$ K. Dashed and solid lines correspond to the PISM-HNC prediction[24] and this work, respectively. Circles correspond to data simulation.

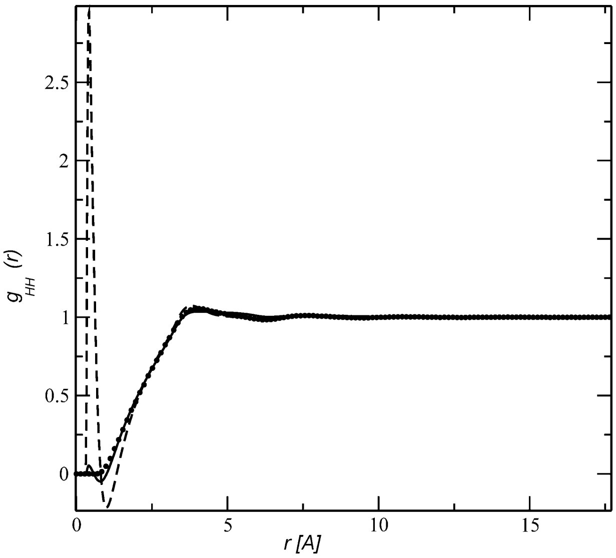


Figure 3. Nonpolar hydrogen chloride: Site−site pair correlation function $g_{HH}(r)$ at density $\rho = 0.018$ Å^{-3}, screened density $\eta = \rho/1.361$, and temperature $T = 210$ K. Dashed and solid lines correspond to the PISM-HNC prediction[24] and this work, respectively. Circles correspond to data simulation.

B. Waterlike Triatomics. The second system analyzed in this paper is that of triatomic molecules characterized by the density $\rho = 0.03345$ Å^{-3}, screened density $\eta = \rho/1.0$,

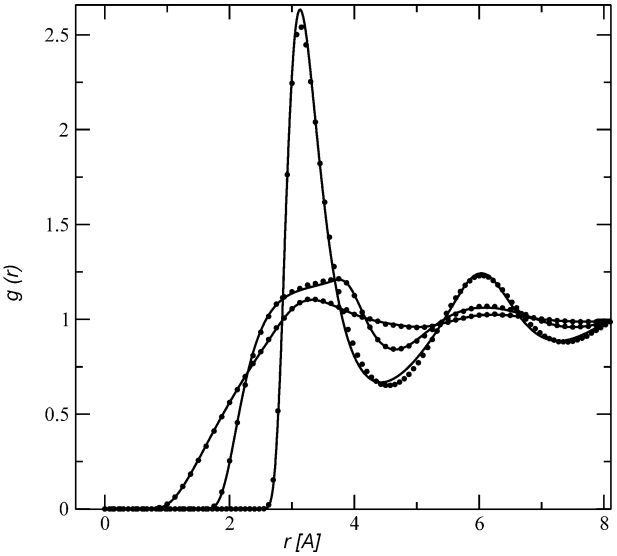


Figure 4. SPC model for water without charges: Site−site pair correlation functions $g_{OO}(r)$, $g_{OH}(r)$, and $g_{HH}(r)$ at density $\rho = 0.03345$ Å^{-3}, screened density $\eta = \rho/1.0$, and temperature $T = 298$ K. Solid line and circles correspond to this work and data simulation, respectively.

and temperature $T = 298$ K. The molecular geometry and Lennard-Jones parameters are that of the SPC model for water given elsewhere.[26,48] Initially, we examined the system without charges, $q_H = q_O = 0$, and later with charges.

These rigid models represent other interesting fluids for testing our approximate theory on nonpolar fluids since the internal structure of the molecule is completely different than the one analyzed previously. In fact, two intramolecular correlations or "bonds" ($L_{HO} = 1.0$ Å and $L_{HH} = 1.633$ Å) are strongly reflected in the behavior of the pair correlation functions. Also for this class of models as originally parametrized, the excluded volume or van der Waals intermolecular interactions (Lennard-Jones terms) act on only one species, namely the oxygen sites ($\sigma_{OO} = 3.1655$ Å and $\epsilon_{OO}/k_B = 78.13$ K).

As has been done before by comparing the structure of this nonpolar fluid with that of more realistic waterlike models with site charges, we are also able to explicitly show how charges, and specially hydrogen-bonds, affect the behavior of correlation functions between hydrogen and oxygen atoms.[49]

We plot in Figure 4 the site−site pair correlation functions predicted for each of $g_{OO}(r)$, $g_{OH}(r)$, and $g_{HH}(r)$. We find remarkable agreement between our approximate solution (solid line) and MD simulation (circle symbols) over the entire range of distances. We found that our prediction reproduces very well not only the location and height of the first peaks but also the phase of oscillation at intermediate distances.

Next, we reanalyzed the same fluid with the addition of charges, $q_H = + 0.41e$ for hydrogen and $q_O = -2q_H$ for oxygen, where e is the elementary charge. Since there would be an attractive pole when a positive and negative charge overlap, we modified the SPC model by introducing a small van der Waals (r^{-12}) repulsive hydrogen-oxygen interaction $u_{OH}(r) = A_{OH}[(kcal\,A^{12})/mol]r^{-12}$ to make a repulsive overlap of hydrogen and oxygen sites of different waterlike molecules for small intersite distances. This is an old patch for these model types.[49] We used two different values for the parameter A_{OH}, namely $A_{OH} = 900$ (Lue and Blankschtein's proposal)[50,51] and $A_{OH} = 225.18$ (Pettitt and Rossky's proposal).[49] The latter was designed to do minimal damage to the thermodynamics and structure of the original model whereas the former was apparently chosen on other grounds.

As a preliminary test of our optimized approach for polar multicomponent fluids, we used the PISM integral equation ($\eta = \rho$) combined with the interpolating closure relationship 10 to compute the structure. As described elsewhere,[26,52] the long-range nature of correlation functions is then approximately handled by dividing the direct correlation functions $c^o_{\alpha\gamma}(r)$ into a short-range and a long-range part and may be handled by renormalization or with other explicit analytical expressions by the method of Ng.[53] Note that, otherwise, expression 4 is no longer useful with long-range interactions. Indeed, the screened density approach also has to be appropriately renormalized in order to lead divergent contributions coming from the integration of those Coulomb terms that appear in the expansion of the partition function. Such extension of the theory is not trivial, and we leave the approach for a future article.

We plotted in Figure 5 Lue and Blankschtein's model and in Figure 6 Pettitt and Rossky's model for the site−site pair correlation functions $g_{OO}(r)$, $g_{OH}(r)$, and $g_{HH}(r)$ predicted by RISM-HNC (dashed line), PISM-HNC (dashed-dotted line), this work (solid line), and MD simulation (symbols). As expected the plots clearly show a significant change between the structure obtained with charges versus the uncharged case. An essential feature of water which is well represented by our prediction is given by the first peak in $g_{OH}(r)$ corresponding to the presence of hydrogen-bonding. Another characteristic of water is the narrow first peak in $g_{OO}(r)$ corresponding to a small coordination number, which is found to be very well reproduced by all of the theoretical predictions analyzed in this article. On the other hand, they do less in describing the long-range structure.

The optimization process is performed, as described in ref 30, by minimizing the approximate analytic expression obtained for the molecular excess chemical potential μ_{ex} 5. The independent parameters satisfying the physical symmetry properties of both models are a^o_{11}, a^o_{12}, a^o_{22}, a^r_{11}, a^r_{12}, a^r_{21}, a^r_{22}, a^b_{11}, a^b_{12}, and a^b_{22}, in which the label number 1 corresponds to hydrogen and the label number 2 corresponds to either chlorine or oxygen, for the first or the second model, respectively. The values of the parameters, obtained from the direct application of this procedure to those models, are the following: $a^o_{HH} = 16.00$, $a^o_{HCl} = 4.85$, $a^o_{ClCl} = 0.015$, $a^r_{HH} = 15.00$, $a^r_{HCl} = 6.00$, $a^r_{ClH} = 6.00$, $a^r_{ClCl} = 0.70$, $a^b_{HH} = 15.00$, $a^b_{HCl} = 1.00$, and $a^b_{ClCl} = 0.70$ for the nonpolar chloride-like model, $a^o_{HH} = 0.00$, $a^o_{HO} = 0.00$, $a^o_{OO} = 0.95$, $a^r_{HH} = 0.00$, $a^r_{HO} = 0.00$, $a^r_{OH} = 0.00$, $a^r_{OO} = 0.00$, $a^b_{HH} = 0.00$, $a^b_{HO} = 0.00$, and $a^b_{OO} = 0.00$ for the SPC model without charges, $a^o_{HH} = 0.025$, $a^o_{HO} = 1.5$, $a^o_{OO} = 1.27$, $a^r_{HH} = 0.5$, $a^r_{HO} = 1.5$, $a^r_{OH} = 1.5$, $a^r_{OO} = 1.27$, $a^b_{HH} = 0.09$, $a^b_{HO} = 1.5$, and $a^b_{OO} = 1.27$ for the modified SPC model for waterlike molecules proposed by Lue and Blankschtein, and

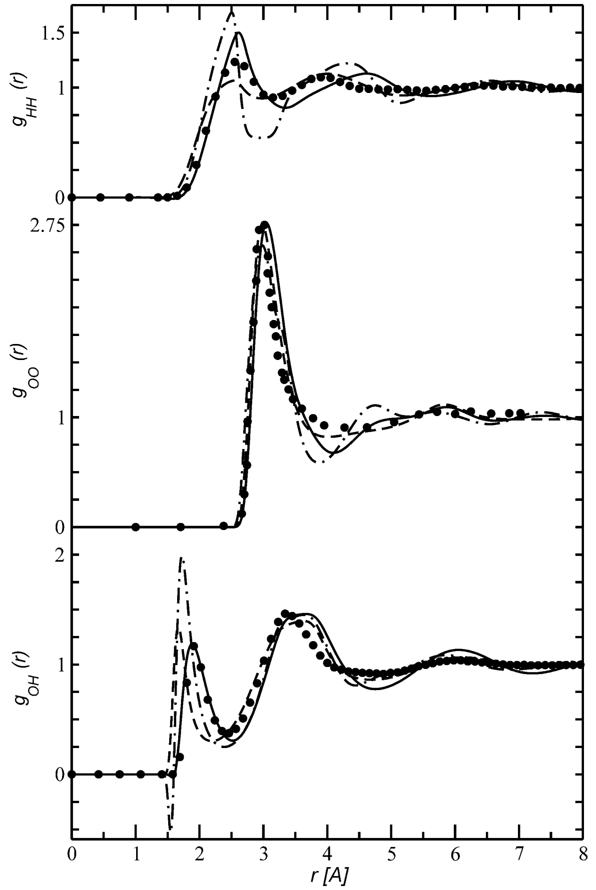


Figure 5. Modified SPC model for water proposed by Blue and Blankschtein: Site−site pair correlation functions $g_{OO}(r)$, $g_{OH}(r)$, and $g_{HH}(r)$ at density $\rho = 0.03345$ Å^{-3}, screened density $\eta = \rho$, and temperature $T = 298$ K. Dashed, dashed-dotted, and solid lines correspond to RISM-HNC, PISM-HNC, and this work, respectively. Circles represent MD simulation.

$a^o_{HH} = 0.07$, $a^o_{HO} = 1.58$, $a^o_{OO} = 0.58$, $a^r_{HH} = 1.01$, $a^r_{HO} = 1.55$, $a^r_{OH} = 1.47$, $a^r_{OO} = 1.15$, $a^b_{HH} = 0.04$, $a^b_{HO} = 1.53$, and $a^b_{OO} = 1.32$ for the modified SPC model for waterlike molecules proposed by Pettitt and Rossky.

Finally, we have tabulated the numerical results for the excess internal energy[40] U^{ex}/Nk_BT predicted by this work, PISM-HNC approximation, and MD simulation. This provides a test of these approximate theories and their solutions on relevant thermodynamic properties of nonpolar molecular fluids. As shown in Table 4, we found that our predictions are in better agreement with simulation data than the older theories.

V. Conclusions

We have introduced a thermodynamically consistent theory which has been successfully tested on representative models of heteronuclear molecular fluids, including water. It is based on an extension of that recently developed for homonuclear fluids. The closure approximation is obtained by analogy to one-component fluids. The approximate theory is completed by coupling the closure approximation with a variant of the diagrammatically proper integral equation introduced recently by this laboratory.

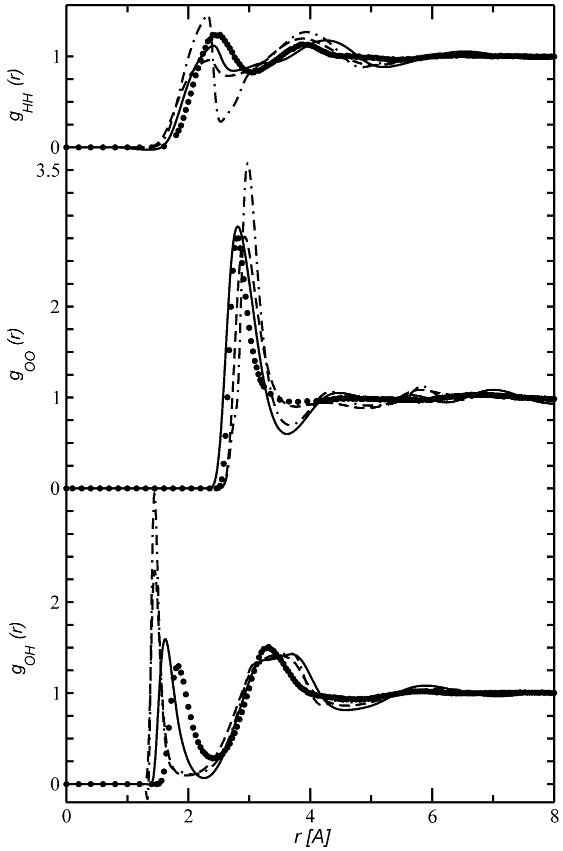


Figure 6. Modified SPC model for water proposed by Pettitt and Rossky: Site−site pair correlation functions $g_{OO}(r)$, $g_{OH}(r)$, and $g_{HH}(r)$ at density $\rho = 0.03345$ Å^{-3}, screened density $\eta = \rho$, and temperature $T = 298$ K. Dashed, dashed-dotted, and solid lines correspond to RISM-HNC, PISM-HNC, and this work, respectively. Circles represent MD simulation.

Table 4. Comparison between MD Simulation and Theoretical Predictions for the Excess Internal Energy U^{ex}/Nk_BT [a]

	PISM-HNC		this work		MD
molecule	U^{ex}/Nk_BT	$\beta\mu_{ex}$	U^{ex}/Nk_BT	$\beta\mu_{ex}$[b]	U^{ex}/Nk_BT
HCl	−82.66	−4, 61	−87.43	−5.0	−88.22
H₂O	−6.71	11.87	−6.70	10.05	−6.45

[a] The excess chemical potential $\beta\mu_{ex}$ predicted by several approximate theories is also included in this table. [b] Results coming from the numerical evaluation of the Morita and Hiroike expression.

The optimal values of the parameters involved in this theory were obtained, as described in ref 30, by minimizing an approximate analytic expression obtained for the molecular excess chemical potential μ_{ex}. The systematic evaluation of correlation functions required by this optimization procedure required better algorithms to solve these integral equations. We introduced a high-performance algorithm to compute iteratively correlation functions, obtaining a significant reduction in the computational cost. Not surprisingly, it is found to be faster and more efficient than the direct Picard iteration.

This numerical method is a modification to the algorithm developed initially by Gillan for atomic fluids.[43] It consists basically of expanding the indirect correlation function on the sine function basis set, a natural expansion within a scheme that already involves a systematic evaluation of sine Fourier transforms. A fast and robust Newton-GMRES algorithm based on Krylov's approach,[47] which approximates the solution for the weight of the projections in each iteration, is initially utilized to solve the set of nonlinear integral equations for the first M sine Fourier components (primary contribution), keeping the remaining components fixed. This is followed by one direct Picard iteration to refine the higher sine Fourier components, keeping those first components fixed. This process is repeated to get convergence.

The remarkable efficiency of this scheme is primarily due to the fact that less than 1% of the sine function basis set was sufficient to provide a good representation of the indirect correlation functions at a coarse level. The Newton−Raphson poor global convergence was notably improved by implementing a nested algorithm, which approximately solves the corresponding nonlinear equation systems using the previous approach on a sequence of meshes, ending with a solution at the target, finest mesh.[47] It was found that this method provides a fast and efficient process by which a sufficiently good initial guess for the target grid is reached at a very low computational cost.

This computational scheme was utilized to analyze the accuracy and efficiency of the theory in predicting structural and thermodynamic properties on two geometrically different polyatomic models. The first fluid is that of nonpolar diatomic molecules of hydrogen chloride (HCL) in which the smaller interaction site (H) is fully enclosed by the larger interaction site (Cl), screening the pair correlation between hydrogen interaction sites. We showed that this optimized theory is capable of successfully describing this phenomenon, eliminating the unphysical behavior exhibited by other known approximate theories based on diagrammatically proper integral equations, including PISM-HNC approximation, in particular the negative region appearing in $g_{HH}(r)$ and $g_{HCL}(r)$ pair correlation functions and a peak predicted at about 0.4 Å in $g_{HH}(r)$.

The second system is the SPC model family for water. It was initially examined without charges since it represents another challenging model for testing our approximate theory on nonpolar fluids. Certainly the geometry and internal structure of this molecule is completely different than the one analyzed previously. Further, the structure of this nonpolar fluid was afterwards compared with that of water-like explicitly showing how charges, and specially hydrogen bonds, affect the behavior of the correlation functions between hydrogen and oxygen atoms. We found that our predictions for this geometrically more intricate model are in excellent agreement with MD simulation.

As a preliminary test of our optimized approach for polar multicomponent fluids, we used PISM integral equation combined with our interpolating closure relationship to describe the structure of waterlike molecular fluid, instead of the aforementioned screened density integral equation, since the latter is no longer valid in the present fashion for long-range interactions. We compared different approximate integral equation theories against simulation, finding that our prediction describes quite well essential features of water such as the first peak in $g_{HO}(r)$ and $g_{OO}(r)$ which represent the hydrogen-bond and the coordination number, respectively. Certainly a notable improvement in the prediction of the PISM integral equation is obtained when properly combined with the interpolating closure approximation proposed in the present article, rather than with HNC or PY approximations. Beyond predicting the first shell of water successfully, this approximate theory fails, as many other approximate theories, in correctly describing the asymptotic behavior of the corresponding correlation functions, which plays a crucial role in the prediction of a reasonable dielectric theory.[80,55] In fact, most of the approximate integral equation theories are not consistent with the zeroth and second moment conditions and yield trivial and incorrect predictions for the static dielectric constant.[52,56,57]

Further work involves the renormalization of the screened density integral equation. As already tested on nonpolar fluids, this variant of the PISM integral equation has been shown to provide an accurate description of structure and thermodynamic properties when it is combined with the interpolating closure approximation. Thus, it represents a promising powerful tool to correctly capture structural, dielectric, and thermodynamic properties of polar multicomponent molecular fluids. It may also provide a proper framework of self-consistency to describe the underlying physics of gas−liquid-phase transition of molecular fluids where most of the approximate integral equations have been shown to predict the incorrect critical temperature as well as isothermal compressibility and constant-volume heat capacity.[58,59]

Another direction we will consider in the future derived from this paper is related to the development of a proper interpolating closure approximation for 3D integral equation theories. It is well-known that many intricate biological systems require more detailed information on the structure of molecular fluids than the one that can be reasonably obtained from using 1D approximate theories.[1] On such 3D grids, a new challenge will surely be the design of a new high performance algorithm based on the ideas proposed in this article. An important future aim will be computing 3D correlation functions as accurately as possible at low computational complexity.

Acknowledgment. M.M. and B.M.P. thank the NIH and the Robert A. Welch foundation for partial support. M.M. is also grateful for financial support from the World Laboratory Center for Pan-American Collaboration in Science and Technology. The work of C.T.K. was supported by NSF grant #DMS-0707220. Drs. Kippi Dyer, John Perkyns, Ka-yiu Wong, and Gillian Lynch are acknowledged for many helpful conversations.

Appendix

The computational scheme described in section III can be summarized as follows:

Algorithm 1: nested_it ($\mathcal{T}^{guess}$, $\Delta\xi_o$, j_{max}, L, m, η_o, η M, IE, CL, $\mathcal{T}^{output}$,$\mathcal{C}^{output}$)

-Define the coarsest grid: $j = 0$; $\Delta\xi = \Delta\xi_o$; $N = L/\Delta\xi + 1$;

- *Call* the numerical_solver ($\mathcal{T}^{guess}$, $\Delta\xi$, N, m, η_o, M, IE, CL, $\mathcal{T}^{output}$,$\mathcal{C}^{output}$) to obtain the approximate solution for

```
the indirect and direct correlation functions $\mathscr{T}^{output}$ and $\mathscr{C}^{output}$ over the coarsest grid, respectively;
  while ($j < j_{max}$) do
    -Define the next level $j = j + 1$;
    -Increase gradually the precision $\eta_j$ in such a way tha $\eta_{jmax} = \eta$;
    -Define the input (initial guess) over the finer grid using the intergrid interpolation
      $\mathscr{T}^{input}_{\mu\nu,2p} = \mathscr{T}^{output}_{\mu\nu,p}$ and $\mathscr{T}^{input}_{\mu\nu,2p+1} = 1/2(\mathscr{T}^{output}_{\mu\nu,p+1} + \mathscr{T}^{output}_{\mu\nu,p})$ for $p = 0, .., (N - 1)$ and $\mu, \nu = 1, .., 2m$;
    -Define the new grid: $\Delta\xi = \Delta\xi/2$; $N = L/\Delta\xi + 1$;
    -Call the numerical_solver ($\mathscr{T}^{input}$, $\Delta\xi$, $N$, $m$, $\eta_j$, $M$, $IE$, $CL$, $\mathscr{T}^{output}$,$\mathscr{C}^{output}$) to obtain the approximate solution for the indirect and direct correlation functions $\mathscr{T}^{output}$ and $\mathscr{C}^{output}$ over the finer grid, respectively;
  end while
  Hereafter $\mu, \nu = 1, .., 2m$; $p = 1, .., N - 1$; $n = 1, .., M$; $k = 1 + M, .., N - 1$ when required;
  Algorithm 2: numerical_solver ($\mathscr{T}^{input}$, $\Delta\xi$, $N$, $m$, $\eta$, $M$, $IE$, $CL$, $\mathscr{T}^{output}$,$\mathscr{C}^{output}$)
  -Use $\mathscr{T}^{input}_{\mu\nu,p}$ to obtain a new estimate for $\mathscr{T}^{output}_{\mu\nu,p}$ and $\mathscr{C}^{output}_{\mu\nu,p}$ from the direct evaluation of eqs 16, 13, 11, and 12, as performed by one Picard iteration;
  -Define the parameter $a = \sqrt{\sum_{\mu,\nu,p}[\mathscr{T}^{output}_{\mu\nu,p} - \mathscr{T}^{input}_{\mu\nu,p}]^2}$;
  while ($a > \sqrt{N-1}\eta$) do
    $\mathscr{T}^{input}_{\mu\nu,p} = \mathscr{T}^{output}_{\mu\nu,p}$
    -Use $\mathscr{T}^{output}_{\mu\nu,p}$ to compute the sine Fourier components $\hat{\mathscr{T}}_{\mu\nu,n}$ using the FFT 13;
    -Use $\hat{\mathscr{T}}_{\mu\nu,n}$ and $\mathscr{T}^{output}_{\mu\nu,p}$ to compute the coarse and fine part as shown by expression 12 and keep the fine part contribution fixed;
    -Use $\mathscr{T}^{output}_{\mu\nu,p}$ to obtain the direct correlation function $\mathscr{C}_{\mu\nu,p}$ from eq 16 and use it to obtain $\hat{\mathscr{C}}_{\mu\nu,p}$ using the FFT 13;
    -Evaluate the Jacobian 17 at $\hat{\mathscr{T}}_{\mu\nu,n}$ and the residual 14 at $\hat{\mathscr{C}}_{\mu\nu,p}$ as described above and then, use them to solve eq 19 to estimate changes on the downhill direction over the first $M$ sine Fourier components $\Delta\hat{\mathscr{T}}_{\mu\nu,n}$;
    -Define $\delta = a/10$
    while ($\sqrt{\sum_{\mu,\nu,n}[\Delta\hat{\mathscr{T}}_{\mu\nu,n}]^2} > \sqrt{M}\delta$) do
      - Calculate the new first $M$ sine Fourier components $\hat{\mathscr{T}}^{new}_{\mu\nu,n} = \Delta\hat{\mathscr{T}}_{\mu\nu,n} - \hat{\mathscr{T}}_{\mu\nu,n}$;
      -Use $\hat{\mathscr{T}}^{new}_{\mu\nu,n}$ to compute the coarse part as shown by expression 12 and then get $\mathscr{T}^{output}_{\mu\nu,p}$ by keeping the fine part contribution fixed;
      - Use $\mathscr{T}^{output}_{\mu\nu,p}$ to obtain the direct correlation function $\mathscr{C}_{\mu\nu,p}$ from eq 16 and use it to obtain $\hat{\mathscr{C}}^{new}_{\mu\nu,p}$ using the FFT 13;
      - Evaluate the Jacobian 17 at $\hat{\mathscr{T}}^{new}_{\mu\nu,n}$ and the residual 14 at $\hat{\mathscr{C}}^{new}_{\mu\nu,p}$ as described above and then, use them to solve eq 19 to estimate new changes on the new downhill direction over the first $M$ sine Fourier components $\Delta\hat{\mathscr{T}}^{new}_{\alpha\gamma,n}$;
      $\hat{\mathscr{T}}_{\mu\nu,n} = \hat{\mathscr{T}}^{new}_{\alpha\gamma,n}$; $\Delta\hat{\mathscr{T}}_{\mu\nu,n} = \Delta\hat{\mathscr{T}}^{new}_{\alpha\gamma,n}$
    end while
    -Use $\hat{\mathscr{T}}_{\mu\nu,n}$ to compute the coarse part as shown by expression (12) and then get $\mathscr{T}^{inter}_{\mu\nu,p}$ by keeping the fine part contribution fixed;
    -Use $\mathscr{T}^{inter}_{\mu\nu,p}$ to obtain a new estimate for $\mathscr{T}^{output}_{\mu\nu,p}$ and $\mathscr{C}^{output}_{\mu\nu,p}$ and the higher sine Fourier components $\hat{\mathscr{T}}_{\mu\nu,k}$ from the direct evaluation of eqs 16, 13, 11, and 12, as performed by one Picard iteration;
    -Use $\hat{\mathscr{T}}_{\mu\nu,k}$ to get the new fine part contribution as shown by eq 12 by keeping the coarse part contribution fixed;
    $a = \sqrt{\sum_{\mu,\nu,p}[\mathscr{T}^{output}_{\mu\nu,p} - \mathscr{T}^{input}_{\mu\nu,p}]^2}$
  end while
```

References

(1) Hirata, F. *Molecular Theory of solvation*; Kluwer Academic: Amsterdam, 2004.

(2) Cramer, C. J.; Truhlar, D. G. Continuum solvation models: classical and quantum mechanical implementations. In *Reviews in computational chemistry;* Lipkowitz, K. B., Ed.; VCH: New York, 1995.

(3) Fantoni, R.; Pastore, G. *J. Chem. Phys.* **2003**, *119*, 3810.

(4) Hansen, J. P.; McDonald, I. R. *Theory of simple Liquids*; Academic Press Inc.: London, 1986.

(5) Lee, L. L. *Molecular Thermodynamics of Nonideal Fluids*; Butterworths: Boston, 1988.

(6) Percus, J. K.; Yevick, G. J. *Phys. Rev.* **1958**, *110*, 1.

(7) Stell, G. *Physica* **1963**, *29*, 517.

(8) Morita, T. *Prog. Theor. Phys.* **1960**, *23*, 829.

(9) Verlet, L. *Nuovo Cimento* **1960**, *18*, 77.

(10) Chandler, D. In *The Liquid State of Matter: Fluids, Simple and Complex;* Montroll, E. W., Lebowitz, J. L., Eds.; North Holland Pub. Co.: Amsterdam, 1982; p 275.

(11) Chandler, D.; Andersen, H. C. *J. Chem. Phys.* **1972**, *57*, 1930.

(12) Ladanyi, B. M.; Chandler, D. *J. Chem. Phys.* **1975**, *62*, 4308.

(13) Chandler, D.; Pratt, L. *J. Chem. Phys.* **1976**, *65*, 2925.

(14) Chandler, D. *Mol. Phys.* **1976**, *31*, 1213.

(15) Chandler, D. *J. Chem. Phys.* **1977**, *67*, 1113.

(16) Hirata, F.; Rossky, P. J.; Pettitt, B. M. *J. Chem. Phys.* **1983**, *78*, 4133.

(17) Pettitt, B. M.; Rossky, P. J. *J. Chem. Phys.* **1983**, *78*, 7296 and references therein.

(18) Jhonson, E.; Hazoume, R. P. *J. Chem. Phys.* **1979**, *70*, 1599.

(19) Monson, P. *Mol. Phys.* **1982**, *47*, 435.

(20) Sullivan, D. E.; Gray, C. G. *Mol. Phys.* **1981**, *42*, 443.

(21) Cummings, P. T.; Stell, G. *Mol. Phys.* **1981**, *44*, 529.

(22) Chandler, D.; Silbey, R.; Ladanyi, B. *Mol. Phys.* **1982**, *46*, 1335.

(23) Chandler, D.; Joslin, C. G.; Deutch, J. M. *Mol. Phys.* **1982**, *47*, 871.

(24) Rossky, P. J.; Chiles, R. A. *Mol. Phys.* **1984**, *51*, 661.

(25) Lupkowski, M.; Monson, P. A. *J. Chem. Phys.* **1987**, *87*, 3618.

(26) Lue, L.; Blankschtein, D. *J. Chem. Phys.* **1995**, *102*, 5427.

(27) Lue, L.; Blankschtein, D. *J. Chem. Phys.* **1995**, *102*, 4203.

(28) Dyer, K. M.; Perkyns, J. S.; Pettitt, B. M. *J. Chem. Phys.* **2005**, *122*, 236101.

(29) Dyer, K. M.; Perkyns, J. S.; Pettitt, B. M. *J. Chem. Phys.* **2005**, *123*, 204512.

(30) Marucho, M.; Pettitt, B. M. *J. Chem. Phys.* **2007**, *78*, 7296.

(31) Lebowitz, J. L.; Percus, J. K. *Phys. Rev.* **1961**, *122*, 1675.

(32) Stell, G. *Mol. Phys.* **1969**, *16*, 209.

(33) Kinoshita, M.; Harada, M. *Mol. Phys.* **1991**, *74*, 443.

(34) Kinoshita, M.; Harada, M. *Mol. Phys.* **1993**, *79*, 145.

(35) Kinoshita, M.; Harada, M. *Mol. Phys.* **1994**, *81*, 1473.

(36) Halminton, T. P.; Pulay, P. *J. Chem. Phys.* **1986**, *84*, 5728.

(37) Kovalenko, A.; Ten-no, S.; Firata, F. *J. Comput. Chem.* **1999**, *20*, 928.

(38) Booth, M. J.; Schlijper, A. G.; Scales, L. E.; Haymet, D. J. *Comput. Phys. Commun.* **1999**, *119*, 122.

(39) Vatamann, J.; Cann, N. M. *J. Chem. Phys.* **2004**, *121*, 6922.

(40) Lue, L.; Blankschtein, D. *J. Chem. Phys.* **1994**, *100*, 3002.

(41) Ryzhik, G.; Jeffrey, A. *Table of Integrals, Series, and Products*, 5th ed.; Academic Press, Inc.: London, 1986.

(42) Abramowitz, M.; Stegun, I. A. *Handbook of Mathematical functions*; Dover: New York, 1972.

(43) Gillan, M. J. *Mol. Phys.* **1979**, *38*, 79.

(44) Labik, S.; Malijevsky, A.; Vonka, P. *Mol. Phys.* **1985**, *56*, 709.

(45) Ichiye, T.; Haymet, D. J. *J. Chem. Phys.* **1988**, *89*, 4315.

(46) Zichi, D. A.; Rossky, P. J. *J. Chem. Phys.* **1985**, *54*, 1712.

(47) Kelley, C. T.; Pettitt, B. M. *J. Comput. Phys.* **2004**, *197*, 491.

(48) Glattli, A.; Daura, X.; Van Gunsteren, W. F. *J. Chem. Phys.* **2002**, *116*, 9811.

(49) Pettitt, B. M.; Rossky, P. J. *J. Chem. Phys.* **1982**, *77*, 1452.

(50) Lue, L.; Blankschtein, D. *J. Chem. Phys.* **1992**, *96*, 8582.

(51) Redy, G.; Lawrence, C. P.; Skinner, J. L.; Yethiraj, A. *J. Chem. Phys.* **2003**, *119*, 13012.

(52) Chandler, D.; Joslin, G. S.; Deutch, J. M. *Mol. Phys.* **1982**, *47*, 871.

(53) Ng, K.-C. *J. Chem. Phys.* **1974**, 61, 2680.

(54) Chandler, D. *J. Chem. Phys.* **1977**, *67*, 1113.

(55) Hoye, J. S.; Stell, G. *J. Chem. Phys.* **1976**, *65*, 18.

(56) Perkyns, J.; Pettitt, B. M. *Chem. Phys. Lett.* **1992**, *190*, 626.

(57) Perkyns, J.; Pettitt, B. M. *J. Chem. Phys.* **1992**, *97*, 7656.

(58) Sarkisov, G.; Lomba, E. *J. Chem. Phys.* **2005**, *122*, 214504.

(59) Peplow, A. T.; Beardmore, R. E.; Bresme, F. *Phys. Rev. E* **2006**, *74*, 1539.

CT700202H

JCTC Journal of Chemical Theory and Computation

On the Nature of the CP Bond in Phosphaalkynes

Maria F. Lucas, Maria C. Michelini, Nino Russo,* and Emilia Sicilia

Dipartimento di Chimica and Centro di Calcolo ad Alte Prestazioni per Elaborazioni
Parallele e, Distribuite-Centro d'Eccellenza MURST, Università della Calabria,
I-87030 Arcavacata di Rende, Italy

Received October 19, 2007

Abstract: In this work, we report results of calculations based on the density functional theory (B3LYP/6-311+G(2d,2p)) of different species containing a terminal cyaphide bond. The chosen species range from small molecules and anions ($C{\equiv}P^-$, $HC{\equiv}P$, *t*$BuC{\equiv}P$, $[(CF_3)_3BC{\equiv}P)]^-$) to large transition-metal containing complexes ($[(dppe)_2Ru(H)(C{\equiv}P)]$, *trans*-$[Pt(PMe_3)_2(Cl)(C{\equiv}P)]$, *trans*-$[Pt(PMe_3)_2(Cl)(CP)Pt(PMe_3)_2]$). A comparative analysis of the description of the C≡P bond obtained by different methodologies is presented. Topological analyses of the electron density in the framework of the theory of atoms in molecules (AIM) and of the electron localization function (ELF) are complemented with the results obtained by natural bond orbital analysis (NBO).

1. Introduction

For many years it was accepted that thermally stable compounds containing multiple bonds would occur only for elements of the second period. The so-called "double-bond rule" stated that compounds with multiple bonds involving heavy main-group elements were unstable. This rule was based on the fact that the σ bonds for heavy elements are relatively long and the increasingly diffuse nature of p orbitals makes for poor overlap to form π bonds. However, early experimental studies clearly established that new compounds having $(p-p)\pi$ bonds could be synthesized provided some criteria were taken into account. In particular, it was shown that this type of π systems could be stabilized by resonance, by reduction of the polarity in the π systems and avoiding the oligomerization reactions.[1] In the last years, the refinement of experimental techniques has permitted the synthesis and characterization of compounds with multiple phosphorus-element bonds and the detailed study, both from experimental and theoretical viewpoints, of their molecular and electronic structure became a very exciting area in organophosphorus chemistry.

In spite of the discovery, which was considered mostly a curiosity, of the phosphaalkyne HCP by Gier in 1961[2] only 20 years later[3] with the synthesis of *t*BuCP, a surprisingly thermally stable compound, the chemistry of phosphaalkynes was firmly established. The synthesis of this compound proved to be the starting point for the rapid development of the chemistry of the C≡P bond. Soon it became clear that phosphaalkynes can be stabilized in several ways, and today a whole range of species containing C≡P triple bonds is known.

Stabilization of the phosphaalkyne group for the synthesis of isolable $R{-}C{\equiv}P$ species is achieved in most cases by steric shielding using large substituents, as the *tert*-butyl one, otherwise the triple C≡P bond polymerizes.[4] As a result, many phosphaalkynes have been synthesized and subsequently incorporated as ligands into transition-metal complexes.[5−8] The ligand was stabilized as a μ_2-bridging ligand in dinuclear complexes in which the carbon atom is coordinated to two platinum or two iron atoms.[6,7] Only recently, the synthesis and characterization of a transition-metal complex[8] with a terminal cyaphide has raised again the challenging problem of the stabilization and, therefore, isolation of transition-metal complexes containing a terminal $C{\equiv}P^-$ ligand.

In addition to some early works on the subject,[9] several theoretical studies have been performed in the last years on species containing the triple C≡P bond.[10] Hübler and Schwerdtfeger have performed a theoretical analysis of the vibrational frequencies and the NMR chemical shifts (^{31}P and ^{13}C), of a range of λ^3-phosphaalkynes.[10a] Kurita and co-workers presented MP2 and B3LYP calculations on several compounds containing single and multiple CP bonds, among which are HCP, CH_3CP, and *t*BuCP.[10b] Pascoli and Lavendy

* Corresponding author e-mail: nrusso@unical.it.

10.1021/ct700277w CCC: $40.75 © 2008 American Chemical Society
Published on Web 02/22/2008

reported a DFT study of C_nP, C_nP^+, and C_nP^- ($n = 1-7$) clusters.[10c] More recently, Mó and collaborators have performed a combined theoretical and experimental work concerning the gas-phase acidity of HCP, CH_3CP, HCAs, and CH_3CAs.[10d]

In the present work, the chemical nature of the C≡P bond in a series of phosphaalkynes was investigated using different bonding analysis methodologies. A comparison of the different bonding descriptions is provided. The molecular structures and vibrational frequencies obtained using density functional theory were compared to the experimental available data.

2. Computational Details

Geometry optimizations as well as frequency calculations for all the examined phosphaalkynes were performed at the Density Functional level of theory as implemented by GAUSSIAN03 code.[11] The Becke's three-parameter hybrid functional[12] combined with the Lee, Yang, and Parr (LYP) correlation functional,[13] denoted as B3LYP, was used. For Pt and Ru LanL2DZ effective core potentials[14] were adopted in conjunction with their split valence basis sets. The standard 6-311+G(2d,2p) basis sets of Pople and co-workers were employed for the rest of the atoms.[15] The same level of theory was also used to obtain the wavefunctions of all the structures.

The bonding features of all the studied species were analyzed by means of Natural Bond Orbital (NBO) and Natural Population Analysis (NPA).[16] We have also analyzed the nature of the bonding by using two different topological methodologies, namely, the topological analysis of the electron localization function (ELF) and the Atoms in Molecules (AIM) approach. ELF analysis is based on the topology of the gradient vector field of the Becke and Edgecombe[17] electron localization function, as implemented by Silvi and Savin.[18] The ELF, $\eta(\mathbf{r})$, can be interpreted as a measure of the electron localization in atomic and molecular systems, namely, as the conditional probability of finding two electrons with the same spin around a reference point. The analysis of the ELF gradient field provides a mathematical model enabling the partition of the molecular position space in basins of attractors (Ω_A), which present in principle a one-to-one correspondence with chemical local objects such as bonds and lone pairs. These basins are either core basins, usually labeled C(A), or valence basins, V(A,...), belonging to the outermost shell and characterized by the number of core basins with which it shares a common boundary, which is called the synaptic order. In this representation the monosynaptic basins correspond to nonbonded pairs of the usual Lewis representation, whereas the di- and polysynaptic basins are related to bonds. The presence of di- or polysynaptic basins is indicative of shared interactions (covalent, dative, metallic bonds), whereas the absence of these basins is indicative of closed-shell interactions (ionic, hydrogen, van der Waals bonds). The electronic population of a synaptic basin, $\bar{N}$ (Ω_A), is obtained as the integral of the one-electron density over the basin. The variance of the basin population, $\sigma^2[\bar{N}(\Omega_A)]$, that is the square of the standard deviation of the population, represents the quantum-mechanical uncertainty of the basin population and is a result of the delocalization of electrons. It has the meaning of an excess in the number of pairs due to the interaction of Ω_A with the other basins and is usually written as the sum of contributions of all other basins.

Within ELF analysis a multiple bond is characterized by a basin population $\bar{N}(\Omega_A)$ higher than 2.0 electrons and a variance $\sigma^2[\bar{N}\ (\Omega_A)]$ less than the corresponding basin population.

The TopMod package was used to analyze the topology of the ELF function.[19]

AIM analysis[20] explores the topology of the electron density, ρ(r), of the molecules revealing insightful information on the nature of the bonds. A (3, −1) critical point of the electron density, ρ(r), located between two atomic centers denotes the presence of a bond. Topologically, this corresponds to a point in the real space where the gradient of ρ(r), $\nabla\rho$(r), is zero and where the curvature of ρ(r), expressed through three eigenvalues of the diagonalized Hessian of ρ-(r), is positive for an eigenvector linking two atomic centers (λ_3) and negative for the two others ($\lambda_{1,}$ λ_2) perpendicular to it. Unequal values of λ_1 and λ_2 at the (3, −1) bond critical points (BCPs) denote an anisotropic spread of electrons quantified through the concept of ellipticity, which is defined as

$$\epsilon = (\lambda_1/\lambda_2) - 1$$

(with $\lambda_1 > \lambda_2$). According to the mathematical definition, values of ϵ greater than zero indicate partial π-character in a bond or electronic distortion away from σ-symmetry along the path.[21] Double bonds are usually characterized by significant ellipticity values, as it is found for C,C double bonds,[21b] whereas in the case of triple bonds, and due to the cylindrical symmetry resulting from the presence of two π-bonds, that values are expected to be very close to zero.

The most used property to evaluate the characteristics of the bond is the Laplacian of the charge density, $\nabla^2\rho$(bcp). When $\nabla^2\rho(\text{bcp}) < 0$, charge is concentrated at the critical point, while when $\nabla^2\rho(\text{bcp}) > 0$, charge is locally depleted.

Within the framework of AIM analysis the variance, σ^2-(Ω_A), can also be spread in terms of the contribution from other basins, the covariance, cov($\Omega_{A,}$ Ω_B), which has a clear relationship with the so-called delocalization index, δ($\Omega_{A,}$ Ω_B)[22]

$$\text{cov}(\Omega_{A,}\ \Omega_B) = -\delta(\Omega_{A,}\ \Omega_B)/2$$

The delocalization index accounts for the electrons delocalized or shared between the basins Ω_A and Ω_B. This index, in the single determinant approach, is exactly the topological bond order defined by Ángayán and co-workers.[23] We must mention, however, that even when for molecular bonds with equally shared pairs a simple relationship between the delocalization index and the formal bond order (number of Lewis bonded pairs) has been generally found,[22a] for polar bonds there is no longer such a simple relationship. It has been shown that the delocalization index tends to decrease with the increased electronegativity difference of the atoms involved in the bond. There has been some discussion in the past regarding the use of this index as a covalent bond order.[24]

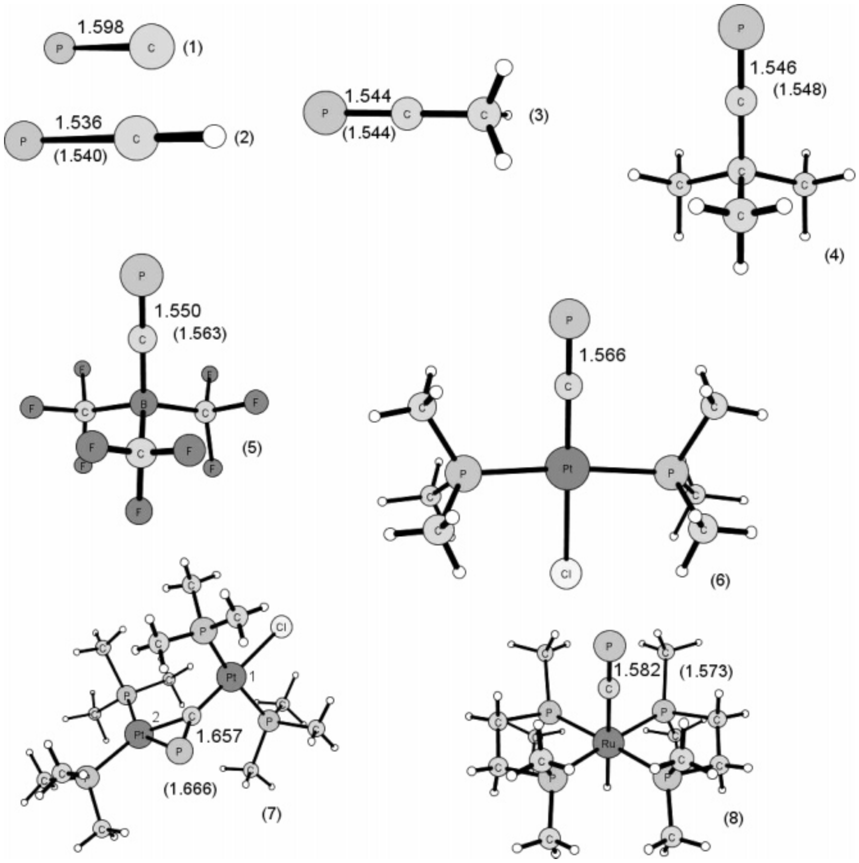


Figure 1. Optimized CP bond length of (1) CP^-; (2) HCP; (3) CH_3CP; (4) *t*BuCP; (5) $[(CF_3)_3BPC)]^-$; (6) *trans*-$[Pt(PMe_3)_2(Cl)(CP)]$; (7) *trans*-$[Pt(PMe_3)_2(Cl)(CP)Pt(PMe_3)_2]$; and (8) $[(dppe)_2Ru(H)(CP)]$. For the sake of comparison, available experimental values are also reported in parentheses.

The BCPs were primarily localized with the EXTREME program (part of the AIMPAC package)[20] and verified with the TopMod program.[19]

3. Results and Discussion

We have initiated the study of the C≡P bond with CP^- (1), HCP (2), and CH_3CP (3). The optimized structures of these compounds are shown in Figure 1, and their Cartesian coordinates are included as Supporting Information.

Previous calculations on (1) show a C≡P bond of 1.609 Å for an optimization at the B3LYP/6-311G(d) level[10c] and 1.604 Å at QCISD/6-311G+(df,p),[10d] while we have a calculated value of 1.598 Å which is in good agreement with previous results. For HCP, at the QCISD /6-311G(d) level, the C≡P bond was previously found to be 1.544 Å,[10d] our calculations indicate 1.536 Å, and for CH_3CP the value is 1.549 Å[10d] in comparison with a value of 1.544 Å calculated by us. Our computations are then consistent with previous calculations as well as with experimental microwave data for HCP (1.5404 Å)[25a] and CH_3CP (1.544(4) Å).[25b]

The frequency analysis also retrieved vibrational frequencies for the C≡P bond stretching consistent with experimental data and with previous calculations. For CP^- the calculated value of 1197 cm^{-1} is similar to the previously calculated[10d] one of 1198 cm^{-1}, and the same takes place for (2) and (3) with values calculated by us of 1335 cm^{-1} and 1591 cm^{-1} in good agreement with previously calculated 1327 cm^{-1} and 1616 cm^{-1} values, respectively.[10d] The lengthening of the C≡P bond as a consequence of deprotonation is expected for CP^-; however, the calculated vibrational frequency (1335 cm^{-1}) is overestimated with respect to the 1265 cm^{-1} assigned experimental value.[2]

For *t*BuCP (4) experimental data are available as well as some theoretical calculations at the RHF/6-31G(d) level.[26] That calculations established a bond distance of 1.519 Å for the CP bond which is close to the experimental diffraction data of 1.548(1)Å.[26] Our calculations assign a C≡P bond distance of 1.546 Å which is very close to the expected experimental value. The vibrational frequency value of 1573 cm^{-1} is consistent with the experimental value (1533 cm^{-1}).[27]

We have also optimized $[(CF_3)_3BC{\equiv}P)]^-$, structure 5 in Figure 1, which is structurally similar to *t*BuCP. The C≡P bond was calculated to be 1.550 Å, and the frequency analysis estimated a stretching mode at 1500 cm^{-1} that is in good agreement with a bond distance of 1.563(10) Å and a vibrational mode at 1468 cm^{-1} from experimental data.[27]

In the present work, we have also investigated the characteristics of the C≡P bond in some metal complexes of platinum and ruthenium. The compound *trans*-$[Pt(PEt_3)_2(Cl)(C{\equiv}P)]$ (6) was the first transition-metal complex containing a terminal C≡P ligand to be synthesized.[7] However, no structural data are available for this compound, since it is too unstable to be isolated, and for this reason theoretical investigations can be very helpful. To save computer time the ethyl (Et) groups were substituted with methyl (Me) ones. We have established that the C≡P bond

Table 1. C≡P Stretching Harmonic Calculated and Experimental Vibrational Frequencies (cm^{-1}) and Natural Atomic and AIM (in Parentheses) Charges on Carbon and Phosphorus Atoms

	calcd freq	exptl freq	C atom charge	P atom charge
CP$^-$	1197	–	−0.83 (−1.29)	−0.17 (0.40)
HCP	1335	1265[a]	−0.73 (−1.17)	0.52 (1.05)
CH$_3$CP	1591	1559[b]	−0.51 (−1.06)	0.49 (0.94)
*t*BuCP	1573	1533[b]	−0.52 (−1.11)	0.52 (0.93)
[(CF$_3$)$_3$BCP)]$^-$	1500	1468[b]	−0.66 (−1.52)	0.41 (0.83)
trans-[Pt(PMe$_3$)$_2$(Cl)(CP)]	1383	–	−0.85 (−1.20)	0.29 (0.71)
trans-[Pt(PMe$_3$)$_2$(Cl)(CP)-Pt(PMe$_3$)$_2$]	1126	–	−1.00 (−1.31)	0.10 (0.54)
[(dppe)$_2$Ru(H)(CP)]	1270	1229[c]	−0.76 (−1.31)	0.12 (0.54)

[a] Reference 2. [b] Reference 27. [c] Reference 8.

is 1.566 Å long, equivalent to the other C≡P bond lengths, and that the C, Pt and Cl atoms lie on the same line. The frequency analysis retrieves a vibration at 1383 cm^{-1} close to the HCP frequency.

Compound 7 in Figure 1 was synthesized[7] from the unstable platinum complex 6 and was studied within this work. We were able to confirm the structural data established experimentally[7a] that determined a bond length of 1.666(6) Å for the C≡P bond in comparison to the 1.657 Å calculated by us.

A ruthenium complex was also investigated, and structural information can be found in Figure 1. We obtained a C≡P bond distance of 1.582 Å which is slightly elongated relative to the experimental value[8] found in literature of 1.573(2) Å. The vibrational frequency is in good agreement, being 1270 cm^{-1} the calculated frequency and 1229 cm^{-1} the experimental one.[8] The complex here studied was modeled with methyl groups instead of the phenyl groups present in the synthesized compound.

All the results concerning frequencies and charges on the carbon and phosphorus atoms of the C≡P bond, obtained from the NPA and AIM analyses, are summarized in Table 1.

NBO analysis clearly confirms the existence of a triple bond between the C and P atoms for all the examined compounds (see below the discussion regarding compound 7). As an example, the σ-bond orbital obtained in the case of *t*BuCP is formed from hybrid orbitals on the C and P atoms: $\sigma(\mathrm{CP}) = 0.81\,(sp)_\mathrm{C} + 0.59\,(sp^{2.57})_\mathrm{P}$, whereas the π-bond orbitals, $\pi\,(\mathrm{CP}) = 0.74\,(p)_\mathrm{C} + 0.67\,(p)_\mathrm{P}$, are formed from pure p atomic orbitals. For the rest of the structures the C≡P bond description offered by NBO differs from the previous one in small variations of the polarization coefficients. We note that the polarization coefficients (0.81 for C and 0.59 for P) indicate that carbon, with the 65%, has the larger percentage of this NBO and gives the larger coefficient of 0.81. We note that in the case of the complex 7 the bonding description obtained by NBO is quite dependent on the basis sets quality. Indeed, some preliminary calculations performed by us at the B3LYP/6-31+G(d) level indicated that the C≡P bond (1.666 Å) was better described as a double bond. However, with the increase of the basis sets (B3LYP/6-311+G(2d,2p)) the NBO analysis characterizes the bond between C and P atoms as a triple bond. The carbon atom is also bonded to the Pt(1) atom interacting with the Cl ligand. As a consequence, the bonding of C≡P with the Pt(2) atom (see Figure 1) can be described in terms of donor−acceptor interactions. NBO second-order perturbation analysis shows that the donation of the C≡P π electrons to the metal *d* orbitals is accompanied by π-backdonation from the metal *d* orbitals to the empty C≡P π* orbital. Since the donation depopulates the π orbital of the ligand and the backdonation populates the ligand antibonding π*, the C≡P bond of the ligand lengthens and its substituent bends away from the metal. Indeed, the C≡P bond is longer (1.657 Å) than in the rest of the examined compounds, and the P−C−Pt(2) bond angle is significantly distorted from linearity (143.3°).

The NPA Charge Analysis gives a −1 charge on CP$^-$ concentrated on the C atom (−0.83 on C versus −0.17 on P). The C≡P bond is in all the other cases polarized as $\mathrm{C}^{\delta-}$−$\mathrm{P}^{\delta+}$. However, the negative charge existing on the C atom is always high, whereas the positive one on P atom significantly decreases when the ligand is coordinated to a metal center in stable complexes.

In Table 1 we have also included the atomic charges calculated within the AIM theory framework. The AIM theory provides a definition of atomic charges that is completely different from any other orbital-based population analysis. Atomic charges are obtained in this case by integration of the electron density within the atomic basins and adding the nuclear charges. Notably different values of the atomic charges were obtained with AIM, which predicts in all cases a larger charge separation. On the basis of the partition scheme used by AIM, we consider that AIM values are more reliable.

Table 2 summarizes the main information concerning C≡P bonds obtained by the topological analysis of the ELF function for all the studied species. In particular, we report the basin populations of the disynaptic V(C,P) and V(C,R) basins, which in terms of ELF analysis represent the CP bond and the second (and third in the case of compound 7) bond formed by that carbon atom, together with the corresponding variances, $\sigma^2(\bar{\mathrm{N}})$. We also report the monosynaptic V(P) basin populations, which represent the P lone pairs.

Figure 2 shows the ELF isosurfaces for CP$^-$, HCP, CH$_3$CP, and *t*BuCP, whereas the rest of the structures are presented in Figure 3.

In the case of the two simplest species studied in this work, CP$^-$ and HCP, the structures are characterized by the presence of a disynaptic valence basin, V(C,P), with an electron population of 2.91 and 4.07 e, respectively, together with monosynaptic V(P) basins with quite high populations (4.22 and 3.45 e, respectively). In both cases, we found strongly polarized disynaptic basins, as the atomic contribution coming from P atom is between 10 and 14% of the total population. Both valence basins are characterized by quite high variances, which indicate a great degree of electron delocalization. With the aim of comparison we have included the same data for CN$^-$ and HCN as a footnote of Table 2.

Table 2. ELF Topological Properties: Electron Population of the V(C,P), V(C,R) and V(P) Valence Basins, Together with the Corresponding Variances, $\sigma^2(\bar{N})$[a]

	V(C,P), $\sigma^2(\bar{N})$	V(C,R), $\sigma^2(\bar{N})$	V(P), $\sigma^2(\bar{N})$
CP^-	2.91, 1.39	2.66, 0.9[b]	4.22,1.55
HCP	4.07, 1.57	2.31, 0.70 (R = H)	3.45, 1.35
CH_3CP	4.26, 2.70	2.15, 1.06 (R = C)	3.48, 1.35
tBuCP	4.19, 2.67	2.22, 1.10 (R = C)	3.53, 1.37
$[(CF_3)_3BCP)]^-$	3.90, 2.66	2.37, 1.10 (R = B)	3.62, 1.40
trans-$[Pt(PMe_3)_2(Cl)(CP)]$	3.73, 2.26	2.21, 1.28 (R = Pt)	3.89, 1.49
trans-$[Pt(PMe_3)_2(Cl)(CP)Pt(PMe_3)_2]$	2.87, 1.47	1.94,1.19 (R = Pt1)	4.68. 1.91
		1.50,1.01 (R = Pt2)	
$[(dppe)_2Ru(H)(CP)$	3.39, 1.55	2.50,1.33 (R = Ru)	4.17, 1.57

[a] All quantities are given in electrons. With the aim of comparison, the same data for CN^- are as follows: V(C,N), $\sigma^2(\bar{N})$ = 3.39, 1.45; V(N), $\sigma^2(\bar{N})$ = 3.54, 1.27; V(C), $\sigma^2(\bar{N})$ = 2.86, 0.98; and for HCN: V(C,N), $\sigma^2(\bar{N})$ = 4.24, 1.54; V(N), $\sigma^2(\bar{N})$ = 3.28, 1.19; V(C,H), $\sigma^2(\bar{N})$ = 2.28, 0.67. [b] This value corresponds to the population and the variance of the monosynaptic V(C) basin.

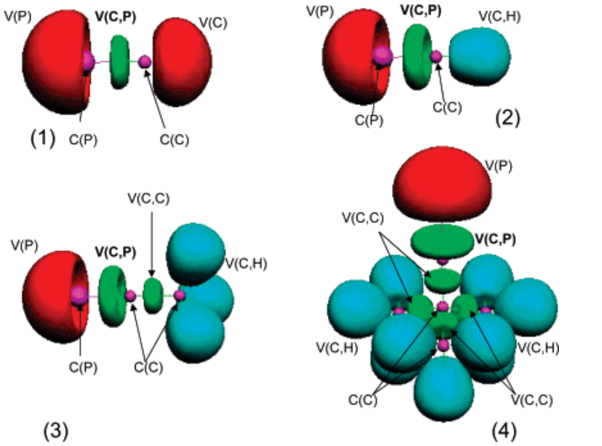


Figure 2. ELF isosurfaces (η(r) =0.8) for the optimized structures of (1) CP^-; (2) HCP; (3) CH_3CP; and (4) *t*BuCP. Core basins are represented in magenta, valence monosynaptic in red, protonated disynaptic in light blue; and valence disynaptic in green.

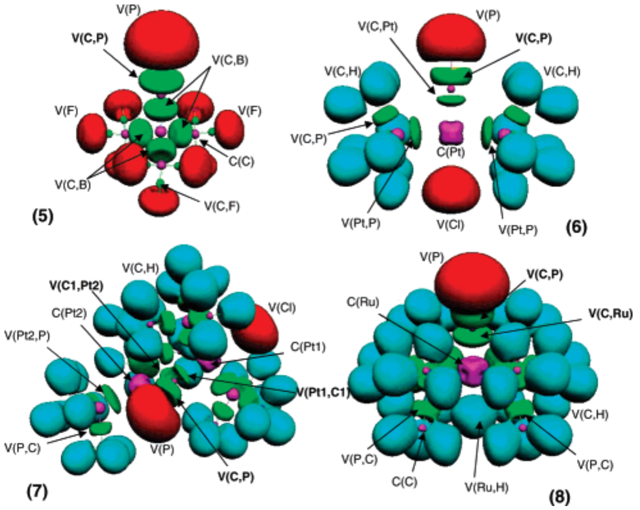


Figure 3. ELF isosurfaces (η(r)=0.75) for the optimized structures of (5) $[(CF_3)_3BCP)]^-$; (6) *trans*-$[Pt(PMe_3)_2(Cl)(CP)]$; (7) *trans*-$[Pt(PMe_3)_2(Cl)(CP)Pt(PMe_3)_2]$; and (8) $[(dppe)_2Ru(H)(CP)]$ (η=0.70) Core basins are represented in magenta, valence monosynaptic in red, protonated disynaptic in light blue, and valence disynaptic in green.

We have also reported the covariance matrix blocks, as obtained from TopMod package, in the Supporting Information.

Several studies present in bibliography have shown how the bond description obtained from ELF populations of multiple bonds involving atoms that contains lone pairs differs from our simplistic traditional pictures.[28−30] Indeed, in the particular case of triple bonds, the bonding population is usually significantly lower than the formal value of six, at the expense of increased lone-pair population. The origin of the so-called lone-pair bond weaking effect (LPBWE) has been largely studied and has demonstrated its importance for all atoms.[31] This subject has been analyzed in the context of Charge-Shift (CS) bonding and has been found that in ELF analysis it is manifested by a depleted basin population with a large variance and negative covariance. Based on Valence Bond Theory and ELF analysis calculations, it has been asserted that *triple bonding is invariably CS bonding*.[31]

The topological representation obtained from ELF analysis is usually interpreted in terms of superposition of mesomeric structures.[28,29] In particular, a detailed description of the N_2 molecule[29] points out that the charge contributions of π orbitals to the V(N,N) bonding basin and V(N) lone pairs are fairly equally shared. The results obtained here for CP^- are in line with the previous study performed by Silvi and co-workers on the CN^- and related series of compounds.[32]

In the next three studied moieties, CH_3CP, *t*BuCP, and $[(CF_3)_3BCP)]^-$, we found disynaptic V(C,P) valence basins with decreasing electron populations of 4.26, 4.19, and 3.90 e, respectively. The contribution to the electron population coming from the P atom is in all cases around 10% of the total electron population. The lower electron population found in the case of $[(CF_3)_3BCP)]^-$ is caused by the fact that the carbon atom involved in the C≡P bond is also engaged in a C−B covalent bond. That bond is mainly formed from the contribution coming from the C atom, as indicated by the presence of a V(C,B) basin with a population of 2.37 e, to which the C atom contributes with 88% of the total population. In all the structures the monosynaptic V(P) basins show electron populations quite high, namely, around 3.50 e. The $\sigma^2(\bar{N})$ values are around 2.67 for V(C,P) basins and 1.37 in the case of V(P).

Table 3. AIM Topological Properties: Charge Density at the C−P Bond Critical Point, ρ(BCP), the Laplacian at the Same Point, $\nabla^2\rho$ (BCP), the Ellipticity, ϵ, and the Delocalization Index, δ(C,P)[a]

	ρ(BCP)[b]	$\nabla^2\rho$(BCP)[b]	ϵ	δ(C,P)[c]
CP^-	0.213	0.32	0.00	2.80
HCP	0.220	0.63	0.00	2.56
CH_3CP	0.214	0.63	2.9×10^{-6}	2.48
tBuCP	0.214	0.63	2.1×10^{-5}	2.46
$[(CF_3)_3BPC)]^-$	0.220	0.59	9.9×10^{-5}	2.56
trans-$[Pt(PMe_3)_2(Cl)(CP)]$	0.210	0.51	2.7×10^{-2}	2.42
trans-$[Pt(PMe_3)_2(Cl)(CP)$-$Pt(PMe_3)_2]$	0.190	0.16	0.09	1.76
$[(dppe)_2Ru(H)(CP)$	0.210	0.43	4.5×10^{-3}	2.44

[a] With the aim of comparison the same data for CN^- are as follows: ρ(BCP) = 0.481, $\nabla^2\rho$(BCP) = −0.77; ϵ = 0.00, δ(C,N) = 2.40 and for HCN: ρ(BCP) = 0.453, $\nabla^2\rho$(BCP) = −0.51, $\epsilon = 9.1\ 10^{-8}$, δ(C,N) = 2.60. [b] Electron density at the bond critical point and its Laplacian, are both in au. [c] The delocalization index, δ(C,P), accounts for the electrons shared between the phosphorus and carbon atoms.

In the last three studied compounds, the C atom involved in the C≡P bond interacts also with a metal atom, as confirmed by the presence of a V(C,Ru) basin with a population of 2.50 e in $[(dppe)_2Ru(H)(CP)]$ and a V(Pt,C) basin with 2.21 e in *trans*-$[Pt(PMe_3)_2(Cl)(CP)]$. In both cases the main contribution to that populations comes from the C atom (around 85%). In $[(dppe)_2Ru(H)(CP)]$ and *trans*-$[Pt(PMe_3)_2(Cl)(CP)$ we found disynaptic V(C,P) basins, with total electron populations of 3.39 and 3.73 e, respectively. As in the previous cases the variances of these basins are quite high (see Table 2).

Finally, in compound 7, we found a V(C,P) valence basin with an electron population of 2.87 e. The electron population of the V(C,P) is quite depleted with respect to the previously described structures. This is due to the fact that according to ELF analysis the C atom interacts with both Pt atoms, as demonstrated by the presence of the V(Pt1,C1) and V(Pt2,C1) valence basins with electron populations of 1.94 and 1.50 e (see Figure 3 and Figure S1 in the Supporting Information), respectively. In both cases, the electron population is mainly brought by the C atom (around 80%). We have, therefore, some contrasting bond descriptions of compound 7. ELF analysis indicates that the central carbon atom is tricoordinated and that the CP bond can be described as a double bond, whereas NBO analysis performed at the B3LYP/6-311+G(2d,2p) level of theory shows that the C atom forms a covalent bond only with the closest Pt atom (the Pt1C bond length is 1.974 Å, whereas the Pt2C distance is 2.092 Å). As shown below, based on the CP bond ellipticity, we conclude that AIM analysis supports the bonding description obtained by ELF calculations.

The properties of the (3, −1) bond critical points are reported in Table 3. In particular, we present the electron density, ρ(BCP), the Laplacian, $\nabla^2\rho$(BCP), the ellipticity, ϵ, and the delocalization index, δ(C,P).

All C≡P BCPs display a significant concentration of electrons, with ρ(BCP) values ranging from 0.190 to 0.220 au. The BCP properties show that the nature of C≡P interactions is similar in all the studied structures. Significant values of ρ(BCP) along with values of $\nabla^2\rho(BCP) > 0$ and $|\lambda_1|/\lambda_3 < 1$ point to an interaction intermediate to closed-shell and shared nature. This description of the CP bond agrees with the atomic contributions to the V(C,P) basins obtained by ELF analysis, which shows a highly polarized bond. For the smaller structures, compounds 1−5, the ellipticity of the C≡P bond is zero or very close to zero (Table 3), whereas in the metal-containing complexes, that values slightly increase with the greatest value found in the case of complex 7, for which that value rises to 0.09.

4. Conclusions

In this paper, we have performed a comparative analysis of the bonding description obtained using different methodologies (ELF, AIM, NBO) of a series of compounds containing a terminal cyaphide bond.

The main conclusions drawn from this study can be summarized as follows:

1. The ELF V(C,P) basin populations obtained for all the studied species are characteristics of multiple bonds in which lone pairs are also involved. Moreover, in all the studied compounds, with the only exception of complex 7, the bonding basins have the axial symmetry characteristic of the triple bonds that yields a unique disynaptic basin which attractor is degenerated in a circle perpendicular to the C_∞ axis. In the case of compound 7 the V(C,P) bonding population is comparable to that obtained for CP^-; however, the shape of the bonding basin resembles more the prolate spheroid basins characteristic of double bonds. Moreover, according to ELF analysis the C atom involved in the compound 7 CP bond interacts with both Pt atoms through the formation of polarized covalent bonds. ELF analysis shows C≡P bonds that are highly polarized in nature.

2. AIM analysis indicates that in all the studied compounds the ρ(BCP) at the C≡P BCP is around 0.2 au. A slightly lower value was found in the case of complex 7. Those values are almost half of the corresponding values in CN^- and HCN, at the same level of theory. All the studied C≡P BCPs are characterized by small positive values of $\nabla^2\rho$(BCP), in contrast to CN^- and HCN which values are small and negative. For all the studied systems, the CP bond ellipticities are zero or very close to zero, with the only exception of compound 7, for which that value rises to 0.09.

3. NBO analysis generally supports the description obtained by ELF indicating the presence of a C≡P triple bond in all the studied species. In all cases, the polarity of the C≡P bond, as evaluated from the NBO polarization coefficients, is generally less marked than that obtained from ELF analysis considering the different atomic contributions to the disynaptic V(C,P) basins.

4. The used level of theory has generally well reproduced the experimental, geometrical, and spectroscopical properties as well as previous theoretical results.

Financial support from the Università degli studi della Calabria is gratefully acknowledged. N.R. thanks the HPC-Europe for the grant IDRIS-62007.

Supporting Information Available: Cartesian coordinates of fully optimized compounds 1−8, covariance matrix blocks of ELF analysis, and an additional image of the ELF basins of compound 7 (Figure S1). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Jutzi, P. *Angew. Chem.* **1975**, *87*, 269−283. Jutzi, P. *Angew. Chem., Int. Ed.* **1975**, *14*, 232−245.

(2) Gier, T. E. *J. Am. Chem. Soc.* **1961**, *83*, 1769−1770.

(3) Becker, G.; Gresser, G.; Uhl, W. *Z. Naturforsch., B: Chem. Sci.* **1981**, *36B*, 16−19.

(4) Regitz, M. *Chem. Rev.* **1990**, *90*, 191−213.

(5) (a) Nixon, J. F. *Chem. Rev.* **1988**, *88*, 1327−1362. (b) Ehlers, A.; Cordaro, J. G.; Stein, D.; Grützmacher, H. *Angew. Chem., Int. Ed.* **2007**, *46*, 7878−7881. (c) Angelici, R. J. *Angew. Chem., Int. Ed.* **2007**, *46*, 330−332.

(6) Weber, L. *Eur. J. Inorg. Chem.* **2003**, 1843−1856.

(7) (a) Jun, H.; Angelici, R. J. *Organometallics* **1994**, *13*, 2454−2460. (b) Jun, H.; Young, V. G., Jr.; Angelici, R. J. *J. Am. Chem. Soc.* **1992**, *114*, 10064−10065.

(8) Cordaro, J. G.; Stein, D.; Ruegger, H.; Grutzmacher, H. *Angew. Chem., Int. Ed.* **2006**, *45*, 6159−6162.

(9) (a) Nguyen, M. T.; Ha, T.-K.; *J. Mol. Struct. (THEOCHEM)* **1986**, *139*, 145−152. (b) Pyykkö, P.; Zhao, Y. *Mol. Phys.* **1990**, *70*, 701−714.

(10) (a) Hübler, K.; Schwerdtfeger, P. *Inorg. Chem.* **1999**, *38*, 157−164. (b) Kurita, E.; Tomonaga, Y.; Matsumoto, S.; Ohno, K.; Matsuura, H. *J. Mol. Struct. (THEOCHEM)* **2003**, *639*, 53−67. (c) Pascoli, G.; Lavendy, H. *J. Phys. Chem. A* **1999**, *103*, 3518−3524. (d) Mó, O.; Yanez, M.; Guillemin, J.-C.; Riague, E. H.; Gal, J.-F.; Maria, P.-C.; Poliart, C. D. *Chem. Eur. J.* **2002**, *8*, 4919−4924.

(11) 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, J. MoC.; 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.

(12) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(13) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623−11627.

(14) Wadt, W. R.; Hay, P. J. *J. Chem. Phys.* **1985**, *82*, 284−298.

(15) (a) Krishnan, R.; Binkley, J. S.; Seeger R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650−654. (b) Blaudeau, J.-P.; McGrath, M. P.; Curtiss, L. A.; Radom, L. *J. Chem. Phys.* **1997**, *107*, 5016−5021.

(16) (a) Reed, A. E.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 1736−1740. (b) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899−926.

(17) Becke, A. D.; Edgecombe, K. E. *J. Chem. Phys.* **1990**, *92*, 5397−5403.

(18) Silvi, B.; Savin, A. *Nature* **1994**, *371*, 683−986.

(19) (a) Noury, S.; Krokidis, X.; F. Fuster, B. Silvi, *TopMod Package*; Paris, 1997. (b) Noury, S.; Krokidis, X.; Fuster, F.; Silvi, B. *Comput. Chem.* **1999**, *23*, 597−604.

(20) Bader, R. F. *Atoms in molecules. A quantum theory*; Clarendon: Oxford, 1990.

(21) (a) Cremer, D.; Kraka, E.; Slee, T. S.; Bader, R. F. W.; Lau, C. D. H.; Nguyen-Dang, T. T.; McDougall, P. J. *J. Am. Chem. Soc.* **1983**, *105*, 5069−5075. (b) Bader, R. F. W.; Slee, T. S.; Cremer, D.; Kraka, E. *J. Am. Chem. Soc.* **1983**, *105*, 5061−5068. (c) Scherer, W; Sirsch, P.; Shorokhov, D . M Tafipolsky, M.; McGrady, G. S.; Gullo, E. *Chem. Eur. J.* **2003**, *9*, 6057−6070.

(22) (a) Fradera, X.; Austen, M. A.; Bader, R. F. W. *J. Phys. Chem. A* **1999**, *103*, 304−314. (b) Fradera, X.; Poater, J.; Simon, S.; Duran, M.; Solà, M. *Theor. Chem. Acc.* **2002**, *108*, 214−224.

(23) Ángayán, J. G.; Loos, M.; Mayer, I. *J. Phys. Chem.* **1994**, *98*, 5244−5248.

(24) Poater, J.; Duran, M.; Solà, M.; Silvi, B. *Chem. Rev.* **2005**, *105*, 3911−3947, and references therein.

(25) (a) Lavigne, J.; Pepin, C.; Cabana, A. *J. Mol. Spectrosc.* **1984**, *104*, 49−58. (b) Kroto, H. W.; Nixon, J. F.; Simmons, N. P. C. *J. Mol. Spectrosc.* **1979**, 77, 270−285.

(26) Antipin, M. Y.; Chernega, A. N.; Lysenko, K. A.; Struchkov, Y. T.; Nixon, J. F. *J. Chem. Soc., Chem. Commun.* **1995**, 505−506.

(27) Finze, M.; Bernhardt, E.; Willner, H.; Lehmann, C. W. *Angew. Chem., Int. Ed.* **2004**, *43*, 4160−4163.

(28) (a) Silvi, B. *Phys. Chem. Chem. Phys*. **2004**, *6*, 256−260. (b) Lepetit, C.; Silvi, B.; Chauvin, R. *J. Phys. Chem. A* **2003**, *107*, 464−473.

(29) Pilme, J.; Silvi, B.; Alikhani, M. E. *J. Phys. Chem. A* **2005**, *109*, 10028−10037.

(30) (a) Silvi, B.; Fourré, I.; Alikhani, M. E. *Monatsh. Chem.* **2005**, *136*, 855−879. (b) Chesnut, D. B. *Heteroat. Chem.* **2000**, *11*, 341−352.

(31) Shaik, S.; Danovich, D.; Silvi, B.; Lauvergnat, D. L.; Hiberty, P. C. *Chem. Eur. J.* **2005**, *11*, 6358−6371, and references therein.

(32) Matito, E.; Silvi, B.; Duran, M.; Solà, M. *J. Chem. Phys.* **2006**, *125*, 024301-9.

CT700277W

JCTC Journal of Chemical Theory and Computation

Full Configuration-Interaction Study on the Tetrahedral Li_4 Cluster

Antonio Monari,*,† Jose Pitarch-Ruiz,‡ Gian Luigi Bendazzoli,†
Stefano Evangelisti,§ and Jose Sanchez-Marin‡

Dipartimento di Chimica Fisica e Inorganica, Università di Bologna, Viale Risorgimento 4, I-40136 Bologna, Italy, Instituto de Ciencia Molecular, Universitat de Valencia, Edificio de Institutos, Campus de Paterna 46980, Valencia, Spain, and Laboratoire de Chimie et Physique Quantiques, Université de Toulouse et CNRS, 118, Route de Narbonne, F-31062 Toulouse CEDEX, France

Received December 4, 2007

Abstract: The Li_4 cluster low lying electronic states were studied. In particular we investigated the tetrahedral geometry at full CI and coupled cluster level, with basis sets of increasing quality. The $^5A_2$ electronic state, characterized by having all the valence electrons unpaired, forming a quite stable no-pair bonding state, was studied in greater detail. In order to compare the energies we also studied the Li_4 rhombus singlet ground state. The ability of coupled cluster with perturbative triples to correctly reproduce energy levels in a quasi-degenerate system was validated with respect to the full CI.

1. Introduction

Alkali metal clusters are very interesting molecular systems which exhibit particular and rather exotic electronic properties. In particular lithium clusters have been extensively studied by Bonačić-Koutecký and co-workers[1−5] as well as by other research groups like for instance Marx's[6] or Goddard's[7] ones. Such systems are known to have bound states in which all the valence electrons have the same spin, giving rise to the so-called “no-paired’’ chemical bonds, a situation which seems to be in contrast with the common chemical bonding model. For these reasons, these systems have been intensively studied by S. Shaik and co-workers using density functional techniques.[8−11] Moreover, these high-spin alkali metal clusters deserve special interest because of the role they play in the field of ultracold molecules. Although the clusters have never been observed as isolated species, at low temperature, highest-spin cluster states are stabilized by helium droplets, with a very high total spin selectivity. This gives rise to aggregates so stable that it was possible to determine experimentally many spectroscopic parameters.[12−19] For an exhaustive recent review see for instance ref 20. In this review the authors state that alkali clusters are likely to reside on the surface of the droplets, and since the probability of desorption directly correlates with the binding energy of the cluster, weakly bound high-spin states are preferentially transported by helium droplets. As an example the authors report how the mass spectroscopy signal for dimers in triplet states is enhanced by a factor of 50.[12,20] In this first paper we restrict the study to the Li_4 cluster, which, due to its small size, can be analyzed at the Full Configuration-Interaction (FCI) as well as the Coupled Cluster (CC) level of theory. The equilibrium geometry of this system is characterized by a rhomboidal structure, with a singlet spin multiplicity. Just at a slightly higher energy we find a manifold of electronic states with tetrahedral geometries, which are characterized by different spin multiplicity culminating in the no-pair bonding quintet. We decided to perform a systematic study of the low lying quasi-degenerate states of Li_4 cluster, at a high level of theory, with particular interest to the somewhat exotic high-spin bound state. Given that FCI is the most reliable method for the description of excited states, we are interested in performing a systematic comparison between FCI and CC to assess the ability of CC to correctly reproduce the energy levels. This is specially important in a situation

* Corresponding author e-mail: amonari@fci.unibo.it.
† Università di Bologna.
‡ Universitat de Valencia.
§ Université de Toulouse et CNRS.

10.1021/ct7003319 CCC: $40.75
Published on Web 02/16/2008

where the spatial symmetry requires more than one determinant to correctly reproduce the wave function.

We performed FCI and CC calculations at the equilibrium geometries of the singlet (rhombus) and quintet (tetrahedron) lowest states.[8−11] The rhomboidal geometrical parameters were obtained by an optimization carried on at the CCSD(T) level with several basis sets of increasing quality. On these optimized geometries, we performed a fixed-geometry FCI computation with all the basis sets.

As far as the tetrahedral geometries are concerned, we computed both at the FCI and CCSD(T) levels the potential energy curve corresponding to the symmetrical expansion of the tetrahedron ("tetrahedron breathing mode'' in spectroscopical terms). From these symmetry-constrained curves we obtained equilibrium bond distances and adjusted harmonic vibrational frequencies. Moreover, we investigated the Basis Set Superposition Error (BSSE) on the computed potential energy curves. The singlet rhomboidal ground state and all the states lying below the $^5A_2$ quintet for the T_d tetrahedral geometry were computed at the FCI and CC levels. Since our CC computations are of single-reference type, we were not able to describe all the multireference levels of this system. In some cases where a single reference CC wave function can be computed, we have a symmetry-breaking problem. In this situation the results are to be taken with caution, as it will be discussed later. This paper is organized as follows: in the second section we present some brief symmetry considerations, in the third section we expose the computational details, while in section four we present our results, drawing final conclusions in section five.

2. Symmetry Considerations

Due to the rather complex nature of the manifold of electronic states in the tetrahedron, some brief symmetry considerations (using the minimal 1s2s basis) can help for the correct analysis of the problem. The symmetry group of tetrahedral Li_4 is T_d. The lowest states are obtained by distributing the four valence electrons into the four valence 2*s* orbitals, while the four doubly occupied 1*s* orbitals form a totally symmetric core (A_1 symmetry) and do not contribute to the total symmetry. The four singly occupied 2*s* atomic orbitals give one a_1 molecular orbital and a 3-fold degenerate t_2 set with a slightly higher energy.

Two electronic configurations can be considered. First, if the a_1 orbital is kept doubly occupied, so that one has the $(a_1)^2(t_2)^2$ case, the following states (including the lowest ones) are possible[21]

$$^3T_1 \oplus {}^1A_1 \oplus {}^1E \oplus {}^1T_2$$

In the second case, we have a $(a_1)^1(t_2)^3$ electronic distribution, so the following states result

$$^{5,3}A_2 \oplus {}^{3,1}E \oplus {}^{3,1}T_1 \oplus {}^{3,1}T_2$$

We can call all the states in the first case as "closed-a_1-shell'' to distinguish them from those of the second case: "open-a_1-shell".

A maximum of six singlets, five triplets, and one quintet states can be obtained from the four valence electrons in the a_1t_2 valence orbitals set.

The following five states are found in the lower energy range: (1) one quintet of A_2 symmetry; (2) a "closed-a_1-shell" triplet of T_1 symmetry; (3) two "closed-a_1-shell" singlets: one of A_1 symmetry and a 2-fold degenerate singlet of E symmetry, and (4) a "closed-a_1-shell" singlet of T_2 symmetry. Notice, however, that the difference between the open-a_1 and closed-a_1 shell singlets is not a rigorous one, and it is here introduced only to make easier the description of the states.

Among all these different states, the triplet is the lowest one. Since there are three degenerate components of the triplets, the system will be Jahn−Teller distorted, giving a triplet minimum that has (as far as the first evidence seems to indicate) D_{2d} symmetry. The study of this distorted Li_4 system will be the object of a future work.

By performing a CAS-CI calculation with the four valence ROHF orbitals of the quintet (the only possible nondegenerate ROHF calculation) and using the cc-pVTZ basis set at the ROHF optimized geometry, one obtains the following energies (in hartrees): (1) $^1A_1$: −29.704539; (2) $^1T_2$: −29.725828; (3) ^1E: −29.735939; (4) $^5A_2$: −29.753167; and (5) $^3T_1$: −29.756536.

These CAS-CI states are the starting point for the FCI and CC computations.

Since our ab initio codes are restricted to use Abelian point groups, our computations were performed in the D_{2h} and C_{2v} symmetries for the rhombus and the tetrahedron, respectively.

3. Computational Details

In this section the basis sets and the computational strategy used in the present study are described in full detail.

3.1. Basis Sets. We performed a systematic study using polarized-valence correlation-consistent basis sets (proposed by Dunning and co-workers[22−24]) of increasing quality, in particular the cc-pVDZ, cc-pVTZ, and cc-pVQZ. The bases were retrieved from the EMSL public database.[25] The use of correlation-consistent basis sets allowed us to perform extrapolation to the infinite basis,[26] leading to a better estimation of the cluster properties. Using the smaller bases, i.e., cc-pVDZ and cc-pVTZ, we computed the FCI potential energy curves for the expansion of the tetrahedron, while with the cc-pVQZ basis, only the regions close to the energy minima were calculated. In this case therefore, the curves were drawn using a limited subset of distance points.

3.2. Computational Methods. The core has been kept frozen in all the FCI computations, performed with the Bologna FCI code.[27] The core 1*s* orbital of each Li atom was doubly occupied and frozen at the SCF level of the $^5A_2$ in tetrahedral geometry cases, while in the rhombus case the 1*s* orbitals were frozen at the SCF level of the singlet ground state. The FCI space amounts to about $134 \cdot 10^6$ symmetry-adapted determinants with the cc-pVQZ basis set in C_{2v} symmetry. Integrals were computed by using the DALTON 2.0 code[28] and subsequently transformed by the Ferrara four-index transformation.[29] These codes were interfaced by using the newly developed Q5Cost format and libraries[30−33] designed for code interoperability. CC to the singles and doubles with noniterative correction to the triples,[34] CCSD(T), were performed by using the MOLPRO 2000 code.[35] In this case both frozen-core (FC) and all-electrons (AE) CCSD(T) computations were carried on. In order to perform CCSD(T) calculations, preliminary ROHF wave functions were computed. The electronic configurations were $1a_1^2 2a_1^2 1b_1^2 1b_2^2 3a_1 4a_1 2b_1 2b_2$, $1a_1^2 2a_1^2 1b_1^2 1b_2^2 3a_1^2 4a_1 2b_1$, and $1a_1^2 2a_1^2 1b_1^2 1b_2^2 3a_1^2 4a_1^2$ in C_{2v} symmetry, for the $^5A_2$, $^3T_1$, and ^1E states, respectively. The corresponding electron configurations in the T_d point group are $1a_1^2 1t_2^6 2a_1 2t_2^3$, $1a_1^2 1t_2^6 2a_1^2 2t_2^2$, and $1a_1^2 1t_2^6 2a_1^2 2t_2^2$, where the last configuration (singlet) includes a doubly occupied t_2 orbital. The implications of this feature on the CCSD(T) calculations will be discussed later. In all cases the orbitals corresponding to the 1s of the Li atoms were frozen in the FC CCSD(T) computation. In case of triplet and singlet states the single-determinant ROHF wave functions are of broken-symmetry type.

3.3. Constrained Geometry Optimization. A symmetry-constrained optimization of the tetrahedral geometry was performed at the FC FCI level for all the states and at the FC and AE CCSD(T) levels for those states dominated by a single determinant. We kept the tetrahedral conformation during the computation, and our results are expressed as a function of the Li−Li distance R. The equilibrium geometry and the energy well depth were obtained by means of an exponential spline interpolation[36] of the potential. The adjusted harmonic frequency ω of the "breathing mode'' of tetrahedral Li_4 was obtained from a fourth-degree polynomial least-square fitting to the energy potential of each state, expressed as a function of the normal coordinate $Q_1 = 2\sqrt{m}\,\delta$. In this expression δ stands for the displacement from the equilibrium position of each Li vertex along the C_3 tetrahedron axis, while m stands for the mass of Li atom. The ^7Li isotope with $m = 7.0160040$ has been assumed. The procedure was as follows: a number of points around the guessed minimum were selected so that both branches of the potential had a similar depth. A fourth-degree polynomial was then fitted to determine the minimum position. Once such position was known, the δ displacement of each point could be calculated. The resulting set of points (E,Q_1) for each state of interest was then fitted to the following polynomial

$$E = V_0 + \frac{1}{2} V_0'' {Q_1}^2 + \frac{1}{3!} V_0''' {Q_1}^3 + \frac{1}{4!} V_0^{IV} {Q_1}^4$$

and the corresponding adjusted harmonic frequency was obtained from the square root of the coefficient V_0'' and then converted to wavenumber units. The rhombus singlet state was optimized at the FC and AE CCSD(T) levels. A single-point FC FCI computation was subsequently performed at the FC CCSD(T) optimized geometry. The equilibrium distances were expressed as the two rhombus diagonals, R_1 and R_2 (with $R_1 > R_2$).

4. Results and Discussion

The main results for the tetrahedron geometry are collected in Tables 1−3. At the FC FCI level the states are the same as those obtained at the CAS-CI level, as reported in section 2, but the energy order is different. At the CC level only the $^3T_1$, ^1E, and $^5A_2$ states could be computed. The $^3T_1$ is the ground state both at the CAS-CI and FCI levels. On the other hand, the $^5A_2$ state, which is the first excited CAS-CI state, becomes the highest, i.e., the fourth excited one in FCI. This fact is certainly due to the small dynamic correlation associated with the quintet states as compared to the singlets and triplets. Among the singlet states the E and the T_2 are interchanged from the CAS to the FCI description. This is due to the presence of an important nondynamic correlation in the case of the open-shell $^1T_2$ state. The $^3T_1$, ^1E, and $^5A_2$ states dissociate to four Li atoms in their ground ^2S states, while the $^1T_2$ and $^1A_1$ multireference states dissociate to a symmetry adapted combination of three Li atoms in their

Table 1. Li_4 cc-pVDZ Tetrahedron Spectroscopic Properties[a]

		R	*E*	*ω*(*err*)
$^3T_1$				
	CCSD(T) FC	5.649	−2.178	316.7 (1.2)
	CCSD(T) AE	5.618	−2.218	316.6 (0.8)
	FCI FC	5.649	−2.238	322.4 (1.1)
$^1T_2$				
	FCI FC	5.722	−1.993	320.7 (1.2)
^1E				
	CCSD(T) FC	5.633	−1.760	312.2 (0.9)
	CCSD(T) AE	5.598	−1.800	306.8 (0.5)
	FCI FC	5.639	−1.838	311.7 (1.2)
$^1A_1$				
	FCI FC	5.761	−1.456	317.6 (1.3)
$^5A_2$				
	CCSD(T) FC	5.777	−1.248	271.7 (1.8)
	CCSD(T) AE	5.742	−1.282	269.2 (0.6)
	FCI FC	5.784	−1.267	271.3 (1.0)

[a] *R* is the equilibrium Li−Li distance in bohr, *E* is the dissociation energy with respect to four Li atoms in their ground states in eV, and *ω*(*err*) is the adjusted harmonic frequencies and fitting error (in brackets) in cm^{-1}.

Table 2. Li_4 cc-pVTZ Tetrahedron Spectroscopic Properties[a]

		R	*E*	*ω*(*err*)
$^3T_1$				
	CCSD(T) FC	5.535	−2.350	329.5 (1.2)
	CCSD(T) AE	5.453	−2.472	323.8 (0.9)
	FCI FC	5.539	−2.404	318.9 ((1.2)
$^1T_2$				
	FCI FC	5.613	−2.170	316.8 (1.4)
^1E				
	CCSD(T) FC	5.506	−1.956	317.3 (1.1)
	CCSD(T) AE	5.423	−2.081	320.6 (0.9)
	FCI FC	5.517	−2.035	315.5 (1.4)
$^1A_1$				
	FCI FC	5.652	−1.632	312.2 (1.1)
$^5A_2$				
	CCSD(T) FC	5.696	−1.347	271.6 (1.1)
	CCSD(T) AE	5.588	−1.464	274.5 (1.1)
	FCI FC	5.702	−1.367	271.1 (1.1)

[a] *R* is the equilibrium Li−Li distance in bohr, *E* is the dissociation energy with respect to four Li atoms in their ground states in eV, and *ω*(*err*) is the adjusted harmonic frequencies and fitting error (in brackets) in cm^{-1}.

Table 3. Li_4 cc-pVQZ Tetrahedron Spectroscopic Properties[a]

		R	*E*	ω(*err*)
$^3T_1$				
	CCSD(T) FC	5.527	−2.382	320.7 (1.2)
	CCSD(T) AE	5.305	−2.745	355.3 (1.2)
	FCI FC	5.532	−2.437	310.8 (1.0)
$^1T_2$				
	FCI FC	5.601	−2.206	309.3 (0.8)
^1E				
	CCSD(T) FC	5.500	−1.996	316.9 (1.2)
	CCSD(T) AE	5.275	−2.361	352.8 (1.5)
	FCI FC	5.511	−2.071	306.0 (1.2)
$^1A_1$				
	FCI FC	5.646	−1.669	305.3 (1.1)
$^5A_2$				
	CCSD(T) FC	5.692	−1.364	271.9 (1.0)
	CCSD(T) AE	5.387	−1.700	305.7 (1.3)
	FCI FC	5.646	−1.383	264.1 (1.2)

[a] *R* is the equilibrium Li−Li distance in bohr, *E* is the dissociation energy with respect to four Li atoms in their ground states in eV, and ω(*err*) is the adjusted harmonic frequencies and fitting error (in brackets) in cm^{-1}.

ground state and one Li atom in a ^2P ($1s^22p^1$) state, as it has been verified by an analysis of the FCI wave function in the dissociation region. This fact has also been confirmed by checking that the tetramer energy at the limit of infinite Li−Li distance is equal to the energy of three lithium atoms, in their ground state, plus the energy of one lithium atom in a ^2P state. The equilibrium properties of the tetrahedral states for all the levels of theory are reported in Tables 1−3 for the cc-pVDZ, cc-pVTZ, and cc-pVQZ basis, respectively. The FC FCI potential energy curves for the tetrahedral states are also reported in Figure 1. In Table 4 we report the equilibrium properties for the rhombus geometry. The Complete Basis Set (CBS) extrapolated values for the two geometries are reported in Table 4 for the rhombus and in Table 5 for tetrahedron geometry, respectively. The CBS-(QT) values were computed by extrapolating to the complete basis set limit the obtained equilibrium properties for the various electronic states, by using the formula proposed in ref 26.

As discussed above, the triplet $^3T_1$ tetrahedron ground state is obtained by putting two electrons in two of the triply degenerate t_2 orbitals. The energy difference between the tetrahedral states and the rhombus ground state is rather small. The energy difference between the tetrahedral triplet and the rhombus singlet is about 0.5 eV, while the corresponding difference for the quintet is about 1.6 eV. The adjusted harmonic frequencies reported in Tables 1−3 reflect the change of curvature of the potential (in the region around the minima) as one goes from the $^3T_1$ ground state to the $^5A_2$ quintet state. The similarity between the FC FCI and FC CCSD(T) values is confirmed, but some interesting features, related to the AE CCSD(T) results, indicate a quite important core effect that we will discuss in greater detail in the next subsection. Finally, we report in Table 6 the dissociation energy for the reaction $2Li_2 \rightarrow 4Li$ computed at FC CCSD (which for this system is equivalent to FC FCI) and AE CCSD(T) with the three different basis sets. From these data one can see that the Li_2 dimer is less stable than the rhombus Li_4 tetramer and also less stable than the tetrahedral $^3T_1$ state. Our results can be compared with the DFT computation recently reported by Shaik et al. in different papers (see ref 8 and references therein). For the $^5A_2$ state the estimated bond dissociation energy is 1.197 eV, and the equilibrium Li−Li distance is 5.5857 bohr.[8] These results were obtained by using the B3PW91 functional and a 6-311G(2d) basis, although some slight variability is found by using different functionals.

4.1. Core Correlation. If we consider the vibrational frequencies reported in Tables 1−3, we may see that the contribution of the core electrons described with the cc-pVDZ basis set has an opposite sign compared to that of the larger basis sets, and this can be certainly related to the poor description of the core. This is a well-known defect of the cc-pVXZ basis sets, and it is quite large for the triple-ζ (about 3−4 cm^{-1}) and very high for the quadruple-ζ basis (about 35 cm^{-1}). This reflects that the bases are far from being saturated for this property, if one wants to account for the core effects. Of course, better adapted basis sets such as the polarized core-valence XZ can be used but, due to the larger basis dimension, the computational cost would be higher. As can be seen from Figure 2, and from the values of equilibrium properties reported in the tables, the FC FCI results are extremely close to those corresponding to the FC CCSD(T). For instance, for the quintet state the energy difference along the potential is usually less than 0.02 eV (cc-pVQZ basis) and never exceeds 0.08 eV. Finally, the effect of core correlation on the value of the energy well depth appears to be important, although not essential. The difference between the FC and the AE CCSD(T) energy well depth is about 0.4 eV, while the equilibrium distance is much less sensitive. For a comparison, at infinite distance the effect of core correlation amounts to 0.365 eV (cc-pVTZ) and 0.466 eV (cc-pVQZ) per Li atom. For these reasons we decided to perform AE and FC CCSD(T) calculations using the polarized core-valence cc-pCVQZ basis set. The corresponding results are collected in Table 7. As can be seen by using such a basis set, the difference among FC and AE CCSD-(T) computed properties diminishes significantly if compared with the cc-pVQZ FC and AE values, thus confirming the rigidity of the core description in the pVXZ basis set series. Unfortunately, due to the larger basis dimension, the price one has to pay in order to use the cc-pCVXZ basis is a higher computational cost, as already stated. Hence, the FCI computations with these basis sets would have been too expensive, and they have not been performed.

4.2. The Quintet State. From a theoretical point of view, the bound quintet $^5A_2$ state is the most interesting one, and a large amount of literature has been produced (for example see refs 8 and 11 and references therein). In Figure 2, we compare the potential energy curve for this state computed at the SCF, FC CCSD(T), AE CCSD(T), and FC FCI levels. The SCF curve shows a pronounced minimum giving a well depth of almost one-half to that obtained in the FC FCI calculation. This situation is in sharp contrast with the behavior of the triplet Li_2 state (see Figure 3), for which the SCF gives a repulsive curve, and only the correlated methods

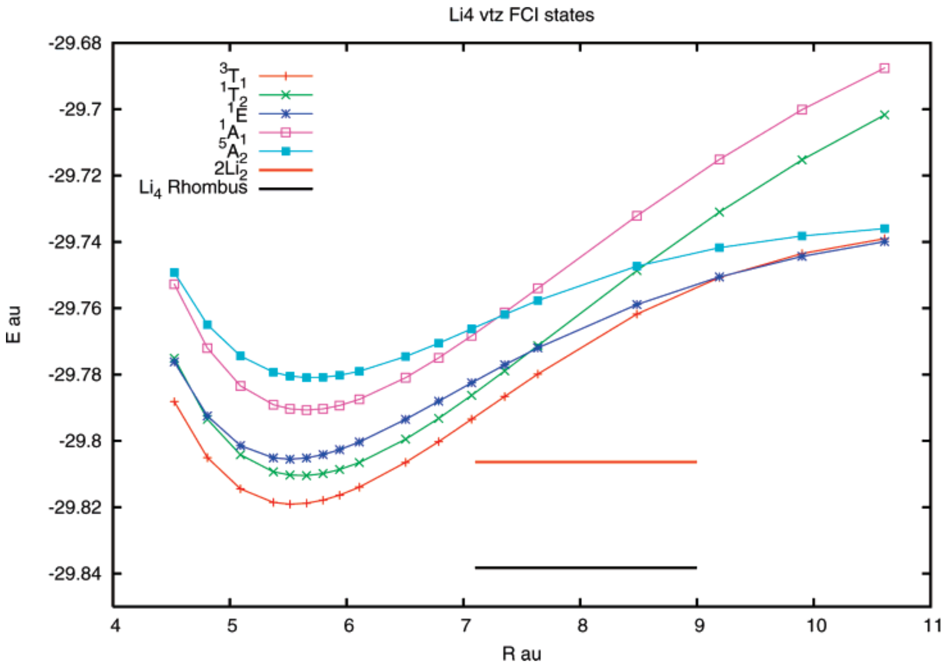


Figure 1. Li_4 FCI cc-pVTZ basis potential energy curves. Li_4 rhombus indicates the energy level of the tetramer in the equilibrium rhombus geometry, while $2Li_2$ is the energy of two isolated Li_2 dimers in the equilibrium $^1\Sigma_g^+$ state. Distances are given in bohrs and energies in hartrees.

Table 4. Li_4 Rhombus Equilibrium Properties[a]

		R_1	R_2	E
cc-pVDZ				
	CCSD(T) FC	10.366	5.160	−2.762
	CCSD(T) AE	10.325	5.113	−2.796
	FCI FC	//	//	−2.783
cc-pVTZ				
	CCSD(T) FC	10.244	5.040	−2.906
	CCSD(T) AE	10.142	4.945	−3.012
	FCI FC	//	//	−2.927
cc-pVQZ				
	CCSD(T) FC	10.238	5.031	−2.940
	CCSD(T) AE	9.902	4.837	−3.237
	FCI FC	//	//	−2.959
infinite basis	set extrapolation			
	CCSD(T) FC	10.235	5.027	//
	CCSD(T) AE	9.782	4.783	//
	FCI FC	//	//	−2.975

[a] R_1 and R_2 are the diagonals of the rhombus (in bohr), and E is the dissociation energy with respect to four Li atoms in their ground states (in eV).

Table 5. Li_4 Tetrahedron Equilibrium Geometry Extrapolated to the Infinite Basis Set[a]

		R	E
$^3T_1$			
	CCSD(T) FC	5.523	−2.248
	CCSD(T) AE	5.231	−2.881
	FCI FC	5.529	−2.454
$^1T_2$			
	FCI FC	5.95	−2.224
^1E			
	CCSD(T) FC	5.497	−2.202
	CCSD(T) AE	5.201	−2.501
	FCI FC	5.508	−2.114
$^1A_1$			
	FCI FC	5.643	−1.688
$^5A_2$			
	CCSD(T) FC	5.690	−1.373
	CCSD(T) AE	5.287	−1.818
	FCI FC	5.618	−1.391

[a] R is the equilibrium Li−Li distance in bohr, and E is the dissociation energy with respect to four Li atoms in their ground states in eV.

Table 6. Li_2 Dimer Dissociation Energy (in eV)[a]

	FC CCSD=FC FCI	AE CCSD(T)
cc-pVDZ	−1.948	−1.968
cc-pVTZ	−2.059	−2.117
cc-pVQZ	−2.085	−2.262

[a] The energy values are given with respect to the dissociation of two dimers (singlet multiplicity) to four Li atoms.

predict a small energy well. By using the cc-pVTZ basis we obtained at the CCSD level, for the Li_2 triplet (note that the FC case is equivalent to the FCI), an equilibrium distance of 3.984 bohr and a well depth of 0.0396 eV only. Instead, in the AE CCSD(T) case, the equilibrium distance amounts to 3.915 bohr and the dissociation energy to −0.0455 eV. The SCF curve of the Li_4 quintet state (see Figure 2) shows a tiny barrier at about 9.5 bohr (9.511 bohr for the cc-pVTZ basis) of about 0.02 hartree (0.024 hartree for the cc-pVTZ basis). This is in agreement with the expected long-range repulsive behavior of the high-spin SCF. These facts, combined with the difference in binding energies for the dimer and the tetramer, indicate that the actual nature of the no-pair bound in these systems is not yet perfectly understood.

As reported by Shaik et al.,[8−11] a key feature of the quintet bound state nature is the strong participation of *p* type orbitals to the molecular orbitals involved in the bond. The key role of the *p* orbitals in explaining the nature of the lithium systems was pointed out also by Marx and Rousseau[6] in their

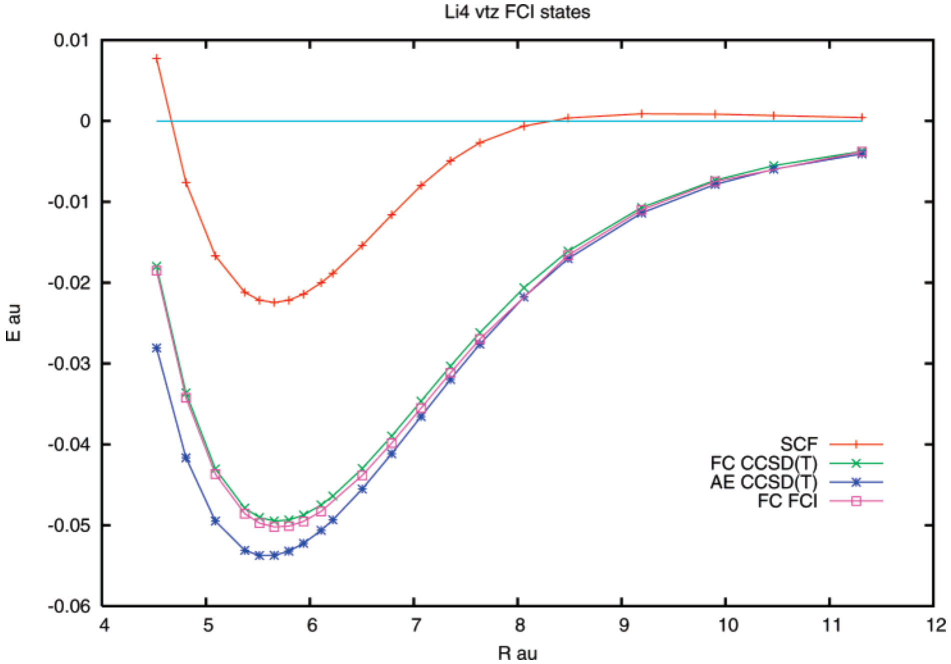


Figure 2. Li_4 $^5A_2$ quintet state potential energy curves at various levels of theory. Distances are given in bohrs and energies in hartrees. Energies are given with respect to the corresponding dissociation limit values.

Table 7. Li_4 cc-pCVQZ CCSD(T) Tetrahedron Spectroscopic Properties[a]

		R	*E*	ω(*err*)
$^3T_1$				
	CCSD(T) FC	5.526	−2.386	321.5 (0.8)
	CCSD(T) AE	5.467	−2.418	325.5 (0.7)
^1E				
	CCSD(T) FC	5.498	−1.999	317.0 (0.7)
	CCSD(T) AE	5.442	−2.025	322.4 (0.7)
$^5A_2$				
	CCSD(T) FC	5.688	−1.366	272.2 (1.0)
	CCSD(T) AE	5.620	−1.368	276.4 (0.8)

[a] *R* is the equilibrium Li−Li distance in bohr, *E* is the dissociation energy with respect to four Li atoms in their ground states in eV, and ω(*err*) is the adjusted harmonic frequencies and fitting error (in brackets) in cm^{-1}.

analysis of the rhombus Li_4 The importance of the *p* orbitals is confirmed, for example, if one performs a SCF computation only using the *s* cc-pVTZ basis without *p* orbitals. In this case a repulsive curve is found. Moreover, the inclusion of *p* orbitals seems to lower the energy gap between the a_1 and the t_2 orbitals, therefore favoring the population of the t_2 shell. With the cc-pVTZ basis at the quintet FC FCI equilibrium geometry (5.702 bohr) one obtains a a_1-t_2 energy gap of 0.09949 hartrees for the complete basis set and a gap of 0.122424 hartrees for the subset comprising only *s* orbitals. Therefore, the elimination of *p* orbitals increases the gap by about 20%. The mixing of *p* orbitals induces a strong distortion in the valence orbitals in the region of the energy minimum. This can be seen if local hybrid orbitals[37] are computed from a ROHF wave function. At long distance the local valence orbitals are spherical, as shown in Figure 4a, where an orbital from an ANO 4s2p basis set,[38] at a Li−Li distance of 12.0 bohr, is represented. In Figure 4b, the corresponding orbital at a distance of 5.5 bohr is plotted. At this last distance one can see how the distorted orbitals allow a migration of the charge toward the inner region of tetrahedron and in particular around the Li−Li bonds. This can explain the stability of the no-pair bond state and confirms the explanation proposed by Shaik et al.[8]

Moreover, these facts, are in agreement with the analysis performed by Gatti et al.[39] using the Atoms in Molecules formalism. In the case of lithium aggregates, they found an unusual maximum of the electronic density at the midpoint of the Li−Li equilibrium distance. At the FC FCI level the quintet wave function has a very strong single-determinant nature, as it was confirmed by the analysis of the wave function in terms of determinant contributions. This can also be seen from the occupation numbers of the natural orbitals reported in Table 8 for the cc-pVTZ basis. The occupation is mainly restricted to the four quasi-degenerated a_1 and t_2 orbitals, with some small contributions from higher orbitals.

4.3. The Symmetry-Breaking Problem. As stated before in the computational details section, the use of Abelian subgroups leads to symmetry breaking for the triplet and the singlet tetrahedral states for our single-reference CCSD(T) computations. FCI is not affected by this problem. This effect is already present at the ROHF level (see the section on Computational Details). One way to investigate the symmetry breaking is the use of the quintet ROHF state (which is symmetry breaking-free) as a reference state for all the subsequent CC calculations. We performed such a test with the cc-pVTZ basis at the Li−Li equilibrium distance of 5.656 bohr. We computed the FC CCSD(T) energy values for the singlet and triplet states starting respectively from the quintet reference state and from the usual symmetry broken ROHF triplet and singlet determinants, subsequently comparing the obtained values among them and with the FC FCI.

For the singlet ^1E state we obtained a FC CCSD(T) energy value of −29.80123 hartrees starting from the quintet reference state. This value can be compared to the FC CCSD-

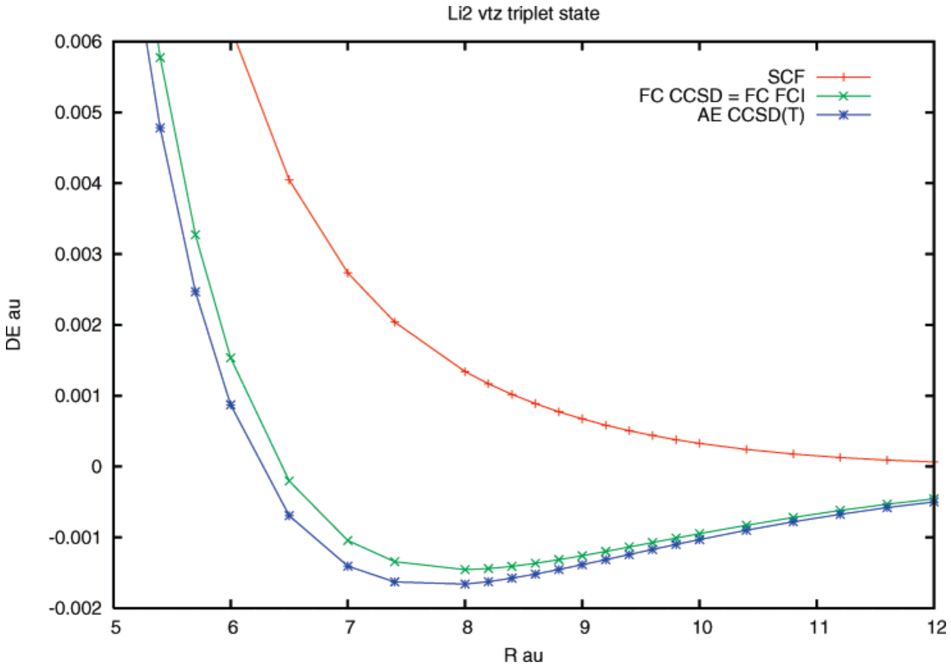


Figure 3. Potential energy curves for the Li_2 triplet $^3\Sigma_u^+$ state at various levels of theory. Distances are given in bohrs and energies in hartrees. Energies are given with respect to the corresponding dissociation limit values.

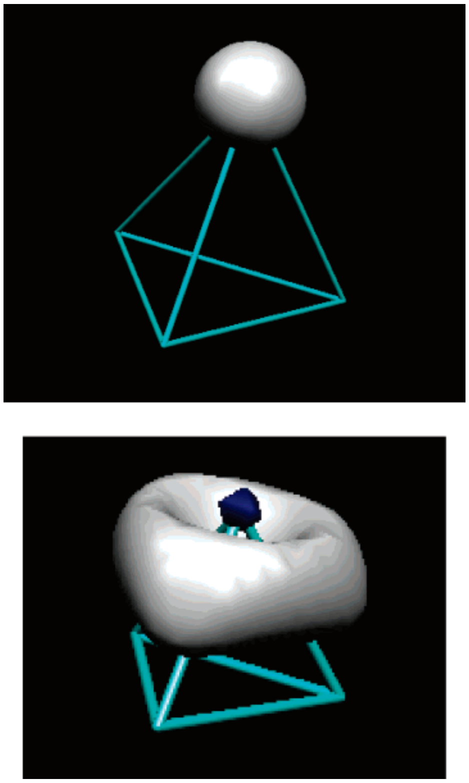


Figure 4. Local valence orbital for the Li_4 $^5A_2$ state at two different distances *R*: (a) *R* = 12.0 bohr and (b) *R* = 5.5 bohr.

Table 8. FC FCI Natural Orbital Symmetries and Occupation Numbers for the $^5A_2$ Li_4 State

orbital multiplet	symmetry	orbital occupation no.
1	a_1	0.9836
2	t_2	0.9493
3	e	0.0294
4	t_2	0.0224
5	a_1	0.0085
6	t_2	0.0067
7	t_1	0.0020
8	t_2	0.0013
9	a_1	0.0009
10	*e*	0.0004
11	t_2	0.0003

(T) value of −29.80218 hartrees obtained using the singlet reference state and to the FC FCI of −29.80514 hartrees.

Correspondingly, for the triplet $^1T_1$ state we obtained a FC CCSD(T) energy value of −29.81668 hartrees using the quintet as the ROHF reference state. This value can be compared with the FC CCSD(T) value of −29.81669 hartrees using the triplet as the reference state and to the FC FCI value of −29.81879 hartrees. These results show effects due to the symmetry breaking that are very small, compared with the corresponding differences with FCI. The excitation energies are almost unaffected. The differences in CC excitation energy are $9.0 \cdot 10^{-4}$ for the ^1E state and $1.0 \cdot 10^{-5}$ for the $^3T_1$ state.

4.4. Basis-Set Superposition Error. The Basis-Set Superposition Error[40] effect was evaluated by using the standard Boys−Bernardi Counterpoise Correction.[41] As we are using size-extensive methods, such a procedure is valid on the whole potential energy curve.[33] The behavior of the counterpoise correction is illustrated in Figures 5 and 6 for the FC and the AE cases, respectively. The effect of the BSSE appears to be small with the bases used in this study. In the FC case the effect decreases by using larger basis sets, and it is practically negligible, even with the smallest cc-pVDZ basis. In the AE case, the behavior is less regular although the effect is still small. For these reasons we do not report the BSSE-corrected potential energy curves and spectroscopic properties. A rather surprising nonmonotone behavior of the counterpoise curves is observed in almost all cases. The

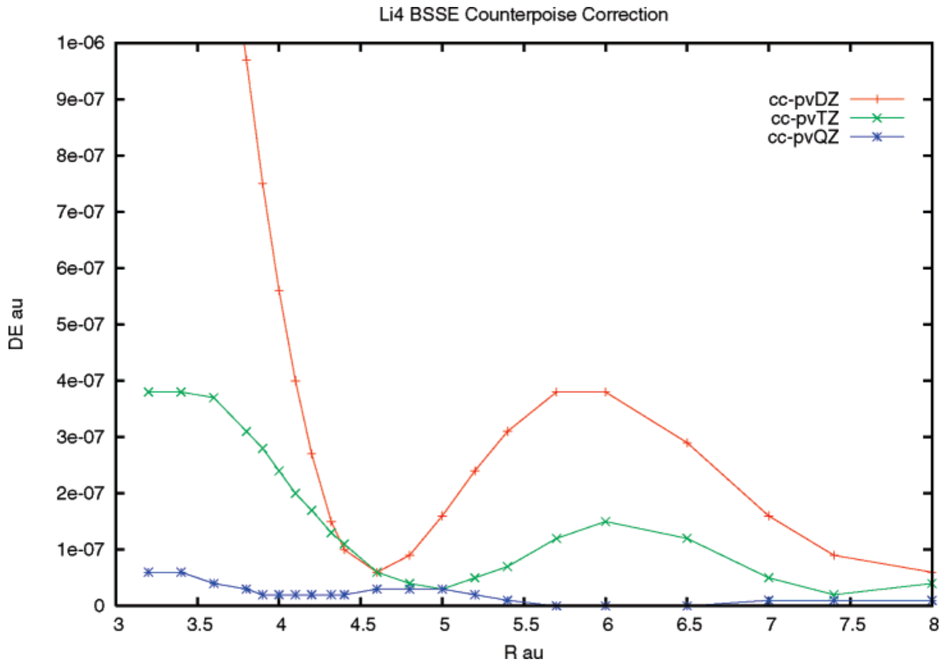


Figure 5. Frozen core BSSE for all states of Li_4 estimated by the counterpoise method. Distances are given in bohrs and energies in hartrees.

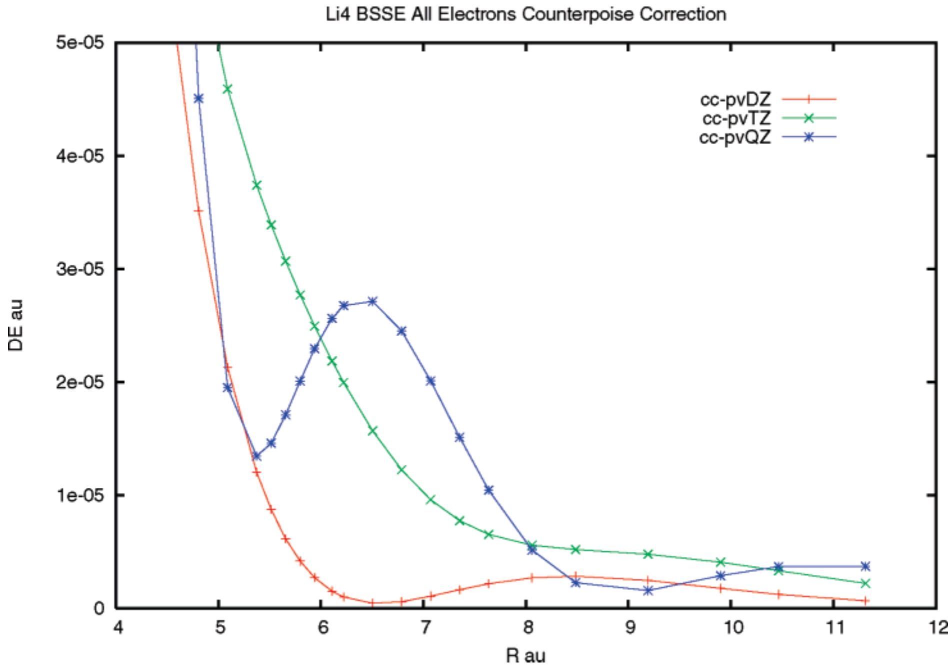


Figure 6. All electrons BSSE for all states of Li_4 estimated by the counterpoise method. Distances are given in bohrs and energies in hartrees.

reason for this behavior is probably due to the rather poor description of the core basis functions of the basis sets (as discussed in subsection 4.1).

5. Conclusions

We present a FCI and coupled cluster benchmark study of some low-lying electronic states of the Li_4 cluster, in particular the rhombus singlet ground and several tetrahedral states with different spin multiplicity. The CCSD(T) method appears to be an extremely reliable tool for the investigation of these systems, as long as no quasi-degenerate states are involved. The cc-pVTZ and a fortiori the cc-pVQZ basis set can be considered quite close to the basis set limit for the frozen-core approach. CCSD(T) computations with these bases can therefore be used to investigate larger high-spin lithium clusters, where FCI would be unfeasible. In particular coupled cluster is able to reproduce the no-pair bonding state with remarkable accuracy, giving values extremely close to the ones in FCI for the equilibrium properties. As far as the nature of the “no-pair bond” of the quintet state is concerned, the explanation proposed by Shaik, on the ground of a VB calculation, has been confirmed by an independent method.

We plan to extend this investigation in two directions: on one hand, we will use CC to study other alkali metal clusters (Na_4, K_4, etc.), looking for further insight into the nature of the bonds in high spin states. On the other hand, it will be interesting to perform a FCI and CC study on the Jahn–Teller distortion of the triplet state in Li_4.

Moreover it will be interesting to perform computation of alkali clusters interacting with helium in order to assess theoretically the experimentally observed stabilization mechanism of the high-spin states.

Acknowledgment. We would like to acknowledge the COST Chemistry Action D37 for partially supporting this work. S.E. would like to thank the French CNRS for financial support; G.L.B. and A.M. wish to thank the Italian Ministry of University, the MUR Research, and the University of Bologna for partly financing this work under the project “PRIN 2006. Molecular Quantum Mechanics: Computational Methods and Analysis of Novel Phenomena”. Support from project CTQ2007-67143-C02/BQU (European FEDER and Spanish MEC) is also acknowledged.

References

(1) Blanc, J.; Bonačić-Koutecký, V.; Broyer, M.; Chevaleyre, J.; Dugourd, Ph.; Koutecký, J.; Scheuch, C.; Wolf, J. P.; Woste, L. *J. Chem. Phys*. **1992**, *96*, 1793−1809.

(2) Dugourd, Ph.; Blanc, J.; Bonačić-Koutecký, V.; Broyer, M.; Chevaleyre, J.; Koutecký, J.; Pittner, J.; Wolf, J. P.; Woste, L. *Phys. Rev. Lett*. **1991**, *67*, 2638−2641.

(3) Plavsic, D.; Koutecký, J.; Pacchioni, G.; Bonačić-Koutecký, V. *J. Phys. Chem*. **1983**, *87*, 1096−1097.

(4) Bonačić-Koutecký, V.; Fantucci, P.; Koutecký, J. *Chem. Phys. Lett*. **1988**, *146*, 518−523.

(5) Fantucci, P.; Bonačić-Koutecký, V.; Koutecký, J. *Z. Phys. D Atoms, Molecules Clusters* **1989**, *12*, 307−314.

(6) Rousseau, R.; Marx, D. *Chem. Eur. J*. **2000**, *6*, 2989−2993.

(7) McAdon, M. H.; Goddard, W. A., III. *J. Phys. Chem*. **1987**, *91*, 2607−2626.

(8) Alikhani, M. E.; Shaik, S. *Theor. Chem. Acc.* **2006**, *116*, 390−397.

(9) de Visser S. P.; Alpert, Y.; Danovich, D.; Shaik, S. *J. Phys. Chem. A* **2000**, *104*, 11223−11231.

(10) Danovich, D.; Wu, W.; Shaik, S. *J. Am Chem Soc.* **1999**, *21*, 3165−3174.

(11) de Visser, S. P.; Danovich, D.; Wu, W.; Shaik, S. *J. Phys. Chem. A* **2002**, *106*, 4961−4969.

(12) Higgins, J.; Callegari, C.; Reho, J.; Stienkeimeier, F.; Ernst, W. E.; Lehmann, K. K.; Gutowski, M.; Scoles, G. *Science* **1996**, *273*, 629−631.

(13) Higgins, J.; Ernst, W. E.; Callegari, C.; Reho, J.; Lehman, K. K.; Scoles, G.; Gutowski, M. *Phys. Rev. Lett.* **1996**, *77*, 4532−4535.

(14) Fioretti, A.; Comparat, D.; Crubellier, A.; Dulieu, D.; Masnou-Seeuws, F.; Pillet, P. *Phys. Rev. Lett.* **1998**, *80*, 4402−4405.

(15) Higgins, J.; Hollebeeck, T.; Reho, J.; Ho, T.-S.; Lehmann, K. K.; Rabitz, H.; Scoles, G.; Gutowski, M. *J. Chem. Phys.* **2000**, *112*, 5751−5761.

(16) Reho, J. H.; Higgins, J.; Nooijen, M.; Lehmann, K. K.; Scoles, G. *J. Chem. Phys.* **2001**, *115*, 10265−10274.

(17) Brühl, F. R.; Miron, R. A.; Ernst, W. E. *J. Chem. Phys.* **2001**, *115*, 10275−10281.

(18) Bodo, E.; Gianturco, F. A.; Yurtsever, E. *J. Low. Temp. Phys.* **2005**, *138*, 259−264.

(19) Bodo, E.; Yurtsever, E.; Yurtsever, M.; Gianturco, F. A. *J. Chem. Phys.* **2006**, *124*, 074320-(1−13).

(20) Tiggensbaumker, J.; Stienkeimer, F. *Phys. Chem. Chem. Phys.* **2007**, 4748−4770.

(21) Hollas, J. M. *High Resolution Spectroscopy*, 2nd ed.; J. Wiley and Sons: Chichester, England, 1998; Table 6.6.

(22) Dunning T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007−1023.

(23) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem Phys.* **1992**, *96*, 6796−6806.

(24) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1994**, *100*, 2975−2988.

(25) Basis sets were obtained from the Extensible Computational Chemistry Environment Basis Set Database, Version 02/25/04, as developed and distributed by the Molecular Science Computing Facility, Environmental and Molecular Sciences Laboratory which is part of the Pacific Northwest Laboratory, P.O. Box 999, Richland, WA 99352, U.S.A., and funded by the U.S. Department of Energy. The Pacific Northwest Laboratory is a multiprogram laboratory operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC06-76RLO 1830. Contact Karen Schuchardt for further information. http://www.emsl.pnl.gov/forms/basisform.html (accessed Jan 14, 2008).

(26) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Koch, E.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243−252.

(27) Bendazzoli, G. L.; Evangelisti, S. *J. Chem. Phys.* **1993**, *98*, 3141−3150.

(28) DALTON, Version 2; DALTON a molecular electronic structure program, Release 2.0. See http://www.kjemi.uio.no/software/dalton/dalton.html (accessed Jan 14, 2008).

(29) Angeli, C.; Cimiraglia, R. private communication.

(30) Angeli, C.; Bendazzoli, G. L.; Borini, S.; Cimiraglia, R.; Emerson, A.; Evangelisti, S.; Maynau, D.; Monari, A.; Rossi, E.; Sanchez-Marin, J.; Szalay, P. G.; Tajti, A. *Int. J. Quantum Chem*. **2007**, *107*, 2082−2091.

(31) Borini, S.; Monari, A.; Rossi, E.; Tajti, A.; Angeli, C.; Bendazzoli, G. L.; Cimiraglia, R.; Emerson, A.; Evangelisti, S.; Maynau, D.; Sanchez-Marin, J.; Szalay, P. G. *J. Chem. Inf. Modell.* **2007**, *47*, 1271−1277.

(32) Rossi, E.; Emerson, A.; Evangelisti, S. *Lect. Notes Comput. Sci.* **2003**, *2658*, 316−323.

(33) Monari, A.; Bendazzoli, G. L.; Evangelisti, S.; Angeli, C.; Ben Amor, N.; Borini, S.; Maynau, D.; Rossi, E. *J. Chem. Theory Comput.* **2007**, *3*, 477−485.

(34) Christiansen, O.; Koch, H.; Jørgensen, P. *J. Chem. Phys.* **1995**, *103*, 7429−7441.

(35) MOLPRO is a package of ab initio programs written by H. J. Werner and P. J. Knowles with contribution from Almlöf, R. D.; Amos, A.; Berning, M. J. O.; Deegan, F.; Eckert, S.

T.; Elbert, C.; Hampel, R.; Lindh, W.; Meyer, A.; Nicklass, K.; Peterson, R.; Pitzer, A. J.; Stone, P. R.; Taylor, M. E.; Mura, P.; Pulay, M.; Schütz, H.; Stoll, Thorsteinsson, T.; Cooper. D. L.

(36) Stoer, J.; Burlisch, R. *Introduction to Numerical Analysis*; Springer-Verlag: New York, Heidelberg, Berlin, 1990.

(37) Maynau, D.; Evangelisti, S.; Guihéry, N.; Malrieu, J. P.; Calzado, C. *J. Chem. Phys*. **2002**, *116*, 10060−10068.

(38) Widmark, P. O.; Malmqvist, P. A.; Roos, B. *Theor. Chim. Acta* **1990**, *77*, 291.

(39) Gatti, C.; Fantucci, P.; Pacchioni, G. *Theor. Chem. Acc*. **1987**, *72*, 433−458.

(40) Liu, B.; McLean, A. D. *J. Chem. Phys*. **1973**, *59*, 4557−4558.

(41) Boys, S. F.; Bernardi, F. *Mol. Phys*. **1970**, *19*, 553−566.

CT7003319

JCTC Journal of Chemical Theory and Computation

Flexible-Boundary Quantum-Mechanical/ Molecular-Mechanical Calculations: Partial Charge Transfer between the Quantum-Mechanical and Molecular-Mechanical Subsystems

Yan Zhang and Hai Lin*

Chemistry Department, University of Colorado Denver, Denver, Colorado 80217-3364

Received November 2, 2007

Abstract: Based on the principle of electronic chemical potential equalization, we propose a flexible-boundary scheme to account for partial charge transfers between the quantum-mechanical (QM) and molecular-mechanical (MM) subsystems in combined QM/MM calculations. The QM subsystem is viewed as an open system with a fluctuating number of electrons and is described by a statistical mixture of ensemble that consists of states of integer number of electrons. The MM subsystem serves as a reservoir that exchanges electrons with the QM subsystem. The electronic chemical potential of the MM subsystem varies whenever charges flow in or out, until equilibrium is established for the electronic chemical potentials between the QM and MM subsystems. Our scheme is demonstrated by calculations of the partial atomic charges for 7 small model systems, each consisting of a singly charged ion and a water molecule, as well as for the Eigen cation, a model system for the solvated structure of hydronium ion in water. Encouraging results are obtained for the partial atomic charges, which are in reasonable agreement with full-QM calculations on those model systems. The averaged mean unsigned deviations between the QM/MM and full-QM calculations are 0.16 e for the partial atomic charges of the entire systems and 0.13 e for the amount of charge transferred between the QM and MM subsystems.

I. Introduction

In combined quantum-mechanical and molecular-mechanical (QM/MM)[1−15] calculations, the entire system (ES) is often partitioned into a small and localized primary system (PS) and its surroundings called secondary system (SS). The PS is treated at the quantum-mechanics (QM) level of theory. The SS, which is modeled at the molecular mechanics (MM) level, interacts with the PS and affects its electronic structure. The PS is also called the QM subsystem, and the SS is also known as the MM subsystem. The QM/MM energy for the entire system (ES) can be formally defined as the sum of the QM energy of the PS, the MM energy of the SS, and the QM/MM interaction energy between them.

$$E(\mathrm{QM/MM;ES}) = E(\mathrm{QM;PS}) + E(\mathrm{MM;SS}) + E(\mathrm{QM/MM;PS|SS}) \quad (1)$$

The inclusion in eq 1 of the interactions between the PS and its surroundings (the SS) is a key issue in the QM/MM methodology.

The interactions between the PS and the SS include bonded interactions, van der Waals interactions, and electrostatic interactions. A given bonded interaction, if presented, is often included at the MM level if it involves at least an SS atom, and the van der Waals interactions are typically evaluated at the MM level. The treatment for electrostatic interactions varies in different QM/MM schemes.[16] The first category is the mechanical-embedding schemes,[16] where the electrostatic interactions between the PS and SS are computed at the MM level, e.g., by Coulomb's law employing atomic charges assigned to both the PS and SS atoms, and the QM

* Corresponding author e-mail: hai.lin@cudenver.edu.

10.1021/ct700296x CCC: $40.75
Published on Web 02/21/2008

calculations for the PS are performed in the gas phase. In the second category of electrostatic-embedding schemes,[16] QM computations for the PS that are carried out with the inclusion of charge distribution of the SS, which is done by including in the QM Hamiltonian the operators that describe the electrostatic interaction between the nuclei and electrons of the PS and the MM partial atomic charges of the SS. The use of the MM partial atomic charges is convenient and popular, but more sophisticated representations of the SS charge density including distributed multipoles and the effective fragment potential[17] have also been developed. Today, most QM/MM implementations are electrostatic-embedding. The third class of embedding schemes is called self-consistent mutual-polarized-embedding schemes[16] or polarized-embedding schemes for short. In the polarized-embedding schemes, the PS and SS will polarize each other until their charge distributions are self-consistent. A number of studies[1,2,16,18−31] have been carried out to develop polarized-embedding QM/MM schemes by combining the commonly used unpolarizable MM potentials (such as AMBER,[32] CHARMM,[33] and OPLS-AA[34−39]) with classical polarization models.[19,40−52] The basic idea is similar to reaction field theory, although the response is now given by a discrete model incorporating the atomic polarizability of individual SS atoms instead of by a continuum. Employing polarization models[19,42−44] based on the principle of electronegativity equalization[53,54] to account for the flexibility of charge redistribution in the SS, we[31] recently developed the polarized-boundary redistributed charge scheme and polarized-boundary redistributed charge and dipole scheme; both schemes permit the mutual polarizations between the PS and SS near the QM/MM boundary.

It is of interest to further develop embedding schemes that permit fractional (or whole) charges flow between the PS and SS. Such treatments, which can be called flexible-boundary embedding schemes, account for both mutual polarization and charge transfer between the PS and SS and are in principle even more realistic than the polarized-boundary embedding schemes. For flexible-boundary embedding calculations, one needs an algorithm that describes the electronic structure of a quantum system with fractional electrons and a prescription that determines how much charge should be transferred between the quantum and classic mechanical subsystems.

Gogonea and Merz (GM)[55,56] have proposed a combined quantum-mechanical/Poisson−Boltzmann equation approach to study the charge transfer between ions and a solvent medium treated as a dielectric continuum. In the GM treatment, the charge being transferred is represented by a surface charge density at the dielectric interface, which modifies the boundary condition for which the Poisson−Boltzmann equation is solved. The ions are described by an effective QM Hamiltonian that resembles Dewar's half-electron method[57,58] but with subtle differences in handling the electron−electron repulsion term. The self-consistent QM calculations are carried out in terms of the density matrix by adding electron density to the LUMO (in the case of charge transferred to ions) or by subtracting electron density from the HOMO (in the case of charge transferred to solvent). The amount of charge being transferred is determined variationally subject to the criterion of the free energy including the environment.

Tavernelli, Vuilleumier, and Sprik (TVS)[59] proposed another scheme that can potentially be adapted to handle fractional charge transfer between the PS and SS in the QM/MM calculations. Their method, which is called the grand-canonical molecular dynamics method, can be traced back to the treatment of fractional particle number of electrons in density function theory by Perdew, Parr, Levy, and Balduz (PPLB).[60] The TVS scheme models the exchange of electrons between a molecule and a reservoir of fixed chemical potential by a modification of the Car−Parrinello[61] method allowing for fluctuating numbers of electrons under constraints of fixed electronic chemical potential. The molecular dynamics simulations involve multiple diabatic potentials energy surfaces where each surface corresponds to a system with a strictly integer number of electrons, e.g., a surface for the reduced state whose charge is 0 and a surface for the oxidized state whose charge is +1 e. Thermochemical properties in a molecular dynamics run were computed by a weighted average of the partition functions for the two oxidation states; in other words, one avoids treating a fractional number of electrons by moving the system on an effective (adiabatic) potential that is a weighted average of diabatic potential surfaces corresponding to integer numbers of electrons. The weights are determined by the chemical potential and the mole fraction of the cations. This provides a more justifiable treatment of electron exchange, but it has been criticized[62] because of the need for a uniform background charge.

Both the GM[55,56] method and the TVS[59] method can in principle be adapted to be used in the flexible-boundary QM/MM methodology. It is conceptually straightforward to replace the Poisson−Boltzmann equation for the solvent medium that is treated as a dielectric continuum in the GM method by the electronegativity equalization models for the SS atoms in the boundary region. That is, the surroundings of the PS are treated explicitly by a discrete model of individual SS atoms in the flexible-boundary QM/MM scheme instead of by a continuum in the GM method. The calculations will provide a single set of molecular orbitals and other quantities (e.g., atomic charges) that are easy to interpret. The drawback is that one needs to modify QM codes so as to implement the half-electron treatments. Moreover, due to its empirical nature, the half-electron treatment is difficult to extend to more advanced QM theories such as coupled-cluster theory.[63] In contrast, a treatment based on the TVS method should not require modification to QM codes, and therefore more advanced QM theories can be used. The artificial uniform background charge is not a requirement in the QM/MM boundary treatment, although it is needed in the TVS method for molecular dynamics simulations of a system with finite net charge employing periodic boundary condition. To make use of the TVS method, one must figure out what the electron reservoir is and how to treat the exchange of electrons between the PS and the reservoir. The disadvantage of this approach is that the picture of the PS fluctuating between two (reduced and oxidized) potential energy surfaces is not so straightforward to grasp.

In the present contribution, we make an attempt to develop the flexible-boundary QM/MM based on the principle of electronic chemical potential equalization. The basic idea is similar to the TVS method, with the difference that the electronic chemical potential of the surroundings (SS) in our treatment varies whenever charges flow in or out. The charge-transfer ceases when equilibrium is established for electronic chemical potentials between the PS and SS, and iterative treatments are required to achieve self-consistence. Our flexible-boundary QM/MM scheme can also be considered as an extension of the classical electronegativity-equalization models, such as the charge equalization (QEq) method proposed by Rappé and Goddard,[44] to the treatment of hybrid quantum-classical systems. In this article, we apply this flexible-boundary treatment to QM/MM calculations on model systems where the QM/MM boundary does not pass through a covalent bond; in particular, we study partial charge transfers between a formally singly charged ion, which is the PS, and a (or several) formally charge-neutral water molecule(s), which is the SS. The cases where the QM/MM boundary does pass through one or more covalent bonds are deferred to future studies. The methodology is described in section II, the computations are carried out in section III, and the results are given in section IV. Discussions are presented in section V, and conclusions are drawn in section VI.

II. Methodology

II.A. Electronegativity Equalization Models. The principle of electronic chemical potential equalization, or electronegativity equalization, has been extensively discussed in literature, and various models have been proposed.[19,42−48,53,54] One of such models is the charge equalization (QEq) method proposed by Rappé and Goddard.[44] In our recent development of the polarized-boundary QM/MM schemes,[31] we employed (with modifications) the QEq model with a shielded Coulomb term (SCT)[44] to account for the charge redistribution within the SS atoms near the QM/MM boundary in response to the electric field generated by the PS. The modified QEq-SCT method is used in the present study of flexible-boundary treatments for the determination of charges for the SS atoms. Below, we give a brief description for our modified QEq-SCT implementation; more details can be found in ref 31 and are not repeated here.

In our polarized-boundary QM/MM implementation,[31] we allow the SS to be separated into two parts. The first part is polarizable, and it normally consists of atoms near the QM/MM boundary (although this is not a requirement). The second part, if presented, normally consists of atoms distance from the PS and is not polarized in the QM/MM calculations. In our study, the first part is called the (polarizable) boundary group, and the second part is called the unpolarized group. The original QEq-SCT method was modified in order to take into account the external electric field generated by the PS and by the unpolarized group of the SS; in the absence of the external electric field, our treatment is identical to the original QEq-SCT scheme. (In ref 31 where the QM/MM boundary passed through covalent bonds, the PS is capped by hydrogen atoms, giving rise to a Capped PS, or CPS.)

The modified atomic potential at atom A of charge Q_A in the (polarizable) boundary group is given by

$$\chi_A(Q_1 \cdots Q_N) = \chi_A^0 + J_{AA}^0 Q_A + U_{A,PS} + \sum_C J_{AC} Q_C + \sum_{A \neq B} J_{AB} Q_B \quad (2)$$

where χ_A^0 is the electronegativity of this isolated atom, J_{AA}^0 is the Coulomb repulsion integral of two electrons residing at the same isolated atom, the electric field at the position of atom A due to the PS is $U_{A,PS}$, J_{AC} is the Coulomb interaction integral between unit charges on centers A and C, C denotes an unpolarized SS atom, the charge at the center C is Q_C, J_{AB} is the Coulomb interaction integral between unit charges on centers A and B, and B is another atom of charge Q_B in the polarizable boundary group. The principle of electronegativity equalization leads to

$$\bar{\chi} = \chi_1 = \cdots = \chi_N \quad (3)$$

where $\bar{\chi}$ is the common value. The principle of charge conservation imposes a constraint on the total charge

$$Q_{tot} = \sum_{i=1}^{N} Q_i \quad (4)$$

The common value of the atomic chemical potential and the atomic charges in the boundary group are computed by solving eqs 3 and 4.

Here, we re-emphasized that our treatments only change the MM background charges in the embedded-QM calculations and do not affect the pure MM calculations, for which the original MM charges are used.[31] Our treatments are, in a sense, optimizations of selected charge parameters in the effective QM Hamiltonian for the operators that describe the electrostatic interaction between the PS and the SS. On the other hand, the MM charges are part of an MM force field, which is parametrized to be used as a whole; the charge parameters are cross correlated with parameters for the other (bonded and van der Waals) interactions. Although the effective SS charge parameters in the embedded-QM calculations and the MM charge parameters in an MM force field share some similarity, and we indeed often use MM charge parameters as effective SS charge parameters in embedded-QM calculations, these two sets of charges are different in nature. For this reason, we feel that it is not appropriate to use charges optimized from the polarized-QM calculations for the MM calculations.

II.B. Partial Charge Transfer between the PS and SS. Following the PPLB argument,[60] we consider the PS as an open system with a fluctuating number of electrons, which is described by a statistical mixture of ensemble that consists of states of integer number of electrons. The SS serves as a reservoir that exchanges electrons with the PS. Unlike the reservoir of fixed chemical potential in the TVS[59] model, the electronic chemical potential of the SS in the flexible-boundary QM/MM scheme varies whenever charges flow in or out. Based on the principle of electronic chemical potential equalization, the charge transfer will continue until the electronic chemical potentials in the PS and SS become equal. Iterative treatments are therefore required to achieve self-consistence.

In the simplest situation, the PS consists of only two states, i.e., a reduced state (X) and an oxidized state (X^+). An example is the neutral sodium atom Na and the sodium cation Na^+. The oxidized state is of charge $q(X^+)$, and its molar fraction is x_+. The reduced state is of charge q(X), and its molar fraction is $x = 1 - x_+$. Let us denote the charges on the ES, PS, and SS as q(ES), q(PS), and q(SS), respectively. Apparently, one has

$$q(\mathrm{PS}) = [x_+\, q(\mathrm{X}^+) + (1 - x_+)\, q(\mathrm{X})] \tag{5}$$

$$q(\mathrm{ES}) = q(\mathrm{PS}) + q(\mathrm{SS}) \tag{6}$$

As the same in the TVS treatment,[59] we assume the existence of equilibrium of PS ionization $X^+ + e^- \leftrightarrow X$. This leads to

$$\mu(\mathrm{X}^+) + \mu(\mathrm{e}^-) = \mu(\mathrm{X}) \tag{7}$$

Here, $\mu(X^+)$ is the chemical potential of X^+, $\mu(e^-)$ is the chemical potential of the free electrons, and μ(X) is the chemical potentials of X. Following the same procedure outlined in ref 59, we separate the chemical potential of a component species A into the contribution of an energetic term μ^0(A) and an ideal gas contribution, and eq 7 can be rewritten as

$$\mu(\mathrm{e}^-) = \mu^0(\mathrm{X}) - \mu^0(\mathrm{X}^+) + k_\mathrm{B}T \ln\left(\frac{x}{x_+}\right) \tag{8}$$

and

$$\mu^0(\mathrm{X}) - \mu^0(\mathrm{X}^+) = -I(X) = -[E(\mathrm{X}^+) - E(\mathrm{X})] \tag{9}$$

where I(X) is the ionization potential of the reduced state X. In our flexible-boundary treatment, I(X) is computed for the PS as the gas-phase energy difference between its reduced state and its oxidized state. Here, we have made an approximation by assuming that I(X) obtained for the PS in the gas-phase is equal to I(X) for the PS in the presence of the SS. Now assuming that the free electrons are in equilibrium of the electrons in the reservoir (SS),

$$\mu(\mathrm{e}^-) = \mu(\mathrm{SS}) = \mu \tag{10}$$

we reach an equation similar to eq 3 in ref 59:

$$\mu(\mathrm{SS}) = -I(\mathrm{X}) + k_\mathrm{B}T \ln\left(\frac{1 - x_+}{x_+}\right) \tag{11}$$

The electronic chemical potential of the SS, μ(SS), is related to the chemical potential for charge transfer, or electronegativity χ, of the SS by

$$\mu(\mathrm{SS}) = -\chi \tag{12}$$

The electronegativity of the SS, χ, can be computed by employing the (modified) QEq-SCT scheme described in section II.A. Equation 11 is the central equation in our flexible-boundary treatment, which must be satisfied when the equilibrium is established for electron exchange between the PS and SS.

In some cases, it is more convenient to denote the reduced state as X^- and the oxidized state as X, such as the Cl^- anion and the Cl atom. Using such notations, one can rewrite eq 11 as follows

$$\mu(\mathrm{SS}) = -A(\mathrm{X}) + k_\mathrm{B}T \ln\left(\frac{x_-}{1 - x_-}\right) \tag{13}$$

where x_- is the molar fraction of X^-, and $A(X) = [E(X) - E(X^-)]$ is the electron affinity of X. For the sake of brevity, unless otherwise indicated, we will focus our discussion on eq 11 with the notations of X for the reduced state and X^+ for the oxidized state.

Two issues need to be addressed here. First, the temperature T in eq 11 is the temperature for electrons, which is not necessarily the same as the temperature that describes nuclear motions (vibration, rotation, and translation). It is probably better to view T in eq 11 as an empirical parameter adjustable to achieve the best agreement with reference data; this is especially true when considering the empirical nature of the calculations of the electronegativity by the QEq-SCT scheme. The second issue is the different zeroes of electronic chemical potentials between QM calculations and QEq-SCT calculations, which require calibration before comparisons can be made. Note that the logarithm term in eq 11 disappears when $x_+ = 0.5$, and the calibration can be done as follows: First, one computes the electronegativity χ_cali by the QEq-SCT method for the PS with a charge of $q(PS) = 0.5\, q(X^+) + 0.5\, q(X)$; for example, because $q(Na) = 0$ and $q(Na^+) = 1$ e, χ_cali will be calculated for the PS of a charge of 0.5 e, or $Na^{+0.5}$. An energy term E_cali is determined by comparing χ_cali and the ionization energy I(X) for the PS:

$$-I(\mathrm{X}) + E_\mathrm{cali} = -\chi_\mathrm{cali} \tag{14}$$

The energy term E_cali is then added to eq 11, yielding

$$\mu(\mathrm{SS}) = -I(\mathrm{X}) + k_\mathrm{B}T \ln\left(\frac{1 - x_+}{x_+}\right) + E_\mathrm{cali} \tag{15}$$

The flexible-boundary treatment needs an iterative procedure for self-consistent calculations. For example, the procedure can start with a guessed molar fraction x_+ for the oxidized state X^+ and enters the cycle of polarized-boundary calculations, where the electrostatic potential at an SS atom A due to the PS, $U_\mathrm{A,PS}$ in eq 2, is computed as an ensemble average of the electrostatic potentials due to the oxidized and reduced states of the PS:

$$U_\mathrm{A,PS} = x_+ U_\mathrm{A,PS}(\mathrm{X}^+) + (1 - x_+) U_\mathrm{A,PS}(\mathrm{X}) \tag{16}$$

After the polarization-boundary calculations converge, one computes the electronic chemical potential of the SS, μ(SS). With the newly obtained μ(SS), the molar fraction x_+ is updated according to eq 11, and a new cycle of polarization-boundary calculations are performed. The loop continues until self-consistency is achieved, i.e., the variations in the amount of charge transferred between the PS and SS are smaller than preset thresholds.

The iterative procedure described above is easy to understand and straightforward to implement but is not

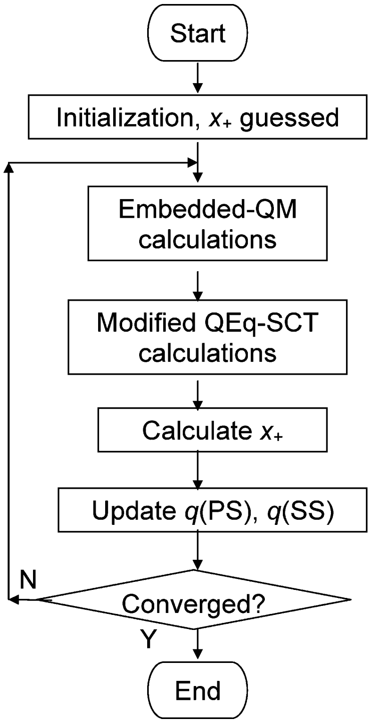


Figure 1. Flowchart for flexible-boundary calculations.

efficient. We have therefore adopted a more efficient procedure where the molar fraction x_+ is updated at every iteration within the cycle of polarization-boundary calculations. (A flow chart is shown in Figure 1.) The self-consistence requires both the convergence in the amount of charge transferred between the PS and SS and the convergence in the partial atomic charges at the SS atoms. Implemented as such, only one cycle of polarized-boundary calculations is needed, and the computational cost for the flexible-boundary calculations is approximately twice the cost for the polarized-boundary calculations (since one has to do embedded-QM calculations for both oxidation states in the flexible-boundary calculations).

III. Computation

The method is demonstrated by studying the partial atomic charges for model systems. First, we study 7 small model systems that each consists of a singly charged ion and a water molecule; these model systems are denoted [A...B], where A = Li^+, Na^+, K^+, NH_4^+, F^-, Cl^-, and HS^- is the PS, and B = H_2O is the SS. Second, we study a larger model system, the Eigen cation $H_9O_4^+$, which is a proposed solvation structure for the hydronium ion H_3O^+ in water; in the Eigen cation, the central H_3O^+ moiety is the PS, and the three hydrogen-bonded neighbor H_2O molecules compose the SS. In total, we have therefore included 8 model systems in the test calculations. The formal charges for the PS are +1 e in [$Li^+\cdots H_2O$], [$Na^+\cdots H_2O$], [$K^+\cdots H_2O$], [$NH_4^+\cdots H_2O$], and the Eigen cation, and are −1 e in [$F^-\cdots H_2O$], [$Cl^-\cdots H_2O$], and [$HS^-\cdots H_2O$], respectively. Note that the PS in [$NH_4^+\cdots H_2O$], the Eigen cation, and [$HS^-\cdots H_2O$] are polyatomic ions.

We compute and compare three types of partial atomic charges: (1) the charges determined by applying the QEq-SCT model to the entire (model) systems, which are denoted QEq-SCT-ES charges, (2) two sets of full-QM calculated charges, which include the charges obtained by fitting to the electrostatic potential (ESP) using the Merz−Singh−Kollman[64,65] scheme, and the charges given by Löwdin[66] population analysis, and (3) two sets of QM/MM charges obtained by the flexible-boundary calculations. The two sets of QM/MM partial atomic charges are identical to each other except for the partial atomic charges for the polyatomic PS. In the first set of QM/MM charges (denoted QM/MM-1), the partial atomic charges for the polyatomic PS are ensemble-averaged ESP charges over the reduced and oxidized states, while in the second set of QM/MM charges (denoted QM/MM-2), the partial atomic charges for the polyatomic PS are ensemble-averaged Löwdin charges. Both QM/MM-1 and QM/MM-2 possess the same charges for monatomic PS, which are computed based on the total charges of the SS and according to the principle of charge conservation (eq 6). In both QM/MM-1 and QM/MM-2, the partial atomic charges for the SS are obtained by the modified QEq-SCT scheme.

It is well-known that partial atomic charges are not experimentally measurable quantities, and there is ambiguity in which set of charges is more “correct” than the other. Indeed, we have found in our calculations that, at the employed level of theory, the partial atomic charges predicted by the QEq-SCT-ES and full-QM calculations sometimes disagree with one’s intuition. For example, the Na center in [$Na^+\cdots H_2O$] is assigned a partial positive charge large than +1 e. However, the partial atomic charge is a very useful concept that provides important information about the charge distributions within a model system, and we feel that it is instructive to make the comparisons between the charges obtained by reference (QEq-SCT-ES and full-QM) calculations and by the flexible-boundary QM/MM calculations. Since this work is the first step toward the full development of the flexible-boundary QM/MM embedding scheme, we aim mainly at achieving a qualitative (or semiquantitative) agreement between the full-QM and QM/MM charges; the methodology is to be refined in the future in order to accomplish more accurate quantitative calculations.

For a given model system and level of theory, the total amount of partial charge transferred between the PS and the SS is computed as

$$q_{\text{trans}} = q_{\text{formal}}(\text{PS}) - q(\text{PS}) \qquad (17)$$

i.e., as the difference between the formal charge of the PS and the actually calculated charge of the PS. Note that QM/MM-1 and QM/MM-2 are identical in q_{trans}.

The QM level of theory is the B3LYP[67−69] density functional model with the 6−31++G(d,p) basis set.[70−74] The MM force field is OPLS-AA.[34−39] Convergence thresholds for the flexible-boundary treatment are as follows: the maximum change in the SS partial atomic charge less than 0.005 e, root-mean-square variation in the SS partial atomic charge less than 0.002 e, and the amount of charge flowing between PS and SS less than 0.005 e. For the present study, the *Gaussian03*[75] program is employed for QM calculations, TINKER[76] is used for MM calculations, and the

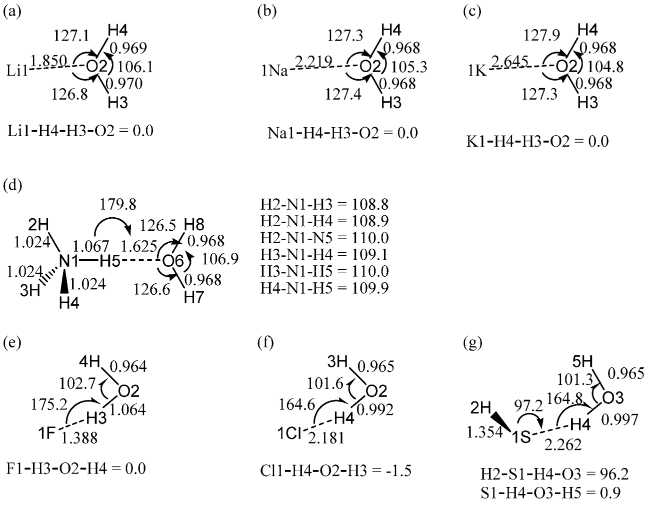


Figure 2. Critical geometric data for small model systems [A···B], where A = Li^+, Na^+, K^+, NH_4^+, F^-, Cl^-, and HS^- is the PS in (a)−(g), respectively, and B = H_2O is the SS. Distances are in Å, and angles and dihedrals are in deg. The geometries are optimized at the B3LYP/6-31++G(d,p) level of theory without symmetry constraints.

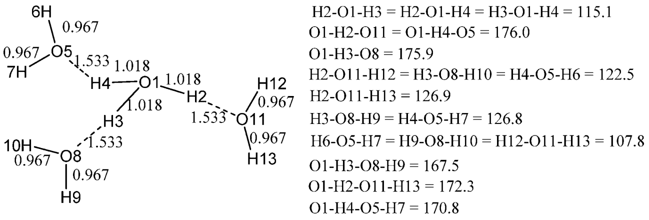


Figure 3. Critical geometric data for the Eigen cation $H_9O_4^+$, where the central H_3O^+ moiety is the PS and the three hydrogen-bonding neighbor H_2O molecules are the SS. Distances are in Å, and angles and dihedrals are in deg. The geometry is optimized at the B3LYP/6-31++G(d,p) level of theory without symmetry constraints.

QMMM[77] program is utilized for QM/MM calculations. The geometries are the full-QM geometries optimized by using the *Gaussian03* program with the (5D, 7F) option for basis sets. No symmetric constraints are imposed in geometry optimizations. The QEq-SCT parameters are taken from ref 44.

IV. Results

The optimized geometries are illustrated in Figure 2 for the small model systems [A···B] and in Figure 3 for the Eigen cation. Table 1 tabulates the partial atomic charges for the small model systems [A···B], while the charges for the Eigen cation are collected in Table 2. No full-QM ESP charge is available for [K^+···H_2O] due to the lack of parameter (the Merz−Kollman atomic radius) for potassium in the ESP charge calculations. Table 3 gives the mean unsigned deviations (MUD) for the partial atomic charges between the QM/MM calculations and the full-QM calculations; for a given temperature and QM/MM charge set, the MUD is averaged over all 8 model systems, except that [K^+···H_2O] is excluded from the calculations for QM/MM-1. The amounts of the charges transferred between the PS and the SS are listed in Table 4, for which the MUD and mean signed deviations (MSD) between the QM/MM calculations and the full-QM calculations are presented in Table 5. Note that in the calculations of the MUD and MSE, we always compare the QM/MM-1 charges with the full-QM ESP charges while compare the QM/MM-2 charges with the full-QM Löwdin charges. The electronic chemical potential $\mu(e^-)$ at T = 30 000 K is shown in Figure 4 for a statistical (Na, Na^+) mixture of ensemble of charge q, along with the molar fractions x of Na and x_+ of Na^+. Also plotted in Figure 4 is the electronegativity χ calculated by the QEq-SCT method for Na^{+q}, where $0 \leq q \leq 1$. In Figure 5, we display $\mu(e^-)$ for the [Na^+···H_2O] model system at three temperatures 10 000 K, 30 000 K, and 50 000 K, all indicated by dashed lines, as well as μ(SS), indicated by a solid line. The convergence of the QM/MM calculated charges is demonstrated in Figure 6 for [Na^+···H_2O] at T = 30 000 K.

V. Discussion

V.A. Case Study for [Na^+···H_2O]. The small model system [Na^+···H_2O] is one of the simplest systems in the test calculations. In this section, we choose it to illustrate the concepts and to examine the difficulties that we have come across in the calculations.

The optimized geometry for [Na^+···H_2O] is planar, as shown in Figure 2(b). As listed in Table 1, the QEq-SCT-ES calculations (for the entire system) imply that some positive charge is transferred from H_2O to Na^+, such that the Na1 center carries a charge of +1.16 e. This sounds somewhat unusual, as the result contradicts one's intuition. We have found that such unusual charges appear when the Na^+ moiety is close to the H_2O moiety. As shown in Table S6 in the Supporting Information, the charge at the Na1 center increases monotonically as the Na^+ and H_2O moieties approach each other and exceeds +1 e when the Na−O distance reduces to 3.5 Å or shorter. The unusual results have also been obtained for [Li^+···H_2O] and [K^+···H_2O] at their optimized geometries, as indicated in Table 1. The reason could be that we have used the simplified SCT treatment to calculate the Coulomb integration in order to reduce computational costs; this SCT treatment was not recommended in the original QEq paper[29] but has been shown to yield quite reasonable charges for many systems.[31,44] On the other hand, we point out that even the full-QM charges are not free of this kind of artifact, either—the full-QM ESP charge at the Na1 center is also larger than +1 e, although to a much smaller extent (by 0.003 e). The full-QM Löwdin charges seem quite reasonable for this system, where +0.84 e is assigned to the Na1 center; but we have noticed that the full-QM Löwdin charge at the K1 center in [K^+···H_2O] is +1.018 e (Table 1), again unusually larger than +1 e. For the H_2O moiety, the partial atomic charges obtained by the QEq-SCT-ES, full-QM ESP, and full-QM Löwdin calculations agree qualitatively with each other.

In the flexible-boundary QM/MM treatment for the [Na^+···H_2O] model system, we consider Na as the reduced state and Na^+ as the oxidized state for the PS, since the Na^{2+} state is likely too high in energy to be involved. (The same consideration has been applied to the other systems whose PS has a formal charge of +1 e.) Therefore, the PS is

Table 1. Partial Atomic Charges for 7 Small Model Systems [A···B][a]

		full-QM		QM/MM		
	QEq-SCT-ES	ESP	Löwdin	$T = 10\,000$	$T = 30\,000$	$T = 50\,000$
$Li^+ \cdots H_2O$						
Li1	1.212	0.896	0.712	0.999	0.918	0.821
O2	−0.759	−0.706	−0.265	−0.782	−0.754	−0.718
H3	0.273	0.405	0.276	0.405	0.430	0.460
H4	0.274	0.405	0.276	0.379	0.405	0.437
$Na^+ \cdots H_2O$						
Na1	1.155	1.003	0.838	0.999	0.920	0.824
O2	−0.693	−1.053	−0.358	−0.765	−0.738	−0.703
H3	0.269	0.525	0.260	0.395	0.420	0.450
H4	0.269	0.525	0.260	0.371	0.397	0.429
$K^+ \cdots H_2O$						
K1	1.237	n/a	1.018	0.999	0.926	0.832
O2	−0.687	n/a	−0.513	−0.716	−0.691	−0.660
H3	0.224	n/a	0.248	0.367	0.391	0.421
H4	0.225	n/a	0.248	0.350	0.375	0.407
$NH_4{}^+ \cdots H_2O$[b]						
N1	−0.581	−0.963	−0.068	−0.722/−0.049	−0.255/−0.068	0.058/−0.077
H2	0.328	0.445	0.239	0.402/0.234	0.259/0.208	0.180/0.203
H3	0.328	0.443	0.239	0.402/0.234	0.260/0.209	0.180/0.204
H4	0.328	0.444	0.239	0.403/0.235	0.262/0.210	0.183/0.205
H5	0.343	0.663	0.252	0.462/0.292	0.244/0.211	0.082/0.150
O6	−0.551	−1.134	−0.416	−0.666	−0.603	−0.571
H7	0.402	0.552	0.257	0.358	0.416	0.445
H8	0.402	0.550	0.257	0.362	0.417	0.441
$F^- \cdots H_2O$						
F1	−0.677	−0.890	−0.812	−0.999	−0.917	−0.803
O2	−0.657	−1.038	−0.634	−0.639	−0.637	−0.634
H3	0.237	0.582	0.262	0.518	0.445	0.345
H4	0.097	0.346	0.184	0.121	0.109	0.092
$Cl^- \cdots H_2O$						
Cl1	−0.706	−0.873	−0.865	−0.994	−0.842	−0.729
O2	−0.641	−0.724	−0.541	−0.519	−0.523	−0.523
H3	0.137	0.337	0.201	0.239	0.206	0.184
H4	0.209	0.260	0.205	0.274	0.158	0.068
$HS^- \cdots H_2O$[b]						
S1	−0.628	−1.045	−0.835	−1.079/−0.934	−0.889/−0.752	−0.809/−0.675
H2	0.007	0.099	0.006	0.148/0.002	0.145/0.007	0.143/0.009
O3	−0.653	−0.946	−0.555	−0.634	−0.656	−0.665
H4	0.164	0.520	0.191	0.390	0.276	0.228
H5	0.109	0.372	0.194	0.176	0.125	0.103

[a] A is the PS, and B is the SS. Geometries and atom labels are given in Figure 2. The QEq-SCT-ES charges are obtained by doing QEq-SCT calculations for the entire model systems. No full-QM ESP charges are available for [$K^+ \cdots H_2O$] due to the lack of parameter (the Merz−Kollman atomic radius) for potassium in the ESP charge calculations. QM/MM charges are obtained by the flexible-boundary calculations, where the atomic charges for polyatomic PS are ensemble-averaged ESP charges in QM/MM-1 and ensemble-averaged Löwdin charges in QM/MM-2. For monatomic PS, QM/MM-1 and QM/MM-2 charges are identical. The QM/MM charges for the SS are determined by employing the modified QEq-SCT procedure. Charges are in e, and temperatures are in K. [b] For the PS, QM/MM charges are given as (QM/MM-1)/(QM/MM-2).

described by a pair of oxidation states (Na, Na^+). Our first question is as follows: for such a statistical (Na, Na^+) mixture of ensemble, how will the electronic chemical potential $\mu(e^-)$ change if the charge q varies from 0 to +1 e. As can be seen in Figure 4, the $\mu(e^-)$ curve computed at $T = 30\,000$ K changes its value sharply when q approaches 0 or +1 e. Such characteristics have already been demonstrated in previous works,[59,60] which approaches the well-known "staircase" shape at $T = 0$ with discontinuity at integer charges. As expected, the molar fractions for Na and Na^+ are linear functions of the charge of the mixture q. Next, we ask how $\mu(e^-)$ differs from χ, the electronegativity calculated by the QEq-SCT method for Na^{+q}, where $0 \le q \le 1$. The $-\chi$ curve, which is also plotted in Figure 3, turns out to be a linear function of q. Besides the different shapes, $-\chi$ and $\mu(e^-)$ differ in values at $x^+ = 0.5$, where Na and Na^+ are equally likely; the difference is about 0.01 au: $-\chi = -0.1889$ au, and $\mu(e^-) = -0.1987$ au. As pointed out in

Table 2. Partial Atomic Charges for the Eigen Cation[a]

		Full-QM		QM/MM		
	QEq-SCT-ES	ESP	Löwdin	$T = 10\,000$	$T = 30\,000$	$T = 50\,000$
O1	−0.643	−0.857	−0.188	−0.651/−0.098	−0.542/−0.123	−0.514/−0.129
H2	0.394	0.609	0.268	0.515/0.328	0.412/0.266	0.386/0.250
H3	0.394	0.594	0.268	0.513/0.326	0.409/0.263	0.382/0.247
H4	0.394	0.608	0.268	0.495/0.315	0.363/0.235	0.330/0.216
O5	−0.575	−1.040	−0.382	−0.643	−0.603	−0.593
H6	0.366	0.529	0.256	0.326	0.353	0.360
H7	0.362	0.525	0.254	0.389	0.403	0.406
O8	−0.575	−1.029	−0.382	−0.680	−0.642	−0.632
H9	0.362	0.521	0.254	0.390	0.403	0.406
H10	0.366	0.527	0.256	0.319	0.342	0.348
O11	−0.575	−1.039	−0.382	−0.697	−0.658	−0.648
H12	0.366	0.529	0.256	0.331	0.354	0.359
H13	0.362	0.523	0.254	0.393	0.406	0.409

[a] Geometry and atom labels are given in Figure 3. The central H_3O^+ moiety is the PS, and the three hydrogen-bonding neighbor H_2O molecules are the SS. QEq-SCT-ES charges are obtained by doing QEq-SCT calculations for the entire system. The QM/MM charges are obtained by the flexible-boundary calculations, where the atomic charges for the PS are ensemble-averaged ESP charges in QM/MM-1 and ensemble-averaged Löwdin charges in QM/MM-2. For the PS, QM/MM charges are given as (QM/MM-1)/(QM/MM-2). The QM/MM charges for the SS are determined by employing the modified QEq-SCT procedure. Charges are in e, and temperatures are in K.

Table 3. Mean Unsigned Deviations for Partial Atomic Charges between QM/MM Calculations and Full-QM Calculations [a]

	$T = 10\,000$	$T = 30\,000$	$T = 50\,000$
QM/MM-1[b]	0.161	0.202	0.238
QM/MM-2[c]	0.126	0.113	0.117
average[d]	0.144	0.158	0.178

[a] Average over the 7 small model systems [A···B] and the Eigen cation, except that $[K^+\cdots H_2O]$ is excluded from the calculations for QM/MM-1. Charges are in e, and temperatures are in K. [b] QM/MM-1 charges versus full-QM ESP charges. [c] QM/MM-2 charges versus full-QM Löwdin charges. [d] Average over QM/MM-1 and QM/MM-2.

section II.B, a calibration step must be taken to account for this difference (different zeros). In the rest of this paper, we only refer to the calibrated $\mu(e^-)$, unless otherwise indicated.

The central ideal of the flexible-boundary treatment is electronic chemical potential equalization, i.e., the PS and the SS should have the same electronic chemical potentials when charge-transfer ceases. This is illustrated in Figure 5 by the crossing between the $\mu(e^-)$ and $\mu(SS)$ curves, both plotted as functions of $q(PS)$. Note that $\mu(SS)$ is determined in the polarized-boundary QM/MM calculations, where the PS and SS polarize each other until self-consistence. Inspection of the graph suggests that the $\mu(SS)$ and $\mu(e^-)$ curves cross at $q(PS) = +1.00$ e, $+0.91$ e, and $+0.82$ e at $T =$ 10 000 K, 30 000 K, and 50 000 K, respectively. It is conceivable that by varying the temperature T, one can obtain $[Na^{+q}\cdots H_2O^{+(1-q)}]$ with the amount of transferred charge $(1 - q)$ in best agreement with reference data, although we have not made such an effort to optimize the temperature parameter in the present study.

We find that the flexible-boundary QM/MM calculations employing the iterative procedure described in section II.B converge within 4 iterations. For example, at $T = 30\,000$ K, the QM/MM charges converge in 4 iterations (Figure 6). Going from iteration 3 to iteration 4, the magnitude of variation in charge is 0.001 e at the Na1 center and less than 0.001 e at the O2, H3, and H4 centers. Such convergence behavior is typical in our calculations for all the model systems.

V.B. Case Study for $[HS^-\cdots H_2O]$. Unlike Na^+, HS^- is a polyatomic anion. The optimized geometry for $[HS^-\cdots H_2O]$ shown in Figure 2(g) reveals that the S1 center is closer to one hydrogen of the water than to the other; in particular, the S1−H4 distance, which is shorter, is 2.262 Å. As listed in Table 1, all three reference calculations assign significant negative charges (−0.63 e ∼ −1.05 e) at the S1 center and small positive charges (+0.01 e ∼ + 0.10 e) at the H2 center for the PS. The charges at the H4 and H5 centers are the same (0.19 e) in the full-QM Löwdin calculations but differ from each other in the QEq-SCT-ES calculations (by 0.05 e) and in the full-QM ESP calculations (by 0.15 e). Both the QEq-SCT-ES and the full-QM calculations suggest that negative charge is transferred from the HS^- moiety to the H_2O moiety.

For the flexible-boundary QM/MM calculations on this system, we consider the PS as a statistical (HS, HS^-) mixture of ensemble, as we feel that the HS^{2-} state is of rather high energy and is unlikely to contribute. (The same consideration has been applied to the other systems where the PS has −1 e formal charge.) The QM/MM partial atomic charges depend on the employed temperature parameter, but overall they agree reasonably well with the reference calculations. For the PS, both QM/MM-1 and QM/MM-2 assign significant negative charges (−0.68 e ∼ −1.08 e) at the S1 center and small positive charge (<+0.15 e) at the H2 center. Interestingly, in both QM/MM-1 and QM/MM-2, the charge at the H2 center varies negligibly (less than 0.01 e) between $T =$ 10 000 K and $T = 50\,000$ K. Turned to the SS, the H4 center, which is the closest to the PS, shows large variations from +0.39 e at $T = 10\,000$ K to +0.23 e at $T = 50\,000$ K. Smaller variations are observed for the H5 center, which is two bonds further away from the PS: the charge decreases from +0.18 e to +0.10 e when T increases from 10 000 K to 50 000 K. The QM/MM charges for the PS and SS suggest

Table 4. Partial Charges Transferred from the PS to the SS[a]

		full−QM		QM/MM		
	QEq-SCT-ES	ESP	Löwdin	*T* = 10 000	*T* = 30 000	*T* = 50 000
$Li^+\cdots H_2O$	−0.212	0.104	0.288	0.001	0.082	0.179
$Na^+\cdots H_2O$	−0.155	−0.003	0.162	0.001	0.080	0.176
$K^+\cdots H_2O$	−0.237	n/a	−0.008	0.001	0.074	0.168
$NH_4^+\cdots H_2O$	0.254	−0.032	0.099	0.053	0.230	0.317
$F^-\cdots H_2O$	−0.323	−0.110	−0.188	−0.001	−0.083	−0.197
$Cl^-\cdots H_2O$	−0.294	−0.127	−0.135	−0.006	−0.158	−0.271
$HS^-\cdots H_2O$	−0.379	−0.054	−0.171	−0.069	−0.256	−0.334
Eigen cation	0.461	0.046	0.384	0.128	0.358	0.416

[a] For the first 7 small model systems [A···B], A is the PS, and B is the SS. For the Eigen cation, the central H_3O^+ moiety is the PS, and the three hydrogen-bonding neighbor H_2O molecules are the SS. The amount of transferred charge is computed as the difference between the formal charge and the actually calculated charge of the PS. Charges are in e, and temperatures are in K.

Table 5. Mean Unsigned Deviations (MUD) and Mean Signed Deviations (MSD) for Partial Charges Transferred from the PS to the SS[a]

	MUD			MSD		
	T = 10 000	*T* = 30 000	*T* = 50 000	*T* = 10 000	*T* = 30 000	*T* = 50 000
QM/MM-1[b]	0.074	0.134	0.212	0.040	0.061	0.066
QM/MM-2[c]	0.167	0.094	0.097	−0.047	−0.026	−0.022
average[d]	0.121	0.114	0.155	−0.004	0.018	0.012

[a] Signed deviations between QM/MM calculations and reference data are determined by q_{trans}(QM/MM) − q_{trans}(reference), where the reference is full-QM ESP and full-QM Löwdin for QM/MM-1 and QM/MM-2, respectively. Mean deviations are averaged over the 7 small model systems [A···B] and the Eigen cation, except that $[K^+\cdots H_2O]$ is excluded from the calculations for QM/MM-1. Charges are in e, and temperatures are in K. [b] QM/MM-1 charges versus full-QM ESP charges. [c] QM/MM-2 charges versus full-QM Löwdin charges. [d] Average over QM/MM-1 and QM/MM-2.

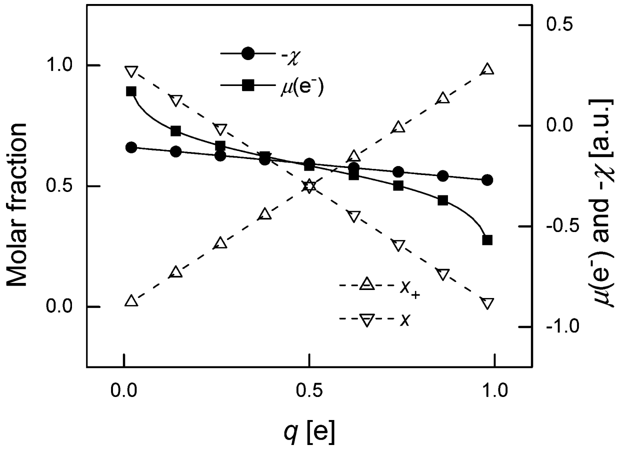


Figure 4. Electronic chemical potential $\mu(e^-)$ at $T = 30\,000$ K, molar fractions x for Na, and molar fraction x_+ for Na^+, all computed for a statistical (Na, Na^+) mixture of ensemble, and electronegativity χ calculated by the QEq-SCT method for Na^{+q}, where $0 \le q \le 1$.

that some negative charge is transferred from HS^- to H_2O, in line with the QEq-SCT-ES and full-QM calculations. The amount of the transferred charge by QM/MM calculations depends on the temperature parameter, which is quite small (−0.07 e) at $T = 10\,000$ K, but increases rapidly to −0.33 e at $T = 50\,000$ K.

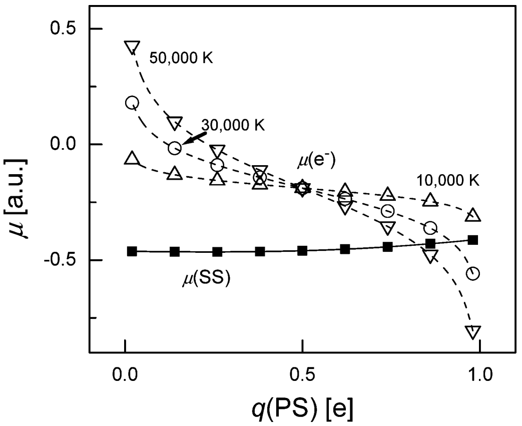


Figure 5. Electronic chemical potentials $\mu(e^-)$ at three temperatures of 10 000 K, 30 000 K, and 50 000 K, and μ(SS) determined by the modified QEq-SCT method in the QM/MM calculations, all computed for the $[Na^+\cdots H_2O]$ model system and expressed as functions of the charge of the PS.

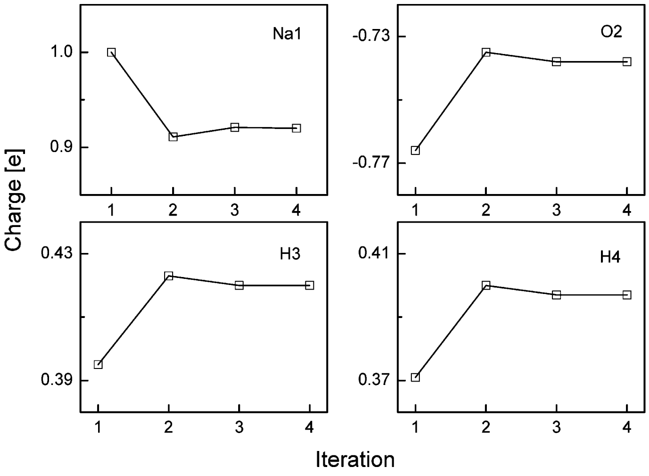


Figure 6. Convergence of the QM/MM calculated charges for $[Na^+\cdots H_2O]$ at $T = 30\,000$ K.

V.C. Overall Assessment. In general, the partial atomic charges are computed quite reasonably by the flexible-boundary QM/MM scheme. The results depend on the employed temperature parameter, but for all the three temperatures (10 000 K, 30 000 K, and 50 000 K) we have tested, the results are qualitatively similar. Furthermore, the QM/MM charges agree reasonably with the reference data.

Not surprisingly, for polyatomic PS, the QM/MM-1 charges resemble the full-QM ESP charges, while the QM/MM-2 charges look similar to the full-QM Löwdin charges; this trend is especially evident in the case of the Eigen cation. As to the partial atomic charges for the SS, the QM/MM calculations and QEq-SCT-ES calculations normally produce quite similar results. As far as the charge transferred between the PS and SS is concerned, the QM/MM calculations predict the direction of charge flow in agreement with our chemical intuitions: for the systems where the PS has a formal charge of +1 e (−1 e), QM/MM calculations predicted that partial positive (negative) charges are transferred to the SS; we however must admit that such predictions are consistent with our choice of the reduced and oxidized states for the PS.

It is not easy to give definite quantitative assessment on the overall performance of the flexible-boundary QM/MM calculations, because the three sets of reference data (QEq-SCT-ES charges, full-QM ESP charges, and full-QM Löwdin charges) vary considerably. Nevertheless, we attempt to address this issue by examining the MUD for the partial atomic charges as well as the MUD for the amount of charge transferred between the PS and SS. Table 3 shows that, for the partial atomic charges, the MUD varies between 0.11 and 0.24 e for the three tested temperatures. Averaging the MUD first over QM/MM-1 and QM/MM-2 and then over the three tested temperatures, we obtain an averaged MUD of 0.16 e, which can be used as a rough indication of the accuracy for the QM/MM-calculated partial atomic charges. In terms of the amount of charge transferred between the PS and SS, as displayed in Table 5, the MUD is in the range of 0.07−0.21 e, and the averaged MUD is 0.13 e. Overall, those deviations are rather moderate, and they imply that our flexible-boundary treatment is able to semiquantitatively describe the charge-transfer cross the QM/MM boundary.

V.D. Involving More Oxidation States. The two-state (reduced and oxidized states) treatment outlined in section II.B can in principle be extended to involve three or more states. In the case of three states, one can assume for the PS the equilibrium between e^-, X, X^+, and X^-,[59] which leads to (a more detailed description is given in the Supporting Information)

$$q(\mathrm{PS}) = (q(\mathrm{X}^+)e^{-\mu+I/kT} + q(\mathrm{X}^-)e^{\mu+A/kT} + q(\mathrm{X}))/(e^{-\mu+I/kT} + e^{\mu+A/kT} + 1) \quad (18)$$

where μ(SS) is denoted by μ, I(X) is denoted by I, and A(X) is denoted by A for short. Note that

$$\mu = \frac{1}{2}\left[-I - A + k_{\mathrm{B}}T\ln\left(\frac{x_-}{x_+}\right)\right] \quad (19)$$

The logarithm term in eq 19 disappears when $x_+ = x_-$. The calibration can be therefore done by comparing the electronegativity χ_{cali} for the PS of the state X with the Mulliken[78−80] absolute electronegativity μ_{M}

$$\mu_{\mathrm{M}} + E_{\mathrm{cali}} = -\chi_{\mathrm{cali}} \quad (20)$$

$$\mu_{\mathrm{M}} = -{}^{1}/_{2}[I(\mathrm{X}) + A(\mathrm{X})] \quad (21)$$

VI. Conclusion

In this work, we propose a flexible-boundary scheme to account for partial charge transfers between the PS and SS for QM/MM calculations. The scheme is based on the principle of electronic chemical potential equalization. The PS, which is described by a statistical mixture of ensemble that consists of states of integer number of electrons, exchanges electrons with the SS, which serves as a reservoir of electrons. The electronic chemical potential of the SS varies when charges flow in or out until equilibrium is established for the electronic chemical potentials between the PS and SS. Our scheme is tested by calculations of the partial atomic charges for 8 model systems. The QM/MM calculated charges agree with full-QM calculations reasonably well. The averaged mean unsigned deviations (over two set of QM/MM charges and three temperatures) between the QM/MM calculations and full-QM calculations are rather moderate: 0.16 e for partial atomic charges for the entire systems and 0.13 e for the amount of charges transferred between the PS and SS.

The flexible-boundary treatment requires embedded-QM calculations for each involved oxidation state. In contrast, a polarized-embedding QM/MM calculation requires embedded-QM calculations for only one specific oxidation state. The flexible-boundary treatment provides enhanced accuracy but is computationally more expensive. Fortunately, in most cases, it is sufficient to consider only two oxidation states, as the other oxidation states are generally much higher in energy and contribute negligibly. Thus, for most applications, the computational costs will increase by approximately a factor of 2, which is still within the acceptable range.

Future work is needed to refine and improve the flexible-boundary treatment. First, analytic gradients should be implemented to facilitate geometry optimizations and molecular dynamics simulations. Second, in our current implementation of the flexible-boundary treatment, the MM parameters such as partial atomic charges and van der Waals parameters are not reoptimized for pure-MM calculations. In future studies, it is desirable to refine those MM parameters for a more self-consistent description of the interactions between the PS and SS. Such a refinement will improve geometry and other molecular properties.

Acknowledgment. This work is supported by the Research Corporation through the award No. CC6725 to Hai Lin. We thank Donald Truhlar for fruitful discussions, and we thank the Advanced Biomedical Computing Center at the National Cancer Institute for providing CPU time and access to the *Gaussian03* program.

Supporting Information Available: Optimized geometries and atomic charges for H_2O, HS^-, NH_4^+, and H_3O^+ in the gas phase (Table S1), Cartesian coordinates for the optimized geometry of the 7 small model systems and of the Eigen cation, along with the absolute energies, HOMO energies, and LUMO energies (Table S2), for the PS the $\mu(e^-)$ at $T = 30\,000$ K, molar fractions of the reduced and oxidized states, and the gas-phase electronegativity χ (Table S3), crossing of $\mu(e^-)$ and μ(SS) for the model systems

(Table S4), convergence of the QM/MM calculated charges (Table S5), QEq-SCT-ES charges calculated for [$Na^+\cdots H_2O$] as functions of the Na−O distance (Table S6), and more details about the extension of the flexible-boundary treatment involving three oxidation states. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227−249.

(2) Singh, U. C.; Kollmann, P. A. *J. Comput. Chem.* **1986**, *7*, 718−730.

(3) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700−733.

(4) Gao, J. *Rev. Comput. Chem.* **1996**, *7*, 119−185.

(5) Friesner, R. A.; Beachy, M. D. *Curr. Opin. Struct. Biol.* **1998**, *8*, 257−262.

(6) *Combined quantum mechanical and molecular mechanical methods;* Gao, J., Thompson, M. A., Eds.; ACS Symp. Ser. 712; American Chemical Society: Washington, DC, 1998.

(7) Ruiz-López, M. F.; Rivail, J. L. In *Encyclopedia of computational chemistry*; Schleyer, P. v. R., Ed.; Wiley: Chichester, 1998; Vol. 1, pp 437−448.

(8) Monard, G.; Merz, K. M., Jr. *Acc. Chem. Res.* **1999**, *32*, 904−911.

(9) Hillier, I. H. *THEOCHEM* **1999**, *463*, 45−52.

(10) Hammes-Schiffer, S. *Acc. Chem. Res.* **2000**, *34*, 273−281.

(11) Sherwood, P. In *Modern Methods and Algorithms of Quantum Chemistry*; Grotendorst, J., Ed.; NIC-Directors: Princeton, 2000; Vol. 3, pp 285−305.

(12) Gao, J.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467−505.

(13) Morokuma, K. *Phil. Trans. R. Soc. London, A* **2002**, *360*, 1149−1164.

(14) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *117*, 185−199.

(15) Senn, H. M.; Thiel, W. *Top. Curr. Chem.* **2007**, *268*, 173−290.

(16) Bakowies, D.; Thiel, W. *J. Phys. Chem.* **1996**, *100*, 10580−10594.

(17) Day, P. N.; Jensen, J. H.; Gordon, M. S.; Webb, S. P.; Stevens, W. J.; Krauss, M.; Garmer, D.; Basch, H.; Cohen, D. *J. Chem. Phys.* **1996**, *105*, 1968−1986.

(18) Thompson, M. A.; Schenter, G. K. *J. Phys. Chem.* **1995**, *99*, 6374−6386.

(19) Bakowies, D.; Thiel, W. *J. Comput. Chem.* **1996**, *17*, 87−108.

(20) Eichler, U.; Kölmel, C. M.; Sauer, J. *J. Comput. Chem.* **1996**, *18*, 463−477.

(21) Field, M. J. *Mol. Phys.* **1997**, *91*, 835−845.

(22) Gao, J. *J. Comput. Chem.* **1997**, *18*, 1061−1071.

(23) Bryce, R. A.; Vincent, M. A.; Malcolm, N. O. J.; Hillier, I. H.; Burton, N. A. *J. Chem. Phys.* **1998**, *109*, 3077−3085.

(24) Aida, M.; Yamataka, H.; Dupuis, M. *Int. J. Quantum Chem.* **2000**, *77*, 199−210.

(25) Sushko, P. V.; Shluger, A. L.; Catlow, C. R. A. *Surf. Sci.* **2000**, *450*, 153−170.

(26) French, S. A.; Sokol, A. A.; Bromley, S. T.; Catlow, C. R. A.; Rogers, S. C.; King, F.; Sherwood, P. *Angew. Chem.* **2001**, *113*, 4569−4572.

(27) Nasluzov, V. A.; Rivanenkov, V. V.; Gordienko, A. B.; Neyman, K. M.; Birkenheuer, U.; Rösch, N. *J. Chem. Phys.* **2001**, *115*, 8157−8171.

(28) Dupuis, M.; Aida, M.; Kawashima, Y.; Hirao, K. *J. Chem. Phys.* **2002**, *117*, 1242−1255.

(29) Jensen, L.; van Duijnen, P. T.; Snijders, J. G. *J. Chem. Phys.* **2003**, *118*, 514−521.

(30) Illingworth, C. J. R.; Gooding, S. R.; Winn, P. J.; Jones, G. A.; Ferenczy, G. G.; Reynolds, C. A. *J. Phys. Chem. A* **2006**, *110*, 6487−6497.

(31) Zhang, Y.; Lin, H.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 1378−1398.

(32) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(33) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III;, 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.

(34) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(35) Jorgensen, W. L.; McDonald, N. A. *THEOCHEM* **1998**, *424*, 145−155.

(36) McDonald, N. A.; Jorgensen, W. L. *J. Phys. Chem. B* **1998**, *102*, 8049−8059.

(37) Rizzo, R. C.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1999**, *121*, 4827−4836.

(38) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474−6487.

(39) Kahn, K.; Bruice, T. C. *J. Comput. Chem.* **2002**, *23*, 977−996.

(40) Sprik, M.; Klein, M. L. *J. Chem. Phys.* **1988**, *89*, 7556−7560.

(41) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141−6156.

(42) Mortier, W. J.; Van Genechten, K.; Gasteiger, J. *J. Am. Chem. Soc.* **1985**, *107*, 829−835.

(43) Mortier, W. J.; Ghosh, S. K.; Shankar, S. *J. Am. Chem. Soc.* **1986**, *108*, 4315−4320.

(44) Rappé, A. K.; Goddard, W. A. *J. Phys. Chem.* **1991**, *95*, 3358−3363.

(45) Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerings, P.; Waroquier, M.; Tollenaere, J. P. *J. Phys. Chem. A* **2002**, *106*, 7887−7894.

(46) York, D. M.; Yang, W. *J. Chem. Phys.* **1996**, *104*, 159−172.

(47) Itskowitz, P.; Berkowitz, M. L. *J. Phys. Chem. A* **1997**, *101*, 5687−5691.

(48) Yang, Z.-Z.; Wang, C.-S. *J. Phys. Chem. A* **1997**, *101*, 6315−6321.

(49) Applequist, J.; Carl, J. R.; Fung, K.-K. *J. Am. Chem. Soc.* **1972**, *94*, 2952−2960.

(50) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341−350.

(51) Stone, A. J. *Mol. Phys.* **1985**, *56*, 1065−1082.

(52) Winn, P. J.; Ferenczy, G. G.; Reynolds, C. A. *J. Comput. Chem.* **1999**, *20*, 704−712.

(53) Parr, R. G.; Donnelly, R. A.; Levy, M.; Palke, W. E. *J. Chem. Phys.* **1978**, *68*, 3801−3807.

(54) Parr, R. G.; Pearson, R. G. *J. Am. Chem. Soc.* **1983**, *105*, 7512−7516.

(55) Gogonea, V.; Merz, K. M., Jr. *J. Chem. Phys.* **2000**, *112*, 3227−3235.

(56) Gogonea, V.; Merz, K. M., Jr. *J. Phys. Chem. B* **2000**, *104*, 2117−2122.

(57) Dewar, M. J. S.; Hashmall, J. A.; Venier, C. G. *J. Am. Chem. Soc.* **1968**, *90*, 1953−1957.

(58) Dewar, M. J. S.; Trinajstic, N. *J. Chem. Soc., Chem. Commun.* **1970**, 646−647.

(59) Tavernelli, I.; Vuilleumier, R.; Sprik, M. *Phys. Rev. Lett.* **2002**, *88*, 213002/1−213002/4.

(60) Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L. *Phys. Rev. Lett.* **1982**, *49*, 1691−1694.

(61) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471−2474.

(62) Jaque, P.; Marenich, A.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. C* **2007**, *111*, 5783−5799.

(63) Bartlett, R. J. *J. Phys. Chem.* **1989**, *93*, 1697−1708.

(64) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129−145.

(65) Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431−439.

(66) Löwdin, P.-O. *J. Chem. Phys.* **1950**, *18*, 365−375.

(67) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098−3100.

(68) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(69) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter* **1988**, *37*, 785−789.

(70) Ditchfield, R.; Hehre, W. J.; Pople, J. A. *J. Chem. Phys.* **1971**, *54*, 724−728.

(71) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; DeFrees, D. J.; Pople, J. A.; Gordon, M. S. *J. Chem. Phys.* **1982**, *77*, 3654−3665.

(72) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 294−301.

(73) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265−3269.

(74) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257−2261.

(75) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, 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.; A., Pople, J. A. *Gaussian03*, Gaussian, Inc.: Pittsburgh, PA, 2003.

(76) Ponder, J. W. TINKER, *Version 4.2*; Washington University: St. Louis, MO, 2004.

(77) Lin, H.; Zhang, Y.; Truhlar, D. G. QMMM, *Version 1.3.5*; University of Minnesota: Minneapolis, MN, 2007.

(78) Mulliken, R. S. *J. Chem. Phys.* **1934**, *2*, 782−793.

(79) Mulliken, R. S. *J. Chem. Phys.* **1935**, *3*, 573−585.

(80) Iczkowski, R. P.; Margrave, J. L. *J. Am. Chem. Soc.* **1961**, *83*, 3547−3551.

CT700296X

JCTC Journal of Chemical Theory and Computation

Effect of Structural Parameters on the Polarizabilities of Methanol Clusters: A Hirshfeld Study

Alisa Krishtal,[†] Patrick Senet,[‡] and Christian Van Alsenoy*[,†]

Chemistry Department, University of Antwerp, Universiteitsplein 1, B2610 Antwerp, Belgium, and Institut Carnot de Bourgogne, UMR 5209 CNRS, Université de Bourgogne, 9 Avenue Alain Savary BP 47870, F-21078 Dijon CEDEX, France

Received November 29, 2007

Abstract: The polarizabilities of fifty methanol clusters $(CH_3OH)_n$, $n = 1$ to 12, were calculated at the B3LYP/6-311++G** level of theory and partitioned into molecular contributions using the Hirshfeld-I method. The resulting molecular polarizabilities were found to be determined by the polarizabilities of the two parts of the molecule, the hydrophilic hydroxyl group and the hydrophobic methyl group, each exhibiting a different dependency upon the local environment. The polarizability of the hydroxyl group was found to be dependent on the number, type, and strength of the hydrogen bonds a methanol molecule makes, whereas the polarizability of the methyl groups is mostly influenced by sterical hindrance. The findings were compared with the results obtained in a previous study on water clusters. The influence of the BSSE correction was investigated and found to increase polarizability values by up to 8.5%.

1. Introduction

One factor determining the impact of liquid methanol on today's chemistry is its very frequent use as a solvent. For this reason, comprehensive studies have investigated the structure of liquid methanol as well as its properties. Although the earlier studies could use only experimental techniques,[1−7] as the power of computational chemistry was yet to be established, nowdays it is common to find experimentalists and theoreticians together tackling this inexhaustable subject together.[8−11]

However, computational studies of liquids still form a challenge, since the system to be investigated has an extensive size and is of a dynamic nature. Therefore, many of the studies employ hybrid quantum mechanical/molecular mechanical (QM/MM) methods or conduct MD simulations.[12−17] Another possibility is to apply conventional electronic structure methods on methanol clusters in the gas phase, which allows for the examination of the properties of the clusters in closer detail.[18−23] Recently, Boyd et al.[23] have published an extensive study on the potential surfaces of methanol clusters ($(CH_3OH)_n$, $n = 2$ to 12), considering various types of isomers for each aggregation number. A selection of fifty of those clusters will be used in this study with the purpose of studying the polarizability of the methanol molecules within the clusters.

For solvents, different properties of the solvent and solute can influence the resulting interaction. As an attempt to understand this complex system, one may first start looking at the pure solvent, where each molecule can be considered as a solute, surrounded by similar solute molecules. Among the different properties that can be studied, electronic properties, and polarizability in particular, are of substance.[24−26] Furthermore, when considering polar solvents as methanol, the relation between the properties of the solvent molecules and the hydrogen bonds that the solvent molecules make with each other and with the solute is of great importance.

Although the structure of a single molecule in a cluster is simple to obtain, the knowledge of its properties usually requires a partitioning method, of which the Hirshfeld method is our method of choice. The Hirshfeld method[27,28] was introduced in 1977 primarily as a method for the partitioning of electron density for the purpose of obtaining atomic charges and dipoles. Later it was extended for the partitioning of properties such as quadrupole moments,[29] similarities,[30,31]

[*] Corresponding author e-mail: kris.vanalsenoy@ua.ac.be.
[†] University of Antwerp.
[‡] Université de Bourgogne.

10.1021/ct700325c CCC: $40.75
Published on Web 02/09/2008

Fukui functions,[32−34] energies,[35] and polarizabilities.[36] In our previous study,[36] the Hirshfeld method was used for the partitioning study of polarizabilities of water molecules in water clusters. In that study the molecular polarizabilities were found to be strongly dependent upon the hydrogen bond network and, moreover, to have highly transferable values between the clusters. Methanol clusters form a greater challenge, possessing a hydrophobic methyl group that disturbs the ordered hydrogen-bond network that can be found in water clusters. The balance between the polar and nonpolar parts of the molecule and the influence of the hydrogen bond network will be a determining factor for eventual polarizability of the solvent molecules and the resulting interactions with the solute. Recently, a new version of the Hirshfeld method has been introduced by Bultinck et al.[38] improving some aspects of the conventional method, such as arbitrariness of the choice of the promolecular density and the restriction of applicability to neutral systems. In this study, the revised method is applied for the first time to the study of polarizabilities. As a result, the obtained numerical results cannot be compared with those obtained in ref 36, yet the trends that are observed using both methods remain unchanged.

2. Method

The partitioning of polarizabilities of clusters into atomic and molecular contributions is accomplished in this study by means of the Hirshfeld method.[27,28] Using this scheme, elements of the total polarizability tensor of the cluster can be reproduced exactly from atomic contributions

$$\alpha_{ij} = \sum_{A=1}^{N} (\alpha_{ij}^{A} + \boldsymbol{q}_{A}^{i} j_{A}) \tag{1}$$

where *i* and *j* stand for the Cartesian directions *x*, *y*, or *z*. In this equation the summation runs over all atoms A in the system, α_{ij}^{A} is referred to as an intrinsic atomic polarizability, while $\boldsymbol{q}_{A}^{i}$ is referred to as a perturbed charge, i.e., it is the atomic Hirshfeld charge calculated using the first-order perturbed density matrix for an electric field oriented along the *i* axis.[36] The second term in the summation in eq 1 can thus be interpreted as a charge delocalization polarizability, which describes the contribution of charge transfer between the atoms to the total polarizability of the system. The intrinsic atomic polarizability α_{ij}^{A} and the perturbed atomic charge $\boldsymbol{q}_{A}^{i}$ are given by

$$\alpha_{ij}^{A} = \int (i - i_{A})\boldsymbol{\omega}_{A}(\vec{r})\boldsymbol{\rho}^{j}(\vec{r})d\vec{r} \tag{2}$$

and by

$$\boldsymbol{q}_{A}^{i} = \int \boldsymbol{\omega}_{A}(\vec{r})\boldsymbol{\rho}^{i}(\vec{r})d\vec{r} \tag{3}$$

respectively. In these i_A is the Cartesian coordinate of atom A in the *i* direction, and $\boldsymbol{\rho}^{i}$ and $\boldsymbol{\rho}^{j}$ are the first-order perturbed density matrices obtained using a coupled perturbed Kohn−Sham procedure for an electric field perturbation in the *i* and *j* direction, respectively. $\omega_A(\vec{r})$ is the Hirshfeld weight function of atom A.

To obtain the intrinsic polarizability of a *molecule* in a cluster, one needs to sum these quantities over the atoms of the molecule:

$$\alpha_{ij}^{int(mol)} = \sum_{A(mol)} \alpha_{ij}^{A} \tag{4}$$

Finally, the total polarizability of a molecule in a cluster is obtained by adding the intramolecular charge delocalization contribution:

$$\alpha_{ij}^{tot(mol)} = \alpha_{ij}^{int(mol)} + \sum_{A(mol)} \boldsymbol{q}_{i}^{A} j_{A} \tag{5}$$

This intramolecular charge delocalization contribution is translationally invariant, when it is defined with respect to the geometrical center of the molecule.

In this study, only the isotropic part of the polarizability, which is independent of the orientation of the system, will be discussed. Therefore the values reported in this paper for the polarizabilities are always the trace of the corresponding polarizability tensor:

$$\alpha = \frac{\alpha_{xx} + \alpha_{yy} + \alpha_{zz}}{3} \tag{6}$$

In the classic version of the Hirshfeld method the weight function is constructed from the free atomic densities $\boldsymbol{\rho}_{A}^{0}$ of the atoms in the system:

$$\boldsymbol{\omega}_{A}(\vec{r}) = \frac{\boldsymbol{\rho}_{A}^{0}(\vec{r})}{\sum_{B=1}^{N} \boldsymbol{\rho}_{B}^{0}(\vec{r})} \tag{7}$$

Recently, Bultinck et al. have revised this method and proposed an iterative version, Hirshfeld-I, which is more in line with information theory.[38] In this revised version, the weight function is constructed in each iteration from the atomic densities $\boldsymbol{\rho}_{A}^{n-1}$ that normalize to the atomic populations that were obtained during the previous iteration:

$$\boldsymbol{\omega}_{A}^{n}(\vec{r}) = \frac{\boldsymbol{\rho}_{A}^{n-1}(\vec{r})}{\sum_{B=1}^{N} \boldsymbol{\rho}_{B}^{n-1}(\vec{r})} \tag{8}$$

This procedure is repeated until convergence of the atomic populations. The converged weight function can then be used for the partitioning of properties such as polarizabilities, according to eqs 2 and 3.

3. Results and Discussion

The polarizabilities of fifty methanol clusters, with aggregation numbers ranging between 2 and 12, were calculated at the DFT level, using the B3LYP/6-311++G(d,p) method and the Gaussian03[39] program. The calculated polarizabilities were subsequently partitioned using the program STOCK.[28] The geometries used were those optimized by Boyd et al.[23] in a study of the energies of the methanol clusters.

Three different types of methanol clusters were used in this study: (1) Chainlike structures, with aggregation numbers ranging between $n = 2$ and $n = 12$, noted as *n*c. For

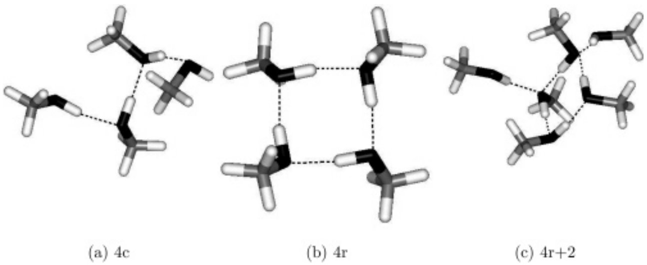


Figure 1. The three different types of clusters that were used in the study.

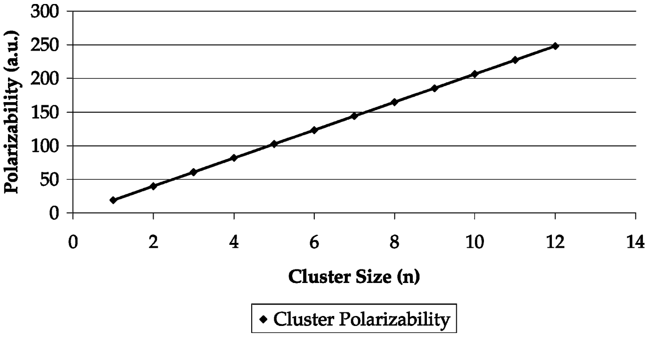


Figure 2. Isotropic part of the cluster polarizability (eq 6) for different aggregation numbers *n*.

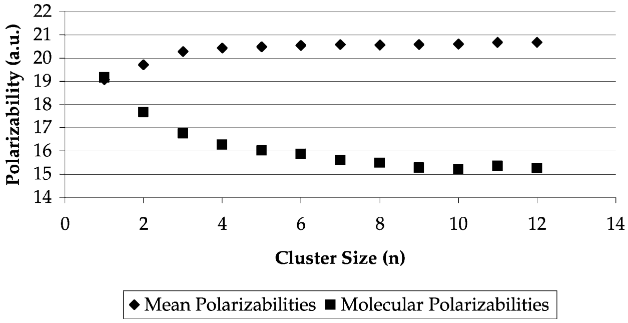


Figure 3. Polarizability of a molecule in a cluster of size *n*. Diamonds are mean polarizabilities and squares are molecular polarizabilities (eq 5).

example, cluster 4c is depicted in Figure 1a. (2) Ring structures, with aggregation numbers ranging between $n = 3$ and $n = 12$, noted as *n*r. For example, cluster 4r is depicted in Figure 1b. (3) Substituted ring structures, where the substituents can be a number of single methanol molecules (denoted as *n*r + *m*), a number of chains of two methanol molecules (denoted as *n*r + *m*d), or a number of chains of three methanol molecules (denoted as *n*r + *m*t). Only in the first of these cases structures were considered with $m > 1$; in all other cases $m = 1$. For example, cluster 4r + 2 is depicted in Figure 1c.

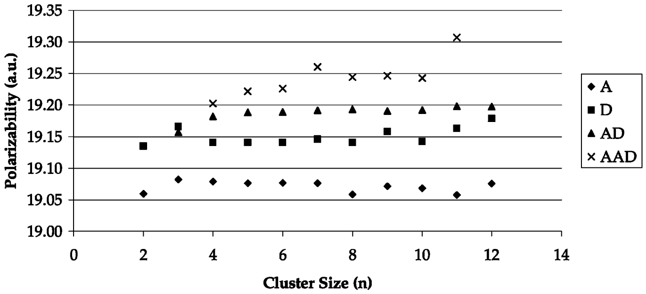


Figure 4. Polarizabilities of isolated molecules for the four types of hydrogen-bonded methanol molecules.

Figure 2 illustrates the isotropic cluster polarizabilities as function of the size of the cluster. The relation between the size of the cluster and the isotropic cluster polarizability appears to be highly linear, with an $R^2 = 1.0000(1)$ for a two parameter fit, suggesting a good transferability of the polarizability values between the molecules in the different clusters. Figure 3 displays both the mean molecular polarizabilities, obtained by dividing the isotropic cluster polarizability by the aggregation number *n*, and the molecular polarizabilities, obtained from eq 5. The mean molecular polarizability increases for the smaller clusters and reaches saturation quicky for clusters with an aggregation of 4−5 at value around 20.7 au, whereas the molecular polarizability decreases with the size of the cluster and stabilizes at a value around 15.3 au only at larger clusters of aggregation 9−10. The difference in behavior between these two properties is due to the charge delocalization contribution, which is dependent on the coordinates and thus on the size of the cluster, as can be seen from eq 1. The charge delocalization contribution increases for larger clusters as result of the extension of the volume of the system, resulting in the increase of mean molecular polarizabilities. However, Figure 3 also illustrates that the polarizability of a methanol molecule in a cluster decreases in going from gas phase to condensed phase, as the values for the molecular polarizability obtained from eq 5 decrease with the size of the cluster. A similar effect was observed for a water molecule in a cluster in ref 36. This behavior can be attributed to different effects, such as the change in geometry of the methanol molecule, the hydrogen-bonding network that the methanol molecules engage in in the clusters, the change of net charge of the molecule in the cluster, steric hindrance, and other local effects.

To investigate the effect of the change in geometry on the polarizability of the methanol molecules in the different clusters, the coordinates of each molecule were used to calculate its polarizability in the absence of the rest of the molecules in that particular cluster. These “isolated” polarizabilities were calculated using the same method and basis set as were used for the calculation of the polarizabilities of the clusters. The average “isolated” polarizability was found to amount to 19.24 au, being 0.17 au above the polarizability of an optimized single methanol molecule (19.07 au). This change in the polarizability on can be ascribed to the geometry deformation due to the formation of hydrogen bonds between the methanol molecules.

A methanol molecule can form up to three hydrogen bonds with neighboring molecules, by either acting as an acceptor through the oxygen atom (A) or by donating a hydrogen atom into the bond (D). In the clusters examined in this study, four different types of methanol molecules are present, namely methanol molecules of type A, type D, type AD, and type AAD. Figure 4 depicts the “isolated” polarizabilities for the four types of methanol molecules as a function of

Table 1. Average "Isolated" Polarizabilities and Average BSSE Corrected "Isolated" Polarizabilities of the Four Types of Methanol Molecules[a]

	"isolated" polarizabilities				BSSE corrected polarizabilities			
n	A	D	AD	AAD	A	D	AD	AAD
2	19.06	19.14			19.71(3.31)	19.81(3.39)		
3	19.08	19.17	19.16		20.10(5.07)	20.03(4.36)	20.22(5.26)	
4	19.08	19.14	19.18	19.20	20.09(5.01)	20.18(5.13)	20.47(6.27)	20.61(6.81)
5	19.08	19.14	19.19	19.22	20.13(5.22)	20.17(5.08)	20.58(6.76)	20.71(7.19)
6	19.08	19.14	19.19	19.23	20.28(5.93)	20.21(5.29)	20.58(6.77)	20.87(7.86)
7	19.08	19.15	19.19	19.26	20.28(5.93)	20.28(5.61)	20.67(7.17)	20.93(7.98)
8	19.06	19.14	19.19	19.24	20.03(4.87)	20.36(6.00)	20.73(7.40)	20.96(8.20)
9	19.07	19.16	19.19	19.25	20.22(5.67)	20.49(6.51)	20.77(7.59)	21.00(8.33)
10	19.07	19.14	19.19	19.24	20.10(5.13)	20.42(6.25)	20.79(7.69)	21.01(8.42)
11	19.06	19.16	19.20	19.31	19.97(4.56)	20.38(5.97)	20.76(7.50)	20.96(7.87)
12	19.08	19.18	19.20		20.21(5.617)	20.48(6.37)	20.78(7.63)	

[a] The values between brackets give the percentage of the contribution of BSSE. All values are in au.

the size of the clusters. Each additional hydrogen bond appears to increase the "isolated" polarizabilities of the methanol molecules by lengthening the bonding distances and thus increasing the volume of the molecule. A D-type hydrogen bond has a larger effect than an A-type hydrogen bond, as the mean "isolated" polarizabilities of the former type of methanol molecules are larger than those of the latter type. The "isolated" polarizabilities are approximately independent of the size of the cluster, although the value for the AAD-type methanol molecule for aggregation number $n = 11$ seems to be exceptionally high. However, keeping in mind that Figure 4 displays only an average value, computed on all molecules of a given type within a cluster, the value for this point is still within the observed deviation (0.04 au). Although values of this magnitude also appear for the lower aggregation numbers, they are eventually averaged out, whereas for aggregation number $n = 11$ this is the only point available.

In order to distinguish the effect of geometry on the polarizability of a molecule from other effects, it is convenient to look at the difference between the "isolated" molecular polarizabilities and the molecular polarizabilities (eq 5) of the methanol molecules in the clusters. As such, the Basis Set Superposition Error (BSSE) can be taken into acount, as the basis functions situated on neighboring methanol molecules in the clusters may influence the polarizability significantly. Furthermore, one cannot assume the BSSE error to be constant because of the wide range of aggregation numbers of the clusters. Indeed, the BSSE error of a cluster with aggregation number $n = 3$ can be expected to be be smaller than the BSSE error of a cluster with aggregation number $n = 10$.

The BSSE error was calculated by means of the counterpoise method.[40,41] The polarizability of a given molecule was calculated by replacing all other atoms in the clusters by ghost atoms. The average values of those BSSE corrected "isolated" polarizabilities are compared with the previously mentioned "isolated" polarizabilities in Table 1, for each type of methanol molecule. The BSSE corrected values are larger, as expected, increasing for the smaller clusters and stabilizing for the larger clusters. The correction also appears to increase with the number of hydrogen bonds, reaching 1.64 au for the AAD-type water molecules, which amounts to approximatively 8.5% of the original value.

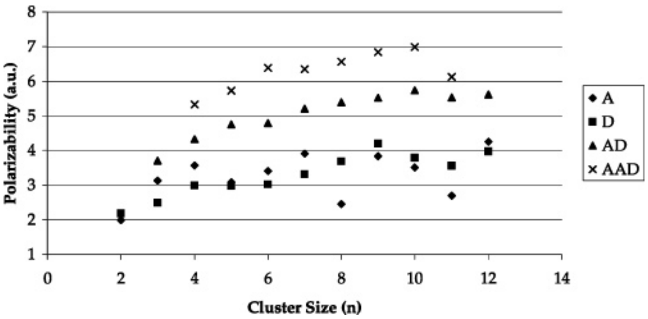


Figure 5. The average difference ($\Delta\alpha^{mol}$) between the BSSE corrected "isolated" polarizabilities and the total molecular polarizabilities for the four types of methanol molecules.

Taking the difference between the BSSE corrected "isolated" polarizabilities and the total molecular polarizabilities, obtained from eq 5, yields the change in the polarizability of the methanol molecule due to the rearrangement of the electron density, that is brought forth by bonding effects, such as hydrogen bonding, and nonbonding effects, such as steric hindrance or other local effects. Figure 5 shows the average change in polarizability ($\Delta\alpha^{mol}$) for the four types of methanol molecules. The values are lowest for the A- and D-type methanol molecules, that appear to be overlapping, and highest for the AAD-type methanol molecules. Note that higher $\Delta\alpha^{mol}$ values imply lower total molecular polarizabilties, in agreement with the values depicted in Figure 3. All the values tend to increase slightly for the smaller clusters and stabilize for the larger clusters. To understand this behavior it is necessary to compute the contributions of the different parts of the molecules to the intrinsic polarizabilities. A methanol molecule consists of a hydrophilic part, namely the hydroxyl group, and a hydrophobic part, namely the methyl group. Since the hydroxyl group can take part in the hydrogen-bonding network, one can expect the intrinsic polarizabilities to be directly influenced by it. On the other hand, the methyl groups do not make hydrogen bonds and can therefore only experience a secondary effect of the hydrogen-bonding network on their intrinsic polarizabilities. Furthermore, the methyl groups are much more voluminous than the hydroxyl group and can therefore be influenced to a greater extent by steric hindrance. Figures 6 and 7 illustrate

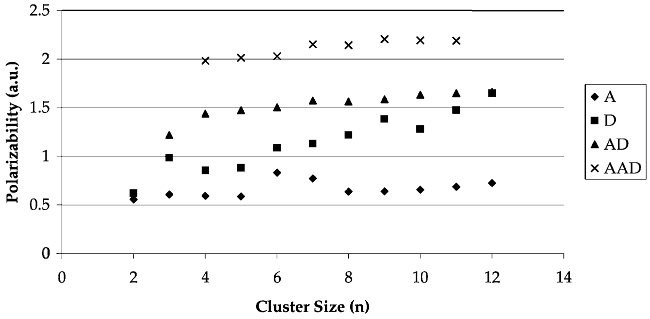


Figure 6. The average difference between the contribution of the hydroxyl groups to the BSSE corrected “isolated” polarizabilities and the molecular polarizabilities ($\Delta\alpha^{OH}$) in the four types of hydrogen-bonded methanol molecules.

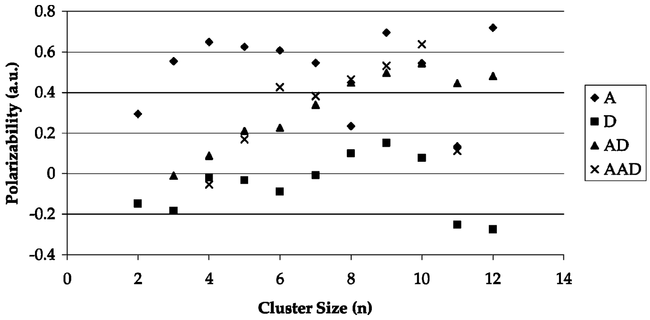


Figure 7. The average difference between the contribution of the methyl groups to the BSSE corrected “isolated” polarizabilities and molecular polarizabilities ($\Delta\alpha^{CH_3}$) in the four types of hydrogen-bonded methanol molecules.

the difference in intrinsic polarizabilities obtained from the BSSE corrected “isolated” molecules and the clusters, for the hydroxyl ($\Delta\alpha^{OH}$, Figure 6) and methyl ($\Delta\alpha^{CH_3}$, Figure 7) groups, respectively. The polarizabilities of the BSSE corrected “isolated” molecules are obtained by applying the same Hirshfeld-I scheme as was done for the clusters. The intrinsic polarizabilities of a group are obtained by summing over the intrinsic polarizabilities of the atoms in the group, analogous to eq 4. The values for the hydroxyl groups are again separated into four different populations, increasing with the number of hydrogen bonds and a D-type hydrogen bond having a greater influence on the values (A < D < AD < AAD), whereas the values for the methyl group exhibit less obvious behavior.

On the other hand, in Figure 7 the A-type molecules have the highest values and the D-type molecules have the lowest values. This difference in behavior from the $\Delta\alpha^{OH}$ causes the overlap in the values of those two types of molecules in Figure 5. The values for the AD- and AAD-type molecules in Figure 7 overlap and are situated between the values for A- and D-types, which may suggest that the polarizabilities of the methyl groups are not significantly influenced by the hydrogen-bonding network and that the primary effect for the change in their polarizability is due to local effects such as steric hindrance.

To investigate the sterical effects further, the structures of the different methanol clusters will be examined closer for the three types of methanol clusters.

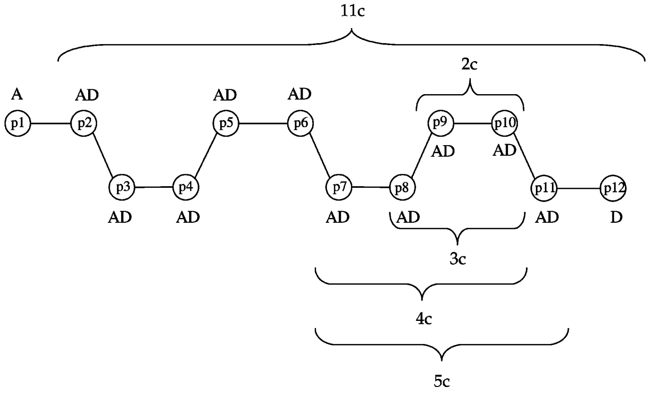


Figure 8. Schematic illustration of the positioning nomenclature of the methanol molecules in the chainlike clusters. The type of hydrogen-bonded methanol molecules is illustrated for the 12c cluster.

3.1. Chains. The chainlike clusters have a structure that allows analysis of the dependence of polarizabilities on the local environment in a straightforward way. The presence of a methyl group in the molecules causes this group of clusters to have a bent structure for the lower aggregation number and a helical structure for the larger aggregation numbers. The chains were constructed in a way that preserves the local environment at a given position throughout the range of the clusters. This local environment is characterized by the length of the hydrogen bonds the molecule makes and the steric hindrance it undergoes from the neighboring molecules.

The chains consist of an A-type molecule at one end and a D-type molecule at the other end, with several AD-type molecules in between connecting them. Each cluster *n*c was constructed by adding a molecule either to the A end or the D end of an ($n - 1$) cluster, preserving the topology of the $n - 1$ molecules of the previous cluster. This way, the structure and the surroundings of a molecule in a certain position in the chain is approximately unchanged throughout all the clusters, even after reoptimization. This allows analysis of the correlation between the polarizability values of the molecules at a certain position and the specific environment at that position.

This concept is schematically illustrated in Figure 8. The two molecules that are grouped together under the label of 2c are the molecules that constitute the starting dimer, the left one being an A-type molecule and the right one being a D-type molecule. The trimer is constructed by adding a molecule to the left side of the dimer. The A-type molecule in the dimer becomes now an AD-type molecule in the trimer. This process is repeated to construct the 4c cluster. For the 5c cluster, a molecule was added to the right D-type side of the 4c cluster. As result, the molecule that was consistently a D-type molecule for clusters 2c to 4c, becomes an AD-type molecule in the 5c cluster. After subsequent addition of another seven molecules to either the A-side or the D-side one finally arrives to the structure of the 12c cluster. The molecules are then assigned numbers according to their position in the 12c cluster, which are also used to label the molecules in the smaller cluster. During the construction of the clusters, the molecules were more

Table 2. Average $\Delta\alpha^{mol}$'s, $\Delta\alpha^{OH}$'s, and $\Delta\alpha^{CH_3}$'s of the Different Types of Molecules in the Chainlike Clusters[a]

	$\Delta\alpha^{OH}$			$\Delta\alpha^{CH_3}$			$\Delta\alpha^{mol}$		
position	A	D	AD	A	D	AD	A	D	AD
p1	1.25			1.1			4.26		
p2	1.21		2.31	0.55		0.64	2.70		4.33
p3	1.19		2.14	1.06		1.00	3.52		5.14
p4	1.17		2.16	1.07		1.28	3.84		5.63
p5	1.16		2.28	0.52		0.45	2.45		4.70
p6	1.28		2.52	1.00		1.13	3.66		5.89
p7	1.05		1.93	1.05		1.47	3.33		5.69
p8	1.11		2.47	0.94		1.31	3.14		6.11
p9	0.88		2.46	0.54		0.85	2.00		5.32
p10		1.26	1.92		0.10	0.90		2.51	4.65
p11		1.53	2.41		0.13	0.44		2.76	4.66
p12		2.20			0.21			3.84	

[a] According to their placement in the chain—all values are in au.

frequently added to the A-side of the previous cluster than to the D-side. As result, the two molecules that originally constituted the dimer are located at the ninth and tenth position in the 12c cluster. This means that the A-type molecules can only occupy positions p1 to p9, while D-type molecules can only occupy positions p10 to p12.

Table 2 summarizes the average $\Delta\alpha^{OH}$'s, $\Delta\alpha^{CH_3}$'s, and the $\Delta\alpha^{mol}$'s of the different types of molecules, according to their position in the molecule. The averaging is performed as follows: the molecules that occupy position p*x* are first separated according to their type (A, D, or AD), and then the values for each type at a given position are averaged. The values for the hydroxyl group follow the already established order A < D < AD. The values of $\Delta\alpha^{OH}$ for the A-type molecules have an average of 1.15 au, with an evident outlier at position p9, which has a value of 0.88 au. This low value can be explained by the relatively longer hydrogen bond in the dimer, 1.9 au, whereas the average hydrogen bond in the larger clusters is around 1.73 au. The hydrogen bond length is defined here as the distance between the hydrogen atom and the oxygen atom that participate in the bond. This effect can also be seen in the average value of $\Delta\alpha^{OH}$ of D-type molecules at position p10 (1.26 au). There are two molecules that contribute to this value, namely the molecules in this position in clusters 2c and 3c. Both molecules take part in longer hydrogen bonds, the value for the 3c cluster being 1.81 au, causing the $\Delta\alpha^{OH}$ of the D-type molecules at this position in the chain to be lower than the average value of 1.68 au. The hydrogen bond at this position is consequently longer throughout the whole range of the clusters, as part of the conservation of the local surroundings of the molecules. As result, also the values of $\Delta\alpha^{OH}$ of the AD-type molecules at position p10 (1.92 au) are lower than the average (2.26 au). It can also be observed that the length of the hydrogen bond has a larger effect on the polarizability of the hydrogen atom than that of the oxygen atom participating in the bond, which will be discussed further at a later stage. Similar arguments can be used to explain the outlier at position p7 for the AD-type molecules. The average hydrogen bond length at this position is 1.77 au, which is higher than the average by 0.04 au.

Table 3. Average $\Delta\alpha^{OH}$'s, $\Delta\alpha^{CH_3}$'s, and $\Delta\alpha^{mol}$'s of the Methanol Molecules in the Clusters with Ring Structures[a]

cluster size	$\Delta\alpha^{OH}$	$\Delta\alpha^{CH_3}$	$\Delta\alpha^{mol}$
3	1.78	0.32	3.65
4	2.06	0.39	4.21
5	2.20	0.69	4.84
6	2.16	0.51	4.46
7	2.25	0.77	4.92
8	2.27	0.89	5.20
9	2.40	0.97	5.42
10	2.42	0.97	5.49
11	2.41	0.97	5.42
12	2.46	1.14	5.80

[a] All values are in au.

The values of the methyl group again follow a less obvious pattern. The D-type molecules have low values, around 0.15 au, indicating that the intrinsic polarizabilities of the atoms in the methyl group are only slightly influenced by the transition from gas phase to a cluster. The methyl groups in the D-type molecules are not influenced by the hydrogen bond formation and undergo little sterical hindrance due to their position at the end of the molecule. The values for the A- and the AD-types molecules are slightly higher, cover a wider range of values than in the case of the hydroxyl groups, and overlap extensively. The overlap of the values indicates that the polarizabilities of the atoms in the methyl groups in those two types of molecules are also not strongly influenced by the type or number of hydrogen bonds the molecule forms. The change in the values in going from single molecules to a cluster must therefore be sought in the steric hindrance effect. Although the A-type molecules are also placed at the end of the chain, they undergo more steric hindrance than the D-type molecules because they are connected to the chain through the oxygen atom, which is situated in the middle of the molecule, and not by the hydrogen atom, which is situated at the end of the molecule. The methyl groups are therefore oriented in such a fashion that they still undergo steric hindrance from the other methyl groups. It is also noticeable from Table 2 that the values for those two types of methanol molecules follow the same pattern. For example, the values at positions p2, p5, and p9 are considerably lower for both types of molecules. Examining the structures closely reveals that the methyl groups at those positions undergo less steric hindrance due to turns in the curve of the chain, resulting in a smaller change in the polarizabilities of those groups.

3.2. Rings. The ring structures, consisting only of AD-type molecules, allow a more thorough analysis of the factors that influence the polarizability within a single type. Table 3 summarizes the average $\Delta\alpha^{OH}$'s, $\Delta\alpha^{CH_3}$'s. and the $\Delta\alpha^{mol}$'s for the molecules in those clusters, according to the aggregation number. The values for the hydroxyl groups and the methyl groups, and as consequence also of the molecules, increase with the aggregation number. Two possible explanations for this behavior, that have been mentioned up till now, are the decreasing hydrogen bond lengths and the increasing steric hindrance. The correlation between the intrinsic polarizability of the hydrogen atoms and the length of the hydrogen bonds those atoms form was found to be 0.66, indicating that some connection must exist between those two properties but that the polarizabilities must also be influenced by other factors. The polarizabilities seem to be influenced by substantial changes in the length of the hydrogen bonds but are less straighforwardly dependent on small changes. Moreover, there seems to be no correlation whatsoever between the length of the hydrogen bond and the intrinsic polarizability of the oxygen atoms, as has been noticed in the previous subsection. This finding is somewhat surprising, in light of the results that were obtained on water clusters in previous work, where correlations of 0.979 and 0.948 were found between the length of the hydrogen bonds and the intrinsic polarizabilities of the hydrogen atoms and oxygen atoms in DA-type water molecules, respectively. A possible explanation in this change of behavior is the more complicated structure of the methanol clusters, where new effects come to light, influencing the polarizabilities of the atoms and therefore reducing the correlation. One of the effects is the steric hindrance, which was found to increase the $\Delta\alpha^{CH_3}$ in the previous section and can therefore be responsible for the increasing trend in the clusters with the ring structures.

Another possible influencing factor on the polarizabilities can be the charge of the molecules. The reorganization of the charge density in going from gas phase to cluster can result in a net charge of the molecule in the cluster and a rearrangement of the charge within the molecule. The atomic charges of the atoms in the methanol clusters were calculated using the Hirshfeld-I scheme. The methanol molecules in the clusters were found to have a negligible net charge, never surpassing 0.01 au, but there seems to be a charge separation within the molecule, between the negatively charged hydroxyl groups and the positively charged methyl groups. The average value of the charge separation, defined as the absolute value of the difference of the charge of the hydroxyl group and the charge of the methyl group, is 0.37 au. A reasonable correlation was found between the $\Delta\alpha^{OH}$'s and the charges of the hydroxyl groups (−0.8932) and between the $\Delta\alpha^{CH_3}$'s and the charges of the methyl groups (0.9001). The polarizabilities of the hydroxyl groups increase with the negative charge, whereas the polarizabilities of the methyl groups increase with the positive charge, resulting in increasing polarizabilities of the molecules with the charge separation.

3.3. Substituted Rings. In this class of structures, the effect of the attachement of methanol molecules through a hydrogen bond was investigated by looking at the $\Delta\alpha^{OH}$'s, $\Delta\alpha^{CH_3}$'s, and $\Delta\alpha^{mol}$'s of a group of clusters nr + m with n = 5 and m varying from 0 to 5. The values for the three different types of methanol molecules are summarized in Table 4.

The values for the hydroxyl groups follow the usual order of D < AD < AAD and seem not to be influenced directly by the number of additional methanol molecules attached to the ring. For the methyl groups the situation is once again different: the D-type molecules have the lowest values that are independent of the size of the cluster or the extent of substitution and the values for the AD- and AAD-type molecules appear to be higher, to overlap and to increase with the extent of substitution. Also in this case, the increase in the values can be ascribed to the rising steric hindrance between the different methyl groups, as more methanol molecules are placed on the ring. The combination of the trends established for the hydroxyl groups and for the methyl groups can be found in the values for the molecule. The values for the different types of molecules are all well separated and increase with the number of hydrogen bonds the molecules form. The values for the D-type molecules are stabilized around an average of 3.45 au, whereas the values for the AD and AAD-types molecules increase with the extent of substitution.

Table 4. Average $\Delta\alpha^{OH}$'s, $\Delta\alpha^{CH_3}$'s, and $\Delta\alpha^{mol}$'s of the Three Types of Methanol Waters in the Different 5r + m Ring Substituted Clusters

	$\Delta\alpha^{OH}$			$\Delta\alpha^{CH_3}$			$\Delta\alpha^{mol}$		
cluster	D	AD	AAD	D	AD	AAD	D	AD	AAD
5r	2.20			0.69			4.84		
5r+1	1.68	2.24	2.95	0.66	0.77	0.66	3.61	4.93	6.40
5r+2	1.57	2.31	2.93	0.41	0.89	1.07	2.99	5.41	6.49
5r+3	1.55	2.35	2.97	0.42	1.11	1.23	3.22	5.70	6.88
5r+4	1.84	2.18	3.00	0.64	1.68	1.31	3.95	6.07	7.07
5r+5	1.79		3.02	0.60		1.56	3.47		7.45

4. Conclusion

In conclusion, the polarizabilities of fifty different methanol clusters containing up to twelve methanol molecules were calculated at the B3LYP/6-311++G** level of theory and subsequently partitioned into atomic and molecular contributions using the Hirshfeld-I method. The obtained molecular polarizabilities were analyzed with respect to the local environment of the molecules in the clusters.

The results demonstrate that in order to understand the trends in the total molecular polarizabilities, the methanol molecules must be considered as a junction of two entities of different nature, namely a hydrophilic hydroxyl group and a hydrophobic methyl group. The polarizabilities of both groups exhibit a behavior that is fundamentally different from each other, but consequently throughout the total collection of the clusters, resulting in a somewhat complicated picture when superpositioned together to form the total polarizability of the molecule.

The polarizabilities of the hydroxyl group is in line with the findings that were obtained in a previous study on water clusters.[36] The values are strongly dependend on the hydrogen bond network that the molecules take part in, decreasing with a rising number of hydrogen bonds. The polarizabilities also tend to decrease with the strength of the hydrogen bond, although the correlation is not as optimal as in the case of the water clusters. This effect emphasizes the complexity of the relation between the local environment and the polarizability, that is far more pronounced in the case of the methanol clusters. In this case the structures are less orderly due to the presence of the methyl group, bringing new parameters into light that influence the polarizability and reduce the high correlation that is present in the case of the water clusters.

On the other hand, the polarizabilities of the methyl groups are not directly influenced by the hydrogen bond network in the cluster and are mostly influenced by steric hindrance effects, that reduce the values. This effect is less extensive numerically, amounting to only 50% of the effect of the hydrogen-bonding network on the polarizabilities of the hydroxyl groups.

As a result of the those deviating behaviors, the polarizabilities of the methanol molecules in the clusters are not as strongly dependent on the hydrogen bond network as was found in the case of water clusters. The effect of steric hindrance contributes to greater deviation in the values within each group of molecules, leading to occasional overlap. The values also tend to decrease with the aggregation number, as steric hindrance is generally more substantial in larger clusters. Another important difference between water and methanol clusters is the absence in the latter of a significant intermolecular charge transfer, which influences the polarizability. The difference between water clusters and methanol clusters will be further explored in future work by examining additional local effects, such as the variation of the local field.

Acknowledgment. A.K. and C.V.A. acknowledge the Flemish FWO for research grant no. G.0629.06. We gratefully acknowledge the University of Antwerp for access to the university's CalcUA supercomputer cluster.

References

(1) Tauer, K. J.; Lipscomb, W. N. *Acta Crystallogr.* **1952**, *5*, 606.

(2) Montague, D. G.; Gibson, I. P.; Dore, J. C. *Mol. Phys.* **1981**, *44*, 1355.

(3) Magini, M.; Paschina, G.; Piccaluga, G. *J. Chem. Phys.* **1982**, *77*, 2051.

(4) Narten, A. H.; Habeschuss, A. *J. Chem. Phys.* **1984**, *80*, 3387.

(5) Tanaka, Y.; Ohtomo, N.; Arakawa, K. *Bull. Chem. Soc. Jpn.* **1984**, *57*, 644.

(6) Torrie, B. H.; Weng, S.-X.; Powell, B. M. *Mol. Phys.* **1989**, *67*, 575.

(7) Sarkar, S.; Joarder, R. N. *J. Chem. Phys.* **1993**, *99*, 2032.

(8) Kashtanov, S.; Augustson, A.; Rubensson, J.-E.; Nordgren, J.; Ȧgren, H.; Guo, J.-H.; Luo, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2005**, *71*, 104205.

(9) Wilson, K. R.; Cavalleri, M.; Rude, B. S.; Schaller, R. D.; Catalano, T.; Nilsson, A.; Saykally, R. J.; Pettersson, L. G. M. *J. Phys. Chem. B* **2005**, *109*, 10194.

(10) Suhara, K.; Fujii, A.; Mizuse, K.; Mikami, N.; Kuo, J. L. *J. Chem. Phys.* **2007**, *126*, 194306.

(11) Larsen, R. W.; Zielke, P.; Suhm, M. A. *J. Chem. Phys.* **2007**, *126*, 194307.

(12) Tsuchida, E.; Kanada, Y.; Tsukada, M. *Chem. Phys. Lett.* **1999**, *311*, 236.

(13) Ladanyi, B. M.; Skaf, M. S. *Annu. Rev. Phys. Chem.* **1993**, *44*, 335.

(14) Krishtal, S.; Kiselev, M.; Kolker, A.; Idrissi, A. *Theor. Chem. Acc.* **2007**, *117*, 297.

(15) Wang, J.; Boyd, R. J.; Laaksonen, A. *J. Chem. Phys.* **1996**, *104*, 7261.

(16) Martin, M. E.; Sánchez, M. L.; Olivares del Valle, F. J.; Aguilar, M. A. *J. Chem. Phys.* **2002**, *116*, 1613.

(17) Morrone, A. J.; Tuckerman, M. E. *Chem. Phys. Lett.* **2003**, *370*, 406.

(18) Pires, M. M.; DeTuri, V. F. *J. Chem. Theory Comput.* **2007**, *3*, 1073.

(19) Mó, O.; Yáñez, M.; Elguero, J. *J. Chem. Phys.* **1997**, *107*, 3592.

(20) Mandado, M.; Graña, A. M.; Mosquera, R. A. *Chem. Phys. Lett.* **2003**, *381*, 22.

(21) Vener, M. V.; Sauer, J. *J. Chem. Phys.* **2001**, *114*, 2623.

(22) Ludwig, R. *Chem. Phys. Chem.* **2005**, *6*, 1369.

(23) Boyd, S. L.; Boyd, R. J. *J. Chem. Theory Comput.* **2007**, *3*, 54.

(24) Wessels, J. M.; Rodgers, J. M. A. *J. Phys. Chem.* **1995**, *99*, 17586.

(25) Kumar, P. V.; Maroncelli, M. *J. Chem. Phys.* **1995**, *103*, 3038.

(26) Gharib, F.; Sadeghi, F. *Appl. Organomet. Chem.* **2007**, *21*, 218.

(27) Hirshfeld, F. L. *Theor. Chim. Acta* (*Berl.*) **1977**, *44*, 129.

(28) Rousseau, B.; Peeters, A.; Van Alsenoy, C. *Chem. Phys. Lett.* **2000**, *324*, 189.

(29) Harrison, J. F. *J. Phys. Chem. A* **2005**, *109*, 5492.

(30) Boon, G.; De Proft, F.; Van Alsenoy, C.; Bultinck, P.; Geerlings, P. *J. Phys. Chem.* **2003**, *107*, 11120.

(31) Geerlings, P.; Boon, G.; Van Alsenoy, C.; De Proft, F. *Int. J. Quantum Chem.* **2005**, *101*, 722.

(32) Ayers, P. W.; Morrison, R. C.; Roy, R. K. *J. Chem. Phys.* **2002**, *116*, 8731.

(33) De Proft, F.; Vivas-Reyes, R.; Peeters, A.; Van Alsenoy, C.; Geerlings, P. *J. Comput. Chem.* **2003**, *24*, 463.

(34) Roy, R. K. *J. Phys. Chem.* **2003**, *107*, 10428.

(35) Mandado, M.; Van Alsenoy, C.; Geerlings, P.; De Proft, F.; Mosquera, R. A. *Chem. Phys. Chem.* **2006**, *7*, 1294.

(36) Krishtal, A.; Senet, P.; Mingli, Y.; Van Alsenoy, C. *J. Chem. Phys.* **2006**, *125*, 034312.

(37) Yang, M.; Senet, P.; Van Alsenoy, C. *Int. J. Quantum Chem.* **2005**, *101*, 535.

(38) Bultinck, P.; Van Alsenoy, C.; Ayers, P. W.; Carbo-Dorca, R. *J. Chem. Phys.* **2007**, *126*, 144111.

(39) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.;

Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez C.; Pople, J. A. *Gaussian 03, Revision B.05*; Gaussian, Inc.: Pittsburgh, PA, 2003.

(40) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *10*, 553.

(41) Simon, S.; Duran, M.; Dannenberg. J. J. *J. Chem. Phys.* **1996**, *105*, 11024.

CT700325C

JCTC Journal of Chemical Theory and Computation

GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation

Berk Hess*

Max-Planck Institute for Polymer Research, Ackermannweg 10,
D-55128 Mainz, Germany

Carsten Kutzner

Department of Theoretical and Computational Biophysics, Max-Planck-Institute of
Biophysical Chemistry, Am Fassberg 11, D-37077 Göttingen, Germany

David van der Spoel

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

Erik Lindahl

Stockholm Center for Biomembrane Research, Stockholm University,
SE-10691 Stockholm, Sweden

Received November 7, 2007

Abstract: Molecular simulation is an extremely useful, but computationally very expensive tool for studies of chemical and biomolecular systems. Here, we present a new implementation of our molecular simulation toolkit GROMACS which now both achieves extremely high performance on single processors from algorithmic optimizations and hand-coded routines and simultaneously scales very well on parallel machines. The code encompasses a minimal-communication domain decomposition algorithm, full dynamic load balancing, a state-of-the-art parallel constraint solver, and efficient virtual site algorithms that allow removal of hydrogen atom degrees of freedom to enable integration time steps up to 5 fs for atomistic simulations also in parallel. To improve the scaling properties of the common particle mesh Ewald electrostatics algorithms, we have in addition used a Multiple-Program, Multiple-Data approach, with separate node domains responsible for direct and reciprocal space interactions. Not only does this combination of algorithms enable extremely long simulations of large systems but also it provides that simulation performance on quite modest numbers of standard cluster nodes.

I. Introduction

Over the last few decades, molecular dynamics simulation has become a common tool in theoretical studies both of simple liquids and large biomolecular systems such as proteins or DNA in realistic solvent environments. The computational complexity of this type of calculations has historically been extremely high, and much research has therefore focused on algorithms to achieve single simulations that are as long or large as possible. Some of the key early work was the introduction of holonomic bond length constraints[1] and rigid-body water models[2,3] to enable longer integration time steps. However, one of the most important general developments in the field was the introduction of parallel molecular simulation implementations during the late

10.1021/ct700301q CCC: $40.75 © 2008 American Chemical Society
Published on Web 02/02/2008

1980s and early 1990s.[4−7] The NAMD program by the Schulten group[8] was the first to enable scaling of large molecular simulations to hundreds of processors, Duan and Kollman were able to complete the first microsecond simulation of a protein by creating a special parallel version of Amber, and more recently Fitch et al. have taken scaling to the extreme with their BlueMatter code which can use all tens of thousands of nodes on the special BlueGene hardware.[9]

On the other hand, an equally strong trend in the field has been the change of focus to statistical properties like free energy of solvation or binding of small molecules and, e.g., protein folding rates. For this class of problems (limited by sampling) the main bottleneck is single-CPU performance, since it is typically always possible to achieve perfect scaling on any cluster by starting hundreds of independent simulations with slightly different initial conditions. This has always been a central theme in GROMACS development and perhaps best showcased by its adoption in the Folding@Home project, where it is running on hundreds of thousands of independent clients.[10] GROMACS achieves exceptional single-CPU performance because of the manually tuned SSE, SSE2, and ALTIVEC force kernels, but there are also many algorithmic optimizations, for instance single-sum virials and strength-reduced algorithms to allow single-precision floating-point arithmetic in all places where it still conserves energy (which doubles memory and cache bandwidth).[11,12] In the benchmark section we show that GROMACS in single precision matches the energy conservation of a double precision package.

Unfortunately it is far from trivial to combine raw single-CPU performance and scaling, and in many cases there are inherent tradeoffs. It is for instance straightforward to constrain all bond lengths on a single CPU, but in parallel it is usually only applied to bonds involving hydrogens to avoid (iterative) communication, which in turn puts a lower limit on the possible time step.

In this paper, we present a completely reworked parallelization algorithm that has been implemented in GROMACS. However, rather than optimizing relative scaling over *N* CPUs we have focused on (*i*) achieving the highest possible absolute performance and (*ii*) doing so on as few processors as possible since supercomputer resources are typically scarce. A key challenge has therefore been to make sure all algorithms used to improve single-CPU performance through longer time steps such as holonomic bond constraints, replacing hydrogens with virtual interaction sites,[13] and arbitrary triclinic unit cells also work efficiently in parallel.

GROMACS was in fact set up to run in parallel on 10Mbit ethernet from the start in 1992[7] but used a particle/force decomposition that did not scale well. The single-instruction-multiple-data kernels we introduced in 2000 made the relative scaling even worse (although absolute performance improved significantly), since the fraction of remaining time spent on communication increased. A related problem was load imbalance; with particle decomposition one can frequently avoid imbalance by distributing different types of molecules uniformly over the processors. Domain decomposition, on the other hand, requires automatic load balancing to avoid deterioration of performance. This load imbalance typically occurs in three cases: The most obvious reason is an uneven distribution of particles in space, such as a system with a liquid−vapor coexistence. A second reason is imbalance due to different interaction densities. In biomolecular systems the atom density is usually nearly uniform, but when a united-atom forcefield is used hydrocarbon segments (e.g., in lipid chains) have a three times lower particle density and these particles have only Lennard-Jones interactions. This makes the computation of interactions of a slab of lipids an order of magnitude faster than a slab of water molecules. Interaction density imbalance is also an issue with all-atom force fields in GROMACS, since the program provides optimized water−water loops for standard SPC/TIP3P/TIP4P waters with Lennard-Jones interactions only on the oxygens.[2,3] (In principle it is straightforward to introduce similar optimization for the CHARMM-style modified TIP water models with Lennard-Jones interactions on the hydrogens too, but since there is no clear advantage from the extra interactions we have not yet done so.) A third reason for load imbalance is statistical fluctuation of the number of particles in a domain decomposition cell. This primarily plays a role when cells only contain a few hundred atoms.

Another major issue for simulation of large molecules such as proteins was the fact that atoms connected by constraints could not be split over processors (holonomic constraints) a problem shared with all other biomolecular simulation packages (the alternative being shorter time-steps, possible coupled with multiple-time-step integration). This issue is more acute with domain decomposition, since even small molecules in general do not reside in a single domain.

Finally, the last challenge was the nonimpressive scaling of the Particle Mesh Ewald (PME) electrostatics[14] as implemented in the previous GROMACS version. Since PME involves two 3D fast Fourier transforms (FFTs), it requires global all-to-all communication where the number of messages scale as the square of the number of nodes involved. There have been several attempts at parallelizing PME using iterative solvers instead of using FFTs. A different algorithm that reduces communication is fast multipole expansion.[15] However, presently none of these methods combine the efficiency of PME using FFTs with good scaling up to many processors.

We have addressed these four issues by devising an eighth-shell domain decomposition method coupled to a full dynamic load-balancing algorithm with a minimum amount of communication, a parallel version of the constraint algorithm LINCS that enables holonomic constraints without iterative communication, and splitting off the PME calculation to dedicated PME processors. These four key advances will be described in the next three sections, followed by a description of other new features and a set of benchmarks to illustrate both absolute performance and relative scaling.

II. Domain Decomposition

Recently, the D. E. Shaw group has performed several studies into general zonal methods[16] for parallelization of particle-based interactions. In zonal (or neutral territory) methods, forces between particles *i* and *j* are not necessarily calculated on a processor where either of particles *i* or *j* resides. Somewhat paradoxically, such methods can be significantly more efficient than traditional domain decomposition methods since they reduce the total amount of data communicated. Two methods achieve the least communication when the domain size is not extremely small compared to the cutoff radius; these two methods were termed *eighth shell*[17] and *midpoint* methods[18] by Shaw and co-workers. In the half shell method, interactions between particle *i* and *j* are calculated in the cell where *i* or *j* resides. The minimum communication required for such a method is half of the volume of a boundary of a thickness equal to the cutoff radius. The eighth shell method improves on this by also calculating interactions between particles that reside in different communicated zones. The communicated volume of the eighth shell method is thus a subset of that of the half shell method, and it also requires less communication steps which helps reduce latency.

The basic eighth shell method was already described in 1991 by Liem et al.,[19] who implemented communication with only nearest neighbors. In GROMACS 4 we have extended this method for communication with multiple cells and staggered grids for dynamic load balancing. The Shaw group has since chosen to use the midpoint method in their Desmond code since it can take advantage of hardware where each processor has two network connections that simultaneously send and receive. After quite stimulating discussions with the Shaw group we chose not to switch to the midpoint method, primarily not only because we avoid the calculation of the midpoint, which has to be determined binary identically on multiple processors, but also because not all hardware that GROMACS will run on has two network connections. With only one network connection, a single pair of send and receive calls clearly causes less latency than two such pairs of calls.

Before going into the description of the algorithm, the concept of charge groups needs to be explained; these were originally introduced to avoid electrostatic artifacts. By grouping several partially charged atoms of a chemical group into a neutral charge group, charge−charge interactions entering and leaving the cutoff are effectively replaced by short-range dipole−dipole interactions. The location of a charge group in GROMACS is given by the (non-mass-weighted) average of the coordinates of the atoms. With the advent of the PME electrostatics method this is no longer an issue. But charge groups can also speed up the neighbor search by an order of magnitude; given a pair of water molecules for instance, we only need to determine one distance instead of nine (or sixteen for a four-site water model). This is particularly important in GROMACS since the neighbor searching is much slower than the force loops, for which we typically use tuned assembly code. Since charge groups are used as the basic unit for neighbor searching, they also need to be the basic unit for the domain decomposition. In GROMACS 4, the domains are rebuilt every time neighbor searching is performed, typically every 10 steps.

The division of the interactions among processors is illustrated in Figure 1. Consider the processor or cell that has the charge groups in zone 0 as home charge groups, i.e., it performs the integration of the equations of motion for the particles in these charge groups. In the eighth shell method each cell should determine the interactions between pairs of charge groups of which, for each dimension, the minimum cell index of the two charge groups corresponds to the index of that cell. This can be accomplished by the following procedure. Cell 0 receives the coordinates of the particles in the dashed zones 1 to 7, by communication only in one direction for each dimension. When all cells dimensions are larger than the cutoff, each zone corresponds to part of a single, neighboring cell. But in general many cells can contribute to one zone. Each processor calculates the interactions between charge groups of zone 0 with zones 0 to 7, of zone 1 with zones 3 to 6, of zone 2 with zone 5, and of zone 3 with zones 5 and 6. If this procedure is applied for all processors, all pair interactions within the cutoff radius are calculated.

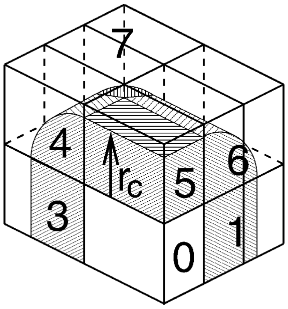


Figure 1. A nonstaggered domain decomposition grid of 3 × 2 × 2 cells. Coordinates in zones 1 to 7 are communicated to the corner cell that has its home particles in zone 0. r_c is the cutoff radius.

Interactions involving three or more atoms cannot be distributed according to the scheme described above. Bonded interactions are distributed over the processors by finding the smallest *x*, *y*, and *z* coordinate of the charge groups involved and assigning the interaction to the processor with the home cell where these smallest coordinates reside—note that this does not require any extra communication between the processors. This procedure works as long as the largest distance between charge groups involved in bonded interactions is not larger than the smallest cell dimension. To check if this is the case, we count the number of assigned bonded interactions during domain decomposition and compare it to the total number of bonded interactions in the system. When there are only two cells in a certain dimension and the corresponding box length is smaller than four times the cutoff distance, a cutoff criterion is required for any pair of particles involved to avoid that bonded interactions are assigned to multiple cells. Unlike the midpoint method, this procedure limits the distances involved in bonded interactions to the smallest cell dimension. For atomistic simulations this is not an issue, since distances in bonded interactions are usually smaller than 0.5 nm, leading to a limit of 10 to 20 atoms per cell, which is beyond the scaling of GROMACS 4. For coarse-grained simulations bonded distances can be larger, but because of the lower interaction density this also does not limit the scaling.

For full dynamic load balancing the boundaries between cells need to be adjusted during the simulation. For 1D

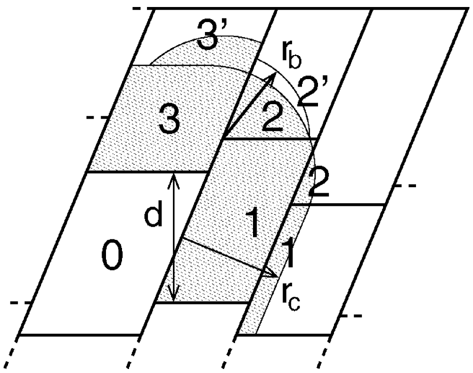


Figure 2. The zones to communicate to the processor of zone 0, see the text for details. r_c and r_b are the nonbonded and bonded cutoff radii, respectively, and *d* is an example of a distance between following, staggered boundaries of cells.

domain decomposition this is trivial, but for a 3D decomposition the cell boundaries in the last two dimensions need to be staggered along the first dimensions to allow for complete load balancing (see the next section for details). Figure 2 shows the communicated zones for 2D domain decomposition in the most general case, namely a triclinic unit cell with dynamic load balancing. Zones 1, 2, and 3 indicate the parts of neighboring cells that are within the nonbonded cutoff radius r_c of the home cell of zone 0. Without dynamic load balancing this is all that would need to be communicated to the processor of zone 0. With dynamic load balancing the staggering can lead to an extra volume 3′ that needs to be communicated, due to the nonbonded interactions between cells 1 and 3 which should be calculated on the processor of cell 0. For bonded interactions, zones 1 and 2 might also require expansion. To ensure that all bonded interaction between charge groups can be assigned to a processor, it is sufficient to ensure that the charge groups within a sphere with a radius r_b, the cutoff for bonded interactions, are present on at least one processor for every possible center of the sphere. In Figure 2 this means we also need to communicate volume 2′. When no bonded interactions are present between charge groups, such volumes are not communicated. For 3D domain decomposition the picture becomes quite a bit more complicated, but the procedure is analogous apart from more extensive bookkeeping. All three cases have been fully implemented for general triclinic cells. GROMACS 4 does not (yet) take full advantage of the reduction in the communication due to rounding of the zones. Currently zones are only rounded in the ‘forward’ directions, for example part 3′ in Figure 2 is replaced by the smallest parallelogram enclosing it.

The communication of the coordinates and charge group indices can be performed efficiently by ‘pulsing’ the information in one direction simultaneously for all cells one or more times. This needs to be repeated for each dimension. The number of pulses n_p in a dimension is given by the cutoff length in that direction divided by the minimum cell size. In most cases n_p will be one or two. Consider a 3D domain decomposition where we decompose in the order *x*, *y*, *z*; meaning that the *x* boundaries are aligned, the *y* boundaries are staggered along the *x* direction, and the *z* boundaries are staggered along the *x* and *y* directions. Each processor first sends the zone that its neighboring cell in $-z$ needs to this cell. This process is done $n_p(z)$ times. Now each processor can send the zone its neighboring cell in $-y$ needs, plus the part of the zone it received from $+z$, that is also required by the neighbor in $-y$. The last step consists of $n_p(x)$ pulses in $-x$ where (parts of) 4 zones are sent over. In this way $n_p(x) + n_p(y) + n_p(z)$ communication steps are required to communicate with $n_p(x) \times n_p(y) \times n_p(z) - 1$ processors, while no information is sent over that is not directly required by the neighboring processors. The communication of the forces happens according to the same procedure but in reversed order and direction.

Another example of a minor complication in the communication is virtual interaction sites constructed from atoms in other charge groups. This is used in some polymer (anisotropic united atom) force fields, but GROMACS can also employ virtual sites to entirely remove hydrogen vibrations and construct the hydrogens in their equilibrium positions from neighboring heavy atoms each time step.[13] Since the constructing atoms are not necessarily interacting on the same node, we have to track the virtual site coordinate dependencies separately to make sure they are both available for construction and that forces are properly communicated back. The communication for virtual sites is also performed with pulses but now in both directions. Here only one pulse per dimension is required, since the distances involved in the construction of virtual sites are at most two bond lengths.

III. Dynamic Load Balancing

Calculating the forces is by far the most time-consuming part in MD simulations. In GROMACS, the force calculation is preceded by the coordinate communication and followed by the force communication. We can therefore balance the load by determining the time spent in the force routines on each processor and then adjusting the volume of every cell in the appropriate direction. The timings are determined using inline assembly hardware cycle counters and supported for virtually all modern processor architectures. For a 3D decomposition with order *x*, *y*, *z* the load balancing algorithm works as follows: First the timings are accumulated in the *z* direction to the processor of cell $z = 0$, independently for each *x* and *y* row. The processor of $z = 0$ sums these timings and sends the sum to the processor of $y = 0$. This processor sums the timings again and sends the sum to the processor of $x = 0$. This processor can now shift the *x* boundaries and send these to the $y = 0$ processors. They can then determine the *y* boundaries, send the *x* and *y* boundaries to the $z = 0$ processors, which can then determine *z* boundaries, and send all boundaries to the processors along their *z* row. With this procedure only the necessary information is sent to the processors that need it and global communication is avoided.

As mentioned in the Introduction, load imbalance can come from several sources. One needs to move boundaries in a conservative fashion in order to avoid oscillations and instabilities, which could for instance occur due to statistical fluctuations in the number of particles in small cells. Empirically, we have found that scaling the relative lengths of the cells in each dimension with 0.5 times the load imbalance, and a maximum scaling of 5%, produced efficient and stable load balancing. For large numbers of cells or inhomogeneous systems two more checks are required: A first restriction is that boundaries should not move more than halfway an adjacent cell (where instead of halfway one could also choose a different value). This prevents cells from moving so far that a charge group would move two cells in a single step. It also prevents load balancing issues when there are narrow zones of high density in the system. A second problem is that due to the staggering, cell boundaries along neighboring rows could shift to such an extent that additional cells would enter the cutoff radius. For load balanced simulations the user can set the minimum allowed cell size, and by default the nonbonded cutoff radius is used. The distance between following, staggered cell boundaries (as indicated by *d* in Figure 2) should not be smaller than this minimum allowed cell size. To ensure this, we limit the new position of each boundary to the old limit plus half the old margin. In this way we make sure that one boundary can move up and independently an adjacent staggered boundary can move down, without extra communication. The neighboring boundaries are communicated after load balancing, since they are needed to determine the zones for communication. When pressure scaling is applied, the limits are increased by 2% to allow the system to adjust at the next domain decomposition before hitting the cutoff restrictions imposed by the staggering.

In practical tests, load imbalances of a factor of 2 on several hundreds of processors were reduced to 2% after a few load balancing steps or a couple of seconds of simulation time.

IV. Parallel Holonomic Constraints

There are two strong reasons for using constraints in simulations: First, a physical reason that constraints can be considered a more faithful representation of chemical bonds in their quantum mechanical ground state than a classical harmonic potential. Second, a practical reason because rapid bond vibrations limit the time step. Removing these vibrations by constraining the bonds thus allows us to increase the time step and significantly improve absolute simulation performance. A frequently used rule-of-thumb is 1 fs without constraints, 1.4 fs with bonds to hydrogens constrained, and 2 fs when all bonds are constrained. Unfortunately, the common SHAKE[1] constraint algorithm is iterative and therefore not very suitable for parallelization—in fact, there has previously not been any efficient algorithm that could handle constraints connected over different processors due to domain decomposition. Most biomolecular packages therefore use constraints only for bonds involving hydrogens.

By default, GROMACS uses a noniterative constraints algorithm called *LINear Constraint Solver* (LINCS), which proved much easier to fully parallelize as hinted already in the original paper.[20] The details of the parallel LINCS algorithm P-LINCS are described elsewhere,[21] so we will only give a brief overview here. In the algorithm, the range of influence of coupled constraints is set by the order of the expansion for the matrix inversion. It is only necessary to communicate a subset of the old and new unconstrained coordinates between neighboring cells before applying the constraints. The atoms connected by up to “one plus the expansion order” bonds away need to be communicated. We can then constrain the local bonds plus the extra bonds. The communicated atoms will not have the final correctly constraint positions (since they interact with additional neighbors), but the local atoms will. The beauty of the algorithm is that normal molecular simulation only requires a first, linear correction and a single iterative step. In both these steps updated positions are communicated and adjustment forces calculated locally. The constraint communication can be accomplished with a single forward and backward pulse of the decomposition grid in each dimension, similar to the domain decomposition communication. The results of P-LINCS in GROMACS are binary identical to those of the single processor version.

In principle a similar method could be used to parallelize other constraint algorithms. However, apart from multiple communication steps for iterative methods such as SHAKE,[1] another problem is that one does not know a priori which atoms need to be communicated, because the number of iterations is not fixed. To our best knowledge, this is the first efficient implementation of an holonomic constraint algorithm for domain decomposition.[21]

The accuracy of the velocities of constrained particles has further been improved both for LINCS and SHAKE using a recently described procedure based on Lagrange multipliers.[22] For SETTLE[23] we have applied the slightly less accurate method of correcting the velocities with the position corrections divided by the time step. These changes significantly improve long-term energy conservation in GROMACS, in particular for single precision simulations.[21] With domain decomposition, SHAKE and SETTLE can only be used for constraints between atoms that reside in the same charge group. SETTLE is only used for water molecules though, which are usually a single charge group anyway.

The virtual interaction sites described earlier require rigid constraint constructs, and the implementation of parallel holonomic constraints was therefore critical to enable virtual sites with parallel domain decomposition. This enables the complete removal of hydrogen angle vibrations, which is normally the next fastest motion after bond length oscillations. Full rotational freedom of CH_3/NH_2/NH_3 groups is still maintained by using dummy mass sites,[13] which enables time steps as long as 5 fs. It has been shown that removing the angle vibrations involving hydrogens has a minor effect on the geometry of intraprotein hydrogen bonds and that properties such as the number of hydrogen bonds, dihedral distributions, secondary structure, and rmsd are not affected.[13] Note that simply constraining all angles involving hydrogens effectively also constrains most of the other angles in a molecule, which would affect the dynamics of molecules significantly.[24] In contrast, replacing hydrogens by virtual interaction sites does not affect the angular degrees of freedom involving heavy atoms. This hydrogen-removal procedure generates uncoupled angle constraints for hydrogens in alcohol groups. These angle constraints converge twice as slow in LINCS as normal constraints. To bring the accuracy of uncoupled angle constraints up to that of bond constraints, the LINCS expansion order for angle constraints has been doubled (see the P-LINCS paper[21] for details). In the benchmark section we show that a time step of 4 fs does not deteriorate the energy conservation.

V. Optimizing Memory Access

The raw speed of processors in terms of executing instructions has increased exponentially with Moore's law. However, the memory access latency and bandwidth has not kept up with the instruction speed. This has been partially compensated by added fast cache memory and smart caching algorithms. But this only helps for repeated access of small blocks of memory. Random access of large amounts of memory has become relatively very expensive. In molecular dynamics simulations of fluid systems, particles diffuse over time. So even when starting out with an ordered system, after some time particles that are close in space will no longer be close in memory. This results in random memory access through the whole coordinate array during the neighbor search, force calculation, and the PME charge and force assignment. Meloni et al. have shown that spatially ordering atoms can significantly improve performance for a Lennard-Jones system.[25]

We have implemented a sorting scheme that improves upon that of Meloni et al. by ordering the charge groups according to their neighbor search cell assignment. Ordering using the neighbor search cell assignment provides the optimal memory access order of atoms during the force calculations. In this way, nearly all coordinates in memory are used along a cell row with a fixed minor index. For major indices there are some jumps, but the number of jumps is now the number of different major row indices instead of the number of charge group pairs. Effectively each part of the coordinate array needs to be read from memory to cache only once, instead M^2 times where M is the total number of charge groups divided by the number of charge groups that fit in cache. This approach requires that the charge groups are resorted at every step where neighbor searching is performed. For optimal performance with PME, the major and minor dimensions for the indexing of the neighbor search cells and the PME grid should match.

A second reason for ordering is to allow for exact rerunning of part of a simulation. Due to the domain decomposition the order of the local charge groups on each processor changes. This order affects the rounding of the least significant bit in the summation of forces. To exactly reproduce part of a simulation the local atom order should be reproducible when restarting at any point in time. To define a unique order, we sort the charge groups within each neighbor search cell according to the order in the topology. Since charge groups only move a short distance between neighbor list updates, few particles cross cell boundaries, and the sorting can be done efficiently with a linear algorithm.

Optimization of memory access becomes particularly important in combination with the assembly kernels, since the SIMD instructions are extremely fast and therefore memory access can be a significant bottleneck. To quantify this we have simulated a 2 M NaCl(aq) solution[26] using SPC/E water[27] with reaction-field and PME electrostatics. The effect of the sorting is shown in Table 1. The sorting ensures a nearly constant performance, independent of the system size. Without sorting there is a 10% performance degradation at 10^4 atoms per core and a factor of 2 at 2−3 × 10^5 atoms. For a Lennard-Jones system of 10^5 atoms the difference is a factor of 4. Note that sorting actually decreases the scaling efficiency with the number of processors, since for low parallelization (more atoms per processor) the absolute performance increases more than for high parallelization, but it obviously always helps absolute performance.

Table 1. Number of MD Steps per Second with and without Spatial Sorting of Charge Groups[a]

		number of atoms per core			
sorting	electrostatics	1705	8525	34100	272800
yes	reaction field	241	48.5	11.9	1.39
no	reaction field	238	44.0	9.6	0.60
yes	PME	102	22.5	5.4	0.61
no	PME	101	20.6	4.8	0.33

[a] As a function of the number of atoms per core for a 2 M NaCl(aq) solution on a 2.2 GHz AMD64 CPU.

VI. Multiple-Program, Multiple-Data PME Parallelization

The typical parallelization scheme for molecular simulation and most other codes today is Single-Program, Multiple-Data (SPMD) where all processors execute the same code but with different data. This is an obvious solution to decompose a system containing hundreds of thousands of similar particles. However, particularly for the now ubiquitous PME algorithm this approach has some drawbacks: First, the direct space interactions handled through classical cutoffs and the reciprocal space lattice summation are really independent and could be carried out in parallel rather than partitioning smaller work-units over more processors. Second, the scaling of PME is usually limited by the all-to-all communication of data during the parallel 3D FFT.[28] While the total bandwidth is constant, the number of messages and latencies grow as N^2, where N is the number of nodes over which the FFT grid is partitioned.

Apart from rewriting and tuning the parallel PME algorithm to support domain decomposition, we have addressed this problem by optionally supporting Multiple-Program, Multiple-Data (MPMD) parallelization where a subset of processors are assigned as dedicated PME processors, while the direct space interactions and integration are domain decomposed over the remaining processors. On most networks the newly added communication step between real and reciprocal space processors is more than compensated by better 3D FFT scaling when the number of nodes involved in the latter is reduced a factor of 3−4. The optimal ratio for real space to reciprocal processors is usually between 2:1 and 3:1. Good load balancing for a given ratio can be reached by moving interactions between direct and reciprocal space to ensure load balance, as long as the real space cutoff and grid cell size are adjusted by the same factor the overall accuracy remains constant.[14] In future versions of GROMACS this procedure may be automated.

We assume that the PME processor count is never higher than the number of real space processors. In general, each PME processor will receive coordinates from a list of real space peers, after which the two sets of nodes start working on their respective (separate) domains. The PME processors communicate particle coordinates internally if necessary, perform charge spreading on the local grid, and then communicate overlapping grid parts with the PME neighbors. The actual FFT/convolution/iFFT is performed the standard way but now involving much fewer nodes. After force interpolation the forces corresponding to grid overlap are communicated to PME neighbors again, after which we synchronize and send communicate all forces back to the corresponding real space processors (energy and virial terms only need to be communicated to one of the processors).

With current multicore processors and multisocket motherboards the MPMD approach is particularly advantageous. The costly part is the redistribution of the 3D FFT grid, which is done twice for the forward and twice for the backward transform. This redistribution requires simultaneous communication between all PME nodes, which occurs when the real space nodes are not communicating, and to make use of this GROMACS interleaves the PME processors with the real space processors on nodes. Thus, on a machine where two cores share a network connection, with MPMD only one PME process uses a single network connection instead of two PME processes, and therefore the communication speed for the 3D FFT is doubled. For a real space to PME processor ratio of 3:1, with four cores sharing a network connection, MPMD quadruples the communication speed for the 3D FFT, while simultaneously decreasing the number of process pairs that need to exchange FFT grid information by a factor 16.

VII. The MD Communication

Previous GROMACS versions used a ring communication topology, where half of the coordinates/forces were sent over half the ring. To be frank, the only thing to be said in favor of that is that it was simple. Figure 3 shows a flowchart of the updated communication that now relies heavily on collective and synchronized communication calls available, e.g., in MPI. Starting with the direct space domain (left), each node begins by communicating coordinates necessary to construct virtual sites and then constructs these. At the main coordinate communication stage, data are first sent to peer PME nodes that then begin their independent work. In direct space, neighboring nodes exchange coordinates according to the domain decomposition, calculate interactions, and then communicate forces. Since the PME virial is calculated in reciprocal space, we need to calculate the direct space virial before retrieving the forces from the PME nodes. Finally, the direct space nodes do integration, parallel constraints (P-LINCS), and energy summation. The reciprocal domain nodes start their work when they get updated coordinates from their peer direct space nodes and exchange data with their neighbors to achieve a clean 1D decomposition of the charge grid. After spreading the charges the overlapping parts are communicated and summed, and 3D FFT, convolution, and 3D inverse FFT are performed in parallel. Finally local forces are interpolated, communicated back to the correct PME processor, and sent back to the direct space processor it came from. Whenever possible we use collective MPI operations, e.g., to enable binary-tree summation, and pulsing operations use combined send-receive operations to fully utilize torus networks present on hardware such as IBM BlueGene or Cray XT4.

VIII. Other New Features

Previously, GROMACS only supported neighbor list updates at fixed intervals, but the use of potentials that are switched exactly to zero at some finite distance is increasing, mainly to avoid cutoff effects. To be sure that no interaction is missed, the neighbor list can be updated heuristically in GROMACS 4. The neighbor list is then updated when one or more atoms have moved a distance of more than half the buffer size from the center of geometry of the charge group they belong to, as determined at the last neighbor search (note that without charge groups this is just the position of the atom at the last neighbor search). Coordinate scaling due to pressure coupling is taken into account.

GROMACS can now also be used very efficiently for coarse-grained simulations (see benchmarks section) or many nonstandard simulations that require special interactions. User defined nonbonded interactions that can be assigned independently for each pair of charge groups were already supported, and we have now additionally implemented user defined bonds, angles, and dihedrals functions. Thus, a user now has full control over functional form as well as the parameters of all interactions. Just as for the tabulated nonbonded interactions, cubic spline interpolation is used, which provides continuous and consistent potentials and forces.

In addition to systems without periodic boundaries and with full 3D periodicity, systems with only 2D periodicity in *x* and *y* are now also supported. The 2D periodicity can be combined with one or two uniform walls at constant-*z* planes. The neighbor searching still uses a grid for dimensions *x* and *y* and with two walls, also in *z*, for optimal efficiency. The walls are represented by a potential that works only in the *z*-direction, which can be, e.g., 9−3, 10−4, or a user defined tabulated potential, with coefficients set individually for each atom type.

Restraining (using an umbrella potential) or constraining the center(s) of mass of a group or groups of atoms can now be done in parallel. One can restrain or constrain absolute positions or relative distances between groups. The center of mass of a group of atoms can be ill-defined in a periodic system. To determine the center of mass a reference atom is chosen for each group. The center of mass of each group relative to its reference atom is then determined, and the position of the reference atom is added to obtain the center of mass position. This provides a unique center of mass, as long as all atoms in the group are within half the smallest box dimension of the reference atom. Since there are no a priori limits on the distances between atoms in a group, global communication is required. There are two global communication steps: one to communicate the reference atom positions and one to sum the center of mass contribu-

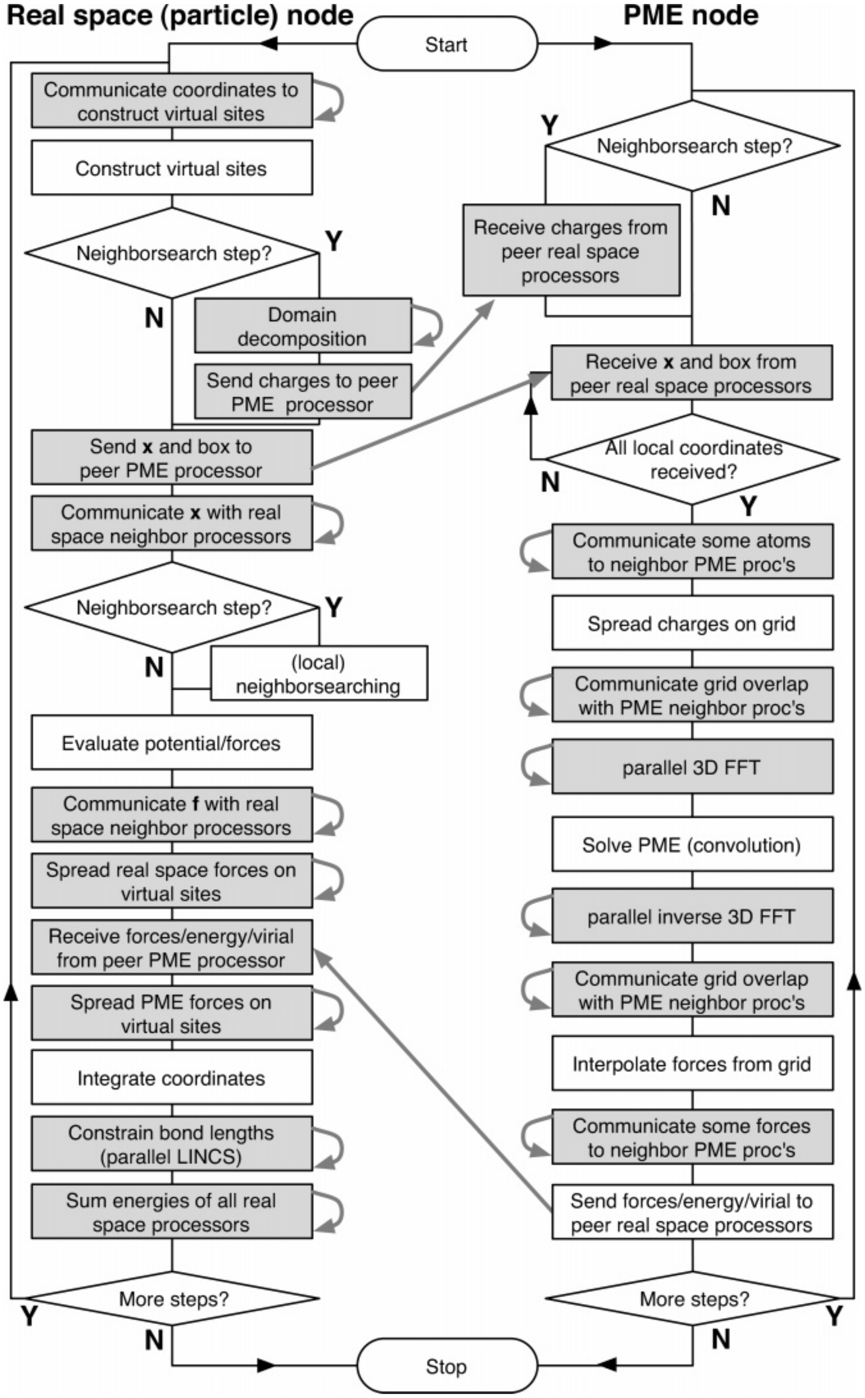

Real space (particle) node
PME node
Start
Communicate coordinates to construct virtual sites
Construct virtual sites
Neighborsearch step?
Y
N
Domain decomposition
Send charges to peer PME processor
Send **x** and box to peer PME processor
Communicate **x** with real space neighbor processors
Neighborsearch step?
Y
N
(local) neighborsearching
Evaluate potential/forces
Communicate **f** with real space neighbor processors
Spread real space forces on virtual sites
Receive forces/energy/virial from peer PME processor
Spread PME forces on virtual sites
Integrate coordinates
Constrain bond lengths (parallel LINCS)
Sum energies of all real space processors
More steps?
Y
N
Neighborsearch step?
Y
N
Receive charges from peer real space processors
Receive **x** and box from peer real space processors
All local coordinates received?
N
Y
Communicate some atoms to neighbor PME proc's
Spread charges on grid
Communicate grid overlap with PME neighbor proc's
parallel 3D FFT
Solve PME (convolution)
parallel inverse 3D FFT
Communicate grid overlap with PME neighbor proc's
Interpolate forces from grid
Communicate some forces to neighbor PME proc's
Send forces/energy/virial to peer real space processors
More steps?
Y
N
Stop

Figure 3. Flowchart for a typical simulation step for both particle and PME nodes. Shaded boxes involve communication, with gray arrows indicating whether the communication only involves similar types of nodes or synchronization between the two domains.

tions over the cells. The restraint or constraint force calculation can then be performed locally.

IX. Benchmarks

The presented benchmarks were performed in the NVT ensemble, using a reversible Nosé-Hoover leapfrog integrator,[29] single precision and dynamic load balancing, unless stated otherwise. Single precision position, velocity and force vectors, combined with some essential variables in double precision is accurate enough for most purposes. In the P-LINCS paper[21] it is shown that with single precision and the constraint velocity correction using the Lagrange multipliers, the energy drift can be reduced to a level unmeasurable over 1 nanosecond. If required, GROMACS can also be compiled in full double precision.

First we will examine the scaling of the basic domain decomposition code, without communication for constraints and virtual sites. To illustrate the basic scaling for all-atom type force fields, we used an OPLS all-atom methanol model,[30] which leads to an interaction density close to that inside a protein. The results for weak scaling, i.e., when the system size grows proportionally with the number of CPUs,

Table 2. Performance in MD Steps per Second for 200 Methanol Molecules (1200 Atoms) per Core[a]

					number of cores				
elec.	prec.	CPU	GHz	cpn	1	2	8	32	128
RF	S	AMD	2.2	8	167	168	166		
	S	Intel	2.33	8	211	216	214		
	S	Intel	3.0	4	274	281	277	265	237
	S	Intel	3.0	2	274	281	284	284	274
RF3	S	Intel	3.0	4	272	274	208	87	44
RF	D	Intel	3.0	4	167	169	159	144	123
	D	Intel	3.0	2	167	169	165	161	153
PME	S	AMD	2.2	8	103	101	98		
	S	Intel	2.33	8	128	127	122		
	S	Intel	3.0	4	172	172	156	150	134
	S	Intel	3.0	2	172	172	162	152	145
PME	D	Intel	3.0	4	112	110	95	90	85
	D	Intel	3.0	2	112	110	100	91	90

[a] With a cutoff of 1 nm, with reaction field (RF), reaction field with GROMACS 3.3 (RF3) and PME with a grid-spacing of 0.121 nm, in single (S) and double (D) precision on AMD64 and Intel Core2 machines with 8 cores per node (cpn) or 4 and 2 cores per node with Infiniband.

are shown in Table 2. With reaction-field electrostatics the computational part of the code scales completely linear. When going from 1 to 2 or 8 cores frequently superlinear scaling can be observed, this is primarily because the charge group sorting is not implemented for single processor simulations. Without PME, the scaling is close to linear, unlike GROMACS 3.3 which already slows down by a factor of 3 on 32 cores. The small drop in performance at 128 processors is caused by the local coordinate and force communication, especially in double precision, and by the global communication for the summation of energies, which is required for temperature coupling. The time spent in the summation increases with the number of processors, since there are more processors to sum over. Unfortunately MPI implementations are often not optimized for the currently typical computing clusters: multiple cores sharing a network connection. With MVAPICH2 on 16 nodes with 4 cores each, the MPI_Allreduce () call takes 120 μs; when we replaced this single call by a two-step procedure, first within each node and then between the nodes, the time is reduced to 90 μs. This global communication is unavoidable for any algorithm that uses global temperature and/or pressure coupling, but the severity depends on the MPI implementation quality. With PME electrostatics linear scaling is impossible, since PME inherently scales as N log(N). However, in practice the scaling of PME is limited more by the communication involved in the 3D-FFT. However, as evident from Table 2, scaling with PME is still very good, particularly when the high absolute performance is taken into account. Furthermore the difference between 2 and 4 cores per node is quite small. This is because for the communication between the PME processes there is no difference in network speed, as in both cases there is only one PME process per node. With 4 cores per node the real space process to PME process communication all happens within nodes. When one puts the real space and PME processes on separate nodes, the performance with 32 processes decreases by 16%, mainly because each PME process needs to communicate over the network with 3 real space processes while sharing its network connection with 3 other PME processes. Without the MPMD PME implementation the scaling would be much worse, since the FFT grid would need to be redistributed over 4 times as many processors. Still, the 3D-FFT algorithm is one of the points we will focus future performance work on. When switching from single to double precision the performance is reduced by a factor of 1.6. This is not due to the higher cost of the floating point operations but more due to doubling of the required memory bandwidth, both for the force computation and the communication. The PME mesh part becomes relatively cheaper in double precision; therefore, one could optimize the simulation setup to obtain a slightly higher performance. This has not been done for this benchmark.

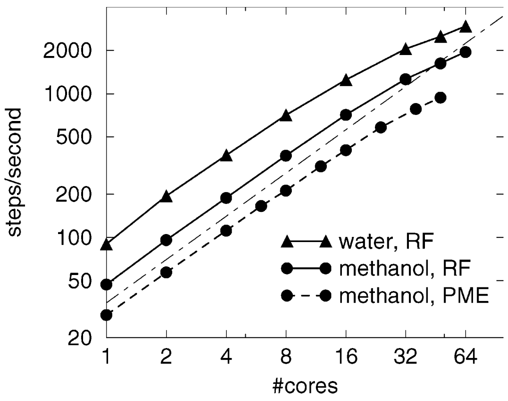


Figure 4. Scaling for a methanol system of 7200 atoms (circles) and an SPC/E water system of 9000 atoms (triangles), with a cutoff 1 nm, with reaction field (solid lines) and PME (dashed line) with a grid-spacing of 0.121 nm (36 × 36 × 36 grid) on a 3 GHz Intel Core2 cluster with Infiniband. The dot-dashed line indicates linear scaling.

To illustrate strong scaling we used the same methanol system mentioned before with 1200 molecules as well as a 3000 SPC/E water[27] system. For water with reaction field the scaling is nearly linear up to 2000 MD steps per second, where there are 200 atoms per core (Figure 4). Without PME, the main bottleneck is the summation of energies over all the processors. For the 3000 water system, the summation of energies over 64 cores takes 17% of the total run time. Water runs about twice as fast as methanol, due to the optimized SSE water loops. With PME, methanol scales in the same way but at about 2/3 of the absolute speed of the reaction-field simulations. In contrast to weak scaling, the relative cost of the latency in the coordinate and force communication increases linearly with the number of processors. However, the summation of the energies is still the final bottleneck, since the *relative* cost of this operation increases faster than linear. Thus, the current limit of about 200 atoms per core is due to the communication latency of the Infiniband network.

It is impossible to quantify the general GROMACS performance for coarse-grained systems, since the different levels and ways of coarse-graining lead to very different types of models with different computational demands. Here, we chose a coarse-grained model for polystyrene that uses

Table 3. Number of Steps per Second for a Coarse-Grained Polystyrene Model[a]

package	thermostat	machine	1	2	8	32	64	96
GROMACS	Nosé-Hoover	3 GHz	126	241	964	2950	4120	
GROMACS	Langevin	3 GHz	106	204	829	2860	4760	6170
GROMACS	Langevin	2.33 GHz	80	155	593			
ESPResSo	Langevin	2.33 GHz	41	85	254			

[a] With 9600 beads as a function of the number of cores on a 3 GHz Intel Core2 cluster with 2 cores per Infiniband connection and an 8 core 2.33 GHz Intel Core2 machine.

nonstandard interactions for the bonded as well as the nonbonded interactions.[31] This model uses 2 beads per repeat unit, which leads to a reduction in particles with a factor of 8 compared to an all-atom model and a factor of 4 compared to a united-atom model. The beads are connected linearly in chains of 96 repeat units with bond, angle, and dihedral potentials. The benchmark system consists of a melt of 50 such chains, i.e., 9600 beads, in a cubic box of 9.4 nm. Since the particle density is 8 times lower and the 0.85 nm neighbor list cutoff shorter than that of an atomistic simulation, the computational load per particle for the nonbonded interactions is roughly 10 times less. For this model, the nonbonded and bonded interactions use roughly equal amounts of computational time. This is the only system for which we did not use dynamic load balancing. Because there are so few interactions to calculate, dynamic load balancing slows down the simulations, especially at high parallelization. The benchmark results with a Nose-Hoover and a Langevin thermostat[32] are shown in Table 3. Also shown is a comparison with the ESPResSo package[33] (Extensible Simulation Package for Research on Soft matter). GROMACS is twice as fast as ESPResSo and shows better scaling. This system scales to more than 6000 MD steps per second. The Langevin integrator used requires four random Gaussian numbers per degree of freedom per integration step. With a simpler integrator, as used by Espresso, the performance increases by 18% one 1 core and by 10% on 96 cores. One can see that at low parallelization Langevin dynamics is less efficient, since generating random numbers is relatively expensive for a coarse-grained system. But above 32 cores, or 300 beads per core, it becomes faster than the Nose-Hoover thermostat. This is because the summation of energies is not required at every step for the local Langevin thermostat. Here one can clearly see that simulations with global thermo- and/or barostats in GROMACS 4 are limited by the efficiency of the MPI_Allreduce() call. With the Langevin thermostat the scaling on an Infiniband cluster is only limited by the latency of the coordinate and force communication.

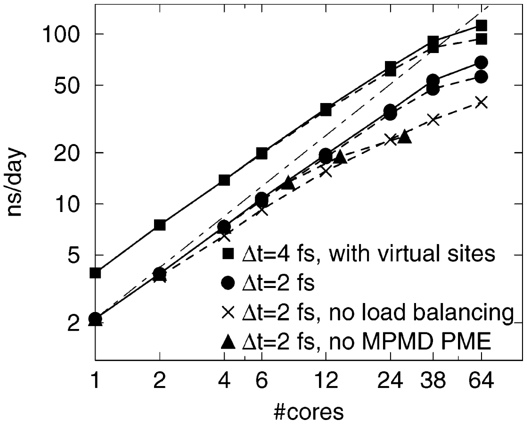


Figure 5. Performance for lysozyme in water (24119 atoms) with OPLS-aa and PME on a 3 GHz dual core Intel Core2 cluster with 2 (solid lines) and 4 (dashed lines) cores per InFIniband interconnect. The dot-dashed line indicates linear scaling.

As a representative protein system, we chose T4-lysozyme (164 residues) and the OPLS all-atom force field. We solvated it in a rhombic dodecahedron (triclinic) unit cell with a minimum image distance of 7 nm, with 7156 SPC/E water molecules and 8 Cl^- ions, giving a total of 24119 atoms. The cutoff was 1 nm, and the neighbor list was updated every 20 fs. For electrostatics we used PME with a grid of 56 × 56 × 56 (0.125 nm spacing). Without virtual sites we used a time step of 2 fs and for LINCS 1 iteration and an expansion order of 4. With virtual sites we used a time step of 4 fs, a single LINCS iteration (expansion order 6). We ran the benchmarks on a 3 GHz Intel Core2 ("Woodcrest") system with Infiniband interconnects. The real space to PME process ratio for this system is 2:1, except for 38 processes (14 PME) and 64 processes (28 PME). This is the only benchmark that actually communicates with more than one cell in each dimension ($n_p = 2$). Results with 2 and 4 cores per Infiniband connection are shown in Figure 5. When all the presented algorithms are used, the scaling is close to linear up to 38 processors. Without dynamic load balancing the performance is reduced by a factor of 1.5 on 38 processors. When all nodes participate in the PME mesh part, the scaling is limited to 14 processors. With a time step of 2 fs a maximum performance of 68 ns/day is reached, and with a time step of 4 fs this increases to 112 ns/day. Up to 12 processors there is no difference between 2 or 4 cores sharing an Infiniband connection, while at 38 processors the difference is 14%. It is worth mentioning that the repartitioning of the domain decomposition, reassigning charge groups to cells, spatial sorting, setting up the zones, assigning the bonded interactionsm and setting up P-LINCS, always takes a negligible amount of time. The percentage of the total run time spent in repartitioning is 2% with a time step of 2 fs and 4−5% with a time step of 4 fs; the difference is mainly due to the difference in neighbor list update frequency.

For a similar sized protein system we performed a comparison to other simulation packages. We chose one of the most commonly used systems: the joint Amber-CHARMM benchmark DHFR (dihydrofolate reductase) of 23558 atoms in a cubic box of 6.2 nm. Choosing the setup for a benchmark that compares different simulation packages is a difficult issue. Different packages support different features, and the parameter settings for optimal performance can differ between packages. One clear example of this is the box shape. GROMACS can use any triclinic box shape without loss of performance, and one would therefore always choose to solvate a spherical protein in a rhombic dodecahedron unit-cell, which reduces the volume by a factor of

Table 4. Parameters for the DHFR Benchmark and the Energy Drift per Degree of Freedom

package	cutoff (nm)	PME grid	PME freq	time step (fs)	constraints	virtual sites	energy drift (k_BT/ns)
GROMACS	0.96	60 × 60 × 60	1 step	1	none	no	0.011
				2.5	H-bonds	no	0.005
				4	all bonds	yes	0.013
Desmond	0.90	64 × 64 × 64	2 steps	1	none	no	0.017
				2.5	H-bonds	no	0.001
NAMD	0.90	64 × 64 × 64	2 steps	1	none	no	0.023

$\sqrt{2}$ compared to a cubic unit-cell with the same periodic image distance. An important aspect of the setup is the nonbonded interaction treatment. The joint Amber-Charmm benchmark uses interactions that smoothly switch to zero at the cutoff combined with a buffer region. Such a setup is required for accurate energy conservation. But it is questionable if such accurate energy conservation is required for thermostatted simulations. GROMACS loses relatively more performance in such a setup than other packages, since it also calculates all interactions with the buffer region, even though they are all zero. Furthermore, we think that the PME settings for this benchmark (see Table 4) are somewhat conservative; this means the PME-mesh code has a relatively high weight in the results. But since determining the sampling accuracy of molecular simulations goes beyond the scope of this paper, we decided to use the same accuracy and aim for energy conservation. Timings for the Desmond and NAMD[34] packages were taken from the Desmond paper.[35] As Desmond, we used the OPLS all-atom force-field with the TIP3P water model.[3] Note that NAMD and Desmond calculate the PME mesh contribution only every second step, while GROMACS does it every step. We chose to increase the cutoff from 0.9 to 0.96 nm and scale the PME grid spacing accordingly, which provides slightly more accurate forces and a real to reciprocal space process ratio of 3:1. The neighbor list was updated heuristically with a buffer of 0.26 nm. The simulation settings and energy drift are shown in Table 4; note that we took the energy drift values for Desmond and NAMD from the ApoA1 system,[35] which uses a 1.2 nm cutoff and should therefore provide comparable or lower drift. The energy drift for GROMACS is 0.01 k_BT/ns per degree of freedom. This is slightly better than NAMD and Desmond without constraints. With constraints the energy drift with Desmond is an order of magnitude smaller. These results show that codes like GROMACS and Desmond that mainly use single precision do not have larger integration errors than NAMD which uses double precision vectors. It also shows that the use of a time step of 4 fs in GROMACS does not deteriorate the energy conservation. Unfortunately we did not have an identical cluster at our disposal. We also ran the GROMACS benchmarks on a dual core cluster with Infiniband but with 3 GHz Intel Core2 nodes instead of 2.4 GHz AMD64 nodes. Timings for DHFR are shown in Figure 6. If we look at the 1 fs time step results, we can see that, per clock cycle, GROMACS is 2 times faster than Desmond and 3−4 times faster than NAMD, even though the benchmark settings are unfavorable for GROMACS. Additionally GROMACS can be another factor 1.5 faster by increasing the time step from 2.5 to 4 fs, which is made possible by constraining all bonds and converting hydrogens to virtual sites. With MPI, Desmond shows similar scaling to GROMACS, whereas NAMD scales worse. With a special Infiniband communication library, Desmond scales much further than GROMACS in terms of number of cores but only slightly further in terms of actual performance. GROMACS would certainly also benefit from such a library.

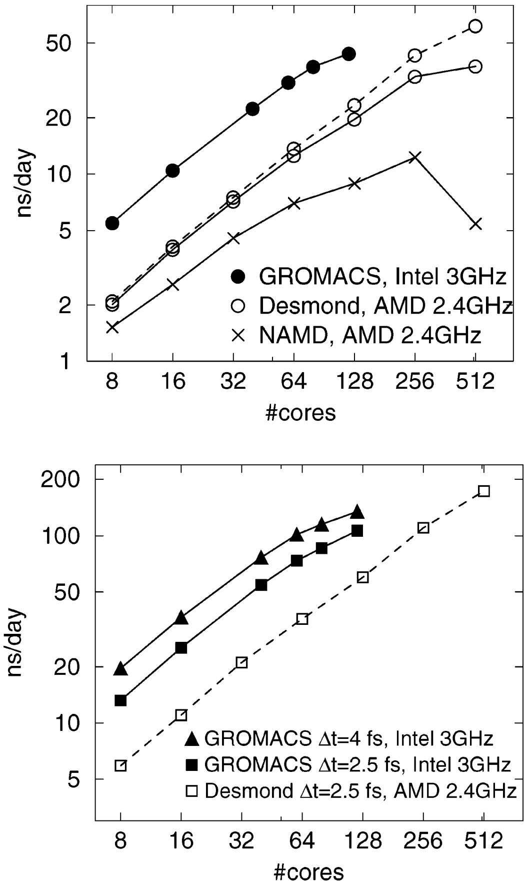


Figure 6. Performance for DHFR in water (23558 atoms) with a 1 fs time step (top panel) and longer time steps (bottom panel) using GROMACS, Desmond, and NAMD. The dashed lines for Desmond show the performance with a tuned Infiniband library.

Finally we show the scaling for a large system, which was somewhat of a weak point in earlier GROMACS versions. The system in question is a structure of the Kv1.2 voltage-gated ion channel[36] placed in a 3:1 POPC:POPG bilayer mixture and solvated with water and ions. The OPLS all-atom force field with virtual site hydrogens is used for the ion channel (18,112 atoms), lipids are modeled with the Berger united-atom force field (424 lipids, 22159 atoms),

Table 5. Simulation Speed in ns/day with GROMACS 4 Domain Decomposition and GROMACS 3.3 Particle Decomposition for the Membrane/Protein System (121449 Atoms)[a]

cores	cpn	4	8	16	32	64	128
GROMACS 4	2	3.1	6.1	11.8	22.3	39.3	65.5
GROMACS 4	4	3.1	6.0	11.6	21.6	38.0	60.1
GROMACS 3.3	2	2.8	4.8	7.7	9.5		
GROMACS 3.3	4	2.7	4.8	7.0	8.4		

[a] With a time step of 5 fs on a 3 GHz Intel Core2 Infiniband cluster with 2 and 4 cores per node/network connection (cpn).

and the total system size is 13 × 13 × 8.8 nm, with 119,913 atoms. We used a cutoff of 1.1 nm and a PME grid of 96 × 96 × 64 (spacing 0.136 nm), giving a real space to PME process ratio of 3:1. Removing the hydrogen vibrations by using virtual sites allows for a time step of 5 fs. The neighbor list was updated every 6 steps (30 fs), since the dynamics in the important membrane region is slower than in water. In Table 5 one can see near linear scaling up to 128 processors, where a performance of slightly more than 60 ns/day is reached. With GROMACS 3.3 the system scales up to 32 processors, where it runs at less than half the speed of the domain decomposition; GROMACS 4 reaches an order of magnitude higher performance. The scaling limitation for this type of system is currently the PME FFT implementation.

X. Conclusions

We have shown that the eighth shell domain decomposition and the dynamic load balancing provide very good scaling to large numbers of processors. Dynamic load balancing can provide a 50% performance increase for typical protein simulations. Another important new feature is the Multiple-Program, Multiple-Data PME parallelization, which lowers the number of processes between which the 3D FFT grid needs to be redistributed, while simultaneously increasing the effective communication speed on systems where multiple cores share a network connection. Since the optimal real space to PME process ratio is often 3:1, the benefit of MPMD is higher with 4 or 8 nodes per core than with 1 or 2. This is advantageous, since having more cores per node decreases the cost and space requirements of computing clusters. MPMD allows simulations with PME to scale to double the number of processors and thereby doubles the simulation speed. The P-LINCS and virtual site algorithms allow a doubling of the time step.

But what makes a biomolecular MD package tick is not just a single algorithm but a combination of many efficient algorithms. If one aspect has not been parallelized efficiently, this rapidly becomes a bottleneck—not necessarily for relative scaling but absolute performance. From the benchmarks above, we believe we have largely managed to avoid such bottlenecks in the implementation described here. Not only do the presented algorithms provide very good scaling to large numbers of processors but also we do so without compromising the high single-node performance or any of the algorithms to extend time steps. Together, these features of GROMACS 4 allow for absolute simulation speed that is an order of magnitude larger than previously.

How good the scaling is depends on three factors: the speed of the computational part in isolation, the efficiency of the parallel and communication algorithms, and the efficiency of the communication itself. The first two factors we have been optimized extensively. The single processor performance of GROMACS is unrivaled. This makes good relative scaling extremely difficult, since communication takes relatively more time. Nevertheless, the benchmarks show that the scaling is now nearly linear over a large range of processor counts. The scaling is usually limited by the third factor, the efficiency of the communication. This is given by the network setup and its drivers. With PME the scaling of GROMACS 4 is limited by communication for the 3D FFT. Without PME the scaling is limited by one single communication call per MD step for summing the energies. For any MD code the latter issue cannot be avoided when a global thermostat or barostat is used every step. As a rough guideline one can say that with modern commodity processors connected by an Infiniband network, GROMACS 4 scales close to linear up to 2000 steps per second for simple liquids without PME, while for complex membrane protein simulations (no optimized water kernels) with PME and constraints it scales up to 500 steps per second. There are still alternatives with even more impressive *relative* scaling,[9] and dedicated-hardware implementation might provide extremely high performance if cost is no issue. However, for all normal cases where resources are scarce and absolute performance is the only thing that matters, we believe the implementation presented here will be extremely attractive for molecular simulations.

Acknowledgment. The authors thank the RRZE in Erlangen and especially Georg Hager for providing computing resources for benchmarks and Vagelis Harmandaris for providing the ESPResSo benchmarks. This work was supported by grants from the Swedish Foundation for Strategic Research and the Swedish Research Council (E.L.).

References

(1) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.

(2) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; D. Reidel Publishing Company: Dordrecht, 1981; pp 331−342.

(3) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

(4) Fincham, D. *Mol. Simul.* **1987**, *1*, 1.

(5) Raine, A. R. C.; Fincham, D.; Smith, W. *Comput. Phys. Commun.* **1989**, *55*, 13.

(6) Clark, T.; McCammon, J. A.; Scott, L. R. In *Proceedings of the Fifth SIAM Conference on Parallel Processing for Scientific Computing*: Dongarra, J., et al., Eds.; SIAM: Philadelphia, 1991; pp 338−344,

(7) Bekker, H.; Berendsen, H. J. C.; Dijkstra, E. J.; Achterop, S.; van Drunen, R.; van der Spoel, D.; Sijbers, A.; Keegstra, H.; Reitsma, B.; Renardus, M. K. R. In *Physics Computing 92*; de Groot, R. A., Nadrchal, J., Eds.; World Scientific: Singapore, 1993; pp 252−256,

(8) Nelson, M.; Humphrey, W.; Gursoy, A.; Dalke, A.; Kalé, L.; Skeel, R.; Schulten, K. *Int. J. High Perform. Comput. Appl.* **1996**, *10*, 251.

(9) Fitch, B.; Germain, R.; Mendell, M.; Pitera, J.; Pitman, M.; Rayshubskiy, A.; Sham, Y.; Suits, F.; Swope, W.; Ward, T.; Zhestkov, Y.; Zhou, R. *J. Parallel Distributed Comput.* **2003**, *63*, 759.

(10) Rhee, Y. M.; Sorin, E. J.; Jayachandran, G.; Lindahl, E.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6456.

(11) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306.

(12) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701.

(13) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. *J. Comput. Chem.* **1999**, *20*, 786.

(14) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577.

(15) Greengard, L.; Rokhlin, V. *J. Comput. Phys.* **1987**, *73*, 325.

(16) Bowers, K. J.; Dror, R. O.; Shaw, D. E. *J. Comput. Phys.* **2007**, 221, 303.

(17) Bowers, K. J.; Dror, R. O.; Shaw, D. E. *J. Phys. Conf. Ser.* **2005**, *16*, 300.

(18) Bowers, K. J.; Dror, R. O.; Shaw, D. E. *J. Chem. Phys.* **2006**, *124* (18), 184109.

(19) Liem, S. Y.; Brown, D.; Clarke, J. H. R. *Comput. Phys. Commun*. **1991**, *67* (2), 261.

(20) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463.

(21) Hess, B. *J. Chem. Theory Comput.* **2008**, *4* (1), 116.

(22) Lippert, R. A.; Bowers, K. J.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Shaw, D. E. *J. Chem. Phys.* **2007**, *126*, 046101.

(23) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952.

(24) van Gunsteren, W. F.; Karplus, M. *Macromolecules* **1982**, *15*, 1528.

(25) Meloni, S.; Rosati, M. *J. Chem. Phys.* **2007**, *126*, 121102.

(26) Weerasinghe, S.; Smith, P. E. *J. Chem. Phys.* **2003**, *119* (21), 11342.

(27) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269.

(28) Kutzner, C.; van der Spoel, D.; Fechner, M.; Lindahl, E.; Schmitt, U. W.; de Groot, B. L.; Grubmuller, H. *J. Comput. Chem.* **2007**, *28*, 2075.

(29) Holian, B. L.; Voter, A. F.; Ravelo, R. *Phys. Rev. E* **1995**, *52* (3), 2338.

(30) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

(31) Harmandaris, V. A.; Reith, D.; van der Vegt, N. F. A.; Kremer, K. *Macromolecules* **2007**, *208*, 2109.

(32) van Gunsteren, W. F.; Berendsen, H. J. C. *Mol. Simul*. **1988**, *1*, 173.

(33) Limbach, H.-J.; Arnold, A.; Mann, B. A.; Holm, C. *Comput. Phys. Commun*. **2006**, *174* (9), 704.

(34) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26* (16), 1781.

(35) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. In *ACM/IEEE SC 2006 Conference (SC'06*); 2006; p 43.

(36) Long, S.; Campbell, E. B.; MacKinnon, R. *Science* **2005**, *309*, 897.

CT700301Q

JCTC Journal of Chemical Theory and Computation

Influence of Density Functionals and Basis Sets on One-Bond Carbon−Carbon NMR Spin−Spin Coupling Constants

R. Suardíaz,[†] C. Pérez,[†] R. Crespo-Otero,[†] José M. García de la Vega,*[,‡] and Jesús San Fabián[‡]

Departmento de Química Física, Facultad de Química, Universidad de la Habana, La Habana 10400, Cuba, and Departmento de Química Física Aplicada, Facultad de Ciencias, Universidad Autónoma de Madrid, 28049 Madrid, Spain

Received December 4, 2007

Abstract: The basis set and the functional dependence of one-bond carbon−carbon NMR spin−spin coupling constants (SSCC) have been analyzed using density functional theory. Four basis sets (6-311G**, TZVP, EPR-III, and aug-cc-pVTZ-J) and four functionals (PBE, PW91, B3LYP, and B3P86) are tested by comparison with 70 experimental values corresponding to 49 molecules that represent multiple types of hybridization of the carbon atoms. The two hybrid functionals B3P86 and B3LYP combined either EPR-III or aug-cc-pVTZ-J basis sets lead to the best accuracy of calculated SSCC. However, a simple linear regression allows for the obtaining of scaled coupling constants that fit much better with the experimental data and where the differences between the different basis sets and/or functional results are significantly reduced. For large molecules the TZVP basis set can be an appropriate election presenting a good compromise between quality of results and computational cost.

1. Introduction

Spin−spin coupling constants (SSCC) represent a valuable source of structural information from nuclear magnetic resonance (NMR) spectroscopy. During the past decade, the use of quantum chemistry methods for the calculation of SSCC has become routine and widespread.[1−3] Predictions from wavefunction-based methods are generally in good agreement with experimental values.[4−6] Nevertheless, the high computational cost required by these methods limits their applicability to small systems. Density functional theory (DFT) methodology combined with analytical linear response techniques is a promising alternative to post Hartree−Fock methods. However, the bibliographic DFT data related to the calculation of SSCC show certain dispersion in the functional and basis sets employed.[6−12] There are various works reporting that B3LYP[13,14] functional yields satisfactory results for SSCC in a small set of molecules.[7−9] On the other hand, Maximoff et al.[6] reported the assessment of 20 different functionals in predicting one-bond carbon−hydrogen. In this work,[6] the best results were reported for PBE[15,16] and PW91[17−19] functionals, both based on the gradient generalized approximation (GGA), and report a good performance[6] for the semiempirical-hybrid B3P86,[13,20] whereas B3LYP resulted in one of the worst. They conclude that meta-GGA and hybrid functionals do not necessarily improve over GGA functionals for this type of couplings. Keal et al.[11] tested functionals B97-2[21] and B97-3[22] with the data set of Maximoff et al.[6] The performance of PBE, B97-2, B97-3, and B3LYP for predicting other kinds of couplings that include N, O, and F elements in 11 alternative molecules were also carried out by Keal et al.[11] They reported that PBE was considerably less accurate than B3LYP for the prediction of those SSCC. Recently, Witanowski et al.[12] studied 257 aromatic carbon−carbon couplings across one, two, and three bonds. They obtained excellent calculated values using the B3PW91/6-311++G(d,p)//B3PW91/6-311++G(d,p) approach where the same functional-basis set combination was employed for geometry optimizations and for coupling

* Corresponding author e-mail: garcia.delavega@uam.es.

[†] Universidad de la Habana.

[‡] Universidad Autónoma de Madrid.

10.1021/ct7003287 CCC: $40.75

Published on Web 02/21/2008

constants computation. Other functionals have been tested with a variety of results.[23]

With regards to basis sets, it is well-known that calculated SSCC strongly depend on the quality of the employed Gaussian basis set functions. The Fermi contact (FC) term usually provides the largest contribution. Therefore, the electronic density at the nucleus should be well described, and, consequently, the selection of the basis set is crucial in SSCC calculations. Several basis sets can be found in the literature[2] to calculate coupling constants. The work of Peralta et al.[24] constitutes, to our knowledge, one of the largest analyses within the DFT framework. These authors analyzed basis set dependence using the B3LYP functional. In that work, some well-known basis set functions were employed, namely IGLO-III,[25] EPR-III,[26,27] aug-cc-pVTZ-J,[28] and Sadlej-J.[29] The authors suggested the combination B3LYP/aug-cc-pVTZ-J as an excellent choice to calculate SSCC.

One-bond carbon−carbon is the most important bond in organic chemistry, and carbon−carbon coupling constants $^1J_{CC}$ possess unique structural information concerning electronic structure, substituent effects, and stereochemical behavior of organic molecules.[30] The hybridization state of the two carbons offers a variety of six different types of bonds and a large range for $^1J_{CC}$. The aim of the present work is to investigate the capabilities of a different combination of functionals and basis sets to predict, with a certain degree of accuracy, the one-bond carbon−carbon coupling constants ($^1J_{CC}$) of organic molecules containing elements of the first and second rows of the periodic table. This goal will be performed using a statistical analysis by comparing the calculated $^1J_{CC}$ with the experimental ones.

2. Computational Details

We have selected four exchange-correlation functionals among the large number available in the literature. The two first GGA functionals were PBE[15,16] and PW91.[17−19] Those functionals had the best performance among 20 other tested by Maximoff et al.[6] for the calculation of $^1J_{CH}$. The two second hybrid functionals were B3P86[13,20] and B3LYP.[13,14] The former functional yielded similar performance to those of PBE and PW91 for $^1J_{CH}$, while the latter has been used successfully by some authors,[7,24] although it has been reported as one of the worst by Maximoff et al.[6]

Computations were performed using four basis sets of contracted Gaussian functions, namely 6-311G**,[31,32] TZVP,[33] EPR-III,[26,27] and aug-cc-pVTZ-J.[28] 6-311G** is a small basis set with a triple-ζ quality plus polarization. TZVP is a DFT-optimized valence triple-ζ basis with promising results in the prediction of hyperfine couplings in combination with the B3LYP functional.[34] EPR-III is larger and has been optimized for the computation of hyperfine coupling constants by DFT methods with the s-part improved to better describe the nuclear region. EPR-III is a triple-ζ basis including diffuse functions, doubled-polarizations, and a single set of f-polarization functions. aug-cc-pVTZ-J is a relatively large basis set, specially designed for the computation of SSCC. aug-cc-pVTZ-J is a recontraction of aug-cc-pVTZ-Juc,[28] that is the triple-ζ aug-cc-pVTZ[35−39] of Dunning completely uncontracted, augmented with four tight s-type functions and without the most diffuse second polarization function. The computational cost of the calculations depends on the complexity of the approximate functional expressions and the basis set dimensions. Thus, computational time estimated by Maximoff et al.[6] in $C_6H_5NO_2$ (514 basis functions) is for the hybrid functionals B3LYP and B3P86 five times larger than for the GGA functionals PBE and PW91. For the molecules calculated in this work, the average computational time for the B3LYP and B3P86 functionals is roughly twice that of the PBE and PW91 ones, when the EPR-III and aug-cc-pVTZ-J basis sets are used. However, the computational time for all functionals is similar when 6-311G** and TZVP are used. Moreover, larger differences are found for the approximated average computational cost of the different basis sets: TZVP is twice as expensive as 6-311G**; EPR-III is between 5 (with PBE or PW91 functionals) and 9 (with B3LYP or B3P86 functionals) times more expensive; and the large basis set aug-cc-pVTZ-J is between 15 (with PBE or PW91 functionals) and 34 (with B3LYP or B3P86 functionals) times more expensive. The computational time for these two large basis set is very dependent on the used functional.

In this study we have considered a set of organic molecules containing first and second row elements. The basic criteria for the selection of the systems have been the rigidity or, at least, the existence of only one populated conformer. We determined $^1J_{CC}$ for these 49 chemically diverse molecules that correspond to 70 experimentally measured one-bond carbon−carbon couplings involving 19 $^1J_{C_{sp3}-C_{sp3}}$, 11 $^1J_{C_{sp3}-C_{sp2}}$, 6 $^1J_{C_{sp3}-C_{sp}}$, 29 $^1J_{C_{sp2}-C_{sp2}}$, 2 $^1J_{C_{sp2}-C_{sp}}$, and 3^1 $J_{C_{sp}-C_{sp}}$. This set was mainly extracted from the reports of Wray and Krivdin et al.[30,40−49] (see the Supporting Information for specific references). Since accurate experimental geometries are only available for a few molecules in this set, we used optimized geometries at the B3LYP/6-31G** level of theory, which is considered sufficiently accurate for the present purpose.[34,50] The set of selected molecules are depicted in Figure 1. Although rovibrational effects can be non-negligible in SSCC[51,52] we do not consider them in this report since their evaluation is computational demanding.[53] Evaluation of solvent effect in small molecules has shown a reduced sensitivity. This effect is mainly due to reaction field effects via the indirect contribution from equilibrium geometry changes.[52,54] The geometrical parameters that more significantly affect the SSCC are the dihedral angles but in selected molecules were essentially constant due to their rigidity. Hence solvent effects are also neglected. All computations were performed using the Gaussian03 package.[55]

3. Results and Discussion

The 70 coupling constants calculated with four functionals and four basis sets have been analyzed by means of statistical methods. An initial exploration makes us withdraw 6 coupling constants from the data set used in the statistics due to their large deviation. For this reason, these 6 calculated values are analyzed at the end of this section. The statistical analysis has been carried out over three sets of couplings: i) **Set-1** formed by the whole set of couplings (64 values);

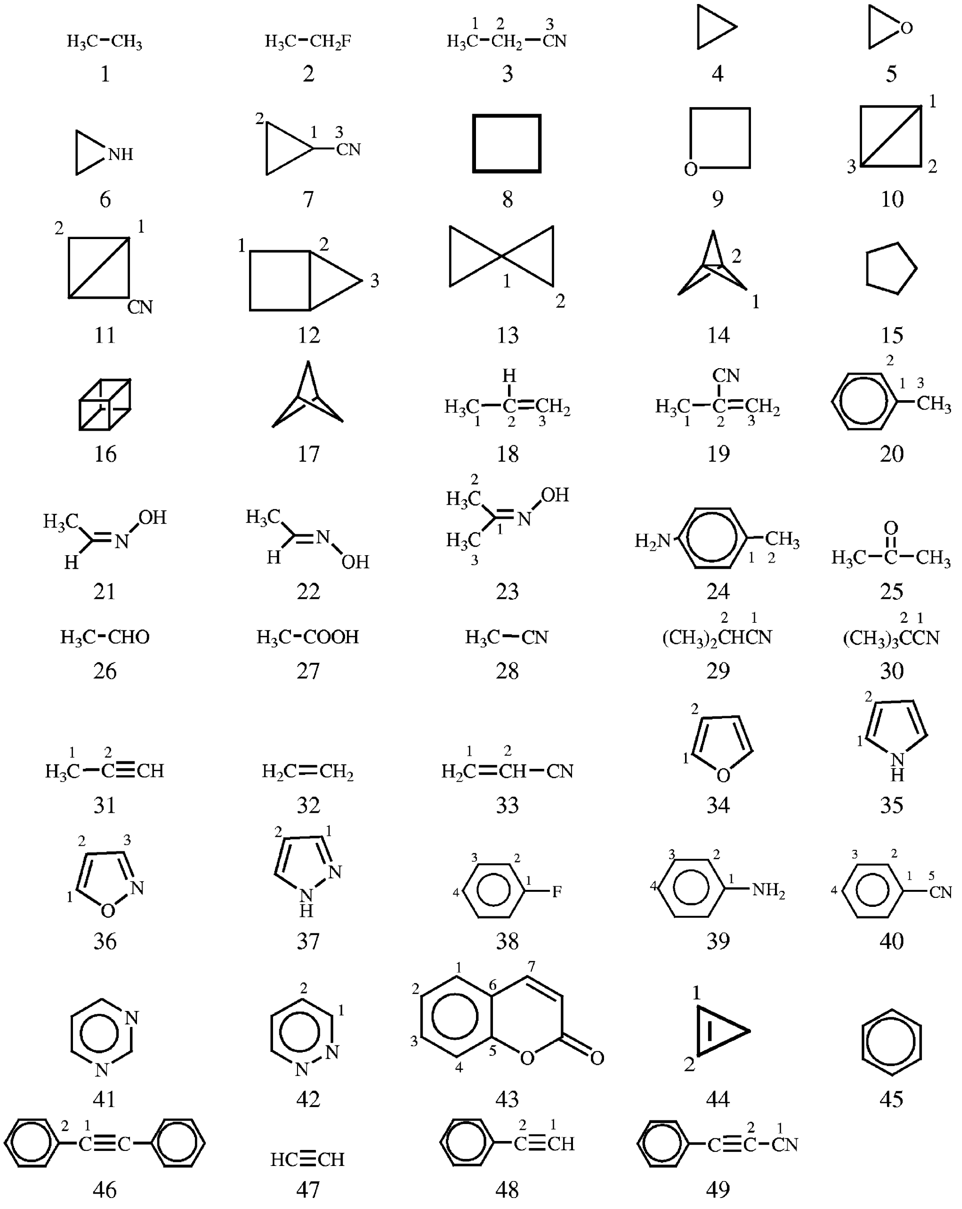


Figure 1. Studied molecules.

ii) **Set-2** with couplings smaller than 46 Hz (26 values), i.e. it includes 18 C_{sp^3}−C_{sp^3} and the 8 smaller C_{sp^3}−C_{sp^2} values; and iii) **Set-3** with couplings larger than 46 Hz (38 values) which include 6 C_{sp^3}−C_{sp}, 27 C_{sp^2}−C_{sp^2}, 2 C_{sp^2}−C_{sp}, and the 3 larger C_{sp^3}−C_{sp^2} values. The election of these data sets of couplings is based on the graphical behavior in which there seems to be a change of the tendency around 46 Hz (see Figure 2).

The statistics for the three data sets presented in Tables 1−3 is based on the values of standard deviation (σ), mean absolute error (MAE), and the minimum (Min) and maximum (Max) deviation

$$\sigma = \sqrt{\frac{\sum({}^1J_{CC}^{exp} - {}^1J_{CC}^{calc})^2}{n-1}}, \quad \text{MAE} = \frac{\sum|{}^1J_{CC}^{exp} - {}^1J_{CC}^{calc}|}{n} \quad (1)$$

Most of the calculated couplings underestimate the experimental values. Therefore, the calculated values can be shifted and/or scaled to obtain better estimations and to detect whether the differences between the results are either merely quantitative or qualitative. Scaled couplings were obtained by a simple linear expression

$${}^1J_{CC}^{scaled} = a + b\cdot{}^1J_{CC}^{calc} \quad (2)$$

The coefficients a and b were calculated by fitting the calculated couplings for each approach (functional/basis set) to the equation ${}^1J_{CC}^{exp} = a + b\ \cdot{}^1J_{CC}^{calc}$. A standard deviation ($\sigma'$), mean absolute error (MAE′), and minimum (Min′) and maximum (Max′) deviation for these fitted values are also considered and included in Tables 1−3.

$$\sigma' = \sqrt{\frac{\sum[{}^1J_{CC}^{exp} - (a + b\cdot{}^1J_{CC}^{calc})]^2}{n-2}}, \quad \text{MAE}' = \frac{\sum(|{}^1J_{CC}^{exp} - (a + b\cdot{}^1J_{CC}^{calc})|}{n} \quad (3)$$

Set-1 includes all calculated couplings (except the six indicated below), and it has an experimental range between 10 and 91 Hz. For this set, the best result is that of the B3P86/

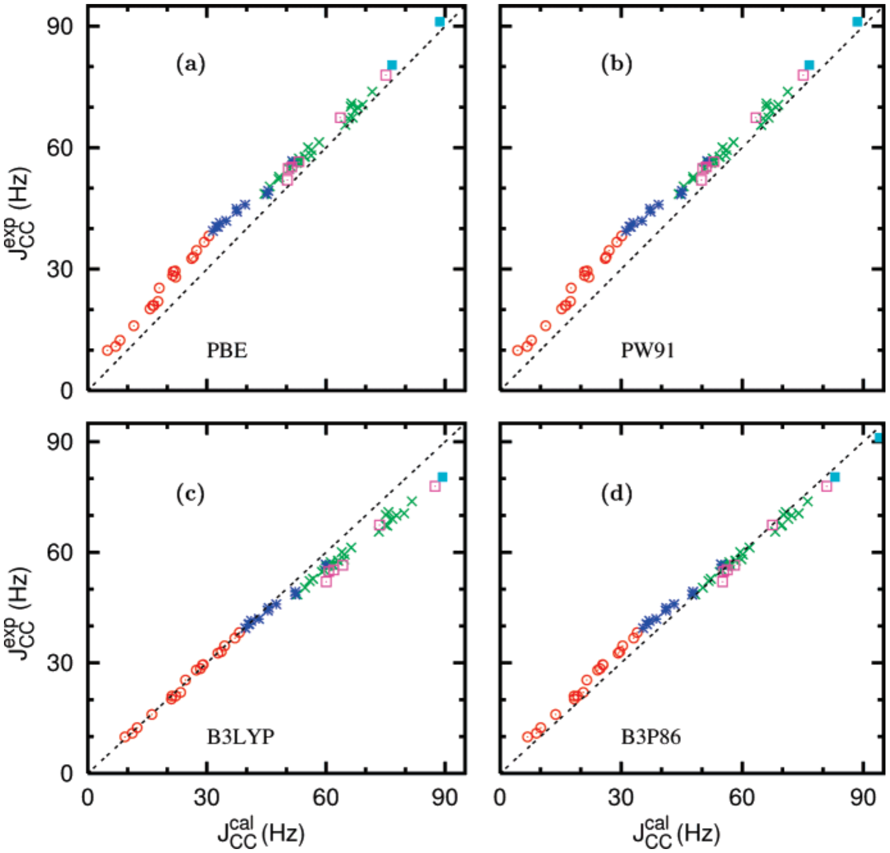


Figure 2. Experimental vs calculated $^1J_{CC}$ couplings. aug-cc-pVTZ-J basis were used for the four indicated functionals: $^1J_{C_{sp3}-C_{sp3}}$ (○), $^1J_{C_{sp3}-C_{sp2}}$ ($*$), $^1J_{C_{sp2}-C_{sp2}}$ (×), $^1J_{C_{sp3}-C_{sp}}$ (□), $^1J_{C_{sp2}-C_{sp}}$ (■).

Table 1. Statistical Results for 64 Data (Set-1) Calculated Using the Indicated Functional/Basis Set (in Hz)

functional/basis set	σ	MAE	Max	Min	a[a]	b[a]	σ'	MAE′	Max′	Min′
PBE/6-311G**	3.53	2.81	4.4	−8.4	7.2(6)	0.87(1)	2.01	1.69	3.9	−4.0
PBE/TZVP	4.58	3.75	1.0	−8.9	8.5(5)	0.89(1)	1.63	1.32	2.9	−3.4
PBE/EPR-III	6.53	6.31	−3.5	−9.3	7.4(4)	0.97(1)	1.43	1.14	2.8	−3.0
PBE/aug-cc-pVTZ-J	4.98	4.54	−0.7	−8.1	7.4(5)	0.93(1)	1.52	1.23	2.9	−3.1
PW91/6-311G**	4.33	3.39	4.1	−10.3	8.4(6)	0.88(1)	2.06	1.71	4.2	−4.2
PW91/TZVP	5.64	4.91	0.1	−10.5	9.5(5)	0.90(1)	1.66	1.33	3.0	−3.6
PW91/EPR-III	6.69	6.46	−3.5	−9.5	7.7(4)	0.97(1)	1.44	1.14	2.8	−3.0
PW91/aug-cc-pVTZ-J	5.25	4.81	−0.9	−8.4	7.8(5)	0.93(1)	1.53	1.23	2.8	−3.1
B3LYP/6-311G**	7.10	5.69	12.9	−2.3	4.3(6)	0.82(1)	1.82	1.48	3.1	−4.1
B3LYP/TZVP	5.98	4.66	12.3	−1.6	4.8(5)	0.83(1)	1.39	1.12	2.4	−3.5
B3LYP/EPR-III	2.92	2.28	6.8	−1.7	3.4(4)	0.90(1)	1.28	1.02	2.2	−3.7
B3LYP/aug-cc-pVTZ-J	4.72	3.63	9.6	−0.7	3.5(4)	0.87(1)	1.30	1.07	2.3	−3.5
B3P86/6-311G**	3.52	3.15	6.0	−6.4	6.6(6)	0.86(1)	1.85	1.54	3.1	−3.7
B3P86/TZVP	3.24	2.64	4.8	−5.9	7.2(5)	0.87(1)	1.45	1.16	2.6	−3.3
B3P86/EPR-III	3.17	2.70	1.3	−5.7	5.4(4)	0.94(1)	1.29	1.04	2.4	−3.5
B3P86/aug-cc-pVTZ-J	2.49	2.03	3.4	−4.4	5.4(4)	0.91(1)	1.34	1.08	2.5	−3.3

[a] The error in the last digit is given between parentheses.

aug-cc-pVTZ-J with a σ deviation of 2.5 Hz (MAE = 2.0 Hz). Other results with σ values smaller than 4 Hz are those of B3LYP/EPR-III (σ = 2.9 Hz), B3P86/EPR-III (σ = 3.2 Hz), B3P86/TZVP (σ = 3.2 Hz), B3P86/6-311G** (σ = 3.5 Hz), and the inexpensive PBE/6-311G** (σ = 3.5 Hz).

When the calculated values are fitted to eq 2, the scaled couplings achieve σ' deviations in a narrow interval (between 1.3 and 2.1 Hz). The best results (σ' smaller than 1.35 Hz) are obtained with the B3P86 and B3LYP functionals and with the EPR-III and aug-cc-pVTZ-J basis sets. The positive intercepts and the slopes that are always smaller than one indicate that the smaller calculated couplings are more underestimated than the larger ones. The intercepts are larger for BPE and PW91 functionals than for B3LYP and B3P86 ones.

For Set-2 (couplings in the range between 10 and 46 Hz) the best results are those obtained with the B3LYP functional that presents σ deviations between 0.7 (when the aug-cc-pVTZ-J basis set is used) and 1.4 Hz (with the 6-311G** basis set). For the coupling constants involving sp^3 carbons

Table 2. Statistical Results for 26 Data (Set-2) Calculated Using the Indicated Functional/Basis Set (in Hz)

functional/basis set	σ	MAE	Max	Min	a[a]	b[a]	σ'	MAE′	Max′	Min′
PBE/6-311G**	5.04	4.65	−1.9	−8.4	2.8(9)	1.07(3)	1.58	1.09	4.1	−2.3
PBE/TZVP	6.76	6.49	−4.4	−8.9	4.9(6)	1.07(2)	1.22	0.99	2.6	−2.1
PBE/EPR-III	7.59	7.26	−4.3	−9.3	4.0(5)	1.14(2)	0.93	0.79	1.6	−1.8
PBE/aug-cc-pVTZ-J	6.65	6.37	−3.9	−8.1	3.8(5)	1.11(2)	0.98	0.82	1.8	−1.7
PW91/6-311G**	6.48	6.08	−3.1	−10.3	4.4(9)	1.07(4)	1.79	1.31	4.6	−2.6
PW91/TZVP	7.91	7.61	−5.3	−10.5	6.2(7)	1.06(3)	1.43	1.20	3.1	−2.2
PW91/EPR-III	7.79	7.47	−4.5	−9.5	4.4(5)	1.13(2)	1.00	0.87	1.8	−1.7
PW91/aug-cc-pVTZ-J	6.96	6.67	−4.2	−8.4	4.3(5)	1.10(2)	1.05	0.91	2.0	−1.6
B3LYP/6-311G**	1.42	1.24	2.4	−2.3	0.2(7)	0.97(2)	1.13	0.80	2.8	−1.9
B3LYP/TZVP	1.03	0.78	2.3	−1.6	1.5(5)	0.95(1)	0.85	0.69	1.5	−1.5
B3LYP/EPR-III	0.98	0.83	0.5	−1.7	0.5(3)	1.01(1)	0.60	0.48	1.0	−1.2
B3LYP/aug-cc-pVTZ-J	0.69	0.52	1.6	−0.7	0.3(4)	0.98(1)	0.63	0.50	0.9	−1.3
B3P86/6-311G**	3.71	3.39	−0.8	−6.4	2.5(7)	1.03(3)	1.33	0.96	3.2	−2.3
B3P86/TZVP	4.52	4.32	−2.4	−5.9	3.8(5)	1.02(2)	1.02	0.84	1.7	−1.8
B3P86/EPR-III	4.41	4.20	−2.2	−5.7	2.2(4)	1.08(1)	0.72	0.58	0.9	−1.5
B3P86/aug-cc-pVTZ-J	3.44	3.26	−1.4	−4.4	2.0(4)	1.05(1)	0.74	0.60	0.9	−1.6

[a] The error in the last digit is given between parentheses.

Table 3. Statistical Results for 38 Data (Set-3) Calculated Using the Indicated Functional/Basis Set (in Hz)

functional/basis set	σ	MAE	Max	Min	a[a]	b[a]	σ'	MAE′	Max′	Min′
PBE/6-311G**	2.02	1.56	4.4	−2.5	4.5(1)	0.91(2)	1.33	1.02	3.0	−2.4
PBE/TZVP	2.21	1.87	1.0	−4.2	7.3(1)	0.91(1)	0.96	0.71	2.0	−2.0
PBE/EPR-III	5.82	5.66	−3.5	−7.4	6.9(1)	0.98(2)	0.95	0.71	1.9	−2.3
PBE/aug-cc-pVTZ-J	3.51	3.29	−0.7	−5.2	6.5(1)	0.94(2)	0.98	0.75	1.7	−2.1
PW91/6-311G**	1.88	1.55	4.1	−4.3	6.6(1)	0.90(2)	1.44	1.07	2.9	−3.4
PW91/TZVP	3.43	3.07	0.1	−5.6	8.9(1)	0.90(2)	1.02	0.76	2.2	−2.1
PW91/EPR-III	5.93	5.77	−3.5	−7.5	7.5(1)	0.97(2)	0.96	0.71	2.0	−2.3
PW91/aug-cc-pVTZ-J	3.77	3.54	−0.9	−5.4	7.2(1)	0.94(2)	1.00	0.76	1.9	−2.2
B3LYP/6-311G**	9.19	8.74	12.9	3.8	2.9(2)	0.83(2)	1.56	1.14	3.0	−3.6
B3LYP/TZVP	7.76	7.31	12.3	3.8	4.6(1)	0.83(2)	1.14	0.81	2.5	−3.3
B3LYP/EPR-III	3.72	3.28	6.8	1.2	3.8(1)	0.89(2)	1.20	0.87	2.3	−3.6
B3LYP/aug-cc-pVTZ-J	6.13	5.76	9.6	2.9	3.3(1)	0.86(2)	1.14	0.85	2.2	−3.2
B3P86/6-311G**	3.43	2.98	6.0	−1.5	5.2(1)	0.88(2)	1.37	1.02	3.2	−3.2
B3P86/TZVP	2.01	1.49	4.8	−2.2	7.1(1)	0.87(1)	0.97	0.67	2.2	−3.0
B3P86/EPR-III	2.00	1.68	1.3	−3.8	5.7(1)	0.93(2)	0.99	0.71	1.9	−3.3
B3P86/aug-cc-pVTZ-J	1.61	1.19	3.4	−1.9	5.3(1)	0.90(2)	0.98	0.73	2.0	−3.0

[a] The error in the last digit is given between parentheses.

the B3LYP functional gives the best results. It should be noted the very good results obtained with the inexpensive TZVP basis set. The values calculated with the B3P86 functional present a deviation between 3.4 and 4.5 Hz, and those obtained with PBE and PW91 show higher deviations (between 5.0 and 7.9 Hz). For $^1J_{C_{sp3}-C_{sp3}}$ couplings, the PBE, PW91, and B3P86 functionals give values smaller than the experimental ones. Using the values scaled with eq 2 the σ' deviations are also smaller for B3LYP results, but for the three other functionals the reduction of the standard deviations is very significant. The σ' values for PW91, PBE, B3P86, and B3LYP with the EPR-III basis set are 1.0, 0.9, 0.7, and 0.6 Hz, respectively.

For Set-3 (couplings in the range between 48 and 91 Hz) the best results are those of B3P86/aug-cc-pVTZ-J ($\sigma = 1.6$ Hz), PW91/6-311G** (1.9 Hz), B3P86/EPR-III (2.0 Hz), B3P86/TZVP (2.0 Hz), PBE/6-311G** (2.0 Hz), and PBE/TZVP (2.2 Hz). The B3P86 seems to yield the best results for this set even with the economic TZVP. On the other hand, the frequently used B3LYP functional presents here larger standard deviations (between 3.8 and 9.2 Hz). Again the use of scaled values reduces significantly the standard deviations. It is worth mentioning that the best results can be obtained with the PBE, PW91, or B3P86 functionals which present a σ' values around 1.0 Hz when one of the three larger basis sets are used. It is also important regarding the large reduction in the standard deviation for the B3LYP results from σ (between 3.8 and 9.2 Hz) to σ' (between 1.2 and 1.6 Hz).

With regards to the basis sets, for the three sets the worse σ' results are obtained by 6-311G**. The σ values for this basis set are also large for the hybrid functionals results; however, these σ deviations are relatively good when the GGA functionals are used, in part, due to a compensation of errors. For the scaled couplings, the basis sets EPR-III and aug-cc-pVTZ-J provide similar results, whereas TZVP presents slightly worse σ' values, except for set-3 in which results similar to those of the two other basis sets are obtained. Taking into account i) the reasonable σ' results for the TZVP basis set, ii) the low σ values for B3LYP/TZVP in set-2 (1.0 Hz) and for B3P86/TZVP and PBE/TZVP in set-3 (2.0 and 2.2 Hz, respectively), and iii) the compu-

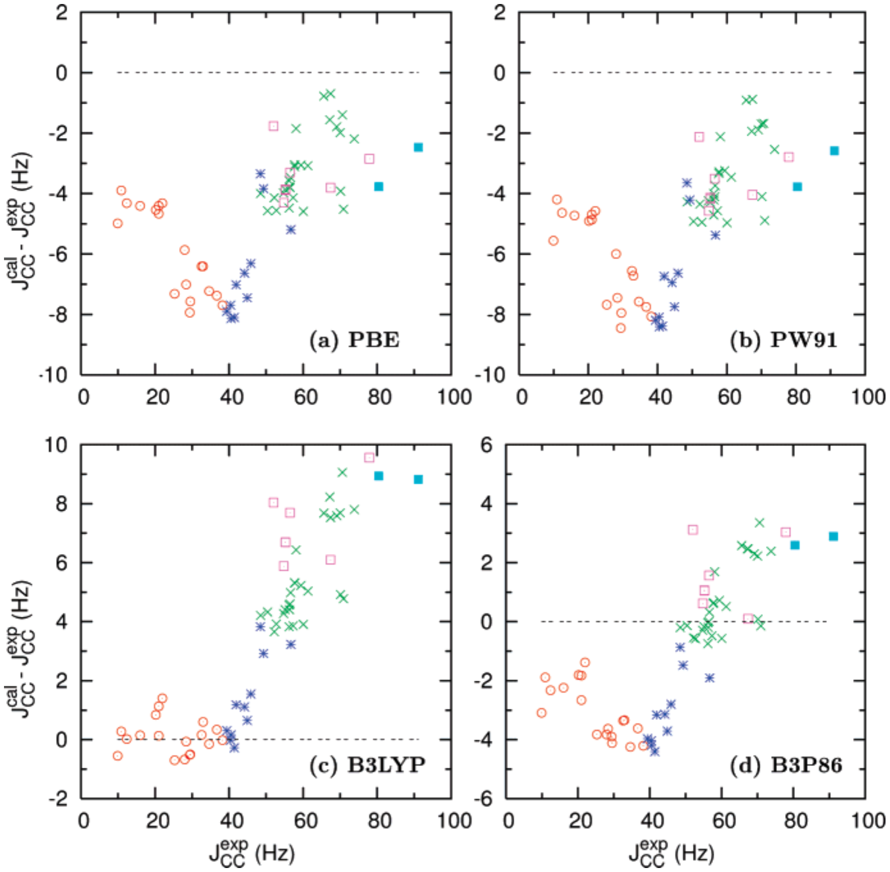


Figure 3. Calculated deviations ($^1J^{cal}_{CC} - {}^1J^{exp}_{CC}$) vs experimental values. aug-cc-pVTZ-J basis were used for the four functionals: $^1J_{C_{sp3}-C_{sp3}}$ (○), $^1J_{C_{sp3}-C_{sp2}}$ (∗), $^1J_{C_{sp2}-C_{sp2}}$ (×), $^1J_{C_{sp3}-C_{sp}}$ (□), $^1J_{C_{sp2}-C_{sp}}$ (■).

tational cost of this basis set, we consider that TZVP is an appropriate election for large molecules.

A more detailed analysis of the functional can be performed representing the differences between the calculated and experimental values (Figure 3). The range of deviations is similar for all functionals, between 0 and −9 Hz for PBE and PW91 results, between −5 and 4 Hz for B3P86, and between −1 and 10 Hz (the largest one) for B3LYP. The calculated PBE and PW91 couplings are always smaller than the experimental, and, as it has already been found for $^1J_{CH}$ coupling,[6] both functionals give similar results. This fact is due to PBE is essentially a simplification of the PW91 where several fundamental constants are imposed on the energy functional.[15] The B3LYP results for couplings smaller than 46 Hz (Set-2) agree satisfactorily with the experimental data with a deviation between −1 and 2 Hz. However, the deviation for coupling larger than 46 Hz is between 2 and 10 Hz (see Figure 3).

It is also interesting to represent the differences between the results of two functionals or two basis sets to prevent possible distortions from the experimental values (see the Supporting Information). As indicated above, PBE and PW91 functionals give similar results, and the differences between them are smaller than 0.6 Hz in magnitude. Accordingly, the figures of the differences between B3LYP (or B3P86) results and those of PBE are similar to the differences of the former functional with PW91. The differences between PBE (or PW91) and B3LYP are in the range of −4 to −13 Hz, and, roughly, they follow a linear relation with the calculated value, i.e., the deviations are larger for large calculated values. The differences between PBE (or PW91) and B3P86 also follow the linear dependence, but now the range of the deviations, between −1 and −6 Hz, is smaller. A similar linear dependence and range of differences are observed between B3P86 and B3LYP. It should be noted that the relative difference ($^1J^{cal}_{CC} - {}^1J^{cal'}_{CC}$)/$^1J^{cal'}_{CC}$ always decreases as the calculated coupling value increases. With regards to the basis sets, in Figure 4 the differences between the results of two basis sets are presented (see also the Supporting Information). The differences between the results of the 6-311G** and any of the three other basis sets are rather scattered. However, the differences between these last basis sets present a linear relation as shown in Figure 4b.

It is interesting to analyze and make some comments about the six above-mentioned experimental couplings that have been removed from the data set used in the statistics. Three of them are the $^1J_{C_{sp}-C_{sp}}$ couplings that present large experimental values (between 155.8 and 175.9 Hz), see Table 4, and introduce a strong distortion in the fits. On the other hand, the number of this type of couplings is too small to get statistical results. However, we tested how they fit with the scaled coupling constants. If these couplings are scaled using eq 2 with the coefficients *a* and *b* of Table 1, they present an average deviation defined as $\sum^{methods}(|{}^1J^{exp}_{CC} - {}^1J^{scaled}_{CC}|)/16$ of −12.6, −8.6, and 3.5 Hz (see Table 4). For

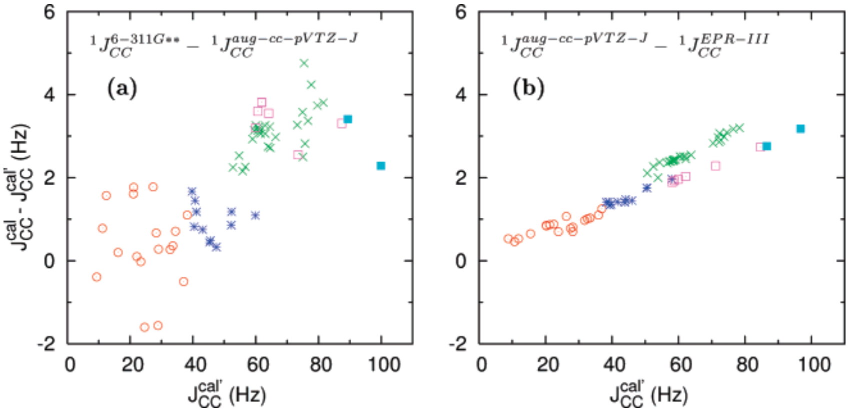
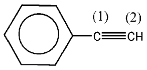
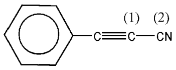
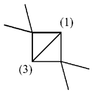
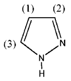
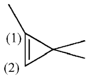


Figure 4. Differences between the results of two different basis sets ($^1J_{CC}^{cal} - {}^1J_{CC}^{cal'}$) vs $^1J_{CC}^{cal'}$). B3LYP functional were used: $^1J_{C_{sp3}-C_{sp3}}$ (○), $^1J_{C_{sp3}-C_{sp2}}$ (∗), $^1J_{C_{sp2}-C_{sp2}}$ (×), $^1J_{C_{sp3}-C_{sp}}$ (□), $^1J_{C_{sp2}-C_{sp}}$ (■).

Table 4. Coupling Constants with the Large Calculated Deviations

				Average deviation^a (Hz)	
No.	Molecule	type	$^1J_{CC}^{exp}$ (Hz)	Initial	Reevaluated
47	HC≡CH	$sp - sp$	170.6	-12.6	—
48		$sp - sp$	175.9	-8.6	—
49		$sp - sp$	155.8	3.5	—
50		$sp^3 - sp^3$	-17.49	-9.2^b	-5.7
37		$sp^2 - sp^2$	58.29	6.5^c	-0.9
51		$sp^2 - sp^2$	59.1	-4.8^d	-0.8

^{a} Defined as $\sum^{methods}\{|^1J_{CC}^{exp} - {}^1J_{CC}^{scaled}|\}/\{16\}$. ^{b} Calculated on unsubstituted bicycle[1.1.0] butane (molecule 10 in Figure 1). ^{c} Considering the static molecule instead of the dynamic tautomerism. ^{d} Calculated on the unsubstituted cyclopropene (molecule 44 in Figure 1) and considering an experimental coupling of 57.1 Hz.

acetylene, this large average deviation (−12.6 Hz) can be explained, in part, with the zero-point vibrational (ZPV) correction (−10.0 Hz) recently calculated by Ruden et al.[53] It is reasonable to ascribe a similar ZPV correction to phenylacetylene that explains also the large and negative deviation for the calculated couplings of this molecule. The deviation for the phenylethynyl cyanide is small (3.9 Hz), albeit, the coupling $^1J_{C_{sp}-C_{sp}}$ is through a single bond. Therefore, it is reasonable to think that the ZPV correction should be different from that of acetylene.

Three additional values were eliminated in the statistics because they show large deviations if comparing with the remaining calculated couplings. For these couplings (see Table 4) the averaged deviations are −9.2, 6.5, and −4.8 Hz. A detailed analysis of these molecules shows that the reported values actually correspond to derivatives. In the case of bicyclo[1.1.0]butane[43] the reported value was derived from that in 2,2,4,4-tetramethylbicyclo[1.1.0]butane.[56] Using this last molecule to calculate the $^1J_{C_1C_3}$ value and scaling them with eq 2 the average deviation reduces to −5.7 Hz. It should be noted that this coupling is the only one that presents a negative value, and this sign is reproduced in all the calculations. The reported value for cyclopropene (57.1 Hz) corresponds to that of 1,3,3-trimethylcyclopropene (59.1 Hz) corrected by 2 Hz for the methyl substitution.[57] When the coupling for 1,3,3-trimethylcyclopropene is calculated and scaled, the deviation reduces to 0.8 Hz. The last molecule that presents a large deviation is the pyrazole. We initially compared the experimental couplings with the values calculated for $^1J_{C_1C_2}$ in the static molecule. However, this molecule presents a dynamic tautomerism, and the experimental coupling is an average between $^1J_{C_1C_2}$ and $^1J_{C_1C_3}$ (see

Table 4). Considering this average calculated coupling and scaling them using eq 2 the average deviation reduces to −0.9 Hz.

It is interesting to note that the deviations for each of these six couplings are in the same direction, independently of the approach used (see Table 5 in the Supporting Information).

4. Conclusions

A set of 70 $^1J_{CC}$ coupling constants has been calculated with four functionals and four basis sets. From this set, 64 $^1J_{CC}$ couplings have been used for the statistical analysis. Compared with the experimental data, the standard deviations for B3P86/aug-cc-pVTZ-J and B3LYP/EPR-III results are 2.5 and 2.9 Hz, respectively, which are excellent considering the range of experimental values (between 10 and 91 Hz) and that the vibrational averaging effects have not been included.

B3LYP/aug-cc-pVTZ-J gives the best agreement between calculated and experimental SSCC smaller than 46 Hz ($\sigma = 0.7$ Hz), while B3P86/aug-cc-pVTZ-J results are better for all calculated values ($\sigma = 2.5$ Hz) and for SSCC larger than 46 Hz ($\sigma = 1.6$ Hz).

The standard deviations for the couplings scaled using either the equation $^1J_{CC} = 3.4 + 0.90{\cdot}^1J_{CC}^{B3LYP/EPR-III}$ or the equation $^1J_{CC} = 5.4 + 0.94{\cdot}^1J_{CC}^{B3P86/EPR-III}$ reduce to 1.3 Hz. It is interesting to note that the scaling of the economical PBE/EPR-III results achieves a standard deviation of 1.4 Hz, suggesting that the main trends on the coupling constants also are correctly represented by this functional/basis set combination. It should be noted that the good agreement with experimental obtained for the SSCC is larger than 46 Hz (set-2) with the GGA functional, PBE ($\sigma = 0.9$ Hz) and PW91 ($\sigma = 0.9$ Hz).

The TZVP basis set is suitable for large molecules due to the reduced computational cost and the reasonable results for scaled and nonscaled couplings. Couplings $^1J_{CC}$ smaller than 46 Hz can be calculated using the combination B3LYP/TZVP ($\sigma = 1.0$ Hz), while larger couplings can be obtained with B3P86/TZVP ($\sigma = 2.0$ Hz) or PBE/TZVP ($\sigma = 2.2$ Hz). The whole set of couplings can be calculated with this basis set using B3P86 ($\sigma = 3.2$ Hz).

Larger deviations found in 6 $^1J_{CC}$ couplings have been analyzed. These deviations in the scaled couplings suggest discrepancies between both calculated and experimental data, which allow for the correction of possible mistakes in the data set.

Acknowledgment. J.M.G.V. and J.S.F. gratefully acknowledge the financial support from the Dirección General de Enseñanza Superior e Investigación Científica of Spain (DGESIC, projects: CTQ2005-04469, CTQ2007-66547. and CTQ2007-63332). R.S. and R.C.O. acknowledge a research fellowship from Universidad Autónoma de Madrid. Computer time provided by the Centro de Computación Científica of Universidad Autónoma de Madrid is gratefully acknowledged.

Supporting Information Available: All calculated NMR coupling constants. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Cremer, D.; Gräfenstein, J. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2791.

(2) Krivdin, L. B.; Contreras, R. H. *Annu. Rep. NMR Spectrosc.* **2007**, *61*, 133.

(3) Vaara, J. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5399.

(4) Helgaker, T.; Jaszuński, M.; Ruud, K. *Chem. Rev.* **1999**, *99*, 293.

(5) Fukui, H. *Prog. Nucl. Magn. Reson. Spectrosc.* **1999**, *35*, 267.

(6) Maximoff, S. N.; Peralta, J. E.; Barone, V.; Scuseria, G. E. *J. Chem. Theory Comput.* **2005**, *1*, 541.

(7) Sychrovský, V.; Gräfenstein, J.; Cremer, D. *J. Chem. Phys.* **2000**, *113*, 3530.

(8) Helgaker, T.; Watson, M.; Handy, N. C. *J. Chem. Phys.* **2000**, *113*, 9402.

(9) Lantto, P.; Vaara, J.; Helgaker, T. *J. Chem. Phys.* **2002**, *117*, 5998.

(10) Keal, T. W.; Tozer, D. J.; Helgaker, T. *Chem. Phys. Lett.* **2004**, *391*, 374.

(11) Keal, T. W.; Helgaker, T.; Sałek, P.; Tozer, D. J. *Chem. Phys. Lett.* **2006**, *425*, 163.

(12) Witanowski, M.; Kamieńska-Trela, K.; Biedrzycka, Z. *J. Mol. Struct.* **2007**, *844−845*, 13.

(13) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(14) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(15) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.

(16) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.

(17) Perdew, J. P. Unified theory of the exchange and correlation beyong the local density approximation. In *Electronic Structure of Solids '91;* Ziesche, P., Eschrig, H., Eds.; Akademie Verlag: Berlin, 1991; pp 11−20.

(18) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Peterson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1992**, *46*, 6671.

(19) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Peterson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1993**, *48*, 4978.

(20) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.

(21) Wilson, P. J.; Bradley, T. J.; Tozer, D. J. *J. Chem. Phys.* **2001**, *115*, 9233.

(22) Keal, T. W.; Tozer, D. J. *J. Chem. Phys.* **2005**, *123*, 121103.

(23) Patchkovskii, S.; Autschbach, J.; Ziegler, T. *J. Chem. Phys.* **2001**, *115*, 26.

(24) Peralta, J. E.; Scuseria, G. E.; Cheeseman, J. R.; Frisch, M. J. *Chem. Phys. Lett.* **2003**, *375*, 452.

(25) Schindler, M.; Kutzelnigg, W. *J. Chem. Phys.* **1982**, *76*, 1919.

(26) Barone, V. *J. Chem. Phys.* **1994**, *101*, 6834.

(27) Barone, V. Structure, Magnetic Properties and Reactivities of Open-Shell Species from Density Functional and Self-Consistent Hybrid Methods. In *Recent Advances in Density Functional Methods Part I*; Chong, D. P., Ed.; World Scientific Publ. Co.: Singapore, 1996; pp 287−334.

(28) Provasi, P. F.; Aucar, G. A.; Sauer, S. P. A. *J. Chem. Phys.* **2001**, *115*, 1324.

(29) Sadlej, J. *Collect. Czech. Chem. Commun.* **1988**, *53*, 1995.

(30) Krivdin, L. B.; Kalabin, G. A. *Prog. NMR Spectrosc.* **1989**, *22*, 293.

(31) McLean, A. D.; Chandler, G. S. *J. Chem. Phys.* **1980**, *72*, 5639.

(32) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.

(33) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. *Can. J. Chem.* **1992**, *70*, 560.

(34) Hermosilla, L.; Calle, P.; García de la Vega, J. M.; Sieiro, C. *J. Phys. Chem. A* **2005**, *109*, 1114.

(35) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.

(36) Peterson, K. A.; Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1994**, *100*, 7410.

(37) Wilson, A. K.; van Mourik, T.; Dunning, T. H., Jr. *J. Mol. Struct. (Theochem)* **1996**, *388*, 339.

(38) Wilson, A. K.; Woon, D. E.; Peterson, K. A.; Dunning, T. H., Jr. *J. Chem. Phys.* **1999**, *110*, 7667.

(39) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1995**, *103*, 4572.

(40) Wray, V. *Prog. NMR Spectrosc.* **1979**, *13*, 177.

(41) Kaski, J.; Lantto, P.; Vaara, J.; Jokisaari, J. *J. Am. Chem. Soc.* **1998**, *120*, 3993.

(42) Krivdin, L. B.; Sauer, S. P. A.; Peralta, J. E.; Contreras, R. H. *Magn. Reson. Chem.* **2002**, *40*, 187.

(43) Krivdin, L. B. *Magn. Reson. Chem.* **2003**, *41*, 91.

(44) Krivdin, L. B. *Magn. Reson. Chem.* **2002**, *41*, 157.

(45) Krivdin, L. B. *Magn. Reson. Chem.* **2003**, *41*, 417.

(46) Krivdin, L. B. *Magn. Reson. Chem.* **2004**, *42*, S168.

(47) Sauer, S. P. A.; Krivdin, L. B. *Magn. Reson. Chem.* **2004**, *42*, 671.

(48) Ruden, T. A.; Helgaker, T.; Jaszuński, M. *Chem. Phys.* **2004**, *296*, 53.

(49) Krivdin, L. B. *Magn. Reson. Chem.* **2005**, *43*, 101.

(50) Suardíaz, R.; Pérez, C.; García de la Vega, J. M.; San Fabián, J.; Contreras, R. H. *Chem. Phys. Lett.* **2007**, *442*, 119.

(51) Oddershede, J.; Geertsen, J.; Scuseria, G. E. *J. Phys. Chem.* **1988**, *92*, 3056.

(52) Contreras, R. H.; Peralta, J. E.; Giribet, C. G.; de Azua, M. C. R.; Facelli, J. C. *Annu. Rep. NMR Spectrosc.* **2000**, *41*, 55.

(53) Ruden, T. A.; Lutnæs, O. B.; Helgaker, T.; Ruud, K. *J. Chem. Phys.* **2003**, *118*, 9572.

(54) Ruud, K.; Frediani, L.; Cammi, R.; Mennucci, B. *Int. J. Mol. Sci.* **2003**, *4*, 119.

(55) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; K. N. K.; 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 D.01;* Gaussian, Inc.: Wallingford, CT, 2004.

(56) Finkelmeier, H.; Lüttke, W. *J. Am. Chem. Soc.* **1978**, *100*, 6261.

(57) Kamieńska-Trela, K.; Bernatowicz, P.; Lüttke, W.; Machinek, R.; Trætteberg, M. *Magn. Reson. Chem.* **2002**, *40*, 640.

CT7003287

JCTC Journal of Chemical Theory and Computation

Strong Two-Photon Circular Dichroism in Helicenes: A Theoretical Investigation

Branislav Jansík

Teoretisk Kemi, Kemisk Institut, Aarhus Universitet, Langelandsgade 140,
8000 Aarhus C, Denmark

Antonio Rizzo*

Istituto per i Processi Chimico-Fisici del Consiglio Nazionale delle Ricerche
(IPCF-CNR), Area della Ricerca, via G. Moruzzi 1, I-56124 Pisa, Italy

Hans Ågren

Theoretical Chemistry, School of Biotechnology, Royal Institute of Technology,
Roslagstullsbacken 15, SE-10691 Stockholm, Sweden

Benoit Champagne

Laboratoire de Chimie Théorique Appliquée, FUNDP, Rue de Bruxelles 61,
B-5000 Namur, Belgium

Received December 10, 2007

Abstract: Using a recently derived origin-invariant quadratic response approach combined with time-dependent density functional theory, four representative helicenes are shown to present a very strong two-photon circular dichroism (TPCD) response, which makes them candidates for the first experimental observation of a TPCD effect. The large response is attributed to the unique combination of chirality and electron delocalization. Comparison with electronic circular dichroism and two-photon absorption (TPA) shows that the three effects exhibit complementary features for unravelling the molecular structures. In particular, for the four (M)-helicenes studied here, the first, i.e., low-energy, dominant Cotton band is always negative, whereas for TPCD it is positive. From an analysis of the frontier orbitals describing most of the one-electron excitation vectors, the largest TPCD response of tetramethoxy-bisquinone-dithia-[7]-helicene has been attributed to the charge-transfer character of the excited state, like for the parent TPA effect. Moreover, the TPCD intensities are found to be mostly governed by the electric and magnetic dipole contributions, while the electric quadrupole terms are, on a relative basis, less important.

I. Introduction

Helicenes are fascinating compounds with unique chiro-optical properties. Like screws, strings, propellers, and other screw-shaped objects do in everyday life, helicenes and other helical systems, including DNA and proteins, play key roles at the molecular or supramolecular levels. Helicenes are made of ortho-fused aromatic rings and combine electron delocalization and helical conformation. The nonplanarity of the π-conjugated network and the associated chirality without stereogenic center results from steric hindrance, which already appears in [4]-helicene and leads to substantial optical

* Corresponding author phone: +39-050-315 2456; fax: +39-050-315 2442; e-mail: rizzo@ipcf.cnr.it.

10.1021/ct700329a CCC: $40.75
Published on Web 02/21/2008

rotation.[1] The first helicene to be obtained in nonracemic form was hexahelicene in 1955.[2] Since then, many homo- and heterohelicenes have been prepared, and synthetic routes have been improved to include functional groups while keeping enantiomeric excess.[3−18] Helicenes have been foreseen for a broad range of applications encompassing chiro-optical photoswitches,[19] enantioselective fluorescence detectors,[20] circularly polarized luminescence for back-lighting in liquid crystals displays,[21,22] or nonlinear optical (NLO) devices.[23,24] Theoretical investigations have been carried out to assess their structures, inversion pathways, aromatic character, and magnetic susceptibility as well as the oscillatory and rotatory strengths.[25−29] These studies further demonstrated that large second-order NLO responses (first hyperpolarizabilities) could be achieved by an appropriate choice of the position and nature of the substituents or by oxidation.[30−34] This unique combination of chirality and electron delocalization undoubtedly generates outstanding properties, including nonlinear circular dichroism, the subject of this paper.

The different interactions of the mirror images of helicenes with left- and right-circularly polarized light is fundamental in determining their handedness as well as to unravel other structural characteristics. More generally, structure-chiro-optics relationships are important for the qualitative and quantitative understanding of chirality, and helicenes appear as nice model compounds to address these features. Among the chiro-optical phenomena, the differential absorption associated with electronic and vibrational circular dichroisms [(E)CD and VCD] for electronic and vibrational transitions are well accepted approaches.[35] Then, vibrational Raman optical activity (VROA) spectroscopy, which probes differential Raman scattering, is receiving increased interest both experimental and theoretically to interpret spectral signatures.[36−39] Two-photon circular dichroism (TPCD)[40] is another chiral sensitive effect, whose interest has recently been revived thanks to the development of new theoretical approaches[41−43] combined with experimental detection improvements.[44] TPCD, the difference in two-photon absorption of left and right circularly polarized light, combines the advantages of two-photon absorption (TPA), i.e. 3D confocality and reduced frequency (and therefore f. ex. reduced damages to biological samples), with the fingerprinting capabilities of circular dichroism. Together with developments of improved measurement tools, progress in using TPCD requires the elaboration of theoretical schemes for simulating and interpreting the TPCD signatures. Since helicenes are among the systems that display the largest rotatory power and rotatory strengths, they turn out to be ideal candidates to study TPCD. In this work four helicenes have been selected (Figure 1) : the classical [6]-helicene (hexahelicene) and dithia-[7]-helicene and its tetramethoxy-bisquinone (TMB) derivative as well as tetrathia-[7]-helicene. For simplicity, only the left-handed enantiomers, known as (M)-helicenes, were considered. Their TPCD spectra are simulated and interpreted at the level of density functional theory (DFT). Key theoretical aspects are presented in section II, while section III describes the computational procedure. Results and Discussion found in section IV highlight the substantial TPCD signatures of helicenes, an incentive for experimental characterization.

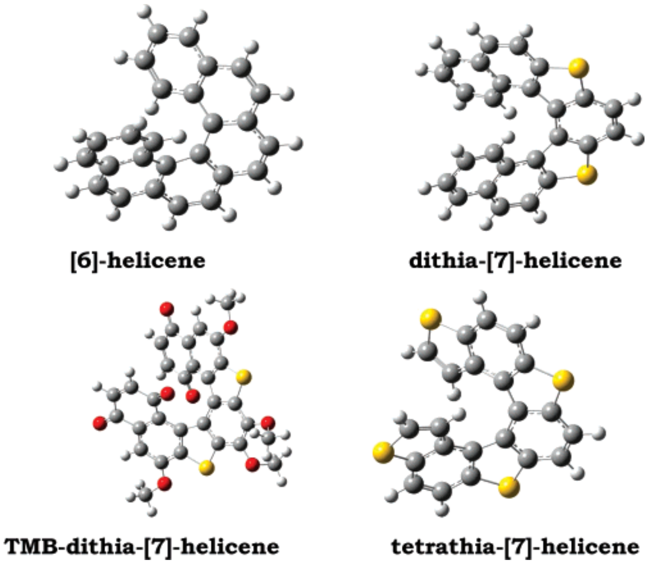


Figure 1. Sketch of the helicenes structures.

II. Theory

Two-photon circular dichroism arises from the difference in two-photon absorption of left (δ_L^{TP}) and right (δ_R^{TP}) circularly polarized light. δ indicates the two-photon absorption coefficient, and its CGS units are cm^4 s mol^{-1} photon^{-1}. The phenomenon, which is part of the vast area of high order optical activity, [45−49] has been theoretically described by Tinoco Jr.,[40] Power,[50] and Andrews[51] in the 1970s. In ref 41 we have given the definitions and discussed our computational approach to the ab initio determination of TPCD spectra. In ref 42 we presented a selection of origin invariant approaches, of which the one based on Tinoco's original formulation,[40] labeled as the "TI" approach, is employed in our present study. In ref 43 a comprehensive study of the ECD, TPA, and TPCD of all natural essential amino acids was carried out. In this section we therefore give only a brief outline the theory; for a detailed derivation we refer to refs 41 and 42.

Two-photon absorption circular dichroism is a differential effect observed when two photons (in our case, of equal frequency ω), one of which at least being of circular polarization, are absorbed inducing a transition from the initial state $|0\rangle$ to the final state $|f\rangle$ ($\hbar\omega_{0f} = 2\omega$ is the energy difference). The difference in absorption can be written, following the original expression of Tinoco Jr.,[40,42] as

$$\delta_L^{TP} - \delta_R^{TP} = \frac{4}{15}\frac{(2\pi)^2\omega^2 g(2\omega)N_A}{{c_0}^3(4\pi\epsilon_0)^2}\,{}^fR^{TP} \qquad (1)$$

$$\approx 4.67299 \times 10^{-32} \times \omega^2 g(2\omega) \times {}^fR^{TP} \qquad (2)$$

where ${}^fR^{TP}$ is the two-photon circular dichroism rotatory strength. In eq 1 $g(2\omega)$ is the normalized line shape, N_A is Avogadro's number, c_0 is the speed of light in vacuo, and ϵ_0 is the vacuum permittivity. Equation 2 yields the dichroism in CGS units from circular frequencies ω, line shapes $g(2\omega)$,

Table 1. Areas of the Portion of TPA, TPCD, and ECD Spectra Yielded by the First Six Excited States for the Four Helicenes Studied Here[a]

	TPA	TPCD × 1000	ECD
[6]-helicene	0.03	1.30 (0.30)	1.00 (−0.67)
dithia-[7]-helicene	0.58	2.27 (1.93)	0.50 (−0.30)
TMB-dithia-[7]-helicene	1.00	4.78 (2.22)	0.32 (−0.02)
tetrathia-[7]-helicene	0.56	2.90 (2.50)	0.51 (−0.15)

[a] Given relative to the area of the TPA spectrum of tetramethoxybisquinone (TMB) derivative of dithia-[7]-helicene (for TPA and TPCD) and to the (absolute) area of the ECD spectrum of hexahelicene (for ECD). For the CD spectra (TPCD and ECD) both the absolute area (obtained by integrating the absolute value of the spectral function over the whole set of frequencies) and (in parentheses) the actual area (with its sign, resulting from the balance of negative and positive portions of the spectrum) are given. Note that the relative area of the TPCD spectra is given in thousands of the reference unit. The calculations have been carried out at the TDDFT/B3LYP/aug-cc-pVDZ level of approximation.

and TPCD rotatory strengths ${}^fR^{TP}$ given in atomic units. When $\delta_L^{TP} - \delta_R^{TP}$ is computed in atomic units through eq 1, a multiplication by the conversion factor ≈1.89679 × 10^{-50} yields the value in cm^4 s mol^{-1} photon^{-1}. In ref 42 we have shown that the two-photon circular dichroism rotatory strength ${}^fR^{TP}$, in the particular formulation which we dubbed as the "TI" approach, is written as

$$ {}^fR^{TP} = -b_1\mathscr{B}_1^{TI}(\omega) - b_2\mathscr{B}_2^{TI}(\omega) - b_3\mathscr{B}_3^{TI}(\omega) \tag{3}$$

$$\mathscr{B}_1^{TI}(\omega) = \frac{1}{\omega^3}\sum_{\rho\sigma} M_{\rho\sigma}^{p,0f}(\omega)\mathscr{P}_{\rho\sigma}^{p,f0}*(\omega) \tag{4}$$

$$\mathscr{B}_2^{TI}(\omega) = \frac{1}{2\omega^3}\sum_{\rho\sigma} \mathscr{T}_{\rho\sigma}^{+,0f}(\omega)\mathscr{P}_{\rho\sigma}^{p,f0}*(\omega) \tag{5}$$

$$\mathscr{B}_3^{TI}(\omega) = \frac{1}{\omega^3}[\sum_{\rho} M_{\rho\rho}^{p,0f}(\omega)][\sum_{\sigma}\mathscr{P}_{\sigma\sigma}^{p,0f}(\omega)] \tag{6}$$

where the tensors $\mathscr{P}_{\alpha\beta}^{p,0f}(\omega_\beta)$, $M_{\alpha\beta}^{p,0f}(\omega_\beta)$, and $\mathscr{T}_{\alpha\beta}^{+,0f}(\omega_\beta)$ are defined through the following sum-over-states expressions (general case $\omega_\alpha + \omega_\beta = \omega_{0f}$)

$$\mathscr{P}_{\alpha\beta}^{p,0f}(\omega_\beta) = \frac{1}{\hbar}\sum_P\sum_{n\neq 0}\frac{(\mu_\alpha^p)^{0n}(\mu_\beta^p)^{nf}}{\omega_\alpha - \omega_{0n}} \tag{7}$$

$$M_{\alpha\beta}^{p,0f}(\omega_\beta) = \frac{1}{\hbar}\sum_P\sum_{n\neq 0}\frac{(\mu_\alpha^p)^{0n}(m_\beta)^{nf}}{\omega_\alpha - \omega_{0n}} \tag{8}$$

$$\mathscr{T}_{\alpha\beta}^{+,0f}(\omega_\beta) = \frac{1}{\hbar}\epsilon_{\beta\rho\sigma}\sum_P\sum_{n\neq 0}\frac{(T_{\alpha\rho}^{+})^{0n}(\mu_\sigma^p)^{nf}}{\omega_\alpha - \omega_{0n}} \tag{9}$$

P takes care of the permutation of the couples (operator/associated frequency), whereas the Levi-Civita $\epsilon_{\beta\rho\sigma}$ tensor in eq 9 implies Einstein summation over repeated indices (ρ and σ). The parameters b_1, b_2, and b_3 in eq 3 depend on the polarization and propagation status of the beam, and they are tabulated for a few combinations in Table 2 of ref 40. Also, the notation $(X_\alpha)^{0n}$ indicates the matrix element $< 0 \mid X_\alpha \mid n >$ of the α Cartesian component of the operator X between the ground $\mid 0 >$ and excited $\mid n >$ electronic states. The operators X appearing in the infinite summations are the velocity operator μ^p

$$\mu_\alpha^p = \sum_i \frac{q_i}{m_i} p_{i\alpha} \tag{10}$$

involving a sum over the linear momentum p_i of all particles of mass m_i and charge q_i; the magnetic dipole operator m

$$m_\alpha = \sum_i \frac{q_i}{2m_i} l_{i\alpha} = \sum_i \frac{q_i}{2m_i}(r_i \times p_i)_\alpha \tag{11}$$

involving the position r_i and angular momentum l_i operators; and the mixed length-velocity form of the quadrupole operator ($T_{\alpha\beta}^{+}$), defined as

$$T_{\alpha\beta}^{+} = \sum_i \frac{q_i}{m_i}(p_{i\alpha}r_{i\beta} + r_{i\alpha}p_{i\beta}) \tag{12}$$

Equation 3, the "TI" equation, can be proven to yield origin invariant results for the observable—the circular dichroism—independent of the completeness of the one-electron basis set employed in the calculation.[42] To end this section we recall that the TPCD rotational strength ${}^fR^{TP}$ is a quantity analogous to the ordinary ECD rotatory strength[52,53] ${}^fR$

$$ {}^fR = \frac{3}{4}\mathscr{T}[\langle 0 \mid \hat{\mu} \mid f\rangle \cdot \langle f \mid \hat{m} \mid 0\rangle] \tag{13}$$

which enters the expression of the linear circular dichroism (written here in terms of the anisotropy of the molar absorptivity ϵ)

$$\Delta\epsilon = \epsilon_L - \epsilon_R = \frac{64\pi^2\omega g(\omega)N_A}{9 \times 1000 \times \ln(10) \times (4\pi\epsilon_0) \times \hbar c_0^2} \times {}^fR \tag{14}$$

$$\approx 2.73719 \times 10^{-2} \times \omega g(\omega) \times {}^fR \tag{15}$$

Equation 15 gives $\Delta\epsilon$ in the usual units of dm^3 mol^{-1} cm^{-1} when, as for eq 2, ω and the rotatory strength ${}^fR$ are in atomic units. Moreover, in the dipole approximation the expression of the two-photon absorption for two photons of equal frequency reads usually as[40,54]

$$\delta^{TP} = \frac{(2\pi)^2\omega^2 g(2\omega)N_A}{c_0^2(4\pi\epsilon_0)^2}\frac{1}{30}\{F[\sum_\rho S_{\rho\rho}^{0f}(\omega)]^2 + (G+H)(\sum_{\rho\sigma} S_{\rho\sigma}^{0f}(\omega)S_{\rho\sigma}^{0f}(\omega))\}$$

$$= \frac{(2\pi)^2\omega^2 g(2\omega)N_A}{c_0^2(4\pi\epsilon_0)^2}\frac{1}{30}\bar{\delta} \tag{16}$$

$$\approx 8.00460 \times 10^{-31} \times \omega^2 g(\omega) \times \bar{\delta} \tag{17}$$

where the $S_{\alpha\beta}^{0f}(\omega_\beta)$ tensor elements given by

Table 2. Excitation Energy $\hbar\omega_{0n}$ (eV), Wavelength λ (nm), Parameters $\mathscr{B}_1^{TI}$ (Eq 4), $\mathscr{B}_2^{TI}$ (Eq 5), and $\mathscr{B}_3^{TI}$ (Eq 6), and Rotational Strengths ${}^fR^{TP}$ (Eq 3), $\bar{\delta}$ (Eq 16), and ${}^fR$ (Eq 13) for Each of the Six Lowest Excited States of the Four Helicenes Studied Here at the TDDFT/B3LYP/aug-cc-pVDZ Level of Approximation[a]

	state (n)	$\hbar\omega_{0n}$ (eV)	λ (nm)	B_1^{TI}	B_2^{TI}	B_3^{TI}	${}^fR^{TP}$	$\bar{\delta}$	$10^3 \times {}^fR$
[6]-helicene[b]	1	3.20	387.1	−536	84	0	3046	572	0.7
	2	3.36	369.5	97	46	−231	−1135	999	4.5
	3	3.62	342.2	−28	1	0	168	1355	−1073.8
	4	3.72	333.3	0	−71	−260	−376	1854	149.3
	5	3.84	323.0	−5	−18	0	65	1270	−109.7
	6	3.93	315.8	50	7	−109	−534	1123	82.6
dithia-[7]-helicene	1	3.14	395.4	15	28	0	−144	2650	−151.2
	2	3.31	374.3	42	−1	481	713	883	115.2
	3	3.65	339.7	−1882	−43	−2331	6719	80475	26.0
	4	3.81	325.2	31	−88	−210	−427	4903	−43.0
	5	3.85	321.7	12	28	0	−130	491	−364.5
	6	3.98	311.6	−204	27	0	1167	49478	−5.2
TMB-dithia-[7]-helicene	1	2.03	610.9	311	−65	890	42	5152	2.5
	2	2.18	568.4	−2326	−155	0	14265	105094	−201.0
	3	2.29	540.8	287	−62	0	−1599	4324	41.7
	4	2.48	499.9	51	7	0	−319	49462	56.3
	5	2.50	495.4	−497	−164	−2168	−1030	47714	−44.0
	6	2.64	469.5	333	25	−120	−2291	27047	115.9
tetrathia−[7]−helicene	1	3.22	384.9	−258	−43	−89	1456	1260	116.6
	2	3.22	384.8	13	−11	0	−56	2417	−276.5
	3	3.78	327.6	148	−68	0	−752	34356	−149.8
	4	3.81	325.5	−1934	354	−2314	6267	65678	−2.8
	5	4.06	305.7	−60	30	0	301	9970	−40.7
	6	4.09	302.8	−650	−54	−499	3014	20138	133.9

[a] The TPA strength and the TPCD rotational strength has been computed for two circularly polarized photons ($b_1 = G + H = 6$, $b_2 = -b_3 = F = 2$). Atomic units where not explicitly specified. [b] For the five lowest lying excitation energies (nm) and corresponding rotational strengths ${}^fR$ ($\times$ 10^3 au, in parentheses) Furche et al.[27] computed (at the TDDFT level, using the BP86 XC functional and with an SV(P)+ basis set): 411. (−2.); 395. (6.); 365. (−457.); 364. (87.); 358. (−291.). The experimental values taken from ref 68 and referring to the measurements of ref 69 are as follows: 412. (2.5); 347. (−137.); 325. (−393.); 292. (−17); 244. (609.). Note that, besides the conversion of units, the rotational strengths taken from refs 27 and 68 are multiplied here by the factor 3/4, which makes them consistent with the convention used in eq 13.

$$S^{0f}_{\alpha\beta}(\omega_\beta) = \frac{1}{\hbar}\sum_P \sum_{n\neq 0} \frac{(\mu_\alpha)^{0n}(\mu_\beta)^{nf}}{\omega_\alpha - \omega_{0n}} \tag{18}$$

are related to those of the $\mathscr{P}^{p,0f}_{\alpha\beta}(\omega_\beta)$ tensor in the limit of a complete one-electron basis set by the relationship

$$\mathscr{P}^{p,0f}_{\alpha\beta}(\omega_\beta) = -\omega_\alpha{}^2 S^{0f}_{\alpha\beta}(\omega_\beta) \tag{19}$$

but are defined using the traditional electric dipole moment operator, μ

$$\mu_\alpha = \sum_i q_i r_{i\alpha} \tag{20}$$

As b_1, b_2, and b_3, see above, the F, G, and H parameters take different values for different polarization and propagation conditions of the two photons. Note that the following relationship holds (cf. eqs 1 and 16 above)

$$\frac{\delta_L^{TP} - \delta_R^{TP}}{\delta^{TP}} = \frac{8}{c_0}\frac{{}^fR^{TP}}{\bar{\delta}} \tag{21}$$

Again, eq 17 above can be used to obtain the TPA rate in the absolute units of cm^4 s mol^{-1} photon^{-1} from all quantities involved—ω, $g(2\omega)$, and $\bar{\delta}$—given in atomic units.

III. Computational Details

The computational schemes employed for the calculation of TPCD, TPA, and ECD have been elaborated and detailed in refs 41 and 42. As shown in ref 41, two-photon circular dichroism can be evaluated via analytical response theory, in a formulation where the summation over intermediate states is replaced by the solution of linear equations, without explicit knowledge of the excited-state wave functions. As a consequence, the properties of interest are obtained in a size-extensive manner, whenever the computational model is itself size-extensive. In the present case, time-dependent density functional theory (TDDFT) with frequency dependent quadratic response theory where both the density and its response are computed employing the B3LYP exchange-correlation functional[55−58] is employed to model TPA and TPCD.

We have calculated the one- and two-photon circular dichroism spectra and the two-photon absorption spectra for the six lowest excited states $|f>$ of the four helicenes. Their structure is shown in Figure 1. The calculations involved neutral, gas-phase, molecules. Note that, with respect to ref 10, the dodecyloxy groups were replaced by methoxy groups in order to reduce the computational cost with only a minor impact on the electronic structure and properties of the helicene. The molecular geometry was taken from B3LYP/6-31G* optimization. The six excited states energies ω_{0f} were obtained from the poles of a linear response function.[59] The two-photon circular dichroism rotatory strength ${}^fR^{TP}$ was calculated within the "TI" formulation defined in ref 42 and

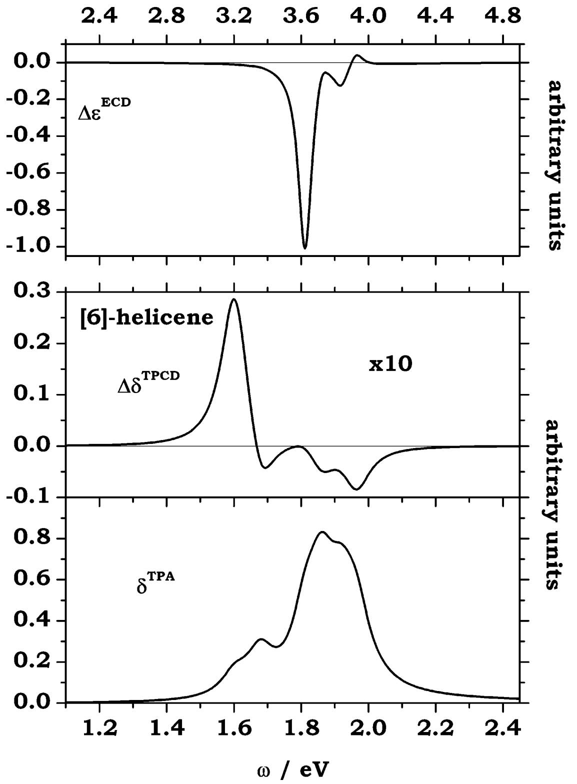


Figure 2. ECD, TPCD, and TPA spectra obtained at the DFT/B3LYP/aug-cc-pVDZ level arising from the lowest six excited electronic states for [6]-helicene.

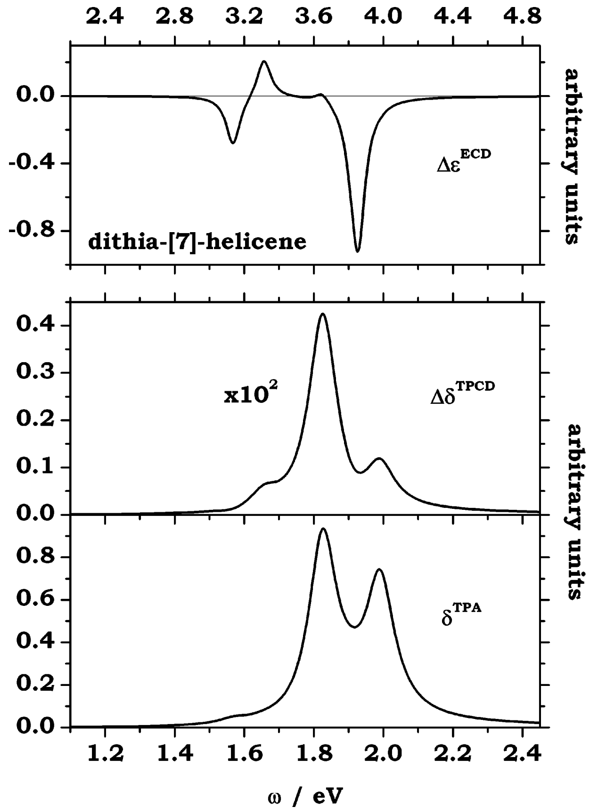


Figure 3. ECD, TPCD, and TPA spectra obtained at the DFT/B3LYP/aug-cc-pVDZ level arising from the lowest six excited electronic states for dithia-[7]-helicene.

briefly outlined in eqs 3−9 above. The $\mathscr{P}^{p,0f}_{\alpha\beta}(\omega_\beta)$, $M^{p,0f}_{\alpha\beta}(\omega_\beta)$, and $\mathscr{T}^{+,0f}_{\alpha\beta}(\omega_\beta)$ tensors were then evaluated as single residues of the appropriate quadratic response functions within the response theory framework, for each final excited state $|f>$ at the frequency $\omega = \omega_{0f}/2$, adopting the procedure described in ref 43 to ensure that phase factors were all properly taken into account. As stated above, all the property calculations were performed employing the B3LYP exchange-correlation functional.[55−57] The aug-cc-pVDZ basis set[60] was used throughout.

Both the absorption and circular dichroism spectra presented here were obtained, according to eqs 16, 14 and 1, assuming a Lorentzian as line shape function—$g(n\omega)$, $n = 1,2$—with a full width at half-maximum Γ of 0.1 eV and with maxima determined so that each Lorentzian, when integration is performed over the whole frequency spectrum, yields the value of the TPA strength ($\bar{\delta}$), TPCD (${}^fR^{TP}$), or ECD (${}^fR$) rotational strength for the given excited state. A Γ of 0.1 eV appears to be a reasonable assumption, considering current spectroscopic spectral resolution capabilities. It is to be noted that the spectral profile may change quite heavily as the value of Γ is varied. The results shown and discussed in the following for the two-photon processes correspond to an experimental setup with two left circularly polarized beams of equal frequency propagating parallel to each other. For this arrangement, $F = -2$, $G + H = 6$, $b_1 = 6$, and $b_2 = -b_3 = 2$.[40]

The intensities of the two-photon spectra presented in the next section are of arbitrary units and normalized in each figure to the area of the two-photon absorption spectrum, thus allowing for an absolute comparison of the TPA and TPCD spectra. For this purpose the prefactor in eq 21, see also ref 41, was included into the scaling of the TPCD spectra. The ECD spectra, obtained applying eq 14 are also given in arbitrary units.

All calculations were carried out using a parallel version of the DALTON 2.0 electronic structure program.[61]

IV. Results and Discussion

TMB-dithia-[7]-helicene appears to be the strongest two-photon absorber within our selection of helicenes, see Table 1, 1−2 orders of magnitude more effective than[6]-helicene and approximately twice as strong as the di- and tetrathia-[7]-helicenes. It is also the most efficient in dichroic response, with its TPCD spectrum covering, in the range of frequencies chosen here, ca. 4.8‰ of its TPA response. These features can be associated with the presence of symmetric substitutions by donors and acceptors.[62−67] The ratios between the TPCD and TPA areas are thus slightly larger for these helicenes than for the amino acids studied in refs 41 and 43, where they amount to up to 2−5‰. Note however the extremely favorable TPCD/TPA ratio in the case of [6]-helicene, where the absolute area of the TPCD spectra is only a factor of ≈30 smaller than that of the corresponding TPA spectrum as a result of its smaller TPA response compared to the other helicenes. [6]-Helicene appears to be the most efficient in the linear dichroism response, its ECD spectrum covering an area ca. twice that of the di- and tetrathia-[7]-helicenes and ca. three times that of the tetramethoxy-

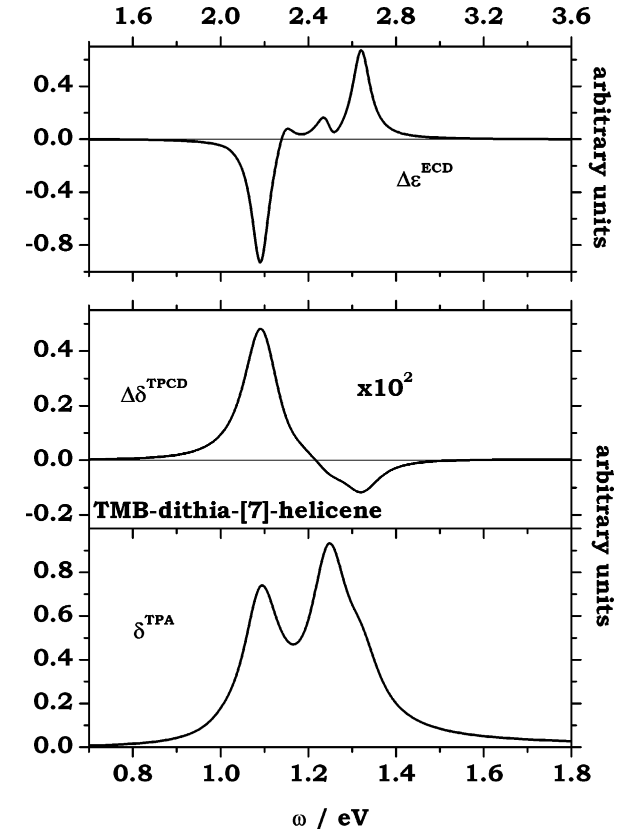


Figure 4. ECD, TPCD, and TPA spectra obtained at the DFT/B3LYP/aug-cc-pVDZ level arising from the lowest six excited electronic states for TMB-dithia-[7]-helicene.

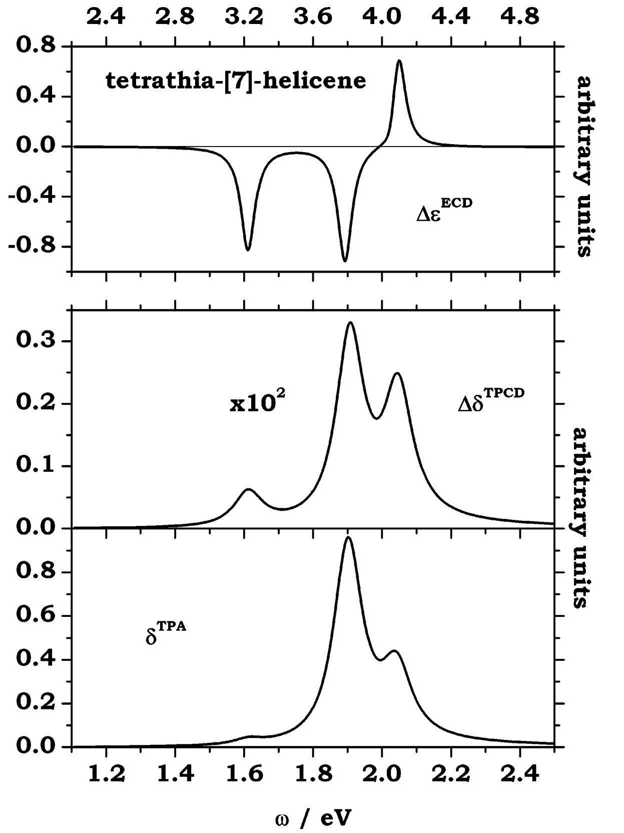


Figure 5. ECD, TPCD, and TPA spectra obtained at the DFT/B3LYP/aug-cc-pVDZ level arising from the lowest six excited electronic states for tetrathia-[7]-helicene.

bisquinone derivative. The largest ECD intensity obtained by integrating the area of the peaks is therefore found for the homohelicene, and it decreases with the number of thiophene rings or with the substituents.

Rotational strengths and spectral characteristics are listed in Table 2, whereas the spectra are displayed in Figures 2−5. From the data in Table 2, with the help of eqs 2, 15, and 17, the spectroscopic properties (TPCD, ECD and TPA rates, respectively) can be obtained in the commonly employed absolute units. Figure 6 shows comparison between our ECD simulated spectra and experiment. Note that, where available, experiment is performed in solution (see caption for details), and the corresponding spectra are reported without further elaboration, with the form and units given in the original references. Our data are given as $\Delta\epsilon$ in the usual units of dm^3 mol^{-1} cm^{-1}. It is beyond the scope of this study to comment in detail on the individual features of the experimental vs our isolated molecule approximation spectra or to speculate on the effect of intermolecular interactions in solution for ECD. Figure 6 should therefore only be intended to provide support to the very general and concise comments given in the following paragraphs.

The four helicenes exhibit excited states with substantial $^{f}R^{TP}$ values, more than 1 order of magnitude larger than for any of the essential proteinogenic amino acids of refs 41 and 43. For instance, $^{f}R^{TP}$for the second excited-state of TMB-dithia-[7]-helicene amounts to more than 14 000 au. As for the ECD intensities, these large chiro-optical responses result from the fact that the π-conjugated electron network is associated with chirality.[1]

In fact, see also refs 27 and 28, the TDDFT approach generally underestimates the excitation energies, though to a lesser extent when using the B3LYP exchange-correlation functional since it contains 20% of exact Hartree−Fock exchange. For [6]-helicene, our TDDFT results are in agreement with the TDDFT data of ref 27 where the main low-energy band is shown to be associated with the 2^1B excited state. On the other hand, with respect to experiment, a blue-shift has to be applied to match the spectra. For TMB-dithia-[7]-helicene the simulated ECD spectrum, which exhibits a negative Cotton effect at a rather large wavelength (568 nm) followed by a small and then a large positive band, also reproduces the sign alternation of the Cotton effects observed in the experimental spectrum,[10] recorded for a dilute solution to avoid the formation of aggregates. Differences of shapes between the simulated and experimental spectra originate from the absence of vibronic treatment of our simulation. Yamada et al.[5] reported the ECD spectrum of tetrathia-[7]-helicene. With the exception of the near degeneracy of the two first transitions, which does not allow for the reproduction of the positive experimentally observed Cotton effect of the lowest-energy excitation, the agreement is also good. Finally, to our knowledge, the experimental ECD spectrum of dithia-[7]-helicene is not known. In these left-handed helicenes, the lowest-energy dominant band always displays a global negative Cotton effect. On the other hand, for the same helicenes, the global TPCD signal due to the six lowest-

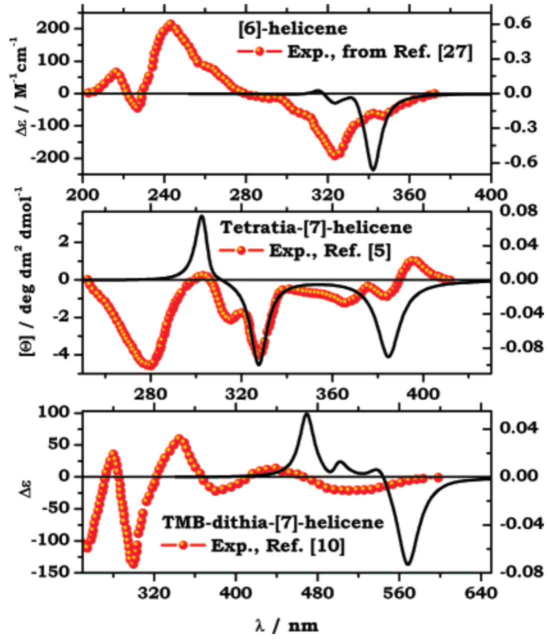


Figure 6. Comparison between experimental and simulated ECD spectra for three of the helicenes studied here. Our spectra, ordinates on the right, show $\Delta\epsilon$ in units of dm^3 mol^{-1} cm^{-1}. Experiment (ordinate on the left) is given in the form and units found in the original papers. In particular: a) for [6]-helicene the spectrum is taken from Figure 5 of ref 27, it comes originally from ref 69 (see also ref 68), and it was taken in a methanol solution; b) for tetrathia-[7]-helicene we report the spectrum found in Figure 2 of ref 5, taken in solution of chloroform; and c) for TMB-dithia-[7]-helicene we refer to Figure 1 of ref 10, in particular to the one recorded on a 2 × 10^{-5} M solution of tetradodecyloxy-helice-bisquinone in dodecane. Note a change of sign with respect to the spectrum in ref 10, which was taken on the P-helicene.

energy excitations is positive, see Table 1, and the first, i.e., low-energy, dominant band is always positive. In the dithia- and tetrathia-[7]-helicene the TPCD profile is positive over the low-excitation energy window, with our choice of line shape width Γ of 0.1 eV (ca 800 cm^{-1}). On the other hand, for the two other compounds, a first positive band at lowest energy is followed by a negative pattern. Summarizing this analysis, ECD and TPCD bring complementary chiro-optical information on the helicenes, a neat example of how nonlinear spectroscopies probe aspects of chirality which are different from those yielded by linear spectroscopies.[70]

The TPA spectra display essentially two major peaks. Their relative intensities change from one compound to another. In particular, the relative intensity of the lowest energy band with respect to the intensity of the second band increases as the number of thiophene rings increases.

The largest contributions to $^fR^{TP}$ come from the $\mathscr{B}_1^{TI}$ and $\mathscr{B}_3^{TI}$ terms, demonstrating the lesser importance of the electric quadrupole terms with respect to the magnetic dipole terms. This substantiates partly the general assumption on the negligible amplitude of the electric quadrupolar effects with respect to the magnetic dipole contributions.[71] On the other hand, the relative importance of the two dominant magnetic dipole terms—as well as their signs—depend much on the excited-state, and no consistent pattern could be pinpointed from Table 2.

Table 3. Largest Contributions to the Excitation Vectors for the Lowest Six Excited States of [6]-Helicene[a]

state (n)	$\hbar\omega_{0n}$ (eV)	excitation		coefficient
1	3.20	HOMO−1	LUMO	−0.510
		HOMO	LUMO+1	−0.476
2	3.36	HOMO−1	LUMO+1	0.338
		HOMO	LUMO	−0.616
3	3.62	HOMO−1	LUMO	−0.465
		HOMO	LUMO+1	0.507
4	3.72	HOMO−3	LUMO	0.307
		HOMO−1	LUMO+1	0.510
		HOMO	LUMO	0.292
5	3.84	HOMO−2	LUMO	0.684
6	3.93	HOMO−3	LUMO	0.322
		HOMO−2	LUMO+1	−0.380
		HOMO−1	LUMO+1	−0.306
		HOMO	LUMO+3	−0.294

[a] In terms of single excitations between the molecular orbitals depicted in Figure 6. Atomic units where not explicitly specified.

Table 4. Largest Contributions to the Excitation Vectors for the Lowest Six Excited States of Dithia-[7]-helicene[a]

state (n)	$\hbar\omega_{0n}$ (eV)	excitation		coefficient
1	3.14	HOMO	LUMO	−0.697
2	3.31	HOMO−1	LUMO	0.671
3	3.65	HOMO−2	LUMO	−0.536
		HOMO	LUMO+1	0.446
4	3.81	HOMO−2	LUMO	−0.426
		HOMO	LUMO+1	−0.534
5	3.85	HOMO−3	LUMO	−0.387
		HOMO−1	LUMO+1	0.520
6	3.98	HOMO−3	LUMO	0.541
		HOMO−1	LUMO+1	0.442

[a] In terms of single excitations between the molecular orbitals depicted in Figure 6. Atomic units where not explicitly specified.

Table 5. Largest Contributions to the Excitation Vectors for the Lowest Six Excited States of TMB-dithia-[7]-helicene[a]

state (n)	$\hbar\omega_{0n}$ (eV)	excitation		coefficient
1	2.03	HOMO	LUMO	0.700
2	2.18	HOMO−1	LUMO	−0.685
3	2.29	HOMO	LUMO+1	0.702
4	2.48	HOMO−4	LUMO	0.283
		HOMO−2	LUMO	−0.594
5	2.50	HOMO−1	LUMO+1	0.625
6	2.64	HOMO−6	LUMO	0.379
		HOMO−5	LUMO	0.306
		HOMO−4	LUMO+1	−0.276
		HOMO−2	LUMO+1	0.237
		HOMO−1	LUMO+1	0.303

[a] In terms of single excitations between the molecular orbitals depicted in Figure 6. Atomic units where not explicitly specified.

Tables 3−6 list the major contributions to the excitation vectors of the lowest six excited states of each helicene, while selected MOs are shown in Figure 7. For all excited states of interest in the tables only coefficients of absolute value larger than 0.2 are reported, whereas the coefficients corre-

Table 6. Largest Contributions to the Excitation Vectors for the Lowest Six Excited States of Tetrathia-[7]-helicene[a]

state (*n*)	$\hbar\omega_{0n}$ (eV)	excitation		coefficient
1	3.22	HOMO−1	LUMO	0.686
2	3.22	HOMO	LUMO	−0.700
3	3.78	HOMO−2	LUMO	−0.681
4	3.81	HOMO−3	LUMO	0.668
		HOMO	LUMO+1	0.208
5	4.06	HOMO−1	LUMO+1	0.641
		HOMO	LUMO+2	0.267
6	4.09	HOMO−1	LUMO+2	0.231
		HOMO	LUMO+1	−0.618

[a] In terms of single excitations between the molecular orbitals depicted in Figure 6. Atomic units where not explicitly specified.

sponding to de-excitations in the paired structure of the excitation vectors[59,72] are neglected. In the case of [6]-helicene, two different and almost "opposite" (in sign) combinations of HOMO − 1 → LUMO and HOMO → LUMO + 1 excitations are responsible for both the dominant bands in ECD and TPCD. Indeed, the first combination leads to the TPCD strong absorbing state at 3.20 eV (${}^{f}R^{TP} = 3046$ au), whereas the second combination yields the ECD intense state at 3.62 eV (${}^{f}R = -675$ au). This is another evidence of the fact that ECD and TPCD are governed by different mechanisms. In the case of dithia-[7]-helicene and tetrathia-[7]-helicene the characters—and therefore the ECD, TPA, and TPCD amplitudes and signs—of the two first transitions are inverted. In dithia-[7]-helicene the first band at 3.14 eV is described by a HOMO → LUMO transition, while the HOMO − 1 → LUMO single excitation governs the second excited state (3.31 eV). Then, for tetrathia-[7]-helicene, as a result of the presence of two additional thiophene rings, the two first transitions are almost degenerate and have excitation energies of 3.22 eV. Nevertheless, the corresponding excited states keep distinct characters as shown by their singly excited determinant vectors. For these thiophene-containing helicenes, the most intense TPCD band, corresponding to the third excited state at 3.65 eV for dithia-[7]-helicene and to the fourth excited state at 3.81 eV for tetrathia-[7]-helicene, see Table 2, are related to HOMO − 2 → LUMO (dithia-[7]-helicene) and HOMO − 3 → LUMO (tetrathia-[7]-helicene) transitions. These dominant TPCD excitations have in fact similar character since the HOMO − 2 of dithia-[7]-helicene and the HOMO − 3 of tetrathia-[7]-helicene look alike, i.e., present much similarities in their nodal structures, see Figure 7. The most ECD-intense excited states are instead the fifth (at 3.85 eV) for the dithia and the second (at 3.22 eV) for the tetrathia-[7]-helicenes. The former is essentially the result of the combination of HOMO − 3 → LUMO and HOMO − 1 → LUMO + 1 transitions, whereas the latter is essentially a clean HOMO → LUMO transition. In the case of the substituted helicene, the strong ECD and TPCD signals are both associated with the second excited state at 2.18 eV, which is mostly described as a HOMO − 1 → LUMO transition. The MOs of the TMB-dithia-[7]-helicene are different from those of the other helicenes, where the frontier orbitals are delocalized over the whole systems. Indeed, the LUMO is localized on the terminal quinone groups, whereas the HOMO − 1 is localized on the thiophene and benzene rings, demonstrating the charge-transfer character of the transition.

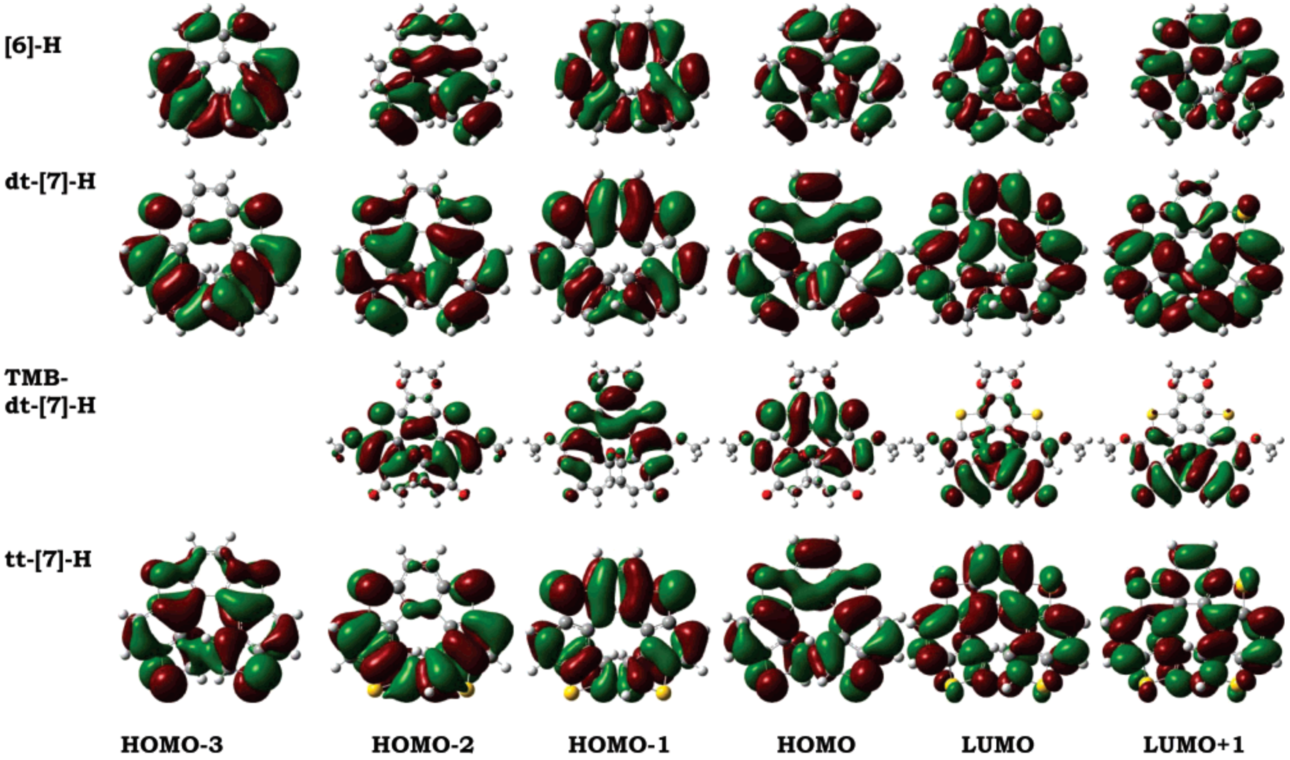


Figure 7. The relevant molecular orbitals of [6]-helicene, dithia-[7]-helicene, TMB-dithia-[7]-helicene, and tetrathia-[7]-helicene, obtained at the B3LYP/aug-cc-pVDZ level.

Conclusions

Using a recently derived origin-invariant quadratic response approach combined with time-dependent density functional theory, here it was demonstrated that helicenes present a strong two-photon circular dichroism (TPCD) response, which makes them candidates for a first experimental observation of TPCD. Indeed the $\Delta\delta^{TPCD}/\bar{\delta}$ ratio amounts to a few per mil, close to the detection limit estimated in ref 44. In that reference Markowitz and co-workers have discussed modified Z-scan techniques which the authors claim should be able to afford a way to measure nonlinear circular birefringences and two-photon circular dichroism. The authors estimated upper limits for the chiral modulations of nonlinear optic effects (the relative difference between the values of a nonlinear coefficient for the left- and right-handed circularly polarized light) that their techniques should be able to detect. In particular, such an upper limit for $\Delta\delta^{TPCD}/\bar{\delta}$, relevant for TPCD, is placed at $\approx 3 \times 10^{-3}$. Moreover, rather recently Li and co-workers[73] have performed studies of gas-phase and solution linear and nonlinear circular dichroism of R-(+)-3-methylcyclopentanone, involving measurements of (2 + 1) resonance-enhanced multiphoton ionization circular dichroism, a process involving a TPCD step.

The large response is attributed to the unique combination of chirality and electron delocalization. Comparison with electronic circular dichroism (ECD) and two-photon absorption (TPA) shows that the three effects exhibit complementary features for unravelling the molecular structures. From an analysis of the frontier orbitals mostly involved in the one−electron excitation vectors, the largest TPCD response of tetramethoxy-bisquinone-dithia-[7]-helicene has been attributed to the charge-transfer character of the excited state, like for the parent TPA effect. The simulated ECD spectral characteristics have further been found to be in good agreement with experimental data, including the fact that the first low-energy dominant Cotton band is negative for (M)-helicenes. On the other hand, for the same handedness, the first dominant TPCD band is positive and mostly described by the magnetic dipole terms rather than by the quadrupole term.

Linear and in particular nonlinear spectroscopic properties as those discussed here, multiphoton absorption and dichroism, involving mixed electric and magnetic transitions, are demanding quantities. Their computational analysis requires care in the choice of approximations, adequate basis sets, an account for the inherent origin dependence of magnetic properties in approximate calculations, and, last but not least, caution in the use of functionals when resorting to DFT. The basis sets employed here are quite reasonable for the rather extended systems studied, and we employ origin independent approaches when needed. We have reasons to believe therefore that the major limitation of the present work might reside in the choice of the exchange correlation (XC) functional, which, in spite of its vast popularity and rather good performance in studies of other mixed electric-magnetic frequency dependent high order properties,[74,75] might be not completely adequate when delocalization effects or charge-transfer excitations become important, as they are proven to be here. It must be noted that several XC functionals, including B3LYP, have been shown to reproduce satisfactorily the ECD (and UV) spectra of helicenes, which might also be affected by the limitations of the functionals to treat long-range effects. Also, the drawbacks observed in DFT with conventional XC functionals are often associated with systems that are far more extended in space than those studied here, or far more conjugated, as polyacetylene and polydiacetylene chains. In recent times it was shown that the Coulomb Attenuating Method B3LYP (CamB3LYP) functional[76] may give good performance when dealing with two-photon absorption.[77] Nevertheless, ongoing studies carried out within our group along these lines on other systems lead us to believe that a different and more proper choice of functional, for instance in the direction of including long-range effects, might improve the agreement between theory and experiment on the position of the absorption peaks and influence the general features of our absorption and dichroism spectra, all the more for higher order processes. However, it should not be able disprove our evidence on the strong responses exhibited by the helicenes studied here.

Acknowledgment. This work has been supported by the European Research and Training Network "Understanding Nanomaterials from a Quantum Perspective" (NANOQUANT), contract No. MRTN-CT-2003-506842. B.C. thanks the FNRS for his Research director position.

References

(1) Martin, R. H. *Angew. Chem., Int. Ed.* **1974**, *13*, 649.

(2) Newman, M. S.; Lutz, W. B.; Lednicer, D. *J. Am. Chem. Soc.* **1955**, *77*, 3420.

(3) Groen, M. B.; Wynberg, H. *J. Am. Chem. Soc.* **1971**, *93*, 2968.

(4) Groen, M. B.; Schadenberg, H.; Wynberg, H. *J. Org. Chem.* **1971**, *36*, 2797.

(5) Yamada, K. I.; Tanaka, H.; Nakagawa, H.; Ogashiwa, S.; Kawazura, H. *Bull. Chem. Soc. Jpn.* **1982**, *55*, 500.

(6) Katz, T. J.; Liu, L.; Willmore, N. D.; Fox, J. M.; Rheingold, A. L.; Shi, S.; Nuckolls, C.; Rickman, B. H. *J. Am. Chem. Soc.* **1997**, *119*, 10054.

(7) Grimme, S.; Harren, J.; Sobanski, A.; Vögtle, F. *Eur. J. Org. Chem.* **1998**, 1491.

(8) Nuckolls, C.; Katz, T. J. *J. Am. Chem. Soc.* **1998**, *120*, 9541.

(9) Rajca, A.; Wang, H.; Pink, M.; Rajca, S. *Angew. Chem., Int. Ed.* **2000**, *39*, 4481.

(10) Phillips, K. E. S.; Katz, T. J.; Jockusch, S.; Lovinger, A.; Turro, N. *J. Am. Chem. Soc.* **2001**, *123*, 11899.

(11) Tanaka, K.; Osuga, H.; Kitahara, Y. *J. Org. Chem.* **2002**, *67*, 1795.

(12) Maiorana, S.; Papagni, A.; Licandro, E.; Annunziata, R.; Paravidino, P.; Perdicchia, D.; Giannini, C.; Bencini, M.; Clays, K.; Persoons, A. *Tetrahedron* **2003**, *59*, 6481.

(13) Field, J. E.; Hill, T. J.; Venkataraman, D. *J. Org. Chem.* **2003**, *68*, 6071.

(14) Rajca, A.; Miyasaka, M.; Pink, M.; Wang, H.; Rajca, S. *J. Am. Chem. Soc.* **2004**, *126*, 15211.

(15) Baldoli, C.; Bossi, A.; Giannini, C.; Licandro, E.; Maiorana, S.; Perdicchia, D. *Synlett* **2005**, 1137.

(16) Bazzini, C.; Brovelli, S.; Caronna, T.; Gambarotti, C.; Giannone, M.; Macchi, P.; Meinardi, F.; Mele, A.; Panzeri, W.; Recupero, F.; Sironi, A.; Tubino, R. *Eur. J. Org. Chem.* **2005**, 1247.

(17) Collins, S. K.; Grandbois, A.; Vachon, M. P.; Côté, J. *Angew. Chem., Int. Ed.* **2006**, *45*, 2923.

(18) Abbate, S.; Bazzini, C.; Caronna, T.; Fontana, F.; Gambarotti, C.; Gangemi, F.; Longhi, G.; Mele, A.; Sora, I. N.; Panzeri, W. *Tetrahedron* **2006**, *62*, 139.

(19) Wigglesworth, T. J.; Sud, D.; Norsten, T. B.; Lekhi, V. S.; Branda, N. R. *J. Am. Chem. Soc.* **2005**, *127*, 7272.

(20) Reetz, M. T.; Sostmann, S. *Tetrahedron* **2001**, *55*, 2515.

(21) Field, J. E.; Muller, G.; Riehl, J. P.; Venkataraman, D. *J. Am. Chem. Soc.* **2005**, *125*, 11808.

(22) Hassey, R.; Swain, E. J.; Hammer, N. I.; Venkataraman, D.; Barnes, M. D. *Science* **2006**, *314*, 1437.

(23) Verbiest, T.; Van Elsocht, S.; Kauranen, M.; Hellemans, L.; Snauwaert, J.; Nuckolls, C.; Katz, T. J.; Persoons, A. *Science* **1998**, *282*, 913.

(24) Verbiest, T.; Sioncke, S.; Persoons, A.; Vylicky, L.; Katz, T. J. *Angew. Chem., Int. Ed.* **2002**, *41*, 3882.

(25) Buss, V.; Kolster, K. *Chem. Phys.* **1996**, *203*, 309.

(26) Schulman, J. M.; Disch, R. L. *J. Phys. Chem. A* **1999**, *103*, 6669.

(27) Furche, F.; Ahlrichs, R.; Wacksmann, C.; Weber, E.; Sobanski, A.; Vögtle, F.; Grimme, S. *J. Am. Chem. Soc.* **2000**, *122*, 1717.

(28) Autschbach, J.; Ziegler, T.; van Gisbergen, S. J. A.; Baerends, E. J. *J. Chem. Phys.* **2002**, *116*, 6930.

(29) Spassova, M.; Asselberghs, I.; Verbiest, T.; Clays, K.; Botek, E.; Champagne, B. *Chem. Phys. Lett.* **2007**, *439*, 213.

(30) Daul, C. A.; Ciofini, I.; Weber, V. *Int. J. Quantum Chem.* **2003**, *91*, 297.

(31) Botek, E.; Champagne, B.; Turki, M.; André, J. M. *J. Chem. Phys.* **2004**, *120*, 2042.

(32) Champagne, B.; André, J. M.; Botek, E.; Licandro, E.; Maiorana, S.; Bossi, A.; Clays, K.; Persoons, A. *Chem. Phys. Chem* **2005**, *5*, 1438.

(33) Botek, E.; Spassova, M.; Champagne, B.; Asselberghs, I.; Persoons, A.; Clays, K. *Chem. Phys. Lett.* **2005**, *412*, 274.

(34) Botek, E.; Spassova, M.; Champagne, B.; Asselberghs, I.; Persoons, A.; Clays, K. *Chem. Phys. Lett.* **2006**, *417*, 282 (Erratum).

(35) Barron, L. D. *Molecular light scattering and optical activity*; Cambridge University Press: Cambridge, 2004.

(36) Hug, W.; Hangartner, G. *J. Raman Spectrosc.* **1999**, *30*, 841.

(37) Pecul, M.; Rizzo, A.; Leszczynski, J. *J. Phys. Chem. A* **2002**, *106*, 11008.

(38) Hug, W.; Haesler, J. *Int. J. Quantum Chem.* **2005**, *104*, 695.

(39) Herrmann, C.; Ruud, K.; Reiher, M. *ChemPhysChem* **2006**, *7*, 7189.

(40) Tinoco, I., Jr. *J. Chem. Phys.* **1975**, *62*, 1006.

(41) Jansík, B.; Rizzo, A.; Ågren, H. *Chem. Phys. Lett.* **2005**, *414*, 461.

(42) Rizzo, A.; Jansík, B.; Bondo Pedersen, T.; Ågren, H. *J. Chem. Phys.* **2006**, *125*, 064113.

(43) Jansík, B.; Rizzo, A.; Ågren, H. *J. Phys. Chem. B* **2007**, *111*, 446. Jansík, B.; Rizzo, A.; Ågren, H. *J. Phys. Chem. B* **2007**, *111*, 2409 (Erratum).

(44) Markowicz, P. P.; Samoc, M.; Cerne, J.; Prasad, P. N.; Pucci, A.; Ruggeri, G. *Opt. Expr.* **2004**, *12*, 5209.

(45) Wagnière, G. *J. Chem. Phys.* **1982**, *77*, 2786.

(46) Meath, W. J.; Power, E. A. *J. Phys. B.: At. Mol. Phys.* **1984**, *17*, 763.

(47) Meath, W. J.; Power, E. A. *Mol. Phys.* **1984**, *51*, 585.

(48) Meath, W. J.; Power, E. A. *J. Phys. B.: At. Mol. Phys.* **1987**, *20*, 1945.

(49) Meath, W. J.; Power, E. A. *J. Mod. Opt.* **1989**, *36*, 977.

(50) Power, E. A. *J. Chem. Phys.* **1975**, *63*, 1348.

(51) Andrews, D. L. *Chem. Phys.* **1976**, *16*, 419.

(52) Condon, E. U. *Rev. Mod. Phys.* **1937**, *55*, 2789.

(53) Craig, D. P.; Thirunamachandran, T. *Molecular Quantum Electrodynamics. An Introduction to Radiation Molecule Interaction*; Dover Publications, Inc.: Mineaol, NY, 1984.

(54) McClain, W. M. *Acc. Chem. Res.* **1974**, *7*, 129.

(55) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(56) Becke, A. D. *Phys. Rev.* **A 1988**, *38*, 3098.

(57) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(58) Sałek, P.; Jonsson, D.; Vahtras, O.; Ågren, H. *J. Chem. Phys.* **2002**, *117*, 9630.

(59) Olsen, J.; Jørgensen, P. In *Modern Electronic Structure Theory, Part II*; Yarkony, D. R., Ed.; World Scientific: Singapore, 1995; Chapter 13, p 857.

(60) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.

(61) DALTON, a molecular electronic structure program, Release 2.0 (2005). See http://www.kjemi.uio.no/software/dalton/dalton.html.

(62) Albota, M.; Beljonne, D.; Brédas, J.-L.; Ehrlich, J. E.; Fu, J.-Y.; Heikal, A. A.; Hess, S. E.; Kogej, T.; Levin, M. D.; Marder, S. R.; McCord-Maughon, D.; Perry, J. W.; Röckel, H.; Rumi, M.; Subramaniam, G.; Webb, W. W.; Wu, X.-L.; Xu, C. *Science* **1998**, *281*, 1653.

(63) Bartholomew, G. P.; Rumi, M.; Pond, S. J. K.; Perry, J. W.; Tretiak, S.; Bazan, G. C. *J. Am. Chem. Soc.* **2004**, *126*, 11529.

(64) Ohta, K.; Kamada, K. *J. Chem. Phys.* **2006**, *124*, 124303.

(65) Norman, P.; Luo, Y.; Ågren, H. *J. Chem. Phys.* **1999**, *111*, 7758.

(66) Luo, Y.; Norman, P.; Macak, P.; Ågren, H. *J. Chem. Phys.* **2000**, *104*, 4718.

(67) Macak, P.; Luo, Y.; Norman, P.; Ågren, H. *J. Chem. Phys.* **2000**, *113*, 7062.

(68) Brickell, W. S.; Brown, A.; Kemp, C. M.; Mason, S. F. *J. Chem. Soc. A* **1971**, 756.

(69) Newman, M. S.; Darlak, R. S.; Tsai, L. *J. Am. Chem. Soc.* **1967**, *89*, 6191.

(70) Shen, Y. R. *The Principles of Nonlinear Optics*; Wiley-Interscience: Wiley Classics Library Ed.: New York, 2003.

(71) Meijer, E. W.; Havinga, E. E.; Rikken, G. L. J. A. *Phys. Rev. Lett*. **1990**, *65*, 37.

(72) Olsen, J.; Jørgensen, P. *J. Chem. Phys.* **1985**, *82*, 3235.

(73) Li, R.; Sullivan, R.; Al-Basheer, W.; Pagni, R. M.; Compton, R. N. *J. Chem. Phys.* **2006**, *125*, 144304.

(74) Baranowska, A.; Rizzo, A.; Jansík, B.; Coriani, S. *J. Chem. Phys.* **2006**, *125*, 054107.

(75) Jansík, B.; Rizzo, A.; Frediani, L.; Ruud, K.; Coriani, S. *J. Chem. Phys.* **2006**, *125*, 234105.

(76) Yanai, Y.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett*. **2004**, *393*, 51.

(77) Peach, M. J. G.; Helgaker, T.; Sałek, P.; Keal, T. W.; Lutnæs, O. B.; Tozer, D. J.; Handy, N. C. *Phys. Chem. Chem. Phys.* **2006**, *8*, 558.

CT700329A

JCTC Journal of Chemical Theory and Computation

Molecular Graphics of Convex Body Fluids

Adrian T. Gabriel,[†] Timm Meyer,[‡] and Guido Germano*,[†]

Department of Chemistry, WZMW, and Department of Mathematics and Computer Science, Philipps-University Marburg, 35032 Marburg, Germany

Received August 1, 2007

Abstract: Coarse-grained modeling of molecular fluids is often based on nonspherical convex rigid bodies like ellipsoids or spherocylinders representing rodlike or platelike molecules or groups of atoms, with site−site interaction potentials depending both on the distance among the particles and the relative orientation. In this category of potentials, the Gay-Berne family has been studied most extensively. However, conventional molecular graphics programs are not designed to visualize such objects. Usually the basic units are atoms displayed as spheres or as vertices in a graph. Atomic aggregates can be highlighted through an increasing amount of stylized representations, e.g., Richardson ribbon diagrams for the secondary structure of proteins, Connolly molecular surfaces, density maps, etc., but ellipsoids and spherocylinders are generally missing, especially as elementary simulation units. We fill this gap providing and discussing a customized OpenGL-based program for the interactive, rendered representation of large ensembles of convex bodies, useful especially in liquid crystal research. We pay particular attention to the performance issues for typical system sizes in this field. The code is distributed as open source.

1. Introduction and Motivation

Generating three-dimensional (3D) pictures of molecular simulation output is useful if not mandatory for understanding the results and for presenting them in publications, talks, and posters. There are hundreds of molecular graphics programs. Freeware examples are MolScript,[1] VMD,[2] Raster3D,[3] Chimera,[4] AtomEye,[5] RasMol,[6] gOpenMol,[7] Jmol,[8] PyMOL,[9] and Molekel.[10] Payware examples are Cerius2,[11] Discovery Studio,[12] SYBYL,[13] and MOLCAD.[14] Inevitably, the basic units of these programs are atoms displayed as spheres or as vertices in a wireframe or "neon tube" graph. However, the Gay-Berne family of anisotropic potentials employs soft ellipsoids as basic modeling units to represent whole rodlike[15] or platelike[16] (and thus usually mesogenic) molecules, in order to speed up Monte Carlo and molecular dynamics[17] calculations by giving up intramolecular detail. Other popular choices for the same purpose are soft spherocylinders;[18] soft biaxial ellipsoids,[19] hard ellipsoids and spherocylinders,[20] and several other site−site variants[21] are employed too. The use of these nonspherical convex rigid bodies has been linked traditionally to liquid crystal research[22−24] but has later been extended to the mesoscopic description of polymers[23] and, more in general, of rigid moieties in larger molecules.[25,26] The attention to "coarse-graining" in molecular simulation has been growing, as shown by a dedicated section in a recent issue of this journal,[27] though in most cases the full potential of a "united atoms" approach is not unleashed because for simplicity researchers too often limit themselves to model functional groups or sets of nearby atoms with one large sphere[28] rather than with other more matching shapes.

Most standard molecular graphics packages can highlight atomic aggregates through an increasing amount of stylized representations, e.g., ribbons or cartoons for the secondary structure of a protein,[29,30] molecular surfaces,[31−34] density maps, etc., but ellipsoids or spherocylinders are not usually implemented. Standard programs are written to process only sets of Cartesian coordinates $\{\mathbf{r}_i\}$ but not orientations $\{\hat{\mathbf{e}}_i\}$ (for the sake of simplicity, we assume axially symmetric bodies, whose orientation is fully determined by a versor,

* Corresponding author e-mail: guido@staff.uni-marburg.de; Web page: www.staff.uni-marburg.de/~germano.

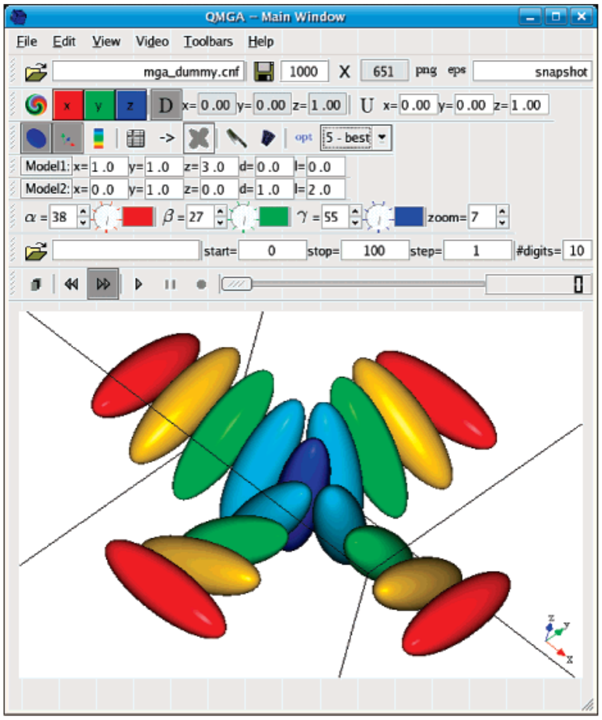
[†] Department of Chemistry and WZMW.

[‡] Department of Mathematics and Computer Science.

10.1021/ct700192z CCC: $40.75
Published on Web 02/21/2008

i.e., a unit vector; the generalization to the biaxial case with an orthogonal orientation matrix[35] or a quaternion[36] is straightforward). An exception is the simplified representation of DNA bases by flat biaxial ellipsoids,[37] recently implemented in some biomolecular graphics programs like Chimera.[38] However, this feature belongs to the above-mentioned category of schematic representations of specific groups of atoms, input as a set of Cartesian coordinates in PDB format. It is inflexible and inefficient when tweaked to display an arbitrary set of ellipsoids used to model a mesophase. Our attempt to employ Chimera in this sense was not satisfactory, though otherwise it is a fine and comprehensive program for its intended purposes. We converted center of mass coordinates and orientations of ellipsoids to a special data file with a much larger number of corresponding atomic coordinates. In addition to being cumbersome, this froze the program when the number of objects was within a typical range used in the study of collective properties of liquid crystalline phases, i.e., 10^4–10^5. For completeness, we mention the ORTEP[39] program that plots thermal ellipsoids for crystal structures, but clearly this is off the track for our aim, so we did not spent any time with it.

Until now, researchers in this niche resort to their own visualization code[40,41] or to workarounds with programs designed for other purposes,[42,43] possibly through conversion steps similar to the one described before. Some of these workarounds, apart from being complicated and time-consuming, preclude a visual feedback before the image is completed, i.e., the system cannot be zoomed, rotated, or sliced interactively in real time. We fill this gap providing a good dedicated molecular graphics program based on OpenGL[44] and available as open source.[45] Its name, QMGA, is an acronym for Qt-based Molecular Graphics Application, and the trailing A stands also for the first name of its principal author. A screenshot of QMGA's main window displaying a test system is shown in Figure 1. We preferred to develop a completely new program tailored for liquid crystal research rather than to extend an existing one burdened by a rich set of features useful in molecular biology, because this allowed us to focus on issues specific to liquid crystals, including the performance needed to display the large amount of objects that must typically be dealt with in this field. Of course we would be glad if our work will spur the future inclusion of QMGA's concepts and features in larger molecular graphics programs meant for general purpose.



QMGA – Main Window

mga_dummy.cnf 1000 X 651 snapshot

D x= 0.00 y= 0.00 z= 1.00 U x= 0.00 y= 0.00 z= 1.00

5 - best

Model1: x= 1.0 y= 1.0 z= 3.0 d= 0.0 l= 0.0

Model2: x= 0.0 y= 1.0 z= 0.0 d= 1.0 l= 2.0

α = 38 β = 27 γ = 55 zoom= 7

start= 0 stop= 100 step= 1 #digits= 10

Figure 1. The QMGA graphical user interface showing a test system.

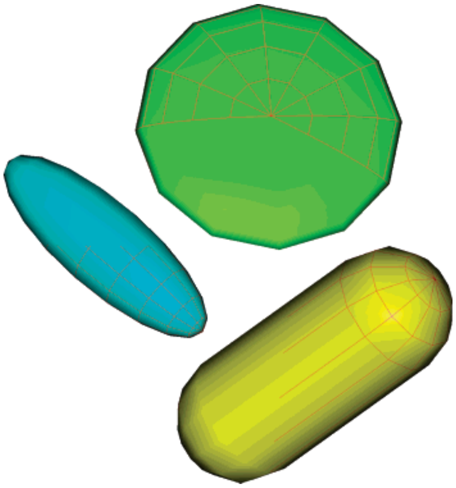


Figure 2. Oblate ($\kappa = 0.2$) and prolate ($\kappa = 3$) ellipsoids as well as a spherocylinder ($L = 2$) with wireframe overlay showing the polygonal surface structure employed by the rendering engine for a medium quality setting. The full range of render quality settings is shown in Figure 7.

2. Program Concepts and Features

In the following we discuss briefly the main concepts and features of our visualization program. The order in which they appear reflects to some extent their importance.

2.1. Fully Rendered View and Simplified View. A rendered picture is obviously the bare basis, since without it nothing is seen. Full rendering consists in drawing each molecule as a space-filling convex body (a sphere, an ellipsoid, or a spherocylinder) approximated by a set of triangles, see Figure 2, in our case a generalized triangle strip. In stick rendering only the molecular axis $\kappa\hat{e}_i$ is drawn. Stick rendering is useful to see through the system for detecting supramolecular structures (or their absence), see Figure 3, and to reduce the computational effort when rotating or zooming a large system.

2.2. Color Coding. In conventional molecular graphics programs, the elementary objects are spheres representing atoms. The latter are usually colored according to their type along the Corey-Pauling-Koltun scheme: white for hydrogen, black or gray for carbon, blue for nitrogen, red for oxygen, etc.[6,47,48]

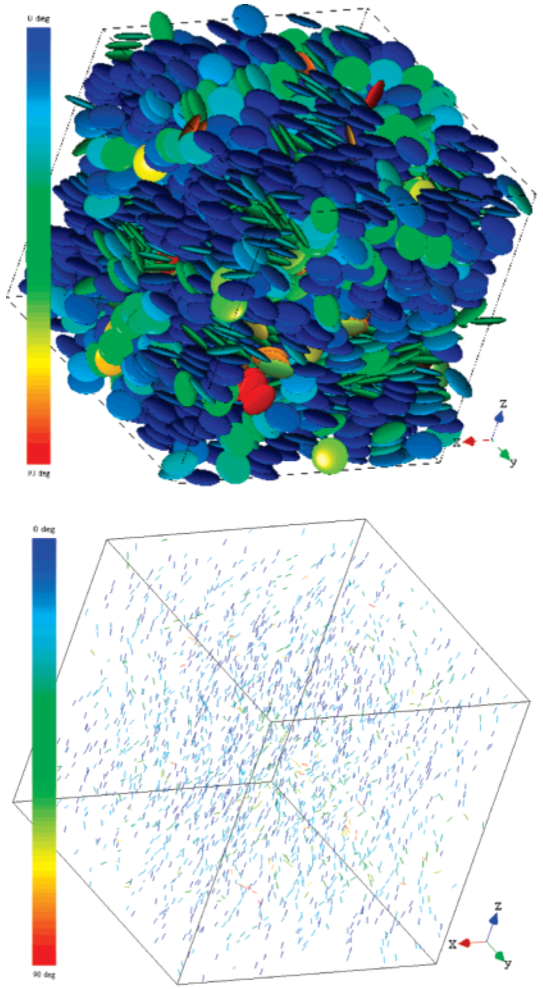


Figure 3. Nematic phase formed by soft oblate ellipsoids interacting with the GBDII potential ($\mu = 1$, $\nu = 2$, $\kappa = 0.2$, $\kappa' = 0.1$, $T^* = 12$, $P^* = 200$),[46] fully rendered and in stick view. The colormap is visible on the left.

Other coloring schemes are based on properties like hydrophobicity, charge, velocity modulus $v_i = |\mathbf{v}_i|$, and, therefore, temperature $T_i = 3m_iv_i^2/k_B$ (because of the equipartition theorem) or, for nonspherical bodies, orientation $\hat{\mathbf{e}}_i$. A color depending on the orientation is particularly useful for liquid crystals to give a first glance impression of the overall order of the phase (one color predominates in a more ordered phase) and has been used at least since the early 1990s.[40,41]

Each molecule i is colored depending on $c_i = |\hat{\mathbf{e}}_i \cdot \hat{\mathbf{n}}| \in [0,1]$, i.e. the absolute value of the scalar product between the individual molecular versor $\hat{\mathbf{e}}_i$ and an overall versor $\hat{\mathbf{n}}$ that is the same for the whole system; $\hat{\mathbf{n}}$ can be set to the director of the mesophase or to a user-defined value. The latter can be one of the three versors $\hat{\mathbf{\imath}}$, $\hat{\mathbf{\jmath}}$, $\hat{\mathbf{k}}$ of the Cartesian reference frame or a particular symmetry axis of the system, e.g., the cylinder axis for a cylindrical pore (see below). The director is the eigenvector corresponding to the eigenvalue with the largest absolute value of the order tensor Q:

$$Q = \frac{3}{2N}\sum_{i=1}^{N} \hat{e}_i \otimes \hat{e}_i - \frac{1}{2}E$$

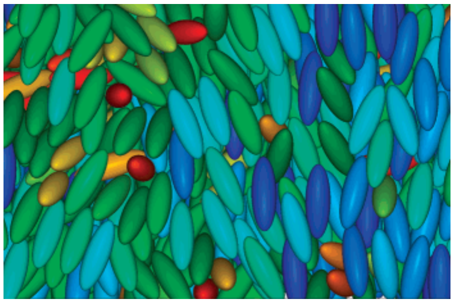


Figure 4. Closeup picture of a nematic phase formed by soft prolate ellipsoids interacting with the Gay-Berne potential ($\mu = 1$, $\nu = 3$, $\kappa = 3$, $\kappa' = 5$, $T^* = 3.45$, $\rho^* = 0.3$).[49]

In general, color coding consists in mapping a variable $x \in [a,b] \subset \mathbb{R}$ that describes a relevant property to a colormap M, i.e. a discrete set of colors, through a function $f_{\mathrm{col}}(x)$: $[a,b] \rightarrow M$. The actual form of f_{col} and M depends on the system. In the present case, $x = c_i$ is mapped to an RGB-encoded rainbowlike spectrum M_{RGB} by the function f_{RGB} (R = red, G = green, B = blue); each color is represented by a tuple of three integers R, G, B between 0 and 255. As shown on the left of Figure 3, M_{RGB} consists of 91 different colors, one for every degree of arccos $c_{\mathrm{i}} \in [0, 90]$. Both the calculated director and a user defined reference versor are shown in the GUI. The user can modify his choice at runtime with an immediate effect on colorization. As an alternative in the case of mixtures, some or all molecules may be colored according to their type.

2.3. User Interface. The 3D representation can be rotated and zoomed with the mouse. The camera position information, i.e., the description from which point and distance in space the user looks upon the system, is shown using three angles and a zoom factor. All three values are continuously updated while zooming and rotating with the mouse. It is also possible to update the render area setting each orientation parameter through the keyboard. This way the user can reproduce exactly a desired viewpoint, e.g., to compare different systems. The hot keys x , y, z and c orient the system axes parallel to the screen axes in a preset manner.

2.4. Printing. A molecular graphics program is useful not only to understand one's own results but also to present them in public. To do so, image files are required and consequently the ability to take screenshots from the render area. With some graphic tools the screenshot picture's resolution depends on the size of the program window and therefore on the resolution of the monitor. As a result, it is not possible to save pictures with a resolution higher than that of the monitor, which leads to problems when these are printed on large scale, e.g., on posters. QMGA allows the user to specify the desired resolution of the picture independently of the output device, which is especially useful on large printouts, choosing at the very least between PostScript and PNG; see Figures 3−6 for examples. The aspect ratio of the picture is automatically taken care of, so that no deformations occur when setting a new resolution value or when resizing the window. Moreover, it is possible to export a screenshot as a POV-Ray[43] script, in order to achieve the final polished characteristics of a ray-traced image as well as many other features of the powerful POV-Ray program.

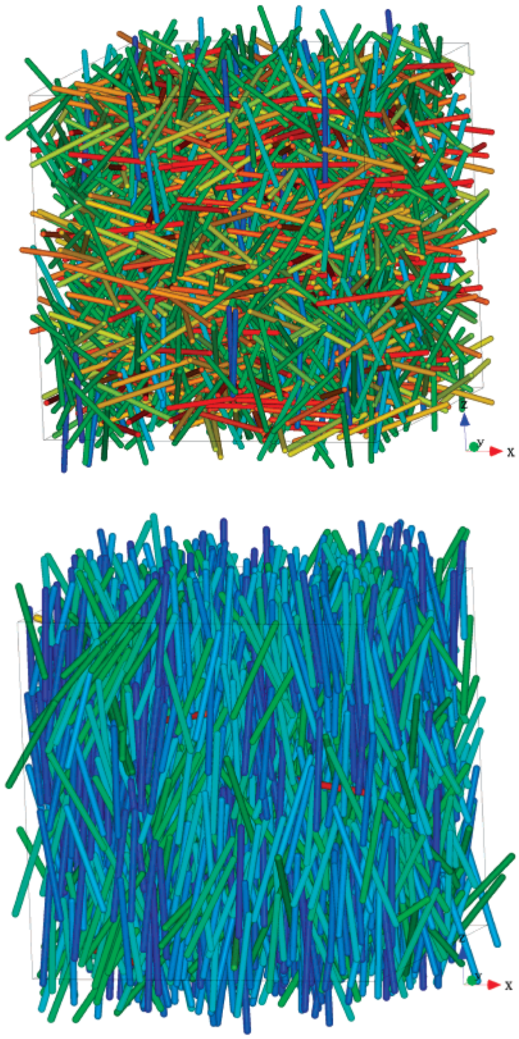


Figure 5. Spherocylinders in an isotropic (top) and nematic (bottom) phase.

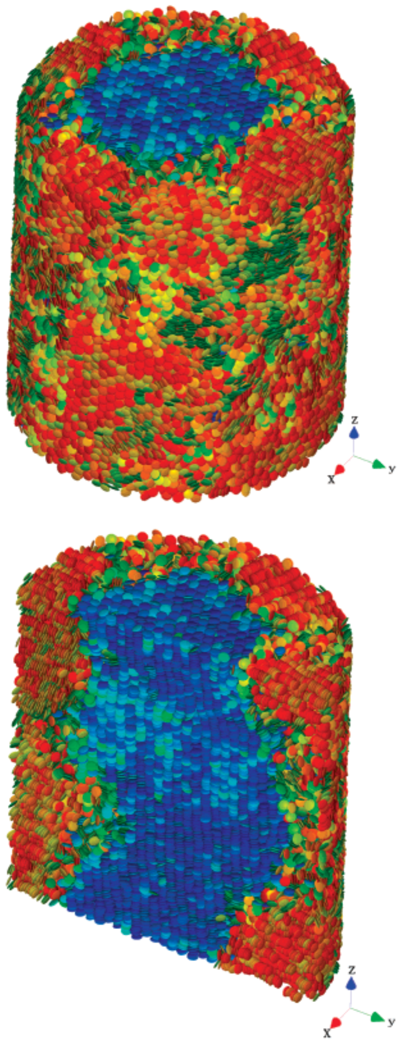


Figure 6. Snapshot from the molecular dynamics simulation of model discotic mesogens confined inside a nanopore.[50] To see that the pore organizes into a core−shell system with a columnar region in the center, the system was cut in half along the cylinder axis using QMGA's slice feature. An alternative is switching to a stick view.

2.5. Slicing. When a system is closely packed, what happens inside is not visible. However, there are cases where the inside is the interesting region: Figure 6 shows the simulation of a discotic liquid crystal inside a nanopore where the pore is sliced in half. Stick view is a possibility to look through a system, but if full render mode is wished, the choice must fall on slicing. QMGA's slice feature displays or hides objects depending on their center, meaning that no objects are truncated: they are either completely displayed or completely hidden. Slicing can take place along any combination of the coordinate axes.

2.6. Video. Since a molecular simulation usually evolves in time, the possibility to view and record animations is convenient. QMGA can load sequentially a number of files and display them one after the other, creating the impression of a motion picture. The interface allows not only for the standard actions expected from a video player (start, stop, and pause) but also forward and backward playback as well as frame capture.

With large systems the load and render times become long, resulting in a stagnant movie. In such a case it is possible (and advisable) to save all frames to disk as images and create a movie file from these. Though the recording of all displayed frames can be done by QMGA, currently there is no functionality to encode them automatically into a movie file, so this has to be done with an external program. A good freeware utility for this purpose is FFmpeg,[51] that produces, e.g., high quality AVI files.

2.7. Mixtures. Whereas many molecular simulations, especially coarse-grained ones of liquid crystals, deal with pure phases, there are also cases with more than one species. To accommodate for this, the internally used molecule class of QMGA has a private member of integer value that is used as a tag to divide the molecules into groups. It is then possible to assign different model parameters to each group. In an extreme case it is possible to give every single molecule its own representation by assigning a different tag to each.

For convenience, two toolbars are shown directly on the main program window to set the size parameters of the first two used objects. Since too many toolbars are confusing and many simulations deal with just one or two different molecular species, only these two were implemented. However, there is an additional window showing all used

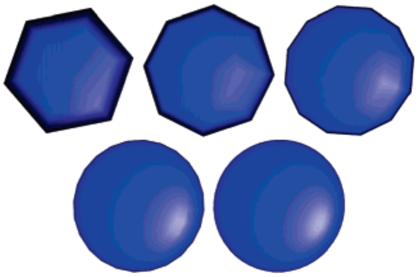


Figure 7. Top view of an oblate ellipsoid showing the effect of the render quality setting. There are five levels ranging from poor to very good. This setting enables the user to speed up the frame rate (if necessary) while working with the program and later on to produce high quality screenshots.

models with all their parameters, so that the parameters not accessible via the toolbars can be adjusted too.

2.8. Periodic Boundary Conditions. Most molecular simulations use periodic boundary conditions (PBC). From the graphical point of view this can be treated in two ways: show all molecules at their positions folded inside the unit box or employ positions without PBC applied on them. While folding is always possible and has been implemented, unfolding requires either the absolute coordinates or a sequence of folded trajectory frames.[52]

2.9. Lighting. It is mostly a mere matter of taste how one prefers the objects to look like, referring to lighting and surface. OpenGL provides the functionality to give the objects, e.g., a beamless or shiny metal or plasticlike finish. All necessary parameters are adjustable from a dialog window, and the resulting changes in colorization are shown immediately. More sophisticated effects can be achieved with the above-mentioned POV-Ray export feature.

2.10. Render Quality. The render speed is a function of parameters like the quality of the video driver, the quality of the video card, and the number of triangles to be drawn. When working with the program, smooth zoom and rotation is more important than a high-level representation. On the other hand, on a printout it is the other way around. The representation must look nice, and, since it is a still picture, render performance is not an issue any more.

To achieve a certain level of adjustability, five presets were implemented to influence the render quality of the shown objects. They range from fairly poor to an excellent, almost perfectly smooth representation. Figure 7 shows the difference on the example of a single oblate ellipsoid.

2.11. Remote File Access. Molecular simulations are often performed on remote supercomputers, while their results are visualized on the screen of a local desktop computer. To simplify file transfer, QMGA uses *ssh* to show the file system of the remote computer in a tree view that can be navigated with the mouse or keyboard. To show a file in the render area, it is copied to a temporary directory on the local machine and opened from there. This is done by a simple double-click, drag and drop or key-press. QMGA shows how much data are stored in the temporary folder and provides a button to purge all files. The remote login and file transfer, realized by *ssh* and *scp*, require a passwordless connection through an entry of the local computer's public key in the *.ssh*/*authorized_keys* file of the remote computer.

2.12. Saving of Options for Restart. More than 70 options are saved when the program closes and are loaded again when it is restarted. They include the last opened file, rotation and zoom settings, lighting settings, which toolbars are shown or hidden, the window position, the render mode (full or stick) and quality, etc.

3. Program Internals

3.1. Structure. QMGA is wholly written in the programming language C++ with a completely object oriented approach, as are most used libraries and toolkits. The window manager is realized with Trolltech's Qt,[53] that provides easy support for elaborate window items. For the 3D part we chose OpenGL[44] rather than the simpler VRML[54] or its successor X3D[55] because of the performance. Moreover, VRML/X3D never became as largely used and as well supported as OpenGL.

Since the output of most molecular simulation programs are text-based files with the variables describing individual molecules, we provide a small library of objects that handles a given system of molecules with respect to visualization. This library consists of three classes: (1) *Molecule* contains position, orientation, size, type, and color information for a single molecule. (2) *Colormap* reads RGB-based color values from a file and contains a function that sets a molecule's color according to a certain rule. (3) *CnfFile* (configuration file) reads all relevant data from a given simulation output file. The main components are a vector containing all *Molecule* objects and the *Colormap* object to be used for colorization.

The program basis of QMGA is given by Qt, that provides the graphical user interface. An OpenGL render area is embedded as a window frame. This area is filled with 3D objects using the information that was loaded into an instance of the *CnfFile* class. At run time it is possible to load and display different systems by overwriting the information stored in the *CnfFile* and sending the new commands to the render area.

3.2. Customization. Obviously some features will need customization to satisfy the pecularities of different users. First of all, different simulation programs will have different output file formats. However, the text parser of QMGA resides in just one single function, named *loadCnfFile()*, that is a member function of the class *CnfFile*. It is easy to modify this part of the code. All that has to be done is parsing the necessary information from the configuration file. In principle any format can be supported, if it provides at least position and orientation information for each molecule. When dealing with spherical objects and therefore there is no orientation information, the related variables can be set to arbitrary numbers.

The current format is the one used by the parallel domain decomposition molecular dynamics program *GBmega*[56] and is structured in the following way:

- header

  int (number of molecules)

double (x side length of the unit box)
double (y side length of the unit box)
double (z side length of the unit box)
2 doubles (for moving boundary conditions)

- molecule information

12 doubles ($\mathbf{r}_i$,$\mathbf{v}_i$,$\hat{\mathbf{e}}_i$,$\mathbf{u}_i$), int (label), int (type tag, optional) where $\mathbf{r}_i$ is the position of molecule i , $\mathbf{v}_i$ is its velocity, $\hat{e}_i$ is its orientation, and $\mathbf{u}_i$ is its orientation velocity. Of these numbers, only the box sides, $\mathbf{r}_i$, and $\hat{e}_i$, are used for visualization purposes. The program can deal also with noncubic unit boxes; in this case, a 3 × 3 matrix must be input to specify it.

For example, the artificially created file displayed in Figure 1 looks like this:

```
17
9.0
9.0
9.0
0.0 0.0
-4.0  0.0  0.0   0 0 0   0.0  1.0  0.0   0 0 0  1
-3.0  0.0  0.0   0 0 0   0.0  2.0  1.0   0 0 0  2
-2.0  0.0  0.0   0 0 0   0.0  1.0  1.0   0 0 0  3
-1.0  0.0  0.0   0 0 0   0.0  1.0  2.0   0 0 0  4
 0.0  0.0  0.0   0 0 0   0.0  0.0  1.0   0 0 0  5
 1.0  0.0  0.0   0 0 0   0.0 -1.0  2.0   0 0 0  6
 2.0  0.0  0.0   0 0 0   0.0 -1.0  1.0   0 0 0  7
 3.0  0.0  0.0   0 0 0   0.0 -2.0  1.0   0 0 0  8
 4.0  0.0  0.0   0 0 0   0.0 -1.0  0.0   0 0 0  9
 0.0 -4.0  0.0   0 0 0   1.0  0.0  0.0   0 0 0 10
 0.0 -3.0  0.0   0 0 0   2.0  0.0  1.0   0 0 0 11
 0.0 -2.0  0.0   0 0 0   1.0  0.0  1.0   0 0 0 12
 0.0 -1.0  0.0   0 0 0   1.0  0.0  2.0   0 0 0 13
 0.0  1.0  0.0   0 0 0  -1.0  0.0  2.0   0 0 0 14
 0.0  2.0  0.0   0 0 0  -1.0  0.0  1.0   0 0 0 15
 0.0  3.0  0.0   0 0 0  -2.0  0.0  1.0   0 0 0 16
 0.0  4.0  0.0   0 0 0  -1.0  0.0  0.0   0 0 0 17
```

Color coding is another aspect likely to be customized. Again, this is simple to do, because the whole relevant instructions that determine which color a molecule shall be given is found in just three rather short functions. One of these is a member function of CnfFile called *colorizeMolecules()* with the main purpose of sending all read molecules successively to another function, that is a member of the *Colormap* class. The name of this function is *setColor()*, and here is the most likely place where a change has to be made. Notice that *setColor()* comes in two different overloaded versions to handle both colorization by axis and colorization by type. Figure 8 shows the code of the currently used version of *setColor()* that realizes the colorization by axis as described earlier.

Last, parts of the GUI are expected to be modified, e.g., to display specific data values. This can be achieved intuitively with the Qt designer, a graphical tool.

4. Performance

Several optimization approaches were used to reduce the workload on the render engine; the most important ones are described below. Benchmarks conclude this section.

4.1. Scene Graph versus Direct Rendering. Initially, the OpenGL render area was realized with little effort resorting to SiM's Coin3D toolkit.[57] Coin3D is an open source library consisting of a collection of objects like ready to use light models, standard forms (sphere, cone, etc.), and a mechanism to assemble everything into a scene graph. Coin3D behaves very much like SGI's OpenInventor,[58] that at the start of the project was still payware, and provides a set of classes that can be used directly to display OpenGL content in a Qt window frame, i.e., the connection between Coin3D and Qt was already built in. The scene shown in the render area was constructed completely relying on Coin3D classes, to a large extent with a single for-loop over all *Molecule* objects inside *CnfFile*. Another example for the help these toolkits provide is how the scene is saved to file. Qt contains a file dialog and Coin3D a number of functions to export the content of the render area to image file in various formats.

```
Molecule* mga::Colormap::setColor(
 Molecule* moleculeTmp, vector<double> &director ) const
{
  if( (moleculeTmp != 0) && (director.size() == 3) )
    {
      double orientationX=moleculeTmp->getOrientationX();
      double orientationY=moleculeTmp->getOrientationY();
      double orientationZ=moleculeTmp->getOrientationZ();
      int numOfMapLines=redVector.size();
      int mapLineNr = int( acos( fabs(
       orientationX*director.at(0) +
       orientationY*director.at(1) +
       orientationZ*director.at(2)   )
      )/M_PI*2*( numOfMapLines ) );
      if( mapLineNr == 90 ) { mapLineNr = 89; }
      moleculeTmp -> setRGB(
        getRed( mapLineNr ),
        getGreen( mapLineNr ),
        getBlue( mapLineNr ) );
    }
  else
    {
       cerr << "Error: no molecule given to setColor,
        or director corrupt." << endl;
     }
  return( moleculeTmp );
}
```

Figure 8. Function *setColor()* of the class *Colormap*. The variables *orientationX*/*Y*/*Z* are the components of the vector describing the orientation of the object; the vector object *director* was calculated previously and represents the director of the mesophase. The colormap itself contains a certain number of RGB coded colors arranged in lines. The line number of the color is calculated by taking the scalar product of the molecular orientation with the director and multiplying the result with the total number of colors in the map.

While such a simple approach based on Coin3D is shared by other molecular graphics programs[10] and works quite well for not too big systems of up to 10^4 molecules, it becomes very slow when large systems of about 10^5 molecules are rendered. Even if the program actually does still run stable with that much workload, the frame rate is too low for an effective use. This is consistent with the experience with Chimera described in the introduction: Chimera uses an even slower scene graph based on VRML. For this reason, we redesigned completely the render area rewriting it from scratch without Coin3D and fitting it to the special purpose of displaying many identical objects. The only limitation is the assortment of supported image output formats, that was reduced to PNG and PostScript because it was too much work to implement the complete list provided by Coin3D (JPEG, TIF, diverse raw pixel formats, etc.) without this library.

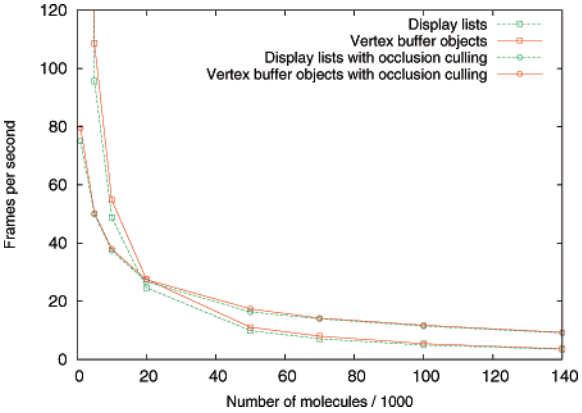


Figure 9. Graphic display of the benchmark results in Table 1. There is little difference between display lists and vertex buffer objects. Notice the crossover for the use of occlusion culling at about 20 000 molecules. Error bars are omitted because, except for 1000 molecules, they are smaller than the symbols used for the data points.

Table 1. Frames per Second for a System of *N* Discotic Ellipsoids in a Columnar Phase, Using Display Lists (DL) or Vertex Buffer Objects (VBO) without or with Occlusion Culling (OC)[a]

N/1000	DL	VBO	DL+OC	VBO+OC
1	485.2	585.7	75.0	79.6
5	95.6	108.7	49.9	50.3
10	48.8	54.8	37.4	38.0
20	24.7	27.6	26.8	27.4
50	9.9	11.1	16.3	17.4
70	7.1	8.0	14.0	14.2
100	4.9	5.4	11.5	11.8
140	3.5	3.6	9.1	9.3

[a] Errors are below 5%.

4.2. Display Lists versus Vertex Buffer Objects. To speed up object rendering we tried both OpenGL's display lists (DLs) and vertex buffer objects (VBOs). An OpenGL DL "precompiles" a model in the graphics memory, so that later it can be drawn by just calling a specific OpenGL function with the index of the DL. Prior to that we move to the specific position and do other transformations to render the model in the desired way. So the main rendering code remains the same, while it becomes easy to replace the molecular model by precompiling another one with a new DL.

An OpenGL VBO can be used similarly in many ways, but VBOs are more lightweight than DLs: They contain by default less information about the object, e.g., no transformations and materials. So the graphics driver can avoid overhead work needed to sort out whether this information is present and must be taken into account. The speed-up achievable by exchanging DLs with VBOs depends on the software and hardware environment. On our test system, VBOs provide just a slightly higher frame rate; see Table 1 and Figure 9. However, VBOs also yield a shorter and cleaner code, while DLs are at risk of being removed from future OpenGL releases, so we preferred VBOs.

4.3. Level of Detail. The next optimization makes use of the common render technique "level of detail" (LOD). Objects near to the camera are rendered with more detail than far away ones. Here we do not use just a linear approach but a self-adjusting one. The user chooses a quality level, and this defines the maximum and minimum rendering quality. If the frame rate drops below a certain level, then the quality of the particles automatically starts dropping from back to front, until either the frame rate becomes high enough again or all particles are drawn with minimum quality. On the other hand, if the frame rate is high enough, then the quality of the drawn particles is enhanced from front to back. It makes sense to use LOD though we employ an orthographic view, because it is still more probable that an object far away from the camera is covered, even if in part, than an object near to the camera.

4.4. Occlusion Query. In a dense system with many particles, it is most likely not necessary to render all of them. If those nearest to the camera are rendered first, then it may be possible to clip many others farther away. For this aim we make use of the OpenGL extension *GL_ARB_occlusion_query*, that asks the graphics card whether the next models must be drawn. Since it does not make much sense to query every single particle, we group them together dividing the bounding box of the system into $n \times n \times n$ smaller cells. Every particle becomes a member of one of these cells according to its center. After every particle has been assigned to one cell, the cells are resized to fit the complete models of the particles and not just their center points. If later we wish to know whether a group must be drawn, then the graphic card can provide an answer. If the answer is negative, then all the particles inside this cell can be discarded. This way, if the cells are drawn from front to back, then the rendering of many particles can be avoided.

4.5. Backface Culling. It is not necessary to render the back of an opaque object. OpenGL can take care of this by itself, if instructed with a simple library call, and so the rendering work is halved. However, this leads to a performance gain of just about 10%, because the library's internal calculations to find out what exactly is the back side of each item in its present orientation to the camera are almost as time-consuming as the avoided rendering.

4.6. Benchmarks. The benchmark results for QMGA presented here were performed on a computer with an AMD Athlon 64 3500+ processor running at 2.2 GHz with 2 GB RAM and an NVidia 8600GT graphic card adapter. The operating system was Fedora Core 7 Linux and the compiler g++ 4.1.2. To achieve meaningful and stable results, a benchmark option was introduced into QMGA. When activated, a series of random rotations is executed while measuring the current and average frame rates. The system used for the measurement consisted of about 140 000 discotic molecules in a columnar phase. To monitor the render speed as a function of the number of particles, the latter were sorted by their distance from the origin and included into larger and larger systems, whose shape was spherical because of the sorting. For each system the benchmark was run several times through 100 rotations with and without occlusion culling (OC). Active OC increases significantly the render performance in systems with more than approximately 20 000 molecules. Below this number the performance is reduced but is still so high that it is safe to leave OC always on.

Table 1 and Figure 9 display how many frames per second (fps) were rendered for test configurations ranging from 1000 to 140 000 molecules. Even with more than 100 000 molecules per configuration, QMGA behaves quite well, and even better when occlusion culling is activated. As few as 8 fps still feel almost completely fluent, and only below about 5 fps some jerkiness starts becoming noticeable.

5. Conclusions

We have filled a gap among molecular graphics programs providing and discussing an open source code for the visualization of large sets of convex bodies like ellipsoids and spherocylinders. This is useful especially not only for the coarse-grained modeling of liquid crystals but also of (bio)polymers and other chemical compounds, with anisotropic site−site potentials belonging to the Gay-Berne family. A rich set of features has been implemented employing easy to use toolkits (Qt) and state of the art libraries (OpenGL). Special attention has been dedicated to performance when displaying large systems of the order of 10^5 molecules. The final result was a useful and fast program fulfilling purposes that previously could be achieved only with difficulty or not at all.

Acknowledgment. We thank Prof. M. P. Allen, University of Warwick, for sharing with us his proprietary molecular graphics program for convex bodies MG and his Ph.D. student P. O'Brien for providing us the spherocylinder configurations shown in Figure 5. We are also grateful to Prof. M. Guthe, Philipps-University Marburg, for help with implementing OpenGL vertex buffer objects. Figures 3 and 6 display results of molecular dynamics simulatons run on the IBM Regatta p690+ parallel computer JUMP of the John von Neumann Institute for Computing in Jülich.

References

(1) Kraulis, P. J. MolScript − A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **1991**, *24*, 946. http://www.avatar.se/molscript (accessed Dec 20, 2007).

(2) Humphrey, W.; Dalke, A.; Schulten, K. VMD − Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33. http://www.ks.uiuc.edu/research/vmd (accessed Dec 20, 2007).

(3) Merritt, E. A.; Bacon, D. J. Raster3D: Photorealistic molecular graphics. *Method. Enzymol.* **1997**, *277*, 505. http://skuld.bmsc.washington.edu/raster3d (accessed Dec 20, 2007).

(4) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera − A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605. http://www.cgl.ucsf.edu/chimera (accessed Dec 20, 2007).

(5) 5. Li, J. AtomEye: an efficient atomistic configuration viewer. *Model. Simul. Mater. Sc.* **2003**, *11*, 173.

(6) RasMol homepage. http://www.umass.edu/microbio/rasmol (accessed Dec 20, 2007).

(7) gOpenMol homepage. http://www.csc.fi/gopenmol (accessed Dec 20, 2007).

(8) Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org (accessed Dec 20, 2007).

(9) Delano, W. L. The PyMOL Molecular Graphics System, DeLano Scientific, Palo Alto, CA, U.S.A. http://pymol-.sourceforge.net (accessed Dec 20, 2007).

(10) Molekel homepage. http://www.cscs.ch/molekel (accessed Dec 20, 2007).

(11) Cerius2 homepage. http://www.accelrys.com/products/cerius2 (accessed Dec 20, 2007).

(12) Discovery Studio homepage. http://www.accelrys.com/products/dstudio (accessed Dec 20, 2007).

(13) SYBYL homepage. http://www.tripos.com (accessed Dec 20, 2007).

(14) Brickmann, J.; Keil, M.; Exner, T.; Marhöfer, R. Molecular graphics − Trends and perspectives. *J. Mol. Mod.* **2000**, *6*, 328. MOLCAD: MOlecular Computer Aided Design. http://www.molcad.com (accessed Dec 20, 2007).

(15) Gay, J. G.; Berne, B. J. Modification of the overlap potential to mimic a linear site-site potential. *J. Chem. Phys.* **1981**, *74*, 3316.

(16) Bates, M. A.; Luckhurst, G. R. Computer simulation studies of anisotropic systems. XXVI. Monte Carlo investigations of a GayBerne discotic at constant pressure. *J. Chem. Phys.* **1996**, *104*, 6696.

(17) Allen, M. P.; Germano, G. Expressions for forces and torques in molecular simulations using rigid bodies. *Mol. Phys.* **2006**, *104*, 3225.

(18) Kihara, T. Convex molecules in gaseous and crystalline states. *Adv. Chem. Phys.* **1963**, *5*, 147.

(19) Berardi, R.; Fava, C.; Zannoni, C. A Gay-Berne potential for dissimilar biaxial particles. *Chem. Phys. Lett.* **1998**, *297*, 8.

(20) Allen, M. P.; Evans, G. T.; Frenkel, D.; Mulder, B. M. Hard convex body fluids. *Adv. Chem. Phys.* **1993**, *86*, 1.

(21) Martinez-Haya, B.; Cuetos, A.; Lago, S.; Rull, L. F. A novel orientation-dependent potential model for prolate mesogens. *J. Chem. Phys.* **2005**, *122*, 024908.

(22) Zannoni, C. Molecular design and computer simulations of novel mesophases. *J. Mater. Chem.* **2001**, *11*, 2637.

(23) Computational soft matter: From synthetic polymers to proteins; Attig, N., Binder, K., Grubmüller, H., Kremer, K., Eds.; Forschungszentrum Jülich: Jülich, 2004. http://www.fz-juelich.de/nic-series/volume23 (accessed Dec 20, 2007).

(24) Wilson, M. R. Progress in computer simulations of liquid crystals. *Int. Rev. Phys. Chem.* **2005**, *24*, 421.

(25) Prampolini, G. Parametrization and validation of coarse grained force-fields derived from ab initio calculations. *J. Chem. Theory Comput.* **2006**, *2*, 556.

(26) Amovilli, C.; Cacelli, I.; Cinacchi, G.; Gaetani, L. D.; Prampolini, G.; Tani, A. Structure and dynamics of mesogens using intermolecular potentials derived from ab initio calculations. *Theor. Chem. Acc.* **2007**, *117*, 885.

(27) Voth, G. A. Introduction: Coarse-graining in molecular modeling and simulation. *J. Chem. Theory Comput.* **2006**, *2*, 463.

(28) Venturoli, M.; Sperotto, M. M.; Kranenburg, M.; Smit, B. Mesoscopic models of biological membranes. *Phys. Rep.* **2006**, *437*, 1.

(29) Richardson, J. S. The anatomy and taxonomy of protein structure. *Adv. Protein Chem*. **1981**, *34*, 167.

(30) Richardson, J. S. Schematic drawings of protein structures. *Method. Enzymol.* **1985**, *115*, 359.

(31) Lee, B.; Richards, F. M. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol*. **1971**, *55*, 379.

(32) Connolly, M. L. Analytical molecular surface calculation. *J. Appl. Crystallogr*. **1983**, *15*, 548.

(33) Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **1983**, *221*, 709.

(34) Connolly, M. L. The molecular surface package. *J. Mol. Graph*. **1993**, *11*, 139.

(35) Goldstein, H.; Poole, C. P.; Safko, J. L. *Classical mechanics*, 3rd ed.; Addison Wesley: San Francisco, 2002.

(36) Altmann, S. L. *Rotations, quaternions and double groups*; Oxford University Press: Oxford, 1986.

(37) Max, N. Hierarchical molecular modelling with ellipsoids. *J. Mol. Graph. Model*. **2004**, *23*, 233.

(38) Couch, G. S.; Hendrix, D. K.; Ferrin, T. E. Nucleic acid visualization with UCSF Chimera. *Nucleic Acids Res*. **2006**, *34*, e29.

(39) Burnett, M. N.; Johnson, C. K. *ORTEP-III: Oak Ridge Thermal Ellipsoid Plot program for crystal structure illustrations*; Oak Ridge National Laboratory Report ORNL-6895; 1996. http://www.ornl.gov/sci/ortep (accessed Dec 20, 2007).

(40) Chiccoli, C.; Pasini, P.; Semeria, F.; Zannoni, C. Three-dimensional visualization of molecular organization and phase transitions in liquid crystal lattice models. *Int. J. Mod. Phys. C* **1992**, *3*, 1209.

(41) Berardi, R.; Emerson, A. P. J.; Zannoni, C. Monte Carlo investigations of a Gay-Berne liquid crystal. *J. Chem. Soc., Faraday Trans.* **1993**, *89*, 4069.

(42) AVS/Express homepage. http://www.avs.com (accessed Dec 20, 2007).

(43) POV-Ray homepage. http://www.povray.org (accessed Dec 20, 2007).

(44) OpenGL homepage. http://www.opengl.org (accessed Dec 20, 2007).

(45) QMGA homepage. http://qmga.sourceforge.net (accessed Dec 20, 2007).

(46) Caprion, D.; Bellier-Castella, L.; Ryckaert, J.-P. Influence of shape and energy anisotropies on the phase diagram of discotic molecules. *Phys. Rev. E* **2003**, *67*, 041703.

(47) Corey, R. B.; Pauling, L. C. Molecular models of amino acids, peptides, and proteins. *Rev. Sci. Instrum*. **1953**, *24*, 621.

(48) Koltun, W. L. Precision space-filling atomic models. *Biopolymers* **1965**, *3*, 665.

(49) Germano, G.; Allen, M. P.; Masters, A. J. Simultaneous calculation of the helical pitch and the twist elastic constant in chiral liquid crystals from intermolecular torques. *J. Chem. Phys*. **2002**, *116*, 9422.

(50) Stillings, C.; Martin, E.; Steinhart, M.; Pettau, R.; Paraknowitsch, J.; Geuss, M.; Schmidt, J.; Germano, G.; Schmidt, H. W.; Gösele, U.; Wendorff, J. H. Nanoscaled discotic liquid crystal/polymer systems: Confinement effects on morphology and thermodynamics. *Mol. Cryst. Liq. Crys*t. **2008**, accepted for publication.

(51) FFmpeg homepage. http://ffmpeg.mplayerhq.hu (accessed Dec 20, 2007).

(52) Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids*, paperback ed.; Oxford University Press: Oxford, 1989.

(53) Qt homepage. http://www.trolltech.com/products/qt (accessed Dec 20, 2007).

(54) VRML97 and Related Specifications. http://www.web3d.org/x3d/specifications/vrml (accessed Dec 20, 2007).

(55) X3D and Related Specifications. http://www.web3d.org/x3d (accessed Dec 20, 2007).

(56) Wilson, M. R.; Allen, M. P.; Warren, M. A.; Sauron, A.; Smith, W. Replicated data and domain decomposition molecular dynamics techniques for simulation of anisotropic potentials. *J. Comput. Che*m. **1997**, *18*, 478.

(57) Coin3D homepage. http://www.coin3d.org (accessed Dec 20, 2007).

(58) OpenInventor homepage. http://oss.sgi.com/projects/inventor (accessed Dec 20, 2007).

CT700192Z

JCTC Journal of Chemical Theory and Computation

Combining Elastic Network Analysis and Molecular Dynamics Simulations by Hamiltonian Replica Exchange

Martin Zacharias*

School of Engineering and Science, Jacobs University Bremen, Campus Ring 1, D-28759 Bremen, Germany

Received September 4, 2007

Abstract: Coarse-grained elastic network models (ENM) of proteins can be used efficiently to explore the global mobility of a protein around a reference structure. A new Hamiltonian-replica exchange molecular dynamics (H-RexMD) method has been designed that effectively combines information extracted from an ENM analysis with atomic-resolution MD simulations. The ENM analysis is used to construct a distance-dependent penalty (flooding or biasing) potential that can drive the structure away from its current conformation in directions compatible with the ENM model. Various levels of the penalty or biasing potential are added to the force field description of the MD simulation along the replica coordinate. One replica runs at the original force field. By focusing the penalty potential on the relevant soft degrees of freedom the method avoids the rapid increase of the replica number with increasing system size to cover a desired temperature range in conventional (temperature) RexMD simulations. The application to domain motions in lysozyme of bacteriophage T4 and to peptide folding indicates significantly improved conformational sampling compared to conventional MD simulations.

Introduction

The limited time scale accessible during conventional classical molecular dynamics (MD) simulations is a major bottleneck to sample relevant conformational states of biomolecules. Conformational transitions between stable states of a biomolecules occur only rarely even on the time scale of tens to hundreds of nanoseconds that are currently possible.[1−12] Parallel tempering or replica exchange molecular dynamics (RexMD) simulations are now frequently used to enhance conformational sampling in Monte Carlo (MC)[13−16] and MD simulations.[5−8,10−12,17−24] During RexMD simulations several copies or replicas of a given system are simulated in parallel using classical MD or MC methods at different simulation temperatures. At preset intervals pairs of replicas (usually neighboring pairs) are exchanged with a specified transition probability. Instead of changing the temperature among the replicas it is also possible to modify the force field or selected force field contributions along the replicas (termed Hamiltonian-RexMD).[24] The exchanges allow conformations trapped in locally stable states (e.g., at a low simulation temperature) to escape by exchanging with replicas at higher simulation temperature (or running with a modified Hamiltonian). The RexMD method has been successfully applied in folding simulations of peptides and miniproteins[5−8,10−12,17−22] as well as for the folding of nucleic acid structural motifs.[25] Unfortunately, efficient exchange between replicas requires sufficient overlap of the energy distributions between neighboring replicas. As a consequence, in order to cover a desired temperature range, the number of required replicas grows approximately with the square root of the number of particles in the system.[24] A larger number of replicas in turn requires also increased simulation times for efficient “travelling” of replicas in the range of different temperatures.

Hybrid explicit/implicit solvent models have been suggested where the simulation of each replica is performed using an explicit solvent description and for each exchange part of the solvent is replaced by a continuum.[26] Another approach employs separate coupling of solute and solvent to different heat baths (target temperatures).[27] Only the solute reference temperatures are varied for each replica. In a further

* Corresponding author phone: ++49-421-200-3541; fax: ++49-421-200-3249; e-mail: m.zacharias@jacobs-university.de.

10.1021/ct7002258 CCC: $40.75 © 2008 American Chemical Society
Published on Web 02/21/2008

extension of this approach the temperature of only selected collective degrees of freedom has been modified along a replica coordinate.[28] These methods reduce the effective system size compared at each attempted replica exchange. However, the artificial temperature gradient at the solute−solvent interface and the inclusion of nonphysical systems as replica runs may cause artefacts in the latter methods.

Alternatively, approaches that scale the Hamiltonian or energy function of the system along a replica-coordinate have been suggested.[24,29−33] Recently, a promising "Hamiltonian"-RexMD method has been developed where the solute−solute, solute−solvent, and solvent−solvent interactions are separately (linearly) scaled for each replica.[31] This approach can be used to "effectively" scale only the solute temperature along the replica coordinate. In case of no scaling of the solvent−solvent interactions the replica exchange probability becomes less dependent on the number of solvent degrees of freedom, and hence fewer replicas are required to cover a desired "effective" temperature range compared to standard temperature replica exchange. A similar approach where the nonbonded (Lennard-Jones and electrostatic) interactions within the solute as well as between solute and solvent have been scaled to various degrees has also been suggested.[32] Another method specifically designed for peptides and proteins employs a biasing potential for the peptide backbone to specifically lower the barriers for backbone dihedral transitions as replica coordinate.[33] The biasing potential is obtained from explicit solvent simulations of a model peptide. The method showed promising results during peptide folding simulations.[33]

In recent years it has been shown that soft normal modes obtained from Elastic network models (ENMs) of proteins frequently overlap with experimentally observed conformational changes in proteins.[34−39] In an Elastic network model a given structure of a protein serves as a reference structure, and the mobility of a residue or protein segment depends on the local density and number of short-range contacts (usually between Cα or heavy atoms of the protein). Collective degrees of freedom can be calculated very rapidly from an ENM model of a protein (within seconds on standard workstation computers) by a normal mode calculation (after diagonalization of the second derivative matrix). Due to the goarse-grained nature, ENMs may indicate directions of possible large scale conformational transitions of a biomolecule. In fact, often very significant overlap of the softest ENM modes with experimentally observed conformational changes in proteins has been found.[37] This has, for example, been explored in efficient flexible docking simulations of proteins.[40,41]

The idea of coupling ENM analysis of proteins and MD simulations has been explored by Zhang et al.[42] by separate temperature coupling of collective ENM degrees of freedom of a molecule and temperature control of the rest of the system in a single simulation. The motion along ENM degrees of freedom is amplified by increasing the temperature "along" the collective degrees of freedom of the molecule. The method allowed enhanced sampling of peptide and protein motion.[42] However, separate temperature coupling of different degrees of freedom corresponds to a nonphysical simulation system, and it is not clear if such a simulation produces conformations compatible with the desired simulation ensemble (e.g., a canonical ensemble). In addition, extended simulation runs with an increased temperature of the soft collective degrees of freedom of a system may lead to sampling of undesired conformations, e.g., unfolding of the protein (if the temperature of a collective degree of freedom is kept above the folding temperature of the protein).

In the present study an alternative "Hamiltonian" replica-exchange method is proposed that includes a biasing or flooding potential compatible with an ENM description of the protein/peptide as a replica coordinate. The purpose of the biasing or penalty potential is to drive the protein or peptide conformation away from its current state in directions compatible with the ENM model of the system. Penalty potentials with the purpose to drive structures away from a given state have already been used in conformational flooding[43] and metadynamics simulations.[44] However, the coupling of ENM derived penalty/biasing potentials and replica exchange simulations has not been tried. The level of biasing is gradually changed along the replicas (one replica runs at the original force field) such that frequent transitions are possible. Since the overall conformation of the biomolecule may change during the simulation, the ENM calculations can be repeated at preset intervals. Since exchanges between replicas depend only on different levels of a very soft potential, the method requires fewer replicas for efficient sampling compared to conventional temperature RexMD. As long as the ENM model is not updated the method simulates and exchanges between replicas of the same system with slightly different force fields not involving any sampling of an artificial nonphysical system (hence sampling the desired ensemble). In the present initial application of the method it has been tested on a peptide and a protein test case of very different size indicating in both cases significantly enhanced sampling compared to standard MD simulations. Possible modifications and improvements of the present initial setup of the method will be discussed.

Computational Methods

Test Systems and Simulation Conditions. The RexMD method with an ENM derived biasing potential (ENM-RexMD) was tested on two different biomolecular systems. In all cases MD simulations were performed employing the *Sander* module of the Amber8 package[45] in combination with the parm03 force field.[46] Studies on peptide folding and domain−domain motions were performed employing a generalized Born implicit solvent model as implemented in Amber8 using the pairwise descreening method by Hawkins et al.[47,48] (corresponding to igb=1 in the input of *Sander*). A Debye-Hückel term as implemented in Sander was used with a salt concentration of 1 M. The Settle algorithm[49] was used to constrain bond vibrations involving hydrogen atoms, which allowed a time step of 2 fs. Folding simulations were performed on a small β-hairpin forming chignolin peptide[50] (sequence: GYDPETGTWG, pdb1UA0). An initial extended structure (independent of the experimental structure) was generated using the *xleap* module of the Amber8 package. Five variants of the start structure were generated using short

(5 ps) MD simulations at 800 K with different initial velocities quenched to 280 K within additional 2 ps simulation time. All subsequent peptide folding simulations were performed at 280 K to stabilize folded structures of the peptide. Conventional MD simulations of up to 25 ns starting from the experimental NMR structure (first structure of the NMR ensemble of pdb1UA0) and employing the GB continuum model showed that at this temperature the folded peptide structure is indeed stable. It remained within 2 Å root-mean square deviation of heavy atoms (Rmsd$_{\text{heavy}}$) from the experimental start structure (not shown). In addition, under the above simulation conditions and employing the parm03 force field all five extended starting structures folded into structures close to experiment within <25 ns of conventional MD simulations.

In case of T4-lysozyme (T4-L) the high-resolution X-ray structure (pdb2LZM)[51] served as the start structure. The structure was first energy minimized (1000 steps) and subsequently heated to 310 K (37 °C) within 0.2 ns and equlibrated within 1 ns simulation time using the same generalized GB model as for the peptide simulations. The equilibrated structure served as the starting structure for conventional and ENM-RexMD simulations.

Distance-Dependent Biasing Potentials for Soft Degrees of Freedom. At the start and at preset time intervals of the RexMD simulations elastic network model (ENM) calculations on the peptide and proteins were performed following the approach by Hinsen.[36] In the ENM a given protein is assumed to be at an equilibrium (reference) state, and it is described as a set of centers (Cα atoms or heavy atoms) that are connected by harmonic springs. The energy change for any deformation is controlled by spring force constants associated with each pair in the structure. In the ENM of Hinsen the spring constant decays with the distance according to a Gaussian function.[36] For the small chigolin peptide the ENM calculations were performed using all heavy atoms, whereas for lysozyme the Cα backbone atoms were used. The calculation of the elastic network modes took only a few seconds and had a neglectable effect on the overall simulation time. The calculated linear independent and orthogonal eigenvectors and associated eigenvalues of the ENM can be used to calculate B-factors and other properties of the structure depending on Cartesian conformational fluctuations around the reference state. It is also possible to calculate average distance fluctuations compatible with deformations in each mode. As outlined below these distance fluctuations were used to construct a biasing (or penalty/flooding) potential that drives the structure away from the current (reference) structure along directions compatible with the ENM model of the structure.

The ENM analysis and recalculation of the penalty potential was performed at intervals of 15−20 ps. The deformability of a structure in normal modes is given by the corresponding eigenvalue. To calculate distance fluctuations the protein or peptide structure was deformed in each mode *i* by a factor proportional to (1/eigenvalue(*i*))$^{0.5}$ followed by calculation of the interatomic distance variance (change of the square of interatomic distances). That is the soft modes (with small eigenvalue) contribute most to the distance variances. Summation over a set of modes gives the average distance fluctuations compatible with the collective motions of the system. The excitation (or deformation) in each mode was scaled such that the average distance fluctuation (summed over all included modes) did not exceed 2 Å. For the present simulations this value was chosen since it corresponds approximately to the motion of atoms in between recalculation of the ENM modes.

From the distance fluctuations (Δd_{ij}) of a given distance between an atom pair *i*, *j* a distance (d_{ij}) dependent penalty potential similar to a Gaussian function (but much less costly to calculate) was constructed:

$$V(d_{ij}) = k([d_{ij} - d_{ij0}]^2 - \Delta d_{ij}^{\;2})^2, \quad \text{if } |d_{ij} - d_{ij0}| \leq \Delta d_{ij}$$

$$V(d_{ij}) = 0, \quad \text{otherwise} \qquad (1)$$

This penalty potential has its maximum at the distance (d_{ij0}) in the reference state (the structure for which the ENM calculation was performed) and decreases both at smaller and larger distances such that it approaches zero when the change in distance approaches the distance fluctuation (Δd_{ij}) derived from the ENM calculation. It was implemented as an optional distance restraining potential in the disnrg.f routine of the Amber8 package.

To limit the number of added distance dependent penalty potentials only pairs with distance fluctuations 50% larger than the average distance fluctuations were included. The penalty potential basically acts as “flooding” potential to drive the structure away from the current state toward regimes that are compatible with the ENM derived collective degrees of the system. Since in the present implementation fluctuations have been calculated by summation over all mode contributions (according to their eigenvalue), the distances in the distance dependent perturbation potential are not projected onto each mode (invariant to the direction of each mode). One could call the relevant distances also soft distances. A projection on one mode direction is, however, possible if one is for example specifically interested in one selected soft mode direction. The advantage of using a distance-dependent restraining potential is that it is invariant under rotation and can therefore be applied directly during the simulation without a special treatment of rotational components of motion.

RexMD Using an ENM Derived Biasing Potential. In standard RexMD, copies or replicas of the system are simulated at different temperatures ($T_0, T_1, T_2, .., T_N$). Each replica evolves independently, and after 500−1000 MD-steps (∼1 ps) an exchange of pairs of neighboring replica is attempted according to the Metropolis criterion:

$$w(x_i \rightarrow x_j) = 1 \quad \text{for } \Delta \leq 0;$$

$$w(x_i \rightarrow x_j) = \exp(-\Delta) \quad \text{for } \Delta > 0$$

where

$$\Delta = (\beta_i - \beta_j)\,[E(r_j) - E(r_i)] \qquad (2)$$

with $\beta = 1/RT$ (*R*: gas constant and *T*: temperature), and *E*(*r*) representing the potential energy of system for a given configuration. Instead of modifying the temperature it is also possible to scale the force field (or part of it) along the replica coordinate. In the present case a distance-dependent potential as described in the last paragraph was added to the force field. Each replica runs at a different level of added biasing potential (the first replica runs with the original force field). Note, that the ENM derived distance dependent perturbation potential was only calculated for the structure of the replica that runs at the original force field. Exchanges at every 750 steps (1.5 ps) between neighboring biasing levels were attempted according to[24]

$$w(x_i \rightarrow x_j) = 1 \quad \text{for } \Delta \leq 0;$$

$$w(x_i \rightarrow x_j) = \exp(-\Delta) \quad \text{for } \Delta > 0$$

where

$$\Delta = \beta[(E^j(r_j) - E^j(r_i)) - (E^i(r_j) - E^i(r_i))] \tag{3}$$

In this case, the Metropolis criterion involves only a single β or temperature and the energy difference between neighboring configurations using the force field for replica j (E^j) minus the same difference using force field for replica i (E^i). Compared to temperature RexMD the energy differences are only affected by the force field term that changes upon going from one replica to another replica run. For all present test cases 5 replicas were used with different levels of the biasing potential. The levels of the biasing potential can be adjusted using the factor k in eq 1. Test calculations indicated that for the replica with highest penalty level a penalty maximum of 20 × RT/(number of distances) and appropriate scaling of the intermediate replicas resulted in an acceptance probability for replica exchanges of ~20−30% for the present systems. However, the scaling needed to be adapted for each system in test calculations.

Results

Application of the ENM-RexMD Method to T4-Lysozyme. Lysozyme from the *Escherichia coli* bacteriophage T4 (T4-lysozyme:T4-L) is one of the best studied proteins with >200 T4-L crystal structures of wild-type and mutants available in the protein data bank.[52] The protein consists of two domains (N-terminal and C-terminal domains) that are both involved in substrate binding in a cleft between the domains. Analysis of different crystal forms[52] and structure determination by NMR spectroscopy[53] as well as computer simulation studies[42,54] indicate that the protein can undergo hinge-bending (opening-closing) motions of the two domains. The ENM-RexMD method was applied to T4-lysozme starting from the experimental crystal structure (pdb2LZM). A GB continuum model was used (see the Methods section). For comparison a conventional MD simulation starting from the same start structure and applying the same simulation conditions was also run.

The conventional MD simulation resulted in a backbone Rmsd of ~3.5 Å from the start structure during ~3.2 ns simulation time. No tendency of unfolding was observed during the simulation time (Figure 1). The Rmsd of the N-terminal and C-terminal domains that encompass the enzyme active site showed a smaller Rmsd of ~2 Å indicating the larger Rmsd of the complete protein is mainly due to a rearrangement of the two domains. In addition to the Rmsd, the distance between centers of mass of atom groups of the N-terminal and C-terminal T4-lysozme domains was also recorded. This distance can be used as a measure of the relative domain motion or the opening and closing of the active site region (Figure 2). During the first 1−2 ns simulation time the domain−domain distance decreased from ~15 Å to ~11.5 Å and stayed at this level throughout the rest of the simulation (Figure 2, green curve). The final distance is slightly smaller than in an X-ray structure of a T4L mutant that is considered to represent one of the most closed forms of the protein (pdb152L).[55] This closing motion is also the reason for the relatively large average Rmsd from the start structure observed during the MD simulation (Figure 1b).

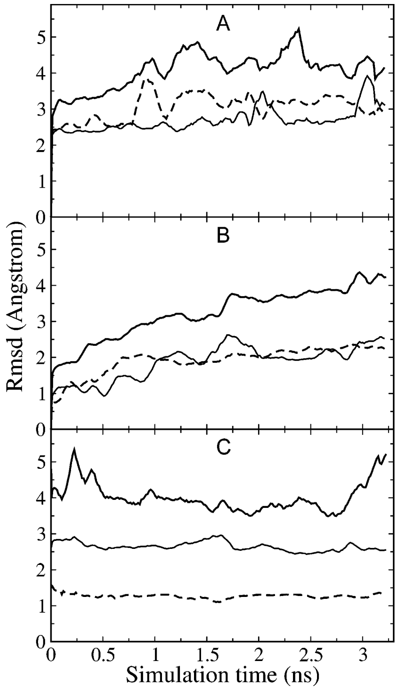


Figure 1. (A) Time course of the T4 lysozyme backbone (Cα) Rmsd (for the replica with the original force field and with respect to the start structure) during the ENM-RexMD (complete structure: continuous bold line) of the N-terminal (residues 15−54: dashed line) and C-terminal (residues 82−152: thine line) domains, respectively, that encompass the ligand binding site. (B) The same Rmsd plots for a conventional MD simulation starting from the same start structure as the ENM-RexMD. (C) Rmsd time courses (same as in A, B) for a conventional MD simulation that started from an open T4 lysozyme structure (obtained during ENM-RexMD after ~1 ns simulation time). Rmsd moving window averages with a window size of 0.1 ns are plotted.

The ENM-RexMD simulation produced an Rmsd time course (for the replica that runs at the original force field) that increased more rapidly at the beginning but with a similar Rmsd toward the end of the simulation. In contrast to the continuous MD simulation, the domain−domain distance flipped many times between several states that caused increased fluctuations of the Rmsd compared to the

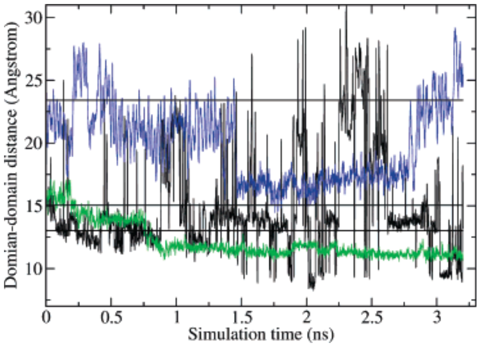


Figure 2. The hinge-bending motion of T4 lysozyme was monitored using the distance (domain−domain distance) between the centers of mass of the Cα backbone atoms of residues 18−26 (belonging to the N-terminal domain) and of residues 138−147 (of the C-terminal domain). These residues form the binding cleft between the N- and C-terminal domains of the enzyme. The distance was recorded for the ENM-RexMD (black line; for the replica with the original force field), during a conventional MD simulation starting from the same start structure as the ENM-RexMD (green line) and during the simulation that started from an open T4 lysozyme structure (blue line). For comparison the corresponding distance in the most closed experimental conformer (pdb152L, smallest distance), the start structure (pdb2LZM), and in one of the most open experimentally determined structures (pdb172L, largest distance) are shown as horizontal (black) lines.

continuous simulation (note, that in Figure 1 these fluctuations are partially damped by plotting a window average of the backbone Rmsd). Similar to the conventional MD simulation the Rmsd of the N- and C-terminal domains was smaller than for the complete protein (Figure 1a). The sampled states included significantly more open conformations than the start structure, states that are close to the start structure, and closed states with a domain−domain distance similar to the states sampled during conventional MD simulations (Figure 2, black line; snapshots shown in Figure 3). Interestingly, several of the sampled open structures showed good agreement (in structure and domain−domain distance) with another X-ray structure of a T4L variant (pdb172L)[52] that has been discussed as a representative conformation for an open T4-lysozyme state.[52,53] A superposition of one selected snapshot on the pdb172L structure (Rmsd(Cα)=2.8 Å) is shown in Figure 3d.

In order to control if the open structures produced during the ENM-RexMD run represented also stable states during conventional MD simulations one such open structure (from a simulation time of ∼1 ns of the ENM-RexMD) was used as a start structure for a conventional MD simulation. During this simulation the Rmsd of the complete protein with respect to the start structure remained at ∼4 Å, and smaller Rmsds of the individual N- and C-terminal domains (∼1 Å and ∼2.5 Å, respectively, Figure 1c) were observed. Several open forms with different domain−domain distances were sampled (blue curve in Figure 2). A transition to a closed form was not observed during the 3.2 ns simulation time. However, an extension of the simulation to 10 ns resulted in a closed structure (not shown).

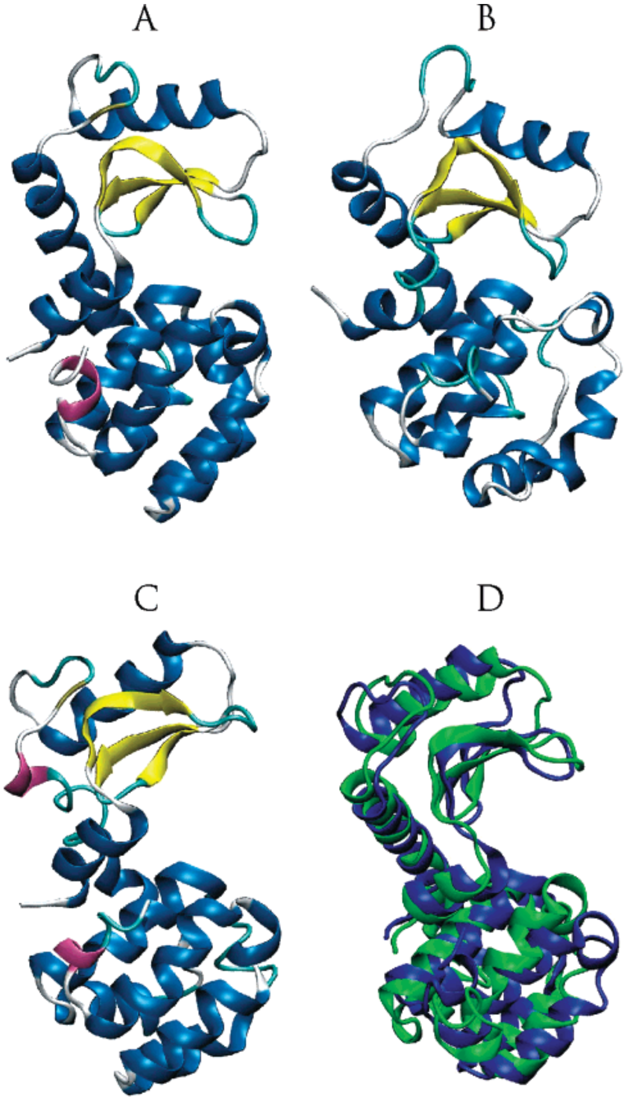


Figure 3. Snapshots of the T4 lysozyme structure observed during the ENM-RexMD simulations: (A) start structure (pdb2LZM), (B) a structure with a more closed cleft between N- and C-terminal domains compared to the start structure, and (C) an open protein structure observed during ENM-RexMD after ∼1 ns simulation time and used as a start structure for a continuous MD simulation. Structures are shown as a cartoon representation with a color coding according to a secondary structure. (D) Superposition of an open T4 lysozyme structure observed during the ENM-RexMD (green cartoon) and the experimental X-ray structure pdb172L (blue) which has been considered as one of the most open available experimental conformers.

The result indicates that the ENM-RexMD shows significantly improved sampling of open and closed states with many (>20) sampled transitions compared to the continuous MD simulations on a relatively short time scale of 3.2 ns. In contrast, not a single complete open−close or close−open transition was sampled during the same simulation time in the conventional MD simulations.

Folding Simulations on a β-Hairpin Forming Peptide. The ENM-RexMD approach was further evaluated on the small 10 residue chignolin β-hairpin forming peptide. The structure of this peptide was recently determined by NMR

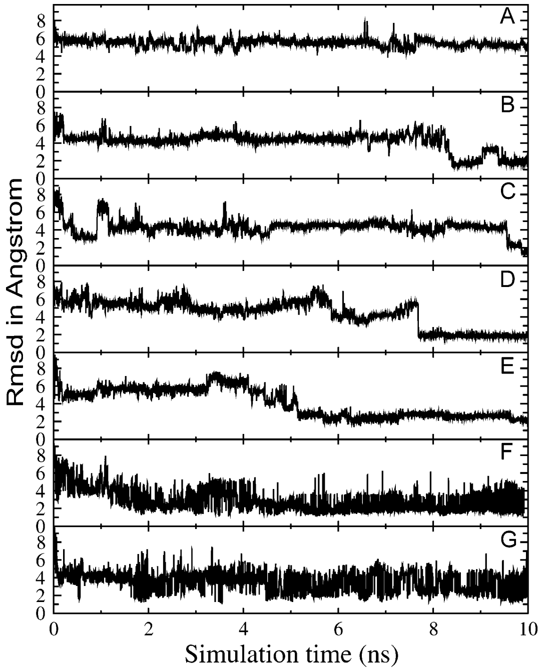


Figure 4. Rmsd (heavy atoms) from the experimental structure of the chignolin peptide (first model of pdb1UA0) observed during standard MD simulations starting from five different extended start structures (A–E). (F) Rmsd (heavy atoms) observed during a 5 replica ENM-RexMD (for the replica with the original force field) starting from five different start structures used in the standard MD simulations A–E. (G) was the same as in (F) but starting from the same start structure in all replicas (the start structure used in the first (A) standard MD simulation).

experiments.[50] It has also been demonstrated that extensive conventional temperature RexMD simulations (using 16 replicas) of more than 100 ns can lead to a folded structure very similar to the experimental NMR structure.[10]

Furthermore, a recent backbone dihedral biasing the potential RexMD approach (termed BP-RexMD) also achieved folding of the peptide to structures in close agreement with experiment within ~10 ns MD simulations in explicit solvent and required only 7 replicas.[33]

Test calculations using an implicit GB continuum model as implemented in Amber8 (igb=1) and the parm03 force field indicated that it is also possible to achieve folding of this peptide in implicit solvent to structures in close agreement with experiment (see also the Methods section). The folded conformation was the most populated structure during 25 ns MD simulations at 280 K staying within 2 Å ($\text{Rmsd}_{\text{heavy}}$) of the experimental start structure (not shown). In order to test the ENM-RexMD approach five different chignolin start structures were generated by short MD simulations at high temperature starting from a fully extended conformation (see the Methods section). As indicated in Figure 4 conventional MD simulations lead to conformations close to experiment after 5–10 ns for four out of five start structures (Rmsd ~2 Å, Figure 4). However, for the first start structure significantly longer simulations (>25 ns) were necessary to achieve a transition to structures in close agreement with experiment (presumably because of an altered Pro backbone conformation compared to the other start structures that corresponded to an altered rotation of the Pro psi dihedral angle).

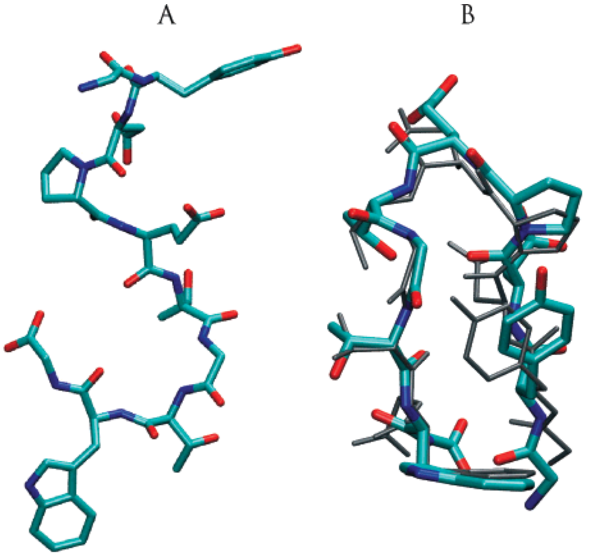


Figure 5. (A) Stick model (atom color-coded) of an extended start structure of the ENM-RexMD simulations. (B) Superposition of a typical near-native structure sampled during ENM-RexMD (stick model, atom color coded) and the experimental NMR structure (first structure of NMR ensemble in pdb1UA0, thin gray stick model). The heavy atom Rmsd between the two structures was ~1.6 Å.

For comparison, two ENM-RexMD simulations with 5 replicas were set up. In contrast to the T4L simulations ENM calculations were performed using all heavy atoms of the peptide structure and were recalculated every 15 ps. One simulation started from all five different start structures (one for each replica and the first most difficult start structure assigned to the replica with the original force field). The second ENM-RexMD simulation started from the first start structure (the most difficult one) assigned as start for all replicas. In both ENM-RexMD simulations transitions to a conformation in good agreement with experiment (Figure 5) were seen already after ~1.5–2 ns (Figure 4, last two plots) significantly faster than in the conventional MD simulations. These structures also quickly evolved to the most populated conformational states (Figure 4).

The mean potential energy of the structures (averages over 0.4 ns) dropped much faster already during the first 2 ns of the ENM-RexMD compared to the conventional MD simulations (Figure 6). Also, the energy probability distribution obtained during the first 10 ns conventional MD simulations (of the 5 different start structures) are all shifted to higher energies compared to the distribution obtained from the ENM-RexMD simulation (Figure 6b). However, the energy distribution obtained after 80 ns conventional MD was in close agreement with the result of 40 ns ENM-RexMD (for the replica run at the original force field, Figure 6c) for each start structure except for the first start structure (thin line in Figure 6c). As mentioned above, in this case transitions to near native conformations (combined with a drop of the potential energy) were observed only after simulation times >25 ns which is a likely reason for the shift of the distribution curve to slightly higher energies. The close

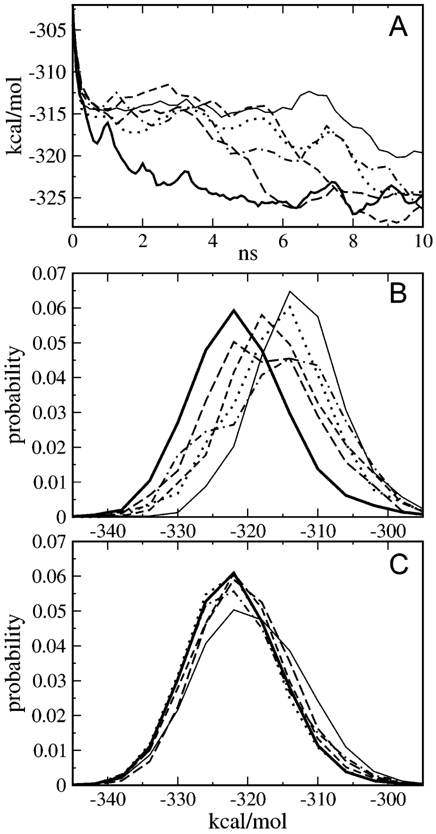


Figure 6. (A) Decrease of the mean potential energy (averages of 0.4 ns) during the first 10 ns of the ENM-RexMD (for the ENM-RexMD that started from 5 different start structures and the replica run with the original force field: bold continuous line). For comparison the decrease in mean potential energy is also shown for 5 conventional MD simulations starting from the 5 different start structures used also for the ENM-Rex-MD (five different line types). (B) Energy probability distribution obtained during the first 10 ns simulation time of the ENM-RexMD (replica run with the original force field: bold continuous line) and of the 5 conventional MD simulations (starting from the same 5 start structures, same line types as in A). (C) Energy probability distribution obtained during entire ENM-RexMD (40 ns; replica run with the original force field: bold continuous line) and 80 ns conventional MD simulation starting from each of the 5 start structures (same lines types as in A, B).

agreement of the final energy distribution of the ENM-RexMD and the conventional MD simulations indicates that the same energetic states are sampled and that the recalculation of the ENM from time to time during the RexMD simulation has only a minor influence on the calculated ensemble of states. The energy distribution obtained during the first 10 ns ENM-RexMD (bold curve in Figure 6b) is already very similar to the complete 40 ns ENM-RexMD (bold curve in Figure 6c) indicating a very rapid relaxation of the sampled ensemble of states. To further check if ENM-RexMD and extensive conventional MD simulations result in similar sampled conformations the Rmsd probability distribution (heavy atoms with respect to experimental structure) of conformations from all 5 conventional MD simulations was compared with those obtained during ENM-RexMD (Figure 7). During the first 10 ns the ENM-RexMD sampled an Rmsd distribution quite similar to the distribution of the entire 40 ns ENM-RexMD (bold curves in Figure 7a,b). However, the Rmsd distribution of the conventional MD simulation was shifted to higher Rmsd values during the initial 10 ns but was very similar to the result of the ENM-RexMD in case of using conformations sampled during all complete conventional MD simulations (5 × 80 ns; Figure 7b). In addition, a cluster analysis with respect to the backbone Rmsd of sampled structures was performed with an Rmsd cutoff of 2 Å and using the *kclust* program of the MMTSB tools.[57] For the cluster analysis 10 000 structures from the ENM-RexMD and 10 000 structures from the final 20 ns of the conventional MD simulations were analyzed. The first 5 clusters with highest population (representing in both cases ~60% of the sampled conformations) showed high overlap. In both cases the highest populated cluster corresponded to structures in close agreement with the experimental structure (Figure 8a). The close correspondence of representative conformations of highly populated clusters from the ENM-RexMD and the conventional MD is shown in Figure 8 indicating that extensive MD simulations starting from different start structures and ENM-RexMD simulation sample similar conformational ensembles.

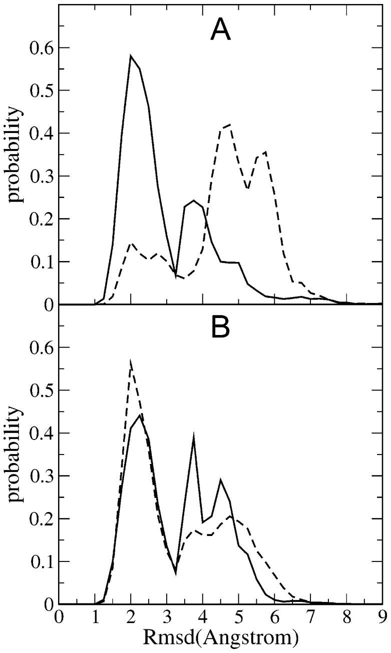


Figure 7. (A) Heavy atom Rmsd probability distribution (Rmsd with respect to experimental reference structure: the first model of pdb1UA0) observed during the first 10 ns of ENM-RexMD for the simulation that started from 5 different structures and for the replica run with the original force field: continuous line) and combined result for all 5 conventional MD simulations (starting from the same 5 different start structures: dashed line). (B) The same result as in (A) but for the entire ENM-RexMD (40 ns) and all conventional MD simulations (5 × 80 ns).

Discussion

To improve conformational sampling RexMD simulations have evolved as an important tool to study biomolecular

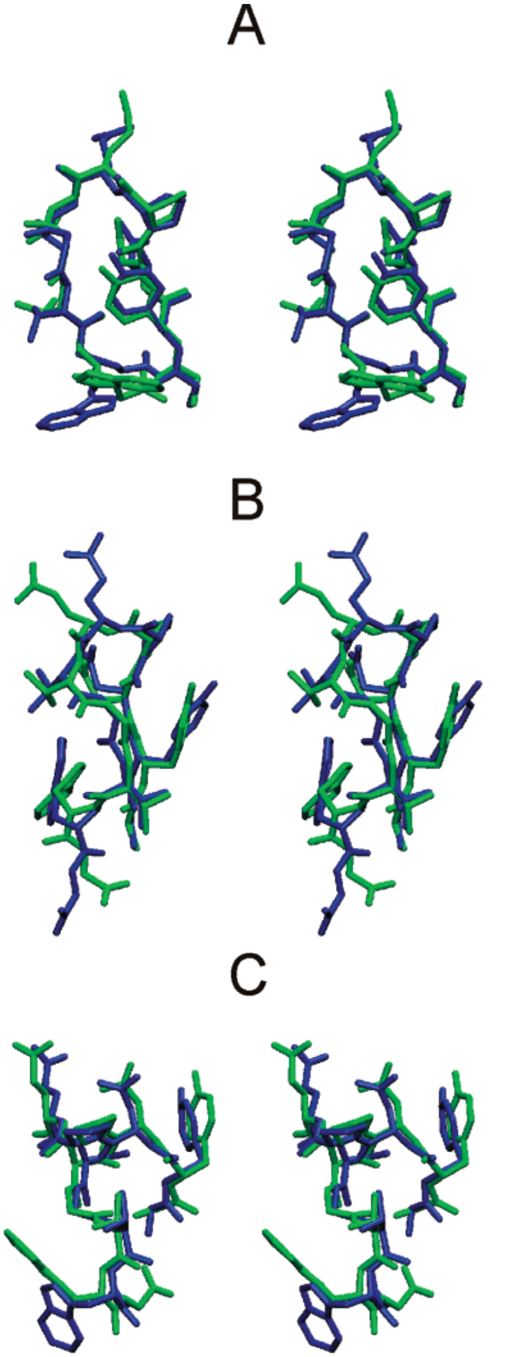


Figure 8. Comparison of highly populated conformational clusters observed during ENM-RexMD (for the replica run with the original force field: green stick models) and during the final 20 ns of all conventional MD simulations (blue structures). Cluster analysis was performed using *kclust* of the MMTSB tools.[56] (A) Superposition (stereoview) of the best representative conformation (closest backbone Rmsd(Cα) to the cluster centroid) for the largest populated cluster observed during ENM-RexMD (green) and conventional MD (blue) with an Rmsd(Cα) of 0.5 Å. (B) Superposition of the best representative of the second most populated cluster of ENM-RexMD and the third most populated cluster of conventional MD (Rmsd(Cα)=0.3 Å). (C) Same as for cluster 5 of ENM-RexMD and cluster 2 of conventional MD (Rmsd(Cα)=0.4 Å).

structures.[5−8,10−33] A major disadvantage of conventional temperature RexMD is the rapid increase of the required number of replicas to cover a desired temperature range.[24] The ratio of the standard deviation of the system potential energy (a measure of the energy fluctuation) vs average energy decreases with the square-root of the system size. Hence, to achieve sufficient overlap of the energy distributions between replicas run at different temperatures (required to achieve a reasonable exchange acceptance ratio) the temperature “spacing” between neighboring replicas is required to decrease with system size. Another drawback of large numbers of replicas is the need to run longer simulations (or more exchanges) to allow sufficient “travelling” or exchanges between high- and low-temperature replicas compared to a small number of replicas. Approaches that circumvent these drawbacks have been developed that include separate temperature coupling during RexMD of solute and solvent[27] or separate temperature coupling of essential degrees of freedom of the system.[28]

In Hamiltonian Rex-MD simulations the potential energy function (Hamiltonian) is scaled along the replicas.[24,25,29−33] A critical issue in the application of such Hamiltonian replica exchange methods is the choice and magnitude of force field energy terms to be scaled along the replicas. Force field terms that drastically change the energy of the simulation system may require many replicas with intermediate levels of the selected force field term to allow efficient exchanges between replicas.

One possibility is to design biasing force field terms that depend specifically on soft degrees of freedom of the system. An advantage is that the magnitude of the biasing potential can be kept small (relative to the total energy of the system) since motions in soft degrees of freedom naturally require smaller driving forces than hard degrees of freedom. Second, the likelihood to observe large scale conformational changes is also enhanced because these often involve motions along soft degrees of freedom of the molecule. The coarse-grained nature of the ENM description has the advantage that it provides soft degrees of freedom of a molecule at a smoother and more long-range level than the atomistic force field description. Hence, the ENM model of a protein molecule can look “beyond barriers” present at the level of a molecular mechanics force field. As already mentioned, the idea to couple ENM analysis to MD simulation has been explored by increasing the effective temperature of motion along soft ENM modes.[42] This resulted in an enhanced conformational sampling, however, with the drawback that the simulation is performed on a nonphysical system (separate temperature coupling of different degrees of freedom of the system).

In the present ENM-RexMD approach an ENM derived distance-dependent penalty or biasing potential is added at various levels for each replica and acts to help to drive the conformation toward regimes compatible with the ENM description. As long as the ENM description is not changing, each replica samples a canonical distribution of conformations compatible with the force field description for each replica. Hence, exchanges are performed between canonical ensembles such that detailed balance conditions are fulfilled (in the present method at no point a nonphysical system is sampled). Since the structure of the peptide or protein is changing, it is necessary from time to time to recalculate the ENM of the system and to adapt the biasing potential (this could also be done using a Monte Carlo scheme to adapt or reject changes in the biasing potential). For the present simulation systems the adaptation was done every 15−20 ps. The initial application of the method resulted in successful folding of a β-hairpin forming peptide in close agreement with experiment and faster than observed during standard MD simulations. For this peptide, folded structures in close agreement with experiment were observed as the most populated state already after <2 ns starting from several different start structures. In order to control if the recalculation of ENM modes from time to time during the ENM-RexMD has an influence on the energies and types of sampled structures, the sampled conformations were compared with trajectories obtained from extended (80 ns) conventional MD simulations. Very good agreement of the potential energy distribution, Rmsd distribution (with respect to the experimental structure), and the most populated conformational clusters between ENM-RexMD and conventional MD at the final stage of the simulations was obtained. This result shows that the updating of ENM-modes has only a small influence on the canonical equilibrium sampling of conformational states. However, much smaller differences in the energy and Rmsd distributions during the early stage vs complete ENM-RexMD simulation were found indicating a much faster relaxation toward an equilibrium distribution compared to conventional MD simulations.

It needs to be emphasized that in the present initial application of the method no extensive optimization of the simulation parameters was performed. It is likely that further simulation parameter optimization can lead to a further improvement of the conformational sampling of the approach. Simulation evaluation and trajectory analysis has been performed only using the unbiased replica run with the original force field (although after proper reweighting the other replicas could also be included).

Application to T4 lysozyme, a two domain protein, resulted in significantly enhanced sampling of hinge-bending motions. During the ENM-RexMD many transitions between open and closed structures (detected by calculating domain−domain distances) were observed. The Rmsd of the individual N- and C-terminal domains showed a deviation of ∼2.5 Å from the experimental structure that was generally smaller than the Rmsd of the complete protein from experiment (∼3.5−4 Å). The sampled domain−domain distance fluctuations reached ∼10−15 Å indicating domain−domain rearrangements that cannot be explained by intramolecular changes within each domain. In addition, sampled open structures were in quite good agreement with the X-ray structures of an open T4L conformation structure (a snapshot with an Rmsd(Cα)=2.8 Å from the open T4L structure in pdb172L is shown in Figure 3d) although the simulation started from a different more closed experimental structure. Conventional MD simulations of the same structures generated only conformations relatively close to the start conformation on the same time scale as the ENM-RexMD runs. This was achieved with a small number of 5 replicas for both the β-hairpin and the T4L system.

In the present implementation the ENM derived distance dependent biasing potential was only calculated for the structure from the replica that runs with the original force field (reference replica). This means that the potential drives the structures in each replica run away from the structure that runs with the original force field (to offer alternative low-energy structures that can then exchange with the reference replica run). In the T4L case the structures in each replica run differ mainly by the degree of opening/closing of the enzyme active site, and the soft mode directions of these structures are presumably quite similar. However, in the peptide folding case the structures in each replica run may differ significantly, and the corresponding mode directions of the structure in the reference replica run may also significantly differ from those of the reference structure. Hence, one may think that in this case the ENM derived potential in each replica acts mainly as a random perturbation potential. However, frequently the structures in each replica run contain at least segments that are similar to segments of the reference structure. Since the biasing potential has been expressed in terms of intramolecular distances, any such segment will be perturbed by the added perturbation potential (in a nonrandom fashion). It should be emphasized that other more sophisticated coupling schemes of ENM and RexMD might be possible that include for example ENM derived biasing potentials calculated separately for the structures of each replica run.

As has been pointed out by Huang et al.[57] a RexMD method that scales only part of the Hamilitonian of a system to reduce the number of required replicas in a RexMD simulation may not cover a large range of different conformations simultaneously (e.g., both unfolded and completely folded structures). It should therefore be emphasized that the current ENM-RexMD method primarily enhances the conformational sampling locally around a reference state. Therefore, it is necessary to recalculate the ENM at frequent intervals to adapt to a new reference state of the system.

A drawback of the ENM-RexMD approach is that for the current setup several parameters need to be adjusted. This includes the scaling of the distance fluctuations obtained from the ENM analysis, the maximum height and levels of the biasing potential along the replicas, and the frequency of recalculating the ENM description of the molecule. Future work will focus on a systematic evaluation of the parameters used during the ENM-RexMD approach and a possible automatic setup for a given simulation system.

Acknowledgment. This work was performed using the computational resources of the CLAMV (Computational Laboratories for Animation, Modeling and Visualization) at Jacobs University Bremen and supercomputer resources of the EMSL (Environmental Molecular Science Laboratories) at the PNNL (Pacific Northwest National Laboratories; grant gc11-2002).

References

(1) Daura, X.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Reversible peptide folding in solution by molecular dynamics simulation. *J. Mol. Biol.* **1998**, *280*, 925−932.

(2) Duan, Y.; Kollman, P. A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **1998**, *282*, 740−744.

(3) Roccatano, D.; Amadei, A.; Di Nola, N.; Berendsen, H. J. C. A molecular dynamics study of the 41−56 β-hairpin from b1 domain of protein G. *Protein Sci.* **1999**, *10*, 2130−2141.

(4) Pande, V. S.; Roshkar, D. S. Molecular dynamics simulations of unfolding and refolding of a β-hairpin fragment of protein G. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9062−9067.

(5) Garcia, A. E.; Sanbonmatsu, K. Y. Exploring the energy landscape of a β-hairpin in explicit solvent. *Proteins: Struct., Funct., Bioinf.* **2001**, *42*, 345.

(6) Zhou, R.; Berne, B. J.; Germain, R. The free energy landscape for β-hairpin folding in explicit water. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931−14937.

(7) Simmerling, C.; Strockbine, B.; Roitberg, A. E. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **2002**, *124*, 11258−11259.

(8) Rao, F.; Caflisch, A. Replica exchange molecular dynamics simulations of reversible folding. *J. Chem. Phys.* **2003**, *119*, 4035−4042.

(9) Roccatano, D.; Nau, W. M.; Zacharias, M. Structural and dynamic properties of the CAGQW peptide in water: A molecular dynamics simulation study using different force fields. *J. Phys. Chem.* **2004**, *108*, 18734−18742.

(10) Seibert, M. M.; Patriksson, A.; Hess, B.; van der Spoel, D. Reproducible Polypeptide Folding and Structure Prediction using Molecular Dynamics Simulations. *J. Mol. Biol.* **2005**, *354*, 173−183.

(11) Nguyen, P.; Stock, G.; Mittag, E.; Hu, C-K.; Li, M. S. Free energy landscape and folding mechanism of a β-hairpin in explicit water: A replica exchange molecular dynamics study. *Proteins* **2006**, *61*, 795−806.

(12) Gnanakaran, S.; Nymeyer, H.; Portman, J.; Sanbonmatsu, K. Y.; Garcia, A. E. Peptide folding simulations. *Curr. Opin. Struct. Biol.* **2003**, *15*, 168−175.

(13) Swendsen, R. H.; Wang, J. S. Replica Monte Carlo simulations of spin glasses. *Phys. Rev. Lett.* **1986**, *57*, 2607−2609.

(14) Okamoto, Y. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *J. Mol. Graphics Modell.* **2004**, *22*, 425−439.

(15) Predescu, C.; Predescu, M.; Ciobanu, C. V. J. On the Efficiency of Exchange in Parallel Tempering Monte Carlo Simulations. *J. Phys. Chem. B* **2005**, *109*, 4189−4196.

(16) Okabe, T.; Kawata, M.; Okamoto, Y.; Mikami M. Replica-exchange Monte Carlo method for the isobaric−isothermal ensemble. *Chem. Phys. Lett.* **2001**, *335*, 435−439.

(17) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(18) Mitsutake, A.; Sugita, Y.; Okamoto, Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers* **2001**, *60*, 96−123.

(19) Sanbonmatsu, K. Y.; Garcia, A. E. Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins: Struct., Funct., Bioinf.* **2002**, *46*, 225.

(20) Zhou, R.; Berne, B. J. Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12777−12782.

(21) Zhou, R. Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins: Struct., Funct., Bioinf.* **2003**, *53*, 148−161.

(22) Nymeyer, H.; Garcia, A. E. Simulation of the folding equilibrium of α-helical peptides: a comparison of the generalized Born approximation with explicit solvent. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13934−13939.

(23) Yoshida, K.; Yamaguchi, T.; Okamoto, Y. Replica-exchange molecular dynamics simulation of small peptide in water and in ethanol. *Chem. Phys. Lett.* **2005**, *41*, 2280−2284.

(24) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: application to protein structure prediction. *J. Chem. Phys.* **2002**, *116*, 9058−9065.

(25) Kannan.; S.; Zacharias, M. Folding of a DNA hairpin loop structure in explicit solvent using replica-exchange molecular dynamics simulations. *Biophys. J.* **2007**, *93*, 3218−3228.

(26) Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C. J. Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model. *J. Chem. Theory Comput.* **2006**, *2*, 420−433.

(27) Cheng, X.; Cui, G.; Hornak, V.; Simmerling, C. Modified Replica Exchange Simulation Methods for Local Structure Refinement. *J. Phys. Chem. B* **2005**, *109*, 8220−8230.

(28) Kubitzki, M. B.; de Groot, B. L. Molecular dynamics simulations using temperature-enhanced essential dynamics replica exchange. *Biophys. J.* **2007**, *92*, 4262−4270.

(29) Jang, S.; Shin, S.; Pak, Y. Replica-exchange method using the generalized effective potential. *Phys. Rev. Lett.* **2003**, *91*, 58305−58309.

(30) Zhu, Z.; Tuckerman, M. E.; Samuelson, S. O.; Martyna, G. J. Using Novel Variable Transformations to Enhance Conformational Sampling in Molecular Dynamics. *Phys. Rev. Lett.* **2002**, *88*, 100201.

(31) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. A. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13749−13754.

(32) Affentranger, R.; Tavernelli, I.; Di Iorio, E. E. A Novel Hamiltonian Replica Exchange MD Protocol to Enhance Protein Conformational Space Sampling. *J. Chem. Theory Comput.* **2006**, *2*, 217−228.

(33) Kannan, S.; Zacharias, M. Enhanced sampling of peptide and protein conformations using replica exchange simulations with a peptide backbone biasing-potential. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 697−706.

(34) Tirion, M. M. Large amplitude elastic motions in proteins from a single-parameter atomic analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905−1908.

(35) Bahar, I.; Atilgan, A. R.; Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding Des.* **1997**, *2*, 173−181.

(36) Hinsen, K. Analysis of domain motions by approximate normal mode calculations. *Proteins: Struct., Funct., Bioinf.* **1998**, *33*, 417−429.

(37) Tama, F.; Sanejouand, Y. H. Conformational change of proteins arising from normal mode calculations. *Protein Eng.* **2001**, *14*, 1−6.

(38) Tobi, D.; Bahar, I. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 18908−18913.

(39) Bahar, I.; Rader, E. J. Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.* **2005**, *15*, 586−592.

(40) May, A.; Zacharias, M. Accounting for global protein deformability during protein-protein and protein-ligand docking. *Biochim. Biophys. Acta* **2005**, *1754*, 225−231.

(41) May, A.; Zacharias, M. Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 794−809.

(42) Zhang, Z.; Shi, Y.; Liu, H. Molecular dynamics simulations of peptides and proteins with amplified collective motions. *Biophys. J.* **2003**, *84*, 3583−3593.

(43) Grubmüller, H. Predicting slow structural transitions in macromolecular systems: conformational flooding. *Phys. Rev. E* **1995**, *52*, 2893−2906.

(44) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *99*, 12562−12566.

(45) Case, D.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsui, V.; Radmer, R. J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *Amber 8*; University of California: San Francisco, CA, 2003.

(46) Duan, Y.; Wu, A.; Chowdhury, C. S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999−2012.

(47) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Pairwise solute descreening of solute charges from a dielectric continuum. *Chem. Phys. Lett.* **1995**, *246*, 122−129.

(48) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Parametrized models of aqueous free energies of solvation based onpairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.* **1996**, *100*, 19824−19839.

(49) Miyamoto, S.; Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952−962.

(50) Honda S.; Yamasaki K.; Sawada Y.; Morii, H. 10 residue folded peptide designed by segment statistics. *Struct. Fold. Des.* **2004**, *12*, 1507−1518.

(51) Weaver, L. H.; Matthews, B. W. Structure of bacteriophage T4 lysozyme refined at 1.7 Angstrom resolution. *J. Mol. Biol.* **1987**, *193*, 189−199.

(52) Zhang, X.-J.; Wozniak, J. A.; Matthews, B. W. Protein flexibility and adaptability seen in 25 crystal forms of T4 lyzozyme. *J. Mol. Biol.* **1995**, *250*, 527−552.

(53) Goto, N. K.; Skrynnikov, N. R.; Dahlquist, F. W.; Kay, L. E. What is the average conformation of bacteriophage T4 lysozyme in solution? A domain orientation study using dipolar couplings measured by solution NMR. *J. Mol. Biol.* **2001**, *308*, 745−764.

(54) de Groot, B. L.; Hayward, S.; van Alten, D. M. F.; Amadei, A.; Berendsen, H. J. C. Domain motions in bacteriophage T4 lysozyme: A comparison between molecular dynamics and crystallographic data. *Proteins: Struct., Funct., Bioinf.* **1998**, *31*, 116−127.

(55) Pjura, P. E.; Matsumura, M.; Wozniak, J. A.; Matthews, B. W. Structure of a thermostable disulfide-bridge mutant of phage T4 lysozyme shows that an engineered cross-link in a flexible region does not increase rigidity of the folded protein. *Nature* **1990**, *345*, 86−89.

(56) Feig, M.; Karanicolas, J.; Brooks, C. L. MMTSB tool set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graphics Modell.* **2004**, *22*, 377−395.

(57) Huang, X.; Hagen, M.; Kim, B.; Friesner, R. A.; Zhou, R.; Berne, B. J. Replica exchange with solute tempering: efficiency in large scale systems. *J. Phys. Chem. B* **2007**, *111*, 5405−5410.

CT7002258

JCTC Journal of Chemical Theory and Computation

Evaluation of Salt Bridge Structure and Energetics in Peptides Using Explicit, Implicit, and Hybrid Solvation Models

Asim Okur,[†] Lauren Wickstrom,[#] and Carlos Simmerling*,[†,#]

Department of Chemistry and Graduate Program in Biochemistry and Structural Biology, Stony Brook University, Stony Brook, New York 11794

Received September 11, 2007

Abstract: Replica exchange or parallel tempering molecular dynamics (REMD) is widely used to enhance the exploration of free energy landscapes for complex molecular systems. However its application to large systems is hampered by the scaling of the number of required replicas with an increasing system size. We recently proposed an improved REMD method where the exchange probabilities were calculated using a hybrid explicit/implicit solvent model. We previously tested this hybrid solvent REMD approach on alanine polypeptides of 1, 3, and 10 residues and obtained very good agreement with fully solvated REMD simulations while significantly reducing the number of replicas required. In this study we continue evaluating the applicability of the hybrid solvent REMD method through comparing the free energy of formation of ion pairs using model peptides. In accord with other studies, pure GB simulations resulted in overstabilized salt bridges, whereas the hybrid models produced free energy profiles in close agreement with fully solvated simulations, including solvent separated minima. Furthermore, the structure of the salt bridge in explicit solvent is reproduced by the hybrid solvent REMD method, while the GB simulations favor a different geometry.

Introduction

Conformational sampling remains one of the biggest challenges in atomistic simulations for biological systems. Rugged and complex energy surfaces often result in simulations being trapped even when a sufficiently accurate Hamiltonian is used, prohibiting complete exploration of the conformational space. Significant effort has been put into developing efficient simulation methods to locate low-energy minima for these complex systems. The challenges in conformational sampling have been discussed in several reviews.[1,2]

One major problem for molecular simulations is quasi-ergodicity where simulations may appear converged when observing some simulation parameters, but in reality large energy barriers may prevent them from sampling important regions of the energy landscape. Another simulation initiated in a different conformation may look converged as well, but comparison may show that only partial equilibration was achieved (see ref 3 as an example for quasi-ergodicity).

One popular approach to overcoming quasi-ergodicity in biomolecular simulation is the replica exchange method.[4−6] In replica exchange molecular dynamics (REMD)[7] (also known as parallel tempering[4]), a series of molecular dynamics simulations (replicas) is performed for the system of interest. In the original form of REMD, each replica is an independent realization of the system, coupled to a thermostat at a different temperature. The temperatures of the replicas span a range from low values of interest (experimentally accessible temperatures such as 280 or 300 K) up to high values (such as 600 K) at which the system is expected to more rapidly overcome potential energy barriers that would otherwise impede conformational transitions on a computationally affordable time scale.

At intervals during the otherwise standard simulations, conformations of the system being sampled at different temperatures are exchanged based on a Metropolis-type

* Corresponding author e-mail: carlos.simmerling@stonybrook.edu.

[†] Department of Chemistry.

[#] Graduate Program in Biochemistry and Structural Biology.

10.1021/ct7002308 CCC: $40.75 © 2008 American Chemical Society
Published on Web 02/21/2008

criterion[8] that considers the probability of sampling each conformation at the alternate temperature (further details are discussed in Methods). In this manner, REMD is hampered to a lesser degree by the local minima problem, since simulations at low temperatures can escape kinetic traps by "jumping" directly to alternate minima being sampled at higher temperatures. Moreover, the transition probability is constructed such that the canonical ensemble properties are maintained during each simulation, thus providing potentially useful information about conformational probabilities as a function of temperature. Due to these advantages, REMD has been widely applied to folding studies of peptides and small proteins.[4,7,9−20]

For large systems, REMD can become intractable since the number of replicas needed to span a given temperature range increases with the square root of the number of degrees of freedom in the system.[21−24] Since the number of accessible conformations also typically increases with system size, the current computational cost for REMD simulations of large systems limits the simulation lengths to tens of nanoseconds per replica, which limits the ability to obtain converged ensembles for large systems. Several promising techniques have been proposed[21,25−35] to deal with this apparent disadvantage of REMD.

Continuum solvent models like the semianalytical Generalized Born (GB) model[36] estimate the free energy of solvation of the solute based on coordinates of the solute atoms. The neglect of explicit solvent molecules can significantly reduce the computational cost of evaluating energies and forces for the system, but a larger effect with REMD can arise from the reduction in the number of replicas due to fewer degrees of freedom. However, these models can also have significant limitations. Since the atomic detail of the solvent is not considered, modeling specific effects of structured water molecules can be challenging. In the case of protein and peptide folding, it appears likely that the current generation of GB models do not have as good a balance between protein−protein and protein−solvent interactions as do the more widely tested explicit solvent models.[37,38] Previous studies on alanine based peptides show that the use of GB models can induce some bias in helical backbone conformations.[29,39] More particularly, it has been reported[14,38,40,41] that ion pairs were frequently too stable in the GB implicit water model, causing salt bridged conformations to be oversampled in MD simulations, thus altering the thermodynamics and kinetics of folding for small peptides. A clear illustration was given by Zhou and Berne[38] who sampled the C-terminal β-hairpin of protein G (GB1) with both a surface-GB (SGB)[42] continuum model and explicit solvent. The lowest free energy state with SGB was significantly different from the lowest free energy state in explicit solvent, with non-native salt bridges formed at the core of the peptide, in place of hydrophobic contacts. Zhou extended this study on GB1 by examining several force field-GB model combinations, with all GB models tested showing erroneous salt bridges.[41] The helical backbone bias and overstabilized salt bridges prevent the use of these GB models for conformational search for peptides and proteins.

Recognizing that a major obstacle in applying REMD with explicit solvent lies in the number of simulations (replicas) required, rather than just the complexity of each simulation, we introduced a new approach in which each replica is simulated in explicit solvent using standard methods such as periodic boundary conditions and inclusion of long-range electrostatic interactions.[28] However, the calculation of exchange probabilities (which determines the temperature spacing and thus the number of replicas) is handled differently. Only a subset of closest water molecules is retained, with the remainder temporarily replaced by a continuum representation. The energy is calculated using the hybrid model, and the exchange probability is determined. The original solvent coordinates are then restored, and the simulation proceeds as a continuous trajectory with fully explicit solvation. This way the perceived system size for evaluation of exchange probability is dramatically reduced and fewer replicas are needed.

Earlier tests of our hybrid solvent REMD method were performed on alanine polymers of 1, 3, and 10 residues, and the performance of hybrid approach was compared to fully solvated explicit solvent REMD simulations.[28] For Ala_{10}, a fivefold reduction in the number of replicas provided similar exchange probability, and good agreement was found for the populations of various minima corresponding to secondary structure types. The explicit inclusion of the first solvation shell eliminated the helical backbone bias introduced by GB and resulted in distributions in close agreement with fully solvated simulations.[28]

A similar approach was developed by Liu et al.[43] where they identified the solute as the central group and separated solute−solvent interaction energies and solvent−solvent energies and made solvent−solvent energies temperature dependent (Replica Exchange with Solute Tempering − REST). On average the temperature-dependent water−water interactions cancel over the replicas, bringing the potential energy distributions closer and providing better overlap with fewer replicas. The REST approach was tested on alanine dipeptide simulations, and significant reduction in the required number of replicas was observed.[43] When REST was applied to larger systems, however, the method was not as efficient.[44] To help overcome this issue they included a subset of water molecules to the central group and calculated their interactions explicitly with the solute, thereby using the increased energy fluctuations from the water interactions to provide the thermal basis for driving solute conformation exchanges. To obtain similar results with fully solvated REMD however a significant number of random central waters had to be added. Since the resulting central group was still smaller than the whole system, some reduction in the number of replicas required was observed.[44]

We previously chose to study polyalanines since the lack of complex side-chain interactions in these peptides enabled direct evaluation of the effect of solvent model on backbone conformation distributions. The results demonstrated that the GB models introduce significant secondary structure bias even in the absence of more complex side-chain functional groups.[39] While the hybrid solvent approach largely corrected these problems, application to more complex systems requires validation with interactions between side chains, particularly charge−charge interactions such as salt bridges, for which GB models have been shown to perform poorly.[45]

In this report we describe further testing of the hybrid approach on peptides with the possibility of interactions between oppositely charged side chains. We calculate the Potential of Mean Force (PMF) of salt bridge formation between Arg and Glu side chains in a small model peptide where the charged residues are separated by 2 alanine residues. The hybrid solvent REMD results are compared to fully TIP3P solvated REMD simulations and GB-REMD simulations on the same system. As we observed with polyalanine peptides, GB models induced a bias resulting in overpopulation of helical backbone conformations. In order to separate the effects of solvent model on backbone conformation from those involving the side chains, we repeated our evaluation of ion pairing in REMD simulations with restrained backbone conformations. With a consistent set of backbone conformations, the GB REMD simulations showed salt bridges that are 2−3 kcal/mol stronger than corresponding TIP3P REMD simulations, and the salt bridge orientation and hydrogen-bonding pattern also differed. The use of hybrid solvent REMD reduced the number of replicas by a factor of 5 compared to fully explicit water REMD while providing the same preferred salt bridge geometry as explicit solvent and also greatly improved the free energy profile for ion pairing compared to pure GB REMD.

To further validate this approach, we applied the hybrid REMD method on a larger system, HP-1, corresponding to the isolated N-terminal helix of villin headpiece helical subdomain HP36. Previous work on HP-1 showed ∼1.5 kcal/mol overstabilization of the α-helix conformation and a stronger salt bridge interaction in GB simulations compared to explicit solvent.[46] Here, we compare melting curves and free energy of salt bridge formation between Lys and Asp residues obtained with the hybrid solvent REMD approach to our previous data obtained using standard REMD in explicit or implicit water. The hybrid solvent REMD simulations showed a significant improvement in the population of helical conformations across a wide range of temperatures. The salt bridge PMFs obtained from hybrid solvent REMD were also in better agreement with explicit solvent including the solvent separated minimum and correct location of the global free energy minimum. The results from both peptides provide further validation of the hybrid solvent REMD approach for application to more complex systems.

Methods

Arg-Ala-Ala-Glu Model Peptide. We simulated a 4 residue model peptide (Arg-Ala-Ala-Glu) with acetylated and amidated N- and C-termini, respectively. All simulations employed Amber ff99SB,[47] a modified version of ff94/ff99[48,49] with corrections to dihedral parameters to improve secondary structure preferences. Explicit solvent and hybrid solvent REMD simulations used the TIP3P water model.[50] The standard REMD simulations in explicit solvent and in pure GB were run using our REMD implementation as distributed in Amber (version 9).[51] The hybrid solvent REMD calculations were performed with a locally modified version of Amber 9. All bonds involving hydrogen were constrained in length using SHAKE.[52] The time step was 2 fs. Temperatures were maintained using weak coupling[53] to a bath with a time constant of 0.5 ps^{-1}.

Explicit Solvent REMD. The model peptide was solvated in a truncated octahedron box with 16 Å buffer using 2286 TIP3P water molecules for a total of 6926 atoms. Such a large solvent box was selected to ensure that the salt bridge distance between images is larger than the maximum distance available for the linear peptide to reduce possible artifacts caused by periodicity.[54] The system was equilibrated at 300 K for 50 ps with harmonic positional restraints on solute atoms, followed by minimizations with gradually reduced solute positional restraints and three 5 ps MD simulations with gradually reduced restraints at 300 K. Long-range electrostatic interactions were calculated using PME.[55] Simulations were run in the NVT ensemble.

Forty-six replicas all starting from a salt bridged conformation were used at temperatures ranging from 296 K to 584 K, which were optimized to give a uniform exchange acceptance ratio of ∼25%. Exchange between neighboring temperatures was attempted every 1 ps, and each REMD simulation was run for 30 000 exchange attempts (30 ns per replica). The first 5000 exchange attempts of the simulation were discarded to remove initial structure bias.

Implicit Solvent REMD. Solvent effects were calculated through the use of the Generalized Born[36] implicit solvent model with the GB^{OBC} [56] implementation in Amber. The intrinsic Born radii were adopted from Bondi[57] with modification of hydrogen.[58] The GB^{OBC} model was employed with mbondi2 radii. Scaling factors were taken from the TINKER modeling package.[59] No cutoff on nonbonded interactions was used. All other simulation parameters were the same as used in explicit solvent.

For the model peptide, the use of the continuum solvent model resulted in a system size of 68 atoms which permitted the use of 6 replicas to cover a temperature range of 300−636 K. Exchanges were attempted every 1 ps, and the REMD simulations were run 30 000 exchange attempts (30 ns). The first 5000 exchanges were again removed. All replicas were initiated with the same initial peptide conformation used for the explicit solvent REMD calculations.

Hybrid Solvent REMD. All simulation parameters in the hybrid solvent REMD simulations were the same as those employed for standard REMD in explicit solvent, with the exception of the number of replicas (8 replicas were used to cover the temperature range from 270 K to 570 K). The hybrid solvent exchange scheme is employed exactly as described in ref 28. At each exchange step during hybrid solvent REMD, the distance between the oxygen atom of each water molecule and all solute atoms was calculated. Water molecules were then sorted by their closest solute distance, and all water molecules except the 75 with the shortest solvent−solute distances were temporarily discarded. Seventy-five water molecules were sufficient to solvate the first shell of the peptide in extended conformation (Figure S.1. in the Supporting Information). The energy of this smaller system was then recalculated using only these close waters, and the remainders were replaced by the GB solvent model. This hybrid solvent energy was used to calculate the exchange probability, and then all waters were restored to their original positions and the simulations were continued. In this manner the simulations using the hybrid solvent model produce continuous trajectories with fully explicit solvent, and the hybrid model was used only during calculation of exchange probabilities. The hybrid exchange scheme significantly reduced the number of replicas required as compared to explicitly including all solvent in the exchange calculation (46 vs 8 replicas) while maintaining a similar exchange success ratio of 20−30% (the observed exchange ratios can be found in Table S.1. in the Supporting Information). We only included the first solvation shell in the present simulations since our previous work[28] suggested that retaining a second shell provided little additional benefit and significantly increased the system size.

Analysis. Salt bridge PMFs were calculated using histogram analysis along a reaction coordinate defined using the distance between the Cζ of Arg1 and Cδ of Glu4 for the model peptide. Using Cζ resulted in a PMF that was less sensitive to the particular H-bond donor of the guanidino group which was analyzed separately. All distances were calculated using the ptraj module in Amber 8. Due to the computational cost of the REMD simulations, particularly standard REMD in explicit water, only a single run was performed for each solvent model, and uncertainties were calculated from the difference between free energy values obtained using the first half of the data set and those obtained only from the second half of the simulation. The convergence of our simulations was further checked by ensuring the salt bridge was formed and broken multiple times for each replica (see Figure S.2. in the Supporting Information for salt bridge distances for sample replicas).

To compare the backbone conformations, cluster analysis over the backbone atoms was performed for temperature trajectories at 300 K obtained from each REMD simulation. The 300 K trajectories were combined, and cluster analysis was performed with Moil-View[60] using backbone atoms as a similarity criterion with average linkage. Clusters were then formed with the bottom-up approach using a similarity cutoff of 1.0 Å. In this approach, each structure was initially assigned to a distinct cluster, average rmsd values between all cluster pairs were calculated, and the cluster pair with the smallest rmsd was merged. This procedure was repeated until the most similar cluster pair exceeded the similarity cutoff. The population of each cluster was then calculated separately for each simulation and plotted against each other for easy comparison. The same clustering scheme was used to investigate salt bridge orientations between different models where the atoms of Arg and Glu side chains were clustered again with a similarity cutoff of 1.0 Å.

Restrained REMD simulations used the same procedure for unrestrained simulations where the backbone conformation was restrained to the representative conformation obtained from the highest populated TIP3P cluster using weak (1.0 kcal/mol*Å) positional restraints. Restrained REMD simulations were run up to 40 000 exchange attempts for each solvation method, and the first 5000 exchange attempts were discarded as equilibration.

HP-1 Model Peptide. The HP-1 REMD simulations were run in a similar manner as the Arg-Ala-Ala-Glu peptide. The system was built from the sequence MLSDEDFKAVFGM which corresponds to the N-terminal helix of the villin headpiece helical subdomain HP36 (pdb code 1VII).[61] We have investigated the structural ensembles of this peptide in an earlier study through well-converged explicit solvent REMD simulations.[46]

Hybrid solvent REMD simulations were performed using 100 explicit water molecules in the exchange calculations. Eight replicas with a temperature range from 272 to 539 K were started from the native helical conformation, and the simulation was run up to 40 ns per replica, where the first 10 ns was discarded for data analysis. The GB^{OBC} REMD simulations were run using the same number of replicas and temperature distributions and were run up to 40 ns. For the GB^{OBC} REMD simulations the first 10 ns was again discarded before analysis.

Melting curves were constructed by calculating the average fraction helicity for each temperature. Helical residues were selected based on DSSP criterion.[62] Salt bridge PMFs were calculated using histogram analysis along a reaction coordinate using the distance between Nζ of Lys48 and Cγ of Asp44. To reduce the effects of different backbone conformations for different solvation schemes, the salt bridge PMF was calculated at 365 K where, based on the melting curve, simulations using each solvent model showed similar helical propensities. Error bars were calculated by comparing the first half and the second half of the data sets.

Results and Discussion

We previously tested hybrid solvent REMD with polyalanine peptides of varying lengths, obtaining good agreement with standard explicit solvent REMD simulations at reduced cost.[28] In simulations using only GB solvation, strong α-helical populations were observed for alanine peptides with 3 and 10 residues, in disagreement with explicit solvent simulations. In contrast, use of hybrid solvent REMD provided secondary structure propensities in near quantitative agreement with standard explicit solvent REMD simulations. Cluster analysis of backbone conformations were also in good agreement between TIP3P and hybrid solvent REMD simulations, but pure GB solvation showed large errors in conformational preferences. These results indicated that GB introduces significant bias in peptide backbone conformations, even in the absence of more complex side-chain interactions.[39] To further validate the use of hybrid solvent REMD on more complex biopolymers we studied the interaction of side-chain ion pairs in a small model peptide Ace-Arg-Ala-Ala-Glu-NH2, with Arg and Glu both modeled in the charged state. Salt bridge strength in the various solvent models was evaluated through calculation of the potential of mean force for the distance between Cζ of Arg and Cδ of Glu as sampled in the simulated ensembles.

The model peptide was simulated with standard REMD using either the GB^{OBC} implicit water model or the TIP3P explicit water model. Hybrid solvent REMD simulations used the same procedure as previously described[28] where the MD portions of the REMD were performed using the same

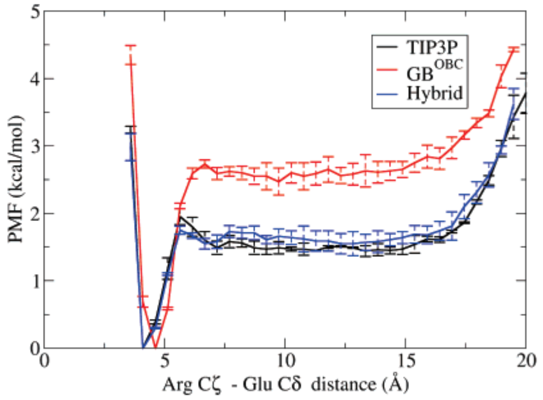


Figure 1. Potentials of mean force for the distance between Cζ of Arg and Cδ of Glu side chains to compare the free energy profiles for salt bridge formation for different methods. The GB^{OBC} method shows more stable salt bridges (∼1.0 kcal/mol) compared to the TIP3P explicit solvent model. The hybrid solvent REMD method shows a profile similar to fully solvated TIP3P REMD.

protocol and system as the standard REMD simulations in explicit water. The only difference was in the calculation of the exchange probability, during which all but the closest 75 water molecules were temporarily removed and replaced with the GB^{OBC} implicit model. All water molecules were restored to their original positions after the exchange calculation, thus the simulations provided continuous trajectories fully explicitly solvated during all MD steps. The use of the hybrid implicit/explicit solvent model during the exchange calculation dramatically reduced the number of replicas (from 46 to 8) required to obtain the desired exchange frequency.

The resulting salt bridge PMF curves for implicit, explicit, and hybrid solvent REMD are shown in Figure 1. The data demonstrate that the GB^{OBC} method produces modestly overstabilized salt bridges (about 1.0 kcal/mol) compared to explicit solvent, in accord with previous studies.[45] The GB^{OBC} profile also shows a free energy minimum at a slightly different value of distance than explicit solvent simulations, which suggests a different side-chain orientation or hydrogen bond pattern between solvent models. The PMF obtained from hybrid solvent REMD is very similar to the full TIP3P simulations, where both curves lie between their respective error bars, and also have their minimum at the same distance value. The curve follows the standard REMD TIP3P PMF closely over all distances sampled, including the prominent solvent separated minimum.

To investigate the differences in depth and location of the free energy minimum between the PMF profiles, we first compared the backbone conformations of the model peptide sampled with each solvation scheme. Knowing that this GB model introduces α-helical bias on the backbone of polyalanine peptides[28,39] we first looked at the backbone conformations from each solvation method to see if the helical bias in GB with polyalanine is also present with more complex side-chain functional groups. As described in the Methods section, we performed cluster analysis over the backbone atoms to identify multiple conformations sampled by each simulation. The temperature trajectories at 300 K from all 3 models were combined and clustered with a backbone rmsd cutoff of 1.0 Å, resulting in 63 clusters. The merging of the trajectory files for clustering allows direct comparison of populations between the different solvation approaches.[63] A comparison of the populations for each cluster in the different simulations is shown in Figure 2.

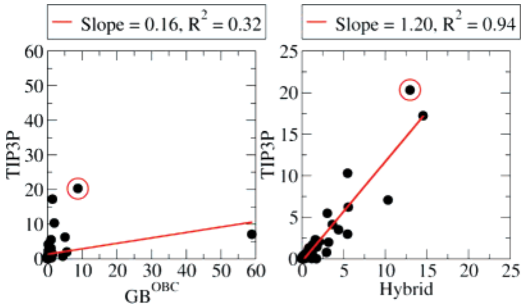


Figure 2. Comparison of backbone conformation populations for (left) standard REMD with GB^{OBC} and TIP3P (right) and standard REMD and hybrid solvent REMD simulations with TIP3P. Note that the scale differs for the right and left graphs, but the X and Y axes match in each graph. The slope and R^2 values for the best fit line are shown on top of each graph. Populations show poor correlation between TIP3P and GB^{OBC} (R^2 = 0.32) where the most populated conformation for GB^{OBC} has predominantly a helical conformation and is only weakly populated in TIP3P. In contrast, the populations of the conformational families show excellent correlation between the ensembles obtained through standard and hybrid solvent REMD (R^2 = 0.94). The most populated TIP3P REMD cluster is shown with red circles on each graph. (The populations for each cluster are provided in Table S.2. in the Supporting Information).

From Figure 2 we can readily observe that the GB and TIP3P solvent models sample very different backbone conformation ensembles. When the most populated conformations for each model are visually inspected, we observe that GB REMD predominantly samples a helical conformation (∼60%) in accord with our previous simulations of alanine peptides.[39,63] Similar with the alanine peptides the extended conformation is the most populated backbone conformation for this model peptide.

Since the preferred backbone conformations are so different, it is difficult to identify the reason for different free energy profiles between GB^{OBC} and TIP3P. It is possible that the strong salt bridge and different orientation could be caused by the helical backbone conformation rather than the intrinsic energy profile of the ion pair. The improvements observed through the use of hybrid solvent REMD could also be because of better backbone sampling than with pure GB. To facilitate a more accurate comparison between the specific effects of solvation on the ion pair, we generated a new set of REMD simulations in which the backbone conformation was restrained to the representative conformation from the most populated TIP3P cluster (red circle in Figure 2). This comparison should eliminate any effects in the PMFs introduced by averaging over different backbone conformations and should give a more clear picture of salt bridge strength and orientation between TIP3P, GB^{OBC}, and hybrid solvent REMD.

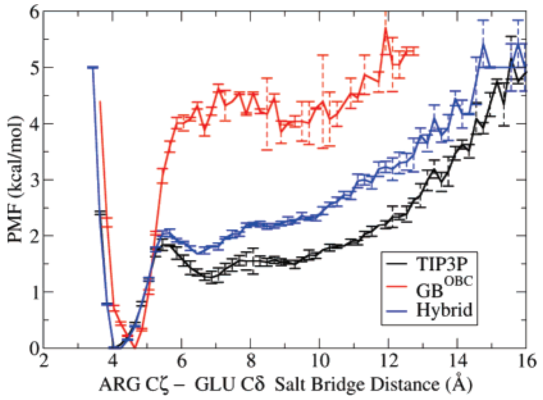


Figure 3. Potentials of mean force for salt bridge formation in the model peptide at 300 K for various solvation methods with restrained backbone conformation. Even with the backbone conformation restrained to that preferred in explicit solvent, GB^{OBC} still shows overstabilized salt bridges. The use of hybrid solvent REMD significantly improves the salt bridge free energy profile and identifies the minimum and solvent separated minimum correctly; however, the curve is shifted about 0.5 kcal/mol higher than with standard explicit solvent REMD.

The restrained REMD simulations were run for all three models for ∼40 000 exchange attempts for each solvation scheme. The resulting PMF curves are shown in Figure 3. The hybrid solvent exchange criterion again provides data in good agreement with that from standard REMD in TIP3P; the hybrid solvent REMD profile is within 0.5−1 kcal/mol of the fully solvated TIP3P results over the entire range of distances. In contrast, the GB simulations show significant inaccuracies. Even after removal of possible bias from different backbone ensembles, the GB^{OBC} REMD data shows stronger salt bridge formation (2.5−3 kcal/mol) compared to the TIP3P standard REMD ensemble. It is interesting to note that this bias is even larger than the ∼1 kcal/mol observed in Figure 1; the bias in backbone conformation with GB in the unrestrained ensemble appears to counteract the ion pair bias, indicating that unrestrained dynamics with GB has significant cancellation of error between inaccuracies in the secondary structure preferences and ion pair preferences that were not apparent without comparison of the backbone-restrained ensembles.

One of the common shortcomings of implicit solvation methods is the difficulty in modeling structured water molecules. In the case of ion pairs, solvent separated minima should be observed where the polar water molecule can hydrogen bond to and bridge the two ions. Such a minimum is clearly present at ∼7 Å in the PMF curves obtained with standard REMD in explicit water (Figure 3; the minimum is not as apparent in Figure 1 due to averaging over profiles from many backbone conformations). The GB model used in the present study is a pairwise descreening model[64] and as such does not use a surface- or volume-based dielectric boundary. Therefore, no solvent separated minimum would be expected for this model, and none is observed in the PMF curves. Figure 3 shows that hybrid solvent REMD can identify the solvent separated minimum correctly, which is expected since the dynamics are carried out with full explicit solvent and the first solvation shell is retained in the exchange probability calculation. Figure 4 shows a typical water bridged conformation obtained from the hybrid solvent REMD simulation. This is a key example showing that the inclusion of the first solvation shell can capture effects of water molecules close to protein surface explicitly without increasing the number of replicas required or the complexity of the GB calculation.

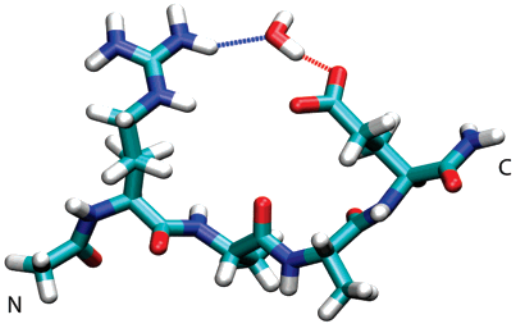


Figure 4. A sample structure snapshot from hybrid solvent REMD showing a water bridging the ion pair that results in a solvent separated minimum in the PMF curve in Figure 2. Other water molecules are present but not shown.

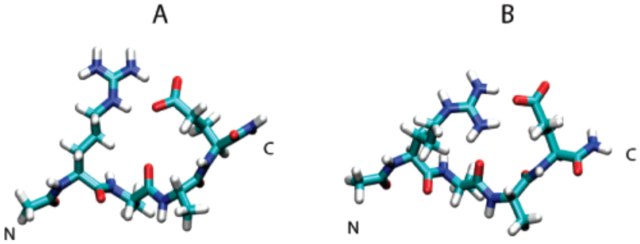


Figure 5. Representative structures of the most populated salt bridge geometries in standard REMD simulations using (A) TIP3P and (B) GB^{OBC}.

The free energy profiles for both restrained and unrestrained REMD simulations show about a 0.5 Å longer salt bridge distance in the GB ensemble as compared to that sampled in either standard or hybrid solvent REMD (Figures 2 and 3). The difference in both restrained and unrestrained backbone ensembles suggests that this shift to longer distances in GB REMD is a direct effect of GB and not a consequence of different salt bridge geometries with different backbone conformations. To investigate differences in geometry arising from GB, conformations with salt bridging present (salt bridge distance <5.5 Å) were subjected to cluster analysis for the heavy atoms participating in salt bridge formation in the backbone-restrained ensembles. Ten conformational clusters were obtained, two of which showed significant populations for all solvation schemes. The most populated cluster for the TIP3P and GB^{OBC} REMD simulations differed; the representative structure for these two clusters is shown in Figure 5.

Figure 5 shows the representative conformations for the most populated clusters. The first cluster (Figure 5A) is the most populated salt bridge conformation for the TIP3P simulations; about 65% of the structures that show a salt bridge adopt this conformation. The hybrid solvent REMD method is in excellent agreement with the standard TIP3P

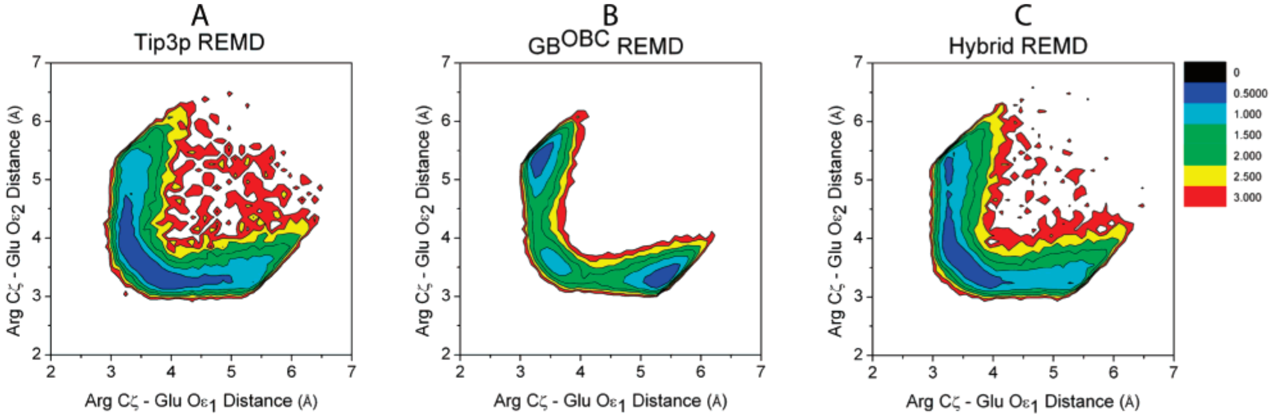


Figure 6. Free energy surfaces describing the geometry of salt bridge formation. The axes show the distance Arg $C\zeta$ and Glu $O\epsilon_1$ versus the distance Arg $C\zeta$ and Glu $O\epsilon_2$. Both standard (A) and hybrid solvent (C) REMD with TIP3P prefer that both oxygens simultaneously hydrogen bond with Arg, resulting in a single free energy minimum near the $x = y$-axis. In the GB simulations (B), there is a strong preference for a single oxygen to hydrogen bond to Arg, while the other remains solvent exposed, resulting a free energy minimum with one short and one long distance. Due to the symmetry of the carboxyl group, two such minima are present.

ensemble, with 52% of the salt bridging structures adopting this geometry. This preferred conformation is also the most populated cluster observed for Arg-Glu pairs in proteins, according to the Atlas of Protein Sidechain Interactions.[65]

In contrast, this geometry is only weakly sampled in the GB^{OBC} ensemble (8% of the salt bridging structures). Instead, the GB^{OBC} REMD simulations prefer the orientation shown in Figure 5B; over 70% of the structures adopt this alternate conformation in which the Arg side chain shows an 180° flip around the χ_4 dihedral angle as compared to the structure preferred in TIP3P. There is also a change in the Glu side chain, where one carboxyl oxygen faces out toward the (implicit) solvent and the other adopts a bifurcated hydrogen bond with Arg hydrogens. This GB-favored geometry between Arg and Glu side chains is not present among the top 6 orientations that are reported in the Atlas of Protein Sidechain Interactions. This conformation is only 23% populated in TIP3P simulations and about 40% in hybrid solvent REMD.

In order to further characterize the change in salt bridge geometry between simulations that do and do not include explicit water at the salt bridge interface, we investigated the hydrogen bond orientation between Arg and Glu side chains by calculating the 2-dimensional free energy profiles for the distances between Arg $C\zeta$ and the two Glu oxygens ($O\epsilon1$ $O\epsilon2$) for salt bridging conformations. In accord with cluster analysis results, the standard and hybrid solvent REMD simulations with TIP3P prefer that both Glu oxygens simultaneously have hydrogen bonds with the Arg. This is seen as one broad minimum in free energy where both oxygens adopt a comparable distance from the Arg $C\zeta$ (Figure 6A and C). However, GB^{OBC} prefers bifurcated hydrogen bonds between Arg and a single Glu oxygen, resulting in a preference for conformations with one of the Arg $C\zeta$ to Glu O distances longer than the other. Due to the symmetry of the Glu carboxyl group, this preference manifests as two free energy minima on the surface, as shown in Figure 6B. A small residual preference of less than 0.5 kcal/mol for the off-diagonal minimum remains with the hybrid solvent approach, although this is within the range of uncertainty in our data since only one of the two symmetric GB-like minima is comparable in free energy to the global free energy minimum. Overall, the hybrid solvent REMD surface is in much better agreement with that from TIP3P REMD than the GB surface. To further investigate this issue we compared the Arg χ_4 dihedral angle distributions between all solvent models. TIP3P and hybrid models show similar distributions in which multiple angles are nearly equally sampled. In contrast, the GB simulation shows a significantly greater preference for adopting a single rotamer (Figure S3).

The combined effect of different Arg−Glu geometries and hydrogen bond preference increases the salt bridge distance in GB ensembles by about 0.5 Å compared to those in standard or hybrid solvent REMD using TIP3P, which results in the difference in locations of free energy minima observed in the 1-dimensional PMF curves in Figures 1 and 3.

Further Testing Using the HP-1 Peptide. After successful tests with alanine peptides and the model peptide we performed a new set of simulations on a more complex peptide. The N-terminal helix of the villin headpiece helical subdomain HP36 (HP-1) is a 13 residue fragment with several charged residues that can form salt bridges. Previously we studied the structure of this fragment through converged explicit solvent REMD simulations which demonstrated that the fragment adopts a nativelike conformation in isolation.[46] In order to evaluate the general nature of observations for the short peptide described above, we performed pure GB^{OBC} and hybrid solvent REMD simulations for HP-1 and compared the results to our previous explicit solvent simulations. Figure 7 summarizes the comparison between each solvation scheme.

First we compared the differences in backbone conformations between the solvation schemes. We calculated the average helicity through DSSP analysis and compared the helical content of the fragment at each temperature (Figure 7A). Explicit solvent REMD shows a maximum ~27% helical content near 300 K. However the GB^{OBC} REMD

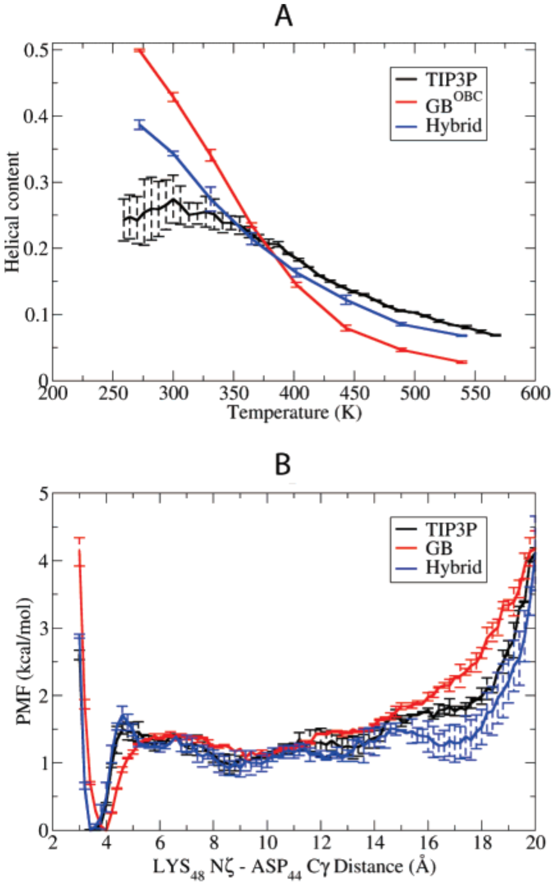


Figure 7. Melting profile (A) and salt bridge PMF (B) for the HP-1 fragment. The melting curve was calculated through average helicity determined through DSSP. At lower temperatures GB^{OBC} shows higher helical content than explicit solvent, while at higher temperatures GB^{OBC} underestimates the helical content. At higher temperatures hybrid solvent REMD is in excellent agreement with standard explicit solvent REMD; at lower temperatures the helical propensity tends to increase as is seen with GB^{OBC}, although the uncertainties in the explicit solvent data below 325 K make it difficult to evaluate this trend. The lower figure shows the PMF for salt bridge formation at 365 K. All solvent models show similar free energy profiles, but the hybrid solvent REMD data are in better agreement with the standard explicit solvent PMF, including the solvent separated minimum. GB^{OBC} also shows the same incorrect location of global minimum as was observed for the model peptide (Figures 1 and 3). This error is corrected in the hybrid model.

simulations show very high helical content at low temperatures (~45% at 300 K) as well as a significantly stronger temperature dependence of the helix population with more helix at low temperature and less helix at high temperature as compared to explicit solvent. Hybrid solvent REMD shows helical populations much closer to explicit solvent at lower temperatures, and the profile follows the explicit solvent profile very closely at higher temperatures. This shows that the hybrid REMD scheme is successful in reducing the backbone conformational bias of GB methods on more complicated sequences. The larger uncertainties in the standard REMD TIP3P data as compared to hybrid solvent REMD may arise from the use of more replicas in the standard REMD simulations and the longer time that it therefore requires for replicas to traverse the entire temperature range. However, even in the standard REMD simulations the largest uncertainties are only ±5%.

To reduce the effect of different backbone geometries we compared the PMFs for the salt bridge between Nζ of Lys48 and Cγ of Asp44 at 365 K (Figure 7B), since this temperature showed the closest agreement in helical content in the thermal profiles (Figure 7A). All solvent models show similar salt bridge free energy profile in terms of relative stability, but GB^{OBC} is again unable to reproduce the position of the global minimum and other details of the profile. The hybrid solvent REMD curve, however, is in very good agreement with standard explicit solvent REMD, successfully identifying the global minimum as well as the solvent separated minimum. Analysis of PMF curves at other temperatures shows similar trends (data not shown).

Conclusions

We studied the performance of the recently developed hybrid solvent REMD method on charged side chains, which have been shown to be problematic with GB implicit solvent models. We used a 4 residue peptide with charged Arg and Glu side chains separated by 2 Ala residues. Standard REMD simulations were performed with TIP3P explicit solvent and also with GB^{OBC} implicit solvent. Similar calculations were performed using hybrid solvent REMD using the same simulation system as the explicit water REMD, with the exception that REMD exchange probabilities were calculated using the 75 closest water molecules along with the GB^{OBC} model. The PMFs of salt bridge distances were calculated and compared for each approach. The GB^{OBC} model showed larger free energies of salt bridge formation compared to full explicit solvent REMD, indicating overstabilized salt bridges as previously observed. The GB simulations also show the same helical bias that we previously reported based on polyalanine simulations.[39,63] In addition to energetics, the GB simulations were unable to reproduce the correct salt bridge geometry; with ion pair and hydrogen bonding orientation is significantly different in GB-REMD simulations compared to TIP3P. Use of the hybrid solvation model in the exchange calculation significantly reduced the computational cost of REMD, while providing backbone conformational sampling, free energy profiles, and salt bridge geometries that were in excellent agreement with data from standard REMD in explicit solvent. Similar improvements in geometry and salt bridge PMFs were observed for the HP-1 model peptide.

As seen from salt bridge PMFs, the hybrid solvent REMD scheme performs very well with charged side chains. In both cases hybrid solvent REMD improves the PMFs with respect to TIP3P REMD simulations and nearly eliminates overstabilization introduced by GB models. However, a slight (0.5−1.0 kcal/mol) bias favoring salt bridge conformations remains in the hybrid model. Likewise, the 2-dimensional free energy surfaces describing salt bridge geometries also suggests a very small residual effect from the GB model, which was only used during the exchange calculation. Further studies will address whether even better agreement between standard explicit solvent and hybrid solvent REMD can be obtained by improving the GB solvent model used in the hybrid approach or by replacing it with a more accurate method such as implicit solvent models based on the Poisson equation. Due to the infrequent need to calculate the hybrid solvent energy (every 500 MD steps in the present case), the additional computational overhead of more accurate implicit solvent models would be expected to have very little impact on the cost of the REMD simulation and still provide very significant savings as compared to standard REMD simulations in explicit solvent (a factor of 6 in the present case).

Acknowledgment. C.S. gratefully acknowledges financial support from the NIH (GM6167803) and supercomputer time at NCSA (NPACI MCA02N028).

Supporting Information Available: Temperature distributions and observed exchange ratios for all REMD simulations, table of populations and average errors shown in Figure 2, distributions of the number of water molecules within the first solvent shell at 300 K, distributions of the Arg Cζ−Glu Cδ salt bridge distance for 8 replicas for TIP3P REMD simulation, and distribution of Arg χ_4 dihedral angle for all solvent models. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Tai, K. Conformational sampling for the impatient. *Biophys. Chem.* **2004**, *107* (3), 213−220.

(2) Roitberg, A.; Simmerling, C. Special issue: Conformational sampling. *J. Mol. Graphics Modell.* **2004**, *22* (5), 317−317.

(3) Smith, L. J.; Daura, X.; van Gunsteren, W. F. Assessing equilibration and convergence in biomolecular simulations. *Proteins: Struct., Funct., Genet.* **2002**, *48* (3), 487−496.

(4) Hansmann, U. H. E. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.* **1997**, *281* (1−3), 140−150.

(5) Swendsen, R. H.; Wang, J. S. Replica Monte-Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.* **1986**, *57* (21), 2607−2609.

(6) Tesi, M. C.; vanRensburg, E. J. J.; Orlandini, E.; Whittington, S. G. Monte Carlo study of the interacting self-avoiding walk model in three dimensions. *J. Stat. Phys.* **1996**, *82* (1−2), 155−181.

(7) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314* (1−2), 141−151.

(8) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21* (6), 1087−1092.

(9) Feig, M.; Karanicolas, J.; Brooks, C. L. MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graphics Modell.* **2004**, *22* (5), 377−395.

(10) Garcia, A. E.; Sanbonmatsu, K. Y. Exploring the energy landscape of a beta hairpin in explicit solvent. *Proteins: Struct., Funct., Genet.* **2001**, *42* (3), 345−354.

(11) Garcia, A. E.; Sanbonmatsu, K. Y. alpha-Helical stabilization by side chain shielding of backbone hydrogen bonds. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (5), 2782−2787.

(12) Karanicolas, J.; Brooks, C. L. The structural basis for biphasic kinetics in the folding of the WW domain from a formin-binding protein: Lessons for protein design? *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (7), 3954−3959.

(13) Kinnear, B. S.; Jarrold, M. F.; Hansmann, U. H. E. All-atom generalized-ensemble simulations of small proteins. *J. Mol. Graphics Modell.* **2004**, *22* (5), 397−403.

(14) Pitera, J. W.; Swope, W. Understanding folding and design: Replica-exchange simulations of "Trp-cage" miniproteins. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (13), 7587−7592.

(15) Roe, D. R.; Hornak, V.; Simmerling, C. Folding cooperativity in a three-stranded beta-sheet model. *J. Mol. Biol.* **2005**, *352* (2), 370−381.

(16) Sugita, Y.; Kitao, A.; Okamoto, Y. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.* **2000**, *113* (15), 6042−6051.

(17) Zhou, R. H.; Berne, B. J.; Germain, R. The free energy landscape for beta hairpin folding in explicit water. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (26), 14931−14936.

(18) Baumketner, A.; Shea, J. E. The structure of the Alzheimer amyloid beta 10−35 peptide probed through replica-exchange molecular dynamics simulations in explicit solvent. *J. Mol. Biol.* **2007**, *366* (1), 275−285.

(19) Paschek, D.; Nymeyer, H.; Garcia, A. E. Replica exchange simulation of reversible folding/unfolding of the Trp-cage miniprotein in explicit solvent: On the structure and possible role of internal water. *J. Struct. Biol.* **2007**, *157* (3), 524−533.

(20) Periole, X.; Mark, A. E. Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent. *J. Chem. Phys.* **2007**, *126* (1).

(21) Cheng, X.; Cui, G.; Hornak, V.; Simmerling, C. Modified Replica Exchange Simulation Methods for Local Structure Refinement. *J. Phys. Chem. B* **2005**, *109* (16), 8220−8230.

(22) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.* **2002**, *116* (20), 9058−9067.

(23) Kofke, D. A. On the acceptance probability of replica-exchange Monte Carlo trials. *J. Chem. Phys.* **2002**, *117* (15), 6911−6914.

(24) Rathore, N.; Chopra, M.; de Pablo, J. J. Optimal allocation of replicas in parallel tempering simulations. *J. Chem. Phys.* **2005**, *122* (2), 024111.

(25) Jang, S. M.; Shin, S.; Pak, Y. Replica-exchange method using the generalized effective potential. *Phys. Rev. Lett.* **2003**, *91* (5), 058305.

(26) Mitsutake, A.; Sugita, Y.; Okamoto, Y. Replica-exchange multicanonical and multicanonical replica-exchange Monte Carlo simulations of peptides. I. Formulation and benchmark test. *J. Chem. Phys.* **2003**, *118* (14), 6664−6675.

(27) Sugita, Y.; Okamoto, Y. Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. *Chem. Phys. Lett.* **2000**, *329* (3−4), 261−270.

(28) Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C. Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model. *J. Chem. Theory Comput.* **2006**, *2* (2), 420−433.

(29) Okur, A.; Roe, D. R.; Cui, G. L.; Hornak, V.; Simmerling, C. Improving convergence of replica-exchange simulations through coupling to a high-temperature structure reservoir. *J. Chem. Theory Comput.* **2007**, *3* (2), 557−568.

(30) Roitberg, A. E.; Okur, A.; Simmerling, C. Coupling of replica exchange simulations to a non-Boltzmann structure reservoir. *J. Phys. Chem. B* **2007**, *111* (10), 2415−2418.

(31) Affentranger, R.; Tavernelli, I.; Di Iorio, E. E. A novel Hamiltonian replica exchange MD protocol to enhance protein conformational space sampling. *J. Chem. Theory Comput.* **2006**, *2* (2), 217−228.

(32) Lyman, E.; Zuckerman, D. M. Resolution exchange simulation with incremental coarsening. *J. Chem. Theory Comput.* **2006**, *2* (3), 656−666.

(33) Li, H.; Li, G.; Berg, B. A.; Yang, W. Finite reservoir replica exchange to enhance canonical sampling in rugged energy surfaces. *J. Chem. Phys.* **2006**, *125* (14), 144902.

(34) Li, H.; Yang, W. Sampling enhancement for the quantum mechanical potential based molecular dynamics simulations: a general algorithm and its extension for free energy calculation on rugged energy surface. *J. Chem. Phys.* **2007**, *126* (11), 114104.

(35) Min, D.; Li, H.; Li, G.; Bitetti-Putzer, R.; Yang, W. Synergistic approach to improve "alchemical" free energy calculation in rugged energy surface. *J. Chem. Phys.* **2007**, *126* (14), 144109.

(36) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112* (16), 6127−6129.

(37) Nymeyer, H.; Garcia, A. E. Simulation of the folding equilibrium of alpha-helical peptides: A comparison of the generalized born approximation with explicit solvent. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (24), 13934−13939.

(38) Zhou, R. H.; Berne, B. J. Can a continuum solvent model reproduce the free energy landscape of a beta-hairpin folding in water? *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (20), 12777−12782.

(39) Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C. Secondary structure bias in generalized born solvent models: Comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation. *J. Phys. Chem. B* **2007**, *111* (7), 1846−1857.

(40) Simmerling, C.; Strockbine, B.; Roitberg, A. E. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **2002**, *124* (38), 11258−11259.

(41) Zhou, R. H. Free energy landscape of protein folding in water: Explicit vs. implicit solvent. *Proteins: Struct., Funct., Genet.* **2003**, *53* (2), 148−161.

(42) Ghosh, A.; Rapp, C. S.; Friesner, R. A. Generalized born model based on a surface integral formulation. *J. Phys. Chem. B* **1998**, *102* (52), 10983−10990.

(43) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (39), 13749−13754.

(44) Huang, X.; Hagen, M.; Kim, B.; Friesner, R. A.; Zhou, R.; Berne, B. J. Replica Exchange with Solute Tempering: Efficiency in Large Scale Systems. *J. Phys. Chem. B* **2007**, *111* (19), 5405−5410.

(45) Geney, R.; Layten, M.; Gomperts, R.; Hornak, V.; Simmerling, C. Investigation of Salt Bridge Stability in a Generalized Born Solvent Model. *J. Chem. Theory Comput.* **2006**, *2* (1), 115−127.

(46) Wickstrom, L.; Okur, A.; Song, K.; Hornak, V.; Raleigh, D. P.; Simmerling, C. L. The Unfolded State of the Villin Headpiece Helical Subdomain: Computational Studies of the Role of Locally Stabilized Structure. *J. Mol. Biol.* **2006**, *360* (5), 1094−1107.

(47) 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: Struct., Funct., Bioinformatics* **2006**, *65* (3), 712−725.

(48) 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 2Nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179−5197.

(49) Wang, J. M.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **2000**, *21* (12), 1049−1074.

(50) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926−935.

(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. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26* (16), 1668−1688.

(52) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327−341.

(53) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular-Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684−3690.

(54) Weber, W.; Hunenberger, P. H.; McCammon, J. A. Molecular dynamics simulations of a polyalanine octapeptide under Ewald boundary conditions: Influence of artificial periodicity on peptide conformation. *J. Phys. Chem. B* **2000**, *104* (15), 3668−3675.

(55) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089−10092.

(56) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Struct., Funct., Bioinformatics* **2004**, *55* (2), 383−394.

(57) Bondi, A. Van Der Waals Volumes + Radii. *J. Phys. Chem.* **1964**, *68* (3), 441−&.

(58) Tsui, V.; Case, D. A. Molecular dynamics simulations of nucleic acids with a generalized born solvation model. *J. Am. Chem. Soc.* **2000**, *122* (11), 2489−2498.

(59) Ponder, J. W.; Richards, F. M. An Efficient Newton-Like Method for Molecular Mechanics Energy Minimization of Large Molecules. *J. Comput. Chem.* **1987**, *8* (7), 1016−1024.

(60) Simmerling, C.; Elber, R.; Zhang, J. MOIL-View - A Program for Visualization of Structure and Dynamics of Biomolecules and STO - A Program for Computing Stochastic Paths. In *Modelling of Biomolecular Structures and Mechanisms*; Pullman et al., A., Ed.; Kluwer Academic Publishers: The Netherlands, 1995; pp 241−265.

(61) McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. NMR structure of the 35-residue villin headpiece subdomain. *Nat. Struct. Biol.* **1997**, *4* (3), 180−4.

(62) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22* (12), 2577−2637.

(63) Okur, A.; Simmerling, C. Hybrid Implicit/Explicit Solvation Methods. In *Annual Reports in Computational Chemistry*; Elsevier: 2006; Vol. 2, pp 97−109.

(64) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Pairwise Solute Descreening of Solute Charges from a Dielectric Medium. *Chem. Phys. Lett.* **1995**, *246* (1−2), 122−129.

(65) Singh, J.; Thornton, J. M. In *Atlas of Protein Side-Chain Interactions*; IRL Press: Oxford, 1992; Vol. I & II.

CT7002308

JCTC Journal of Chemical Theory and Computation

Computational Electrochemistry of Ruthenium Anticancer Agents. Unprecedented Benchmarking of Implicit Solvation Methods[†]

Ion Chiorescu,[‡] Dirk V. Deubel,*[,‡,§] Vladimir B. Arion,[‡] and Bernhard K. Keppler[‡]

Institute of Inorganic Chemistry, Faculty of Chemistry, University of Vienna, A-1090 Vienna, Austria, and Laboratory of Physical Chemistry, D-CHAB, ETH Zurich, CH-8093 Zurich, Switzerland

Received September 24, 2007

Abstract: Two ruthenium(III) complexes {(HIm)[*trans*-$RuCl_4$(DMSO)(Im)] (NAMI-A) and (HInd)-[*trans*-$RuCl_4(Ind)_2$] (KP1019), DMSO = dimethyl sulfoxide, Im = imidazole, Ind = indazole} have been tested in phase I clinical trials as potential anticancer drugs. Ru(III) anticancer agents are likely activated in vivo upon reduction to their Ru(II) analogs. Aiming at benchmarking implicit solvation methods in DFT studies of ruthenium pharmaceuticals at the B3LYP level, we have calculated the standard redox potentials (SRPs) of Ru(III/II) pairs that were electrochemically characterized in the literature. 80 SRP values in four solvents were calculated using three implicit solvation methods and five solute cavities of molecular shape. Comparison with experimental data revealed substantial errors in some of the combinations of solvation method and solute cavity. For example, the overall mean unsigned error (MUE) with the PCM/UA0 combination, which is the popular default in Gaussian 03, amounts to 0.23 V (5.4 kcal/mol). The MUE with the CPCM/UAKS combination, which was employed by others for recent computational studies on the hydrolysis of NAMI-A and *trans*-$[RuCl_4(Im)_2]^-$, amounts to 0.30 V (7.0 kcal/mol) for all compounds and to 0.60 V (13.9 kcal/mol) for a subset of compounds of the medicinally relevant type, *trans*-$[RuCl_4(L)(L')]^-$. The SRPs calculated with the PCM or CPCM methods in Gaussian 03 can be significantly improved by a more compact solute cavity constructed with Bondi's set of atomic radii. Earlier findings that CPCM performs better than PCM cannot be confirmed, as the overall MUE amounts to 0.19 V (4.3−4.4 kcal/mol) for both methods in combination with Bondi's set of radii. The Poisson−Boltzmann finite element method (PBF) implemented in Jaguar 7 together with the default cavity performs slightly better, with the overall MUE being 0.16 V (3.7 kcal/mol). Because the redox pairs considered in this study bear molecular charges from +3/+2 to −1/−2 and the prediction of solvation free energies is most challenging for highly charged species, the present work can serve as a general benchmarking of the implicit solvation methods.

Introduction

Due to the success of cisplatin as an anticancer drug,[1] the search for new metallopharmaceuticals has continued and extended to non-platinum complexes,[2] most notably ruthenium[3] and rhodium.[4] Two Ru(III) complexes, NAMI-A[5] and KP1019,[6] with the common lead structure, *trans*-$[RuCl_4(L)(L')]^-$ (Figure 1), successfully completed phase I clinical trials.[7] Recent developments of organometallic Ru(II) anticancer complexes are promising as well.[8,9] Ru(III) complexes are believed to be activated in vivo upon reduction to their Ru(II) analogs.[10] A selective reduction in cancer cells occurs likely due to the reducing environment caused by deficiency of molecular oxygen in tumors (*hypoxia*),[11] as the blood flow to the rapidly growing tumor is insufficient. The standard redox potential (SRP) of a Ru(III) complex is believed to be crucial to its anticancer activity.[12] If it is too low, then

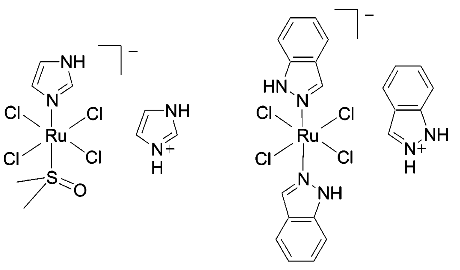
[†] Quantum Chemical Studies of Metals in Medicine, IX.

* Corresponding author e-mail: metals-in-medicine@phys.chem.ethz.ch.

[‡] University of Vienna.

[§] ETH Zurich.

10.1021/ct700247g CCC: $40.75
Published on Web 02/02/2008



NAMI-A KP1019 (FCC14a)

Figure 1. NAMI-A and KP1019 (FCC14a), two Ru(III) anticancer complexes in clinical trials.

the complex is not reduced and remains inactive. If it is too high, then the complex is reduced too easily and the selectivity to cancer cells is lost.

The prediction of SRPs can rationalize and accelerate the search for new drugs, because an in silico screening can be performed before relevant compounds are selected, synthesized, and tested in vitro and in vivo. So far, SRPs were often predicted using an empirical increment system[13] derived from experimentally measured SRPs. This empirical approach is particularly successful if strongly related complexes with the same stereochemical arrangement of ligands are compared to one other. Note that geometric isomers (e.g., *cis* vs *trans*; *mer* vs *fac*) require different parameter sets. A more general approach for the prediction of SRPs is based on modern quantum chemical calculations.[14] The accurate prediction of solvation free energies poses a true challenge in these calculations. The objective of the present work is benchmarking various implicit solvation methods[15] and molecule-shaped solute cavities employed for density functional theory (DFT) studies at the B3LYP[16,17] level.

Recent computational studies of other groups are highly complementary to the present work. *First*, benchmarking studies were carried out to assess the performance of different implicit solvation protocols for organic and main group element compounds.[18,19] *Second*, computation of SRPs of group 8 metal complexes mostly in organic solvents demonstrated that the approach is valid for a wide range of SRPs.[20] *Third*, the SRP of the aqueous $Ru^{3+}_{aq}/Ru^{2+}_{aq}$ pair was calculated using a variety of quantum chemical methods, effective core potentials (ECPs), and basis sets, implicit solvent models, and cavities.[21] The influence of the first and second solvation shells on the SRP of the $Ru^{3+}_{aq}/Ru^{2+}_{aq}$ pair was investigated as well.[21] DFT studies on Ru(III) anticancer agents were recently reported, but validation of the method by predicting their SRPs was not taken into account.[22,23] In the present work, we consider 80 SRPs of 61 ruthenium complexes in four solvents.[24,25] While various types of ruthenium complexes are included, an emphasis is placed on anticancer complexes in aqueous solution, the SRPs of which were measured recently[25−27] and fall into a fine biologically relevant window (from −0.4 to +0.8 V vs NHE).[28] We believe that the results of this benchmarking work will improve the accuracy and credibility of future computational studies of ruthenium pharmaceuticals and other metal complexes.

Methods

The geometries of the molecules were optimized at the gradient-corrected DFT level using the 3-parameter fit of exchange and correlation functionals of Becke[16] (B3LYP), which includes the correlation functional of Lee, Yang, and Parr (LYP),[17] as implemented in Gaussian 03.[29] The Stuttgart-Dresden scalar-relativistic energy-consistent small-core ECPs and the corresponding valence-basis sets were used for the Ru (MWB28 ECP together with an (8s7p6d)[6s5p3d] basis set)[30] and I atoms (MWB46 ECP together with a (4s5p)-[2s3p] basis set),[31] and the 6-31G(d,p) basis sets were used for the other atoms.[32] This basis-set combination is denoted M. Vibrational frequencies were also calculated at B3LYP/M; all structures reported herein are minima on the potential energy surfaces. Improved total energies were calculated at the B3LYP level using for Ru the MWB28 ECP[30] and the valence-basis set augmented with two sets of f functions and one set of g functions to obtain an (8s7p6d2f1g)[6s5p3d2f1g] basis set,[33] using for I the MWB46 ECP[31] together with an (14s10p3d1f)[3s3p2d1f] basis set,[33] with the 6-311+G(3d) on S, Cl, and Br atoms, and the 6-311+G(d,p) basis sets at the other atoms. Note that large basis sets are required for obtaining accurate energies of S[34] and I[35] compounds. This basis-set combination is denoted XL. Free energies in vacuo (G^1) were calculated by adding corrections from unscaled zero-point energy (ZPE), thermal energy, work, and entropy evaluated at the B3LYP/M level at 298.15 K, 1 atm to the energies calculated at the B3LYP/XL//M level.

Solvation free energies (ΔG_{solv}) of the structures optimized in vacuo at the B3LYP/M level were calculated using three implicit solvation methods:[15] The Polarizable Continuum model (PCM)[36] in its integral equation formalism (IEF)[37] as implemented in Gaussian 03 and the Conductor-like[38] Polarizable Continuum Model (CPCM)[39] in Gaussian 03[29] and the Poisson−Boltzmann finite element method (PBF)[40] in Jaguar 7.[41] These methods belong to the class of self-consistent reaction field (SCRF) methods. The solute is embedded in a continuum dielectric with a dielectric constant ϵ representing the solvent.[42] The solute charge distribution polarizes the continuum dielectric, and the potential arising from the solvent polarization in turn modifies the solute Hamiltonian. The calculation of the solute wave function is carried out iteratively until self-consistency. The PCM method describes the solvent reaction potential in terms of apparent surface charges localized on tesserae at the continuum boundary. The CPCM method uses apparent surface charges as well but describes the solvent first as a conductor ($\epsilon = \infty$) and then rescales the charges for a finite value of the dielectric constant. The PBF method uses finite elements for numerical solution of the Poisson−Boltzmann differential equation. The PCM and CPCM calculations were done at the B3LYP/M level, while the PBF calculations were done at the B3LYP/LACVP** level, which includes a scalar-relativistic energy-consistent small-core ECP[43] and basis set at the metal and the 6-31G(d,p) basis sets at the other atoms.

The three solvation methods were used together with solute cavities of molecular shape; the PCM and CPCM methods together with the cavities based on the UA0, UAHF, UAKS,

Table 1. Radii (Å) Used for Constructing the Solute Cavities Considered in This Work

	UA0	UAHF	UAKS	Bondi	PBF
H				1.200	1.150
C				1.700	1.900
C[e]	1.925	1.725	1.725		
CH[e]	2.125	1.905	1.905		
CH_2[e]	2.325	2.193	2.193		
CH_3[e]	2.525	1.950	1.950		
N				1.550	1.600
N[a]	1.830	1.551	1.461		
NH[a]	1.930	1.641	1.551		
NH_2[c]	2.030	1.680	1.680		
NH_3[b]	2.130	1.770	1.770		
O				1.520	1.600
O[a]	1.750	1.569	1.479		
OH[d]	1.780	1.590	1.590		
OH_2[b]	1.950	1.569	1.569		
S	2.017[a]	1.959[a]	1.959[a]	1.800	1.900
Cl	1.973[a]	1.959[a]	1.959[a]	1.750	1.974
Br	2.094[c]	2.080[c]	2.080[c]	1.850	2.095
Ru	1.482	1.482	1.482	1.482	1.481
I	2.250[f]	2.350[f]	2.350[f]	1.980	2.250

[a] In *trans*-$[RuCl_4(DMSO)(Im)]^-$. [b] In *cis*-$[Ru(NH_3)_4(OH_2)_2]^{3+}$. [c] In *trans*-$[Ru(NH_3)_4Br(isonicotinamide)]^{2+}$. [d] In $[Ru(NH_3)_5(OH)]^{2+}$. [e] In *mer,trans*-$[RuCl_3(Et_2S)(Ind)_2]$. [f] In *cis*-$[Ru(NH_3)_4I(py)]^{2+}$.

and Bondi radii and the PBF method in Jaguar 7 together with the default set of radii. In the United Atom (UA) topological models,[44] the cavity is obtained from spheres centered on non-hydrogen atoms. The radius of each sphere depends on the atom type, its connectivity, and the number of hydrogen atoms attached; typical values are listed in Table 1. The UA0 radii are based on the Universal Force Field (UFF).[45] Variants of these radii were optimized at the Hartree−Fock level (UAHF)[46] and the Perdew-Burke-Ernzerhof (PBE)[47] Kohn−Sham DFT level (UAKS).[29] In contrast, Bondi's[48] set of radii consider hydrogen atoms explicitly, and the values depend on atom identity rather than on its connectivity or hybridization. The radii are used in several ways for the construction of cavity in the calculation of the solvation free energy $\Delta G_{solv}{}^{\epsilon}$, which is decomposed into solute−cavity-formation[49] (ΔG_{cav}), electrostatic (ΔG_{elst}), and dispersion and repulsion ($\Delta G_{dis-rep}$) contributions ($\Delta G_{solv} = \Delta G_{elst} + \Delta G_{dis-rep} + \Delta G_{cav}$). In the PCM and CPCM implementations, the nonelectrostatic contributions $\Delta G_{dis-rep}$ and ΔG_{cav} are calculated using a solvent accessible surface (SAS),[50] whereas ΔG_{elst} is calculated using a similar surface constructed by the radii scaled by a factor[51] and additional spheres to smoothen the surface. In the PBF method, ΔG_{elst} is calculated using an SAS constructed by a set of standard radii including explicit hydrogen atoms (Table 1). An empirical formula depending linearly on the SAS area is employed for the remaining terms, $\Delta G_{dis-rep} + \Delta G_{cav}$. Our attempts to further refine this set of radii or change the basis sets indicated that the PBF protocol had been reasonably optimized.

Standard redox potentials (SRPs) were calculated on the basis of the thermodynamic cycle shown in Figure 2,[14,52] which is similar to that proposed for the prediction of acidity constants.[53] The SRP is related to the reaction free energy in solution (ΔG^{ϵ})

$$\begin{tikzcd}
Ox_{vac} \arrow[d, "\Delta G_{solv}"'] & + & e^- \arrow[r, "\Delta G^1"] \arrow[d, "\Delta G=0"] & Red_{vac} \arrow[d, "\Delta G_{solv}"] \\
Ox_{solv} & + & e^- \arrow[r, "\Delta G^{\varepsilon}"] & Red_{solv}
\end{tikzcd}$$

Figure 2. Thermodynamic cycle for calculating redox potentials.

$$SRP = -(\Delta G^{\epsilon}/zF) - \Delta SRP_{NHE}$$

with

$$\Delta G^{\epsilon} = -\Delta G_{solv}(Ox) + \Delta G^1 + \Delta G_{solv}(Red)$$

according to Figure 2. ΔG^1 is the reaction free energy in vacuo, $\Delta G_{solv}(Ox)$ and $\Delta G_{solv}(Red)$ are the solvation free energies in the oxidized and reduced form, $F = 96\,485$ C/mol $= 23.061$ kcal/(Vmol) is the Faraday constant, and $z = 1$ for one-electron reductions. The SRPs are shifted by $\Delta SRP_{NHE} = 4.28$ V to align the results to the scale of the Normal Hydrogen Electrode (NHE).[54]

Results and Discussion

We start with the *default* solvation protocol in Gaussian 03, which consists of the Polarizable Continuum Model (PCM)[28] together with the default definition of the solute cavity on the basis of the united atom topological model UA0.[32] This approach is denoted as Protocol I. Figure 3a shows the calculated (PCM/UA0) vs experimental SRPs. The results are displayed as follows: (i) The *solvent* (ACN, DMF, DMSO, water) is indicated by the color of symbol. (ii) The *molecular charge* of the redox pairs ranging from +3/+2 to −1/−2 is indicated by the shape of symbol. (iii) The *medicinal relevance* of the aqueous SRPs for the complexes of the type *trans*-$[RuCl_4(L)(L')]^-$ is indicated by filled symbols. The two drug candidates in the clinics, KP1019 and NAMI-A, belong to this category (Figure 1). To analyze the results in a quantitative manner, mean unsigned errors (MUE) are listed in Table 2, sorted by solvent, molecular charge, and medicinal relevance. The overall MUE of Protocol I for the complete set of data (all compounds in all solvents) is 0.23 V (5.4 kcal/mol). Similar MUE values are obtained for subsets of complexes in water and medicinally relevant complexes, 0.26 V (6.0 kcal/mol) and 0.29 V (6.8 kcal/mol), respectively.[55]

Protocol II combines the PCM method with the UAHF cavity, as implemented in Gaussian 03. This approach has been frequently used, as it is recommended for the prediction of solvation free energies by comparing solution- and gas-phase free energies.[29] Figure 3b displays the calc. vs exp. SRPs, and Table 2 includes the MUEs. The calculations reveal that this protocol performs worse than Protocol I: The MUE for the complete set of data (all compounds in all four solvents) is 0.32 V (7.4 kcal/mol). Considering the molecular

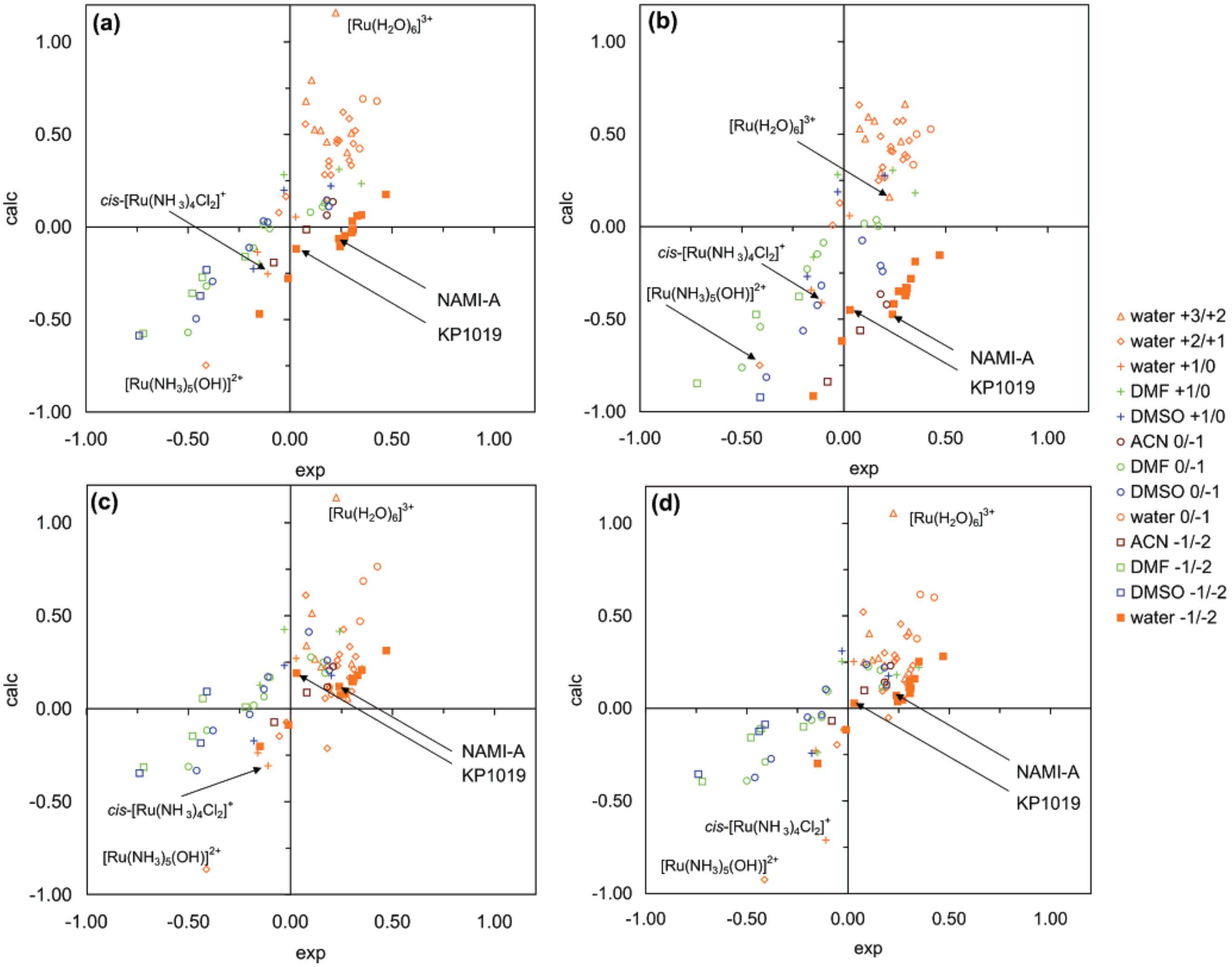


Figure 3. Calculated vs experimental SRPs. (a) Protocol I, PCM/UA0, (b) Protocol II, PCM/UAHF, (c) Protocol IV, PCM/Bondi, (d) Protocol VII, PB. The data are sorted by solvent (color of symbol) and charge (shape of symbol). Aqueous SRPs of the medicinally relevant complexes with the lead structure, *trans*-$[RuCl_4(L)(L')]^-$, are displayed as filled symbols.

Table 2. Mean Unsigned Errors (MUE) in the Calculation of SRPs Using Protocols I−VII[b]

		I	II	III	IV	V	VI	
		PCM	PCM	PCM	PCM	CPCM	CPCM	VII
compds	no.	UA0	UAHF	UAKS	Bondi	UAKS	Bondi	PBF
all	80	0.23	0.32	0.31	0.19	0.30	0.19	0.16
solvent								
ACN	4	0.10	0.64	0.65	0.02	0.64	0.03	0.02
DMF	16	0.10	0.16	0.16	0.25	0.15	0.26	0.16
DMSO	14	0.34	0.38	0.36	0.21	0.35	0.22	0.17
water	46	0.26	0.33	0.31	0.18	0.31	0.17	0.17
charge								
+3/+2	8	0.45	0.30	0.27	0.27	0.27	0.27	0.23
+2/+1	18	0.19	0.18	0.16	0.15	0.16	0.15	0.13
+1/0	10	0.10	0.15	0.13	0.19	0.13	0.19	0.19
0/−1	21	0.26	0.24	0.24	0.18	0.23	0.19	0.11
−1/−2	23	0.23	0.59	0.57	0.20	0.57	0.20	0.19
lead[a]	14	0.29	0.63	0.61	0.14	0.60	0.14	0.16

[a] Aqueous SRPs of the compounds with the lead structure *trans*-$[RuCl_4(L)(L')]^-$. [b] All values are in volt (V).

charges of the redox couples, the most unsatisfactory results are obtained for the highly charged complexes, as the +3/+2 pairs and −1/−2 redox pairs have MUEs of 0.30 V (7.0 kcal/mol) and 0.59 V (13.5 kcal/mol), respectively. An even higher MUE of 0.63 V (14.6 kcal/mol) is obtained for the medicinally relevant subset. Protocol III uses the same PCM method together with the UAKS radii, as implemented in Gaussian 03; this definition of the solute cavity is recommended for DFT calculations.[29] The calculations show only a very minor improvement in comparison with the UAHF approach, as the MUEs are similar to those of Protocol II (Table 2).

To analyze the performance of the methods further and to identify systematic errors, we have also calculated the mean signed errors (MSE, see Table 3). For protocols I−III, the MSEs are positive for *cationic* redox pairs charge (+3/+2), but they are negative for *anionic* redox pairs (−1/−2). The relatively large errors for these highly charged complexes remained undiscovered in the past, because typical benchmarking studies used pK_a values and SRPs involving only species that have a molecular charge between +1 and −1. Because SRPs are derived from the differential free energies of the reduced forms (Figure 2), the MSEs translate to an underestimation of solvation free energies that is strongest for the species bearing the highest (positive or negative)

Table 3. Mean Signed Errors (MSE) in the Calculation of SRPs Using Protocols I−VII[b]

		I	II	III	IV	V	VI	
		PCM	PCM	PCM	PCM	CPCM	CPCM	VII
compds	no.	UA0	UAHF	UAKS	Bondi	UAKS	Bondi	PBF
all	80	0.01	−0.17	−0.17	0.08	−0.17	0.08	0.03
solvent								
ACN	4	−0.10	−0.64	−0.65	−0.01	−0.64	0.00	0.00
DMF	16	0.06	−0.11	−0.11	0.23	−0.10	0.24	0.11
DMSO	14	−0.18	−0.33	−0.33	0.21	−0.33	0.22	0.15
water	46	0.06	−0.09	−0.10	−0.01	−0.10	−0.01	−0.03
charge								
+3/+2	8	0.45	0.29	0.25	0.20	0.25	0.20	0.20
+2/+1	18	0.14	0.14	0.12	−0.07	0.11	−0.07	−0.05
+1/0	10	0.03	−0.01	−0.02	0.10	−0.02	0.10	−0.02
0/−1	21	−0.11	−0.22	−0.20	0.18	−0.19	0.19	0.09
−1/−2	23	−0.15	−0.59	−0.57	0.04	−0.57	0.05	−0.01
lead[a]	14	−0.29	−0.63	−0.61	−0.12	−0.60	−0.11	−0.16

[a] Aqueous SRPs of the compounds with the lead structure *trans*-$[RuCl_4(L)(L')]^-$. [b] All values are in volt (V).

molecular charge q. Bearing the simple Born model[56] for ions in spherical cavities (ΔG_{solv} proportional $-q^2r^{-1}$) in mind, one would expect that a more compact molecular cavity improves the results, because of the reciprocal dependence of the solvation free energy ΔG_{solv} on the sphere radius r. While our attempts to manually adjust the cavity within the united atom approach did not improve the results, we have also considered the PCM model combined with a more compact cavity based on Bondi radii (Table 1); this combination is denoted as Protocol IV. SRPs are plotted in Figure 3c, and the MUE and MSE are listed in Tables 2 and 3. The calculations reveal that Protocol IV significantly improves the agreement with the experimental values. The overall MUE for the complete set of data (all compounds in all solvents) is 0.19 V (4.3 kcal/mol). Remarkably, the MUEs are now relatively similar for all charges of redox pairs from +3/+2 to −1/−2 (Table 2). The MSEs indicate that the systematic errors have been largely eliminated (Table 3). The performance of this protocol is convincing, as is indicated by MUEs of 0.18 V (4.0 kcal/mol) for the complexes in aqueous solution and 0.14 V (3.3 kcal/mol) for the medicinally relevant set (Table 2).

On the basis of pK_a predictions for neutral and monocationic organic molecules, it was convincingly shown[18] that the conductor-like screening approach (CPCM) performs better than the polarizable continuum model (PCM). Without further benchmarking calculations for metal complexes, the CPCM method together with the UAKS cavity was recently employed for a computational studies on the hydrolysis of NAMI-A and *trans*-$[RuCl_4(Im)_2]^-$.[23] Hence, we have included the CPCM/UAKS approach in our benchmarking (Protocol V). The calculations reveal that Protocol V performs as disappointingly as does Protocol II (PCM/UAKS). For example, the MUE for the set of medicinally relevant compounds of the *trans*-$[RuCl_4(L)(L')]^-$ type amounts to 0.60 V (13.9 kcal/mol). To compare the PCM and CPCM performance further, we have also considered the CPCM method and the Bondi radii (Protocol VI); these results are very similar to those of Protocol III (PCM/Bondi). In summary, there is a striking agreement between the PCM and CPCM approaches, which is consistent with a benchmarking study of the SRPs of nitrogen oxides.[19] An appropriate definition of the solute cavity appears to be at least as important as the choice of the solvent model.

We have also employed a finite element method for the numerical solution of the Poisson−Boltzmann equation (PBF) as implemented in Jaguar 7, together with the default set of radii for setting up the solute cavity (Protocol VII). This approach has been used in our series *Quantum chemical studies of metals in medicine*[57] and in other studies[58] of medicinally relevant metal complexes. Figure 3d visualizes the SRPs, and Tables 2 and 3 list the errors. The overall MUE for the complete set of data (all compounds in all solvents) is 0.16 V (3.7 kcal/mol), which is the best value among all solvation protocols considered in this work. The approach performs well for all four solvents and all molecular charges of redox pairs from +3/+2 to −1/−2. The MUE is 0.17 V (4.0 kcal/mol) for the complexes in water and 0.16 V (3.7 kcal/mol) for the medicinally relevant set. The performance of Protocol VII (PB) is very similar to the PCM and CPCM methods with the Bondi radii.

Despite the overall success of the PBF method, we have identified a few cases where this approach predicts unsatisfactory results, as compared to the experimental values. The aqueous SRP of the hydrated Ru ion redox couple, $[Ru(OH_2)_6]^{3+/2+}$, is computationally overestimated by +0.83 V (Figure 3d).[59] In contrast, the SRPs of $[Ru(NH_3)_5(OH)]^{2+/+}$ and *cis*-$[Ru(NH_3)_4Cl_2]^{+/0}$ are computationally underestimated by −0.52 V and −0.60 V, respectively (Figure 3d). The other protocols perform better for the latter compound (see Figure 3 and Table S-3), but all protocols using the Bondi and UA0 cavities overestimate the SRP of $[Ru(OH_2)_6]^{3+/2+}$ by +0.91−0.93 V. The fact that the aqueous Ru(III/II) redox couple poses a challenge to computational chemistry was also pointed out in a recent article that focuses entirely on this redox pair.[21] These authors made extensive use of their computationally efficient solvation model termed SM6[60] and predicted at B3LYP a SRP that is +0.77 V higher than the experimental value. As the $[Ru(OH_2)_6]^{3+}$ is highly charged and contains six aqua ligands, the predicted SRP depends strongly on the radius employed for the oxygen and hydrogen atoms. We find that shrinking the H atom radius or H and O atom radii in the PBF method to 1.02 Å or 1.02 and 1.52 Å, respectively, reduces the error in the calculated SRP to +0.46 V or +0.39 V but increases the errors for the other two problem cases. The surface charge density of $[Ru(OH_2)_6]^{3+}$ can be lowered by the second hydration shell, which was represented in the recent study[21] by a symmetric arrangement of 12 additional water molecules. Remarkably, this approach led at B3LYP to a SRP that is −0.23 V lower than the experimental value.

Finally, we would like to emphasize that there is some arbitrariness in the prediction of absolute SRPs because of the choice of ΔSRP_{NHE}. Several studies[19,20] used an electrochemical estimate of the absolute ΔSRP_{NHE} of 4.43 V.[61] Alternatively, ΔSRP_{NHE} may be calculated using a thermodynamic cycle for the reaction, $^1/_2H_2 \rightarrow H^+ + e^-$, analogue to that in Figure 2: ΔSRP_{NHE} is the sum of the gas-phase free energy ΔG^1_{NHE} of this reaction and the solvation free energy of H^+, $\Delta G_{solv}(H^+)$. The authors of ref 54 chose $\Delta G^1_{NHE} = 362.59$ kcal/mol obtained from thermochemical data[62] and $\Delta G_{solv}(H^+) = -263.98$ kcal/mol extrapolated from cluster ion solvation data[63] to arrive at $\Delta SRP_{NHE} = 4.28$ V,[54] which is the value employed for the present work. Considering a former experimental $\Delta G_{solv}(H^+)$ value of -259.5 kcal/mol[64] determined electrochemically led to $\Delta SRP_{NHE} = 4.44$ V.[65,66] Because of this uncertainty, one may simply treat ΔSRP_{NHE} as a free parameter to be obtained in a fitting procedure. Given that the mean signed error (MSE) of the PBF method (Protocol VII) amounts to only 0.03 V (Table 3), however, we believe that $\Delta SRP_{NHE} = 4.28$ V suggested in ref 54 is an excellent choice.

In conclusion, we recommend the Poisson−Boltzmann finite element solver (PBF) implemented in Jaguar (Protocol VII) and the PCM or CPCM method together with the Bondi radii implemented in Gaussian (Protocols IV and VI) for future studies. However, caution is advised if the solute is highly charged and contains many hydrogen bond donors in the first shell. Our recommendation is not limited to ruthenium complexes, as the metal center in the compounds considered in this work is surrounded by an octahedral ligand environment and not directly exposed to solvent. In addition, the wide range of the molecular charges of the redox pairs from +3/+2 to −1/−2 has made the calculation of solvation free energies very challenging. Consideration of highly charged species has led to the identification of systematic errors in the solvation free energies that were difficult to find in previous benchmarking studies involving compounds with molecular charges from +1 to −1. Hence, the results of the present work may serve as an unprecedented benchmarking of implicit solvation methods for density functional studies of the reactions of metal complexes involved in catalysis, biology, and medicine.

Acknowledgment. We thank the Gaussian and Jaguar developers for helpful discussions, the Austrian Science Foundation (FWF) and the Swiss National Science Foundation (SNF) for financial support, and the Swiss National Computing Center and the Schroedinger Computing Center at the University of Vienna for computing time.

Supporting Information Available: Mean unsigned errors (MUE) and mean signed errors (MSE) of calculated SRPs, in kcal/mol, calc. vs exp. SRPs of individual compounds, and plot of calc. vs exp. SRPs for Protocols III, V, and VII. This material is available free of charge via http://pubs.acs.org.

References

(1) Jung, Y.; Lippard, S. J. *Chem. Rev.* **2007**, *107*, 1387.

(2) Clarke, M. J.; Zhu, F.; Frasca, D. *Chem. Rev.* **1999**, *99*, 2511.

(3) Clarke, M. J. *Coord. Chem. Rev.* **2003**, *236*, 209.

(4) Chifotides, H. T.; Dunbar, K. R. *Acc. Chem. Res.* **2005**, *38*, 146.

(5) (a) Sava, G.; Capozzi, I.; Clerici, K.; Gagliardi, R.; Alessio, E.; Mestroni, G. *Clin. Exp. Metastasis* **1998**, *16*, 371. (b) Mestroni, G.; Alessio, E.; Sava, G.; Pacor, S.; Coluccia, M. In *Metal Complexes in Cancer Chemotherapy*. Keppler, B. K., Ed.; VCH: Weinheim, Germany, 1993, p. 157. (c) Sava, G.; Alessio, E.; Bergamo, A.; Mestroni, G. *Top. Biol. Inorg. Chem.* **1999**, *1*, 143. (d) Alessio, E.; Mestroni, G.; Bergamo, A.; Sava, G. In *Metal Ions in Biological Systems*. Sigel, A.; Sigel, H., Ed.; Marcel Dekker: New York, 2004, Vol. 42, p 323.

(6) (a) Keppler, B. K.; Rupp, W.; Juh, U. M.; Enders, H.; Niebl, R.; Balzer, W. *Inorg. Chem.* **1987**, *26*, 4366. (b) Keppler, B. K.; Berger, M. R.; Heim, M. H. *Cancer Treatments Rev.* **1990**, *17*, 261.

(7) (a) Galanski, M.; Arion, V. B.; Jakupec, M. A.; Keppler, B. K. *Curr. Pharm. Des.* **2003**, *9*, 2078. (b) Rademaker-Lakhai, J. M.; van den Bongard, D.; Pluim, D.; Beijnen, J. H.; Schellens, J. H. M. *Clin. Cancer Res.* **2004**, *10*, 3717. (c) Hartinger, C. G.; Zorbas-Seifried, S.; Jakupec, M. A.; Kynast, B.; Zorbas, H.; Keppler, B. K. *J. Inorg. Biochem.* **2006**, *100*, 891.

(8) Yan, Y. K.; Melchart, M.; Habtemariam, A.; Sadler, P. J. *Chem. Commun.* **2005**, *38*, 4764.

(9) Ang, W. H.; Dyson, P. J. *Eur. J. Inorg. Chem.* **2006**, *20*, 4003.

(10) (a) Kelman, A. D.; Clarke, M. J.; Edmonds, S. D.; Peresie, H. J. *J. Clin. Hematol. Oncol.* **1977**, *7*, 274. (b) Clarke, M. J. In *Metal Complexes in Cancer Chemotherapy.* Keppler, B. K. Ed.; VCH, Weinheim, 1993, p. 129.

(11) Brown, J. M.; Giaccia, A. J. *Cancer Res.* **1998**, *58*, 1408.

(12) Reisner, E.; Arion, V. B.; Keppler, B. K.; Pombeiro, A. J. L. *Inorg. Chim. Acta* **2007**, doi:10.1016/j.ica.2006/12/005.

(13) Lever, A. B. P. *Inorg. Chem.* **1990**, *29*, 1271.

(14) Wheeler, R. A. *J. Am. Chem. Soc.* **1994**, *116*, 11048.

(15) (a) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161. (b) Tomasi, J.; Menucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.

(16) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(17) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(18) Takano, Y.; Houk, K. N. *J. Chem. Theory Comput.* **2005**, *1*, 70.

(19) Dutton, A. S.; Fukuto, J. M.; Houk, K. N. *Inorg. Chem.* **2005**, *44*, 4024.

(20) Baik, M.-H.; Friesner, R. A. *J. Phys. Chem. A* **2002**, *106*, 7407.

(21) Jaque, P.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. C* **2007**, *111*, 5783.

(22) Besker, N.; Coletti, C.; Marrone, A.; Re, N. *J. Phys. Chem. B* **2007**, *111*, 9955.

(23) (a) Chen, J.; Chen, L.; Liao, S.; Zheng, K.; Ji, L. *J. Phys. Chem. B* **2007**, *111*, 7862. (b) Chen, J.; Chen, L.; Liao, S.; Zheng, K.; Ji, L. *Dalton Trans.* **2007**, *32*, 3507.

(24) (a) Marchant, J. A.; Matsubara, T.; Ford, P. C. *Inorg. Chem.* **1977**, *16*, 2160. (b) Yee, E. L.; Cave, R. J.; Guyer, K. L.; Tyma, P. D.; Weaver, M. J. *J. Am. Chem. Soc.* **1979**, *101*, 1131. (c) Costa, G.; Balducci, G.; Alessio, E.; Tavagnacco, C.; Mestroni, G. *J. Electroanal. Chem.* **1990**, *296*, 57. (d) Alessio, E.; Balducci, G.; Lutman, A.; Mestroni, G.; Calligaris, M.; Attia, W. M. *Inorg. Chim. Acta* **1993**, *203*, 205. (e) Mestroni, G.; Alessio, E.; Sava, G.; Pacor, S.; Coluccia, M.; Boccarelli, A. *Metal Based Drugs* **1993**, *1*, 41. (f) Dhubhghaill, N.; Orla M.; Hagen, W. R.; Keppler, B. K.; Lipponer, K.-G.; Sadler, P. J. *J. Chem. Soc., Dalton Trans.*

1994, 3305. (g) Clarke, M. J.; Bailey, V. M; Doan, P. E.; Hiller, C. D.; LaChance-Galang, K. J.; Daghlian, H.; Mandal, S.; Bastos, C. M.; Lang, D. *Inorg. Chem.* **1996**, *35*, 4896. (h) Mura, P.; Camalli, M.; Messori, L.; Piccioli, F.; Zanello, P.; Corsini, M. *Inorg. Chem.* **2004**, *43*, 3863. (i) Eskelinen, E.; Da Costa, P.; Haukka, M. *J. Electroanal. Chem.* **2005**, *579*, 257.

(25) (a) Reisner, E.; Arion, V. B.; Guedes da Silva, M. F. C.; Lichtenecker, R.; Eichinger, A.; Keppler, B. K.; Kukushkin, V. Yu.; Pombeiro, A. J. L. *Inorg. Chem.* **2004**, *43*, 7083. (b) Egger, A.; Arion, V. B.; Reisner, E.; Cebrian-Losantos, B.; Shova, S.; Trettenhahn, G.; Keppler, B. K. *Inorg. Chem.* **2005**, *44*, 122. (c) Reisner, E.; Arion, V. B.; Eichinger, A.; Kandler, N.; Giester, G.; Pombeiro, A. J. L.; Keppler, B. K. *Inorg. Chem.* **2005**, *44*, 6704.

(26) Ravera, M.; Baracco, S.; Cassino, C.; Zanello, P.; Osella, D. *Dalton. Trans.* **2004**, 2347.

(27) Note that there is a relatively small but noticeable dependence of the experimental SRPs on ionic strength and pH, with typical changes amounting to ~0.02 V.[24b,26] Most complexes in aqueous solution were measured at pH ~ 7, but some of them had to be measured at low pH (e.g., $[Ru(NH_3)_5(OH_2)]^{3+}$) or high pH (e.g., $[Ru(NH_3)_5(OH)]^{2+}$),[24b] due to likely changes in the protonation state of the ligands upon reduction at neutral pH. A change in the protonation state upon reduction certainly alters the measured SRP, as was described, e.g., for *trans*-$[RuCl_2(trz)_4]Cl$, in ref 25c.

(28) Kirlin, W. G.; Cai, J.; Thompson, S. A.; Diaz, D.; Kavanagh, T. J.; Jones, D. P. *Free Rad. Biol. Med.* **1999**, *27*, 1208.

(29) 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 D.01, Gaussian, Inc., Wallingford, CT, 2004. www.gaussian.com (accessed November 9, 2007).

(30) Andrae, D.; Haeussermann, U.; Dolg, M.; Stoll, H.; Preuss, H. *Theor. Chim. Acta* **1990**, *77*, 123.

(31) Bergner, A.; Dolg, M.; Kuechle, W.; Stoll, H.; Preuss, H. *Mol. Phys.* **1993**, *80*, 1431.

(32) (a) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257. (b) Binkley, J. S.; Pople, J. A.; Hehre, W. J. *J. Am. Chem. Soc.* **1980**, *102*, 939.

(33) Martin, J. M. L.; Sundermann, A. *J. Chem. Phys.* **2001**, *114*, 3408.

(34) Deubel, D. V. *J. Org. Chem.* **2001**, *66*, 2686.

(35) Deubel, D. V. *J. Am. Chem. Soc.* **2004**, *126*, 996. (See Supporting Information).

(36) (a) Miertus, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117. (b) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.

(37) Cances, M. T.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032. (b) Cossi, M.; Scalmani, G.; Rega, N.; Barone, V. *J. Chem. Phys.* **2002**, *117*, 43.

(38) Klamt, A.; Schuurmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799.

(39) Barone, V.; Cossi, M. *J. Phys. Chem. A* **1998**, *102*, 1995.

(40) (a) Tannor, D. J.; Marten, B.; Murphy, R. B.; Friesner, R. A.; Sitkoff, D.; Nicholls, A.; Ringnalda, M. N.; Goddard, W. A., III; Honig, B. *J. Am. Chem. Soc.* **1994**, *116*, 11875. (b) Marten, B.; Kim, K.; Cortis, C.; Friesner, R. A.; Murphy, R. B.; Ringnalda, M. N.; Sitkoff, D.; Honig, B. *J. Phys. Chem.* **1996**, *100*, 11775. (c) Edinger, S. R.; Cortis, C.; Shenkin, P. S.; Friesner, R. A. *J. Phys. Chem.* **1997**, *101*, 1190. (d) Friedrichs, M.; Zhou, R. H.; Edinger, S. R.; Friesner, R. A. *J. Phys. Chem. B* **1999**, *103*, 3057.

(41) Jaguar, version 7.0, Schrödinger, LCC, New York, NY, 2007. www.schrodinger.com (accessed November 9, 2007).

(42) Dielectric constants ϵ of solvents in Gaussian 03: acetonitrile (ACN) 36.64, dimethyl formamide (DMF) 36.7, dimethyl sulfoxide (DMSO) 46.7, water 78.39. In Jaguar 7: ACN 37.5, DMF 36.7, DMSO 47.24, water 80.37.

(43) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299.

(44) Ben-Naim, A.; Marcus, Y. *J. Chem. Phys.* **1984**, *81*, 2016.

(45) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddart, W. A., III; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024.

(46) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210.

(47) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.

(48) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441.

(49) Pierotti, R. A. *Chem. Rev.* **1976**, *76*, 717.

(50) Lee, B.; Richards, F. M. *J. Mol. Biol*. **1971**, *55*, 379.

(51) 1.2 for *water* and *DMF*; 1.4 for *ACN* and *DMSO*.

(52) For an alternative approach, see: (a) Tavernelli, I.; Vuilleumier, R.; Sprik, M. *Phys. Rev. Lett.* **2002**, *88*, 213002-1. (b) Blumberger, J.; Sprik, M. *J. Phys. Chem. B* **2005**, *109*, 6793. (c) Blumberger, J.; Sprik, M. *Theor. Chem. Acc.* **2006**, *115*, 113.

(53) Jorgensen, W. L.; Briggs, J. M.; Gao, J. *J. Am. Chem. Soc.* **1987**, *109*, 6857.

(54) Lewis, A.; Bumpus, J. A.; Truhlar, D. G.; Cramer, C. J. *J. Chem. Ed.* **2004**, *81*, 596 incl. Addition/Correction.

(55) One referee commented that the PCM approach to deriving the dielectric radii is fundamentally flawed and the theoretical justification is inconsistent with the statistical mechanics of solvation, which might be reflected in the relatively poor results obtained with this approach.

(56) Born, M. *Z. Physik* **1920**, *1*, 45.

(57) (a) Deubel, D. V. *J. Am. Chem. Soc.* **2004**, *126*, 5999. (b) Lau, J. K.-C.; Deubel, D. V. *Chem. Eur. J.* **2005**, *11*, 2849. (c) Lau, J. K.-C.; Deubel, D. V. *J. Chem. Theory Comput.*

2006, *2*, 103. (d) Lau, J. K.-C.; Deubel, D. V. *Chem. Comm.* **2006**, 1654. (e) Deubel, D. V. *J. Am. Chem. Soc.* **2006**, *128*, 1654. (f) Deubel, D. V.; Chifotides, H. T. *Chem. Commun.* **2007**, 3438. (g) Deubel, D. V. *J. Am. Chem. Soc.* **2008**, *130*, 665.

(58) (a) Baik, M.-H.; Friesner, R. A.; Lippard, S. J. *J. Am. Chem. Soc.* **2002**, *124*, 4495. (b) Baik, M.-H.; Friesner, R. A.; Lippard, S. J. *J. Am. Chem. Soc.* **2003**, *125*, 14082. (c) Mantri, Y.; Lippard, S. J.; Baik, M.-H. *J. Am. Chem. Soc.* **2007**, *129*, 5023.

(59) The experimental SRP is 0.23 V; we report in the text the calculated deviations from this value. Vanýsek, P. In *CRC Handbook of Chemistry and Physics*, 87th ed.; Lide, D. R., Ed.; CRC Taylor and Francis: Boca Raton, FL, 2006; p. 8-24 and 8-26.

(60) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 1133.

(61) Reiss, H;, Heller, A. *J. Phys. Chem.* **1985**, *89*, 4207.

(62) NIST Chemistry Webbook. http://webbook.nist.gov (accessed November 10, 2007).

(63) Tissandier, M. D.; Cowen, K. A.; Fendg, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R. *J. Phys. Chem. A* **1998**, *102*, 7787.

(64) Farrell, J. R.; NcTigue, P. *J. Electroanal. Chem.* **1982**, *37*, 139.

(65) Trasatti, S. *Pure Appl. Chem.* **1986**, *58*, 955.

(66) Considering $\Delta G^{1}_{\mathrm{NHE}} = 364.27$ kcal/mol, calculated at the B3LYP/XL//M level, and $\Delta G_{\mathrm{solv}}(\mathrm{H}^{+}) = -263.98$ kcal/mol[63] results in $\Delta \mathrm{SRP}_{\mathrm{NHE}} = 4.35$ V.

CT700247G

JCTC Journal of Chemical Theory and Computation

On the Dielectric Boundary in Poisson−Boltzmann Calculations

Harianto Tjong and Huan-Xiang Zhou*

Department of Physics and Institute of Molecular Biophysics and School of Computational Science, Florida State University, Tallahassee, Florida 32306

Received November 21, 2007

Abstract: In applying the Poisson−Boltzmann (PB) equation for calculating the electrostatic free energies of solute molecules, an open question is how to specify the boundary between the low-dielectric solute and the high-dielectric solvent. Two common specifications of the dielectric boundary, as the molecular surface (MS) or the van der Waals (vdW) surface of the solute, give very different results for the electrostatic free energy of the solute. With the same atomic radii, the solute is more solvent-exposed in the vdW specification. One way to resolve the difference is to use different sets of atomic radii for the two surfaces. The radii for the vdW surface would be larger in order to compensate for the higher solvent exposure. Here we show that radius reparametrization required for bringing MS-based and vdW-based PB results to agreement is solute-size dependent. The difference in atomic radii for individual amino acids as solutes is only 2−5% but increases to over 20% for proteins with ∼200 residues. Therefore two sets of radii that yield identical MS-based and vdW-based PB results for small solutes will give very different PB results for large solutes. This finding raises issues about two common practices. The first is the use of atomic radii, which are parametrized against either experimental solvation data or data obtained from explicit-solvent simulations on small compounds, for PB calculations on proteins. The second is the parametrization of vdW-based generalized Born models against MS-based PB results.

I. Introduction

The Poisson−Boltzmann (PB) equation is widely used for modeling electrostatic effects and solvation of biomolecules.[1−30] The calculated electrostatic free energy of a solute molecule depends on the permanent partial charges on the atoms of the solute and the boundary of the low-dielectric solute and the high-dielectric solvent. Even when the radii of the atoms are given, there is still considerable freedom in specifying the dielectric boundary. In particular, two choices widely used in PB calculations are the van der Waals (vdW) surface and the molecular surface (MS) (see Figure 1). The vdW surface consists of the exposed surfaces of the spheres representing the solute atoms. The MS, introduced by Richards,[31] relies on a spherical solvent probe. According to the MS, the atomic spheres and all crevices inaccessible to the solvent probe are all treated as part of the solute dielectric (the MS hence has also been referred to as the solvent-exclusion surface). The added crevices reduce the exposure of the solute charges to the solvent. Since solute charges have strong interactions with the solvent, the cumulative effects of the reduced solvent exposure of all the solute atoms can lead to a significant change in the electrostatic solvation energy. As a result, the electrostatic interaction free energy between an oppositely charged protein−protein pair or protein-RNA pair can change from negative to positive when the dielectric boundary is switched from vdW to MS.[13,20,29,30,32] The electrostatic contribution of even a single mutation to the folding stability of a protein or the binding stability of a protein−protein or protein-RNA complex can be predicted very differently by the two choices of the dielectric boundary.[8,10,13,20,32] One possible way to

* Corresponding author phone: (850)645-1336; fax: (850)644-7244; e-mail: zhou@sb.fsu.edu.

10.1021/ct700319x CCC: $40.75
Published on Web 02/21/2008

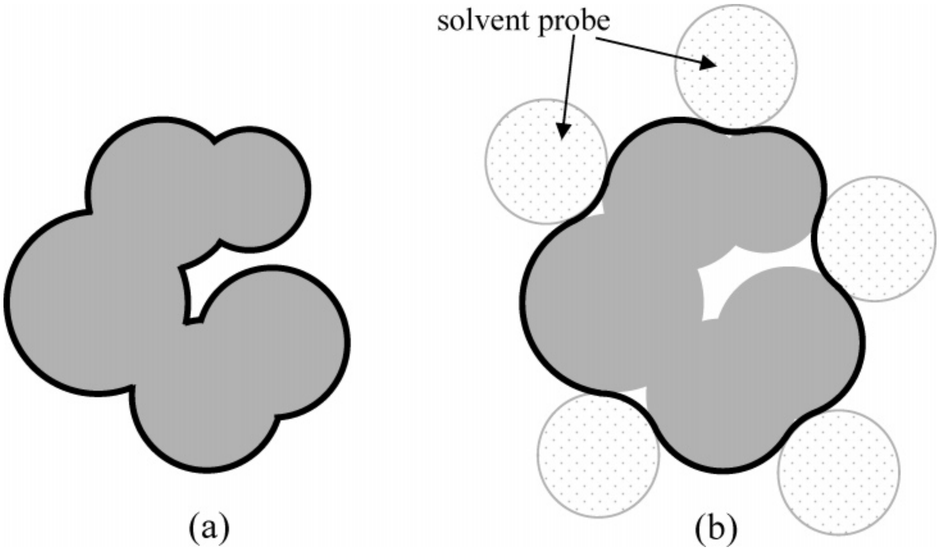


Figure 1. Definitions of (a) the van der Waals surface and (b) the molecular surface. In this two-dimensional illustration, atoms are represented by gray disks. In (a), the exposed boundaries of the disks, shown in dark, constitute the van der Waals surface. In (b), a spherical solvent probe is rolled around the solute molecule. In addition to the van der Waals spheres, small crevices inaccessible to the solvent probe are now part of the solute region. The boundary of this filled-up solute region, shown in dark, is the molecular surface.

reconcile the differences in calculated electrostatic free energy is to use different sets of atomic radii for the two choices of the dielectric boundary.[5,26] Specifically, to compensate for the higher solvent exposure by the vdW specification, atomic radii would be increased relative to those in the MS specification. We carried out such radius reparametrization and found that the changes in atomic radii are very dependent on the solute size. The difference in atomic radii for individual amino acids as solutes is only 2−5% but increases to over 20% for proteins with ∼200 residues.

There is a widely held perception that, between vdW and MS, the latter is a better choice for the dielectric boundary, though a convincing argument has not been laid out. To the contrary, it has been shown that PB calculations with the vdW choice consistently give better agreement with experimental results for mutational effects on protein folding and binding stability[8,10,13,20,32] and for electrostatic contributions to protein binding rates.[29,30] This paper does not aim to settle the difference between MS and vdW. Rather, the significance of our finding lies in its implications for two common practices in PB calculations. The first is parametrization of atomic radii using either experimental solvation data or explicit-solvent simulations, which are restricted to small solute molecules.[19,23,26,33,34] Our finding would suggest that, on these solute molecules, the values of atomic radii obtained using either vdW or MS as the dielectric boundary differ very little (e.g., <5%). However, when these radii are then used for PB calculations on proteins, the electrostatic solvation energies will be very different depending on whether vdW or MS is specified as the dielectric boundary. The uncertainty on calculated solvation energies for proteins thus diminishes the value of experimental and explicit-solvent data on small solutes for parametrizing the PB model.

The second common practice occurs in developing generalized Born (GB) methods[35] as a fast substitute of the PB model. In some GB methods, the MS specification of the dielectric boundary is directly implemented, and the GB results are benchmarked against MS-based PB results.[36−39] In many other GB methods,[40−46] the vdW specification of the dielectric boundary is implemented, and the resulting GB results are then benchmarked against the MS-based PB results through additional parametrization. Our finding suggests that the parametrization required for matching vdW-based GB and MS-based PB is protein dependent, and the use of a uniform set of parametrization introduces a new source of error for individual proteins.

II. Calculation Details

We carried out different sets of PB calculations over 55 test proteins. One set, used as the target, had MS as the dielectric boundary and Bondi radii[47] for the protein atoms. All the other sets had vdW as the dielectric boundary and the atomic radii increased by various percentages (denoted as $\%\Delta r$) from the Bondi values. The aim of the variation in $\%\Delta r$ was to find the value which would lead to agreement in the electrostatic solvation energy, ΔG_{solv}, between the MS-based and vdW-based calculations for a particular protein. In the end, a collection of 55 “optimized” $\%\Delta r$ values was obtained for the test proteins.

The 55 test proteins have been used in our previous studies to find an empirical dependence of ΔG_{solv} on solute and solvent dielectric constants[21] and to develop GB methods as substitutes of the linearized and full PB equation.[48,49] These proteins were collected from the Protein Data Bank (http://www.rcsb.org/pdb) using the following criteria: sequence identity less than 10%, resolution better than 1.0 Å, and number of residues less than 250. For protein structures without hydrogen atoms, hydrogen atoms were added with the LEAP module in the AMBER package,[50] and then energy-minimized in vacuum with heavy atoms fixed. The

Table 1. Number of Atoms, Net Charge, and MS and vdW Solvation Energy (in kcal/mol) for 55 Test Proteins

PDB	N_{atom}	Q	ΔG_{solv}^{MS}	$\Delta G_{solv}^{vdW}(0\%\Delta r)$
1a6m	2435	2	−1893.3	−2716.3
1aho	967	−2	−943.3	−1299.3
1byi	3383	−4	−2408.9	−3578.1
1c75	987	−4	−1094.4	−1398.5
1c7k	1929	−5	−1672.0	−2439.7
1cex	2867	1	−1863.7	−2873.1
1eb6	2572	−15	−4062.5	−5094.0
1ejg	678	0	−356.4	−574.7
1etl	145	0	−213.1	−288.2
1exr	2240	−25	−8081.8	−9253.6
1f94	982	1	−858.6	−1206.0
1f9y	2535	−5	−2018.1	−2915.9
1g4i	1842	−1	−1659.2	−2402.0
1g66	2794	−2	−1628.6	−2945.2
1gqv	2143	7	−1768.6	−2561.5
1hje	179	1	−221.7	−275.3
1iqz	1171	−17	−4149.4	−4598.3
1iua	1207	−1	−873.0	−1289.8
1j0p	1597	8	−2242.3	−2810.4
1k4i	3253	−6	−2696.4	−3888.8
1kth	894	0	−1104.7	−1454.4
1l9l	1230	11	−2684.4	−3084.0
1m1q	1265	−7	−1945.0	−2379.0
1nls	3564	−7	−2927.5	−4680.5
1nwz	1912	−6	−2015.0	−2728.9
1od3	1900	−3	−1307.3	−2026.4
1ok0	1076	−5	−1153.9	−1546.3
1p9g	529	4	−556.0	−745.6
1pq7	3065	4	−1484.9	−2574.2
1r6j	1230	0	−972.9	−1337.4
1ssx	2750	8	−1674.4	−2623.6
1tg0	1029	−12	−2815.9	−3191.5
1tqg	1660	−7	−2373.2	−2903.5
1tt8	2676	1	−1655.7	−2604.9
1u2h	1526	4	−1521.1	−2036.2
1ucs	997	0	−705.1	−1021.8
1ufy	1926	−3	−1679.0	−2293.7
1unq	1966	−3	−2635.0	−3410.4
1vb0	921	3	−794.7	−1107.3
1vbw	1058	8	−1476.3	−1805.0
1w0n	1756	−5	−1685.6	−2417.1
1wy3	560	1	−600.6	−768.9
1x6z	1741	0	−1511.5	−2153.4
1x8q	2815	−1	−2325.5	−3550.2
1xmk	1268	1	−1151.3	−1589.0
1yk4	770	−8	−1578.3	−1874.2
1zzk	1252	1	−1202.8	−1591.7
2a6z	3432	−3	−2363.5	−3636.6
2bf9	560	−2	−763.8	−911.8
2chh	1624	−3	−1523.6	−2128.3
2cws	3400	−3	−1936.4	−3208.1
2erl	573	−6	−983.5	−1167.2
2fdn	731	−8	−1410.3	−1702.1
2fwh	1830	−6	−1629.1	−2251.1
3lzt	1960	8	−1866.9	−2587.4

PDB codes, total number of atoms (N_{atom}), and net charge (Q) for each of the 55 test proteins are listed in Table 1.

PB results for ΔG_{solv} were obtained by using the UHBD program.[51] The dielectric boundary was chosen as MS or vdW by the presence or absence of the “nmap 1.4, nsph 500” option in the UHBD input file. By default dielectric smoothing was applied to both choices of the dielectric boundary. UHBD calculations on all the test proteins used a coarse grid with a 1.5-Å spacing followed by a fine grid with a 0.5-Å spacing. The dimensions of the coarse and fine grids were 160 × 160 × 160 and 200 × 200 × 200, respectively. The solute and solvent dielectric constants were set to 1 and 78.5, respectively. No salt was present in the solvent.

For investigating the dependence of optimized %Δr on solute size, we carried out corresponding PB calculations on individual amino acids as solutes. For each of the 20 types of amino acids, 10 conformations were randomly carved out of the 55 test proteins. The UHBD calculations were done on the individual amino acids, with a coarse grid with a 50 × 50 × 50 dimension and a 1.0-Å spacing followed by a fine grid with a 60 × 60 × 60 dimension and a 0.25-Å spacing. For each type of amino acid, the average of optimized %Δr values over the 10 conformations is reported. Results from averaging over 20 conformations for each amino acid were essentially unchanged.

Areas of the dielectric boundary according to the two choices were calculated. For vdW and MS, the respective programs used were Naccess v2.1.1 (http://www.bioinf-.manchester.ac.uk/naccess/) with a probe radius of 0 and dms (http://www.cgl.ucsf.edu/chimera/docs/UsersGuide/midas/dms1.html) with a probe radius of 1.4 Å.

III. Results and Discussion

The electrostatic solvation energies of the 55 test proteins, calculated using Bondi radii and either the MS or vdW choice for the dielectric boundary, are listed in Table 1 and displayed in Figure 2a. It can be seen that the magnitudes of ΔG_{solv} are consistently larger with the vdW dielectric boundary, due to the resulting higher solvent exposure of solute charges. When the atomic radii are increased in vdW calculations, the magnitudes of ΔG_{solv} decrease and hence move toward those of the MS results. However, as Figure 2b shows, with a uniform increase of 6% in atomic radii, vdW results still consistently show larger magnitudes than the MS targets.

Figure 3 displays the optimized %Δr values for the 20 types of amino acids as solutes. The increases in atomic radii required to achieve consistency between vdW-based results for ΔG_{solv} and the MS-based target values are small, falling in the narrow range of 2% to 5%. The small changes in atomic radii are expected. With small solutes, all the atoms are well exposed to the solvent. Hence there are only limited chances that the MS will enclose small crevices outside the vdW surface. Interestingly, even within the narrow range of optimized %Δr values among the 20 types of amino acids, a positive correlation between optimized %Δr and N_{atom} is apparent. Linear regression analysis gave $R^2 = 0.65$.

On the 55 test proteins, the optimized %Δr values increase to at least 10%. As Figure 4a shows, there still seems to be a positive correlation between optimized %Δr and N_{atom}, but the data now exhibit much greater scatter. R^2 for linear

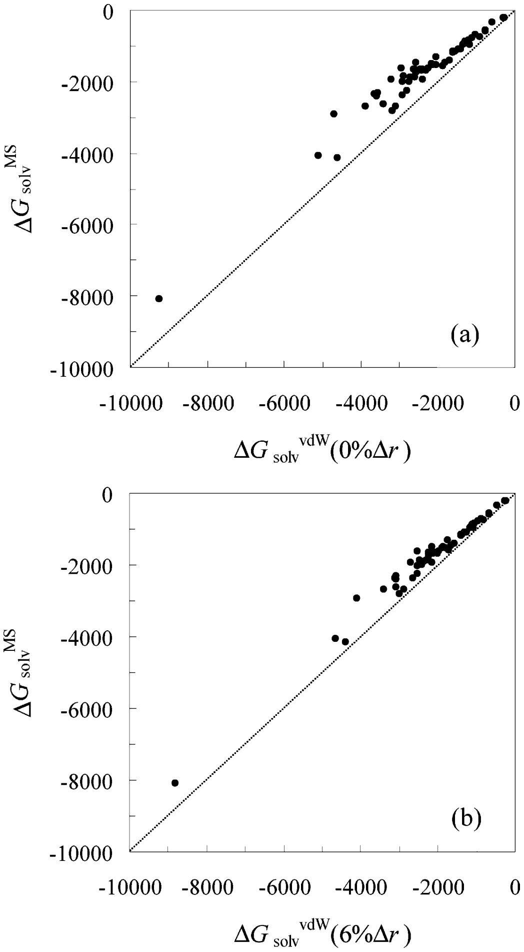


Figure 2. Comparison of the electrostatic solvation energies of the 55 test proteins from MS-based and vdW-based PB calculations. For MS-based PB calculations, the Bondi radii are always used: (a) $\Delta G_{\rm solv}{}^{\rm vdW}$ calculated with Bondi radii and (b) $\Delta G_{\rm solv}{}^{\rm vdW}$ calculated with atomic radii increased by 6% from the Bondi values.

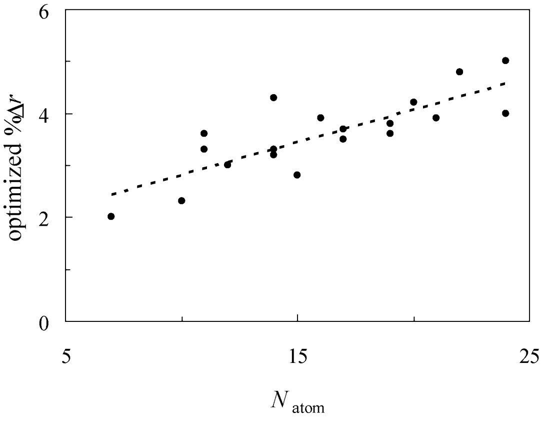


Figure 3. The percentage increase in atomic radii from the Bondi values required for $\Delta G_{\rm solv}{}^{\rm vdW}$ to match with $\Delta G_{\rm solv}{}^{\rm MS}$ for 20 types of amino acids as solutes.

correlation is now at 0.58. The variations in optimized $\%\Delta r$ within the 20 types of amino acids and within the 55 test proteins as well as between the two collections of solute molecules point to the accumulation of crevices that are outside the vdW surface but inside the MS as the major reason for the increase in optimized $\%\Delta r$.

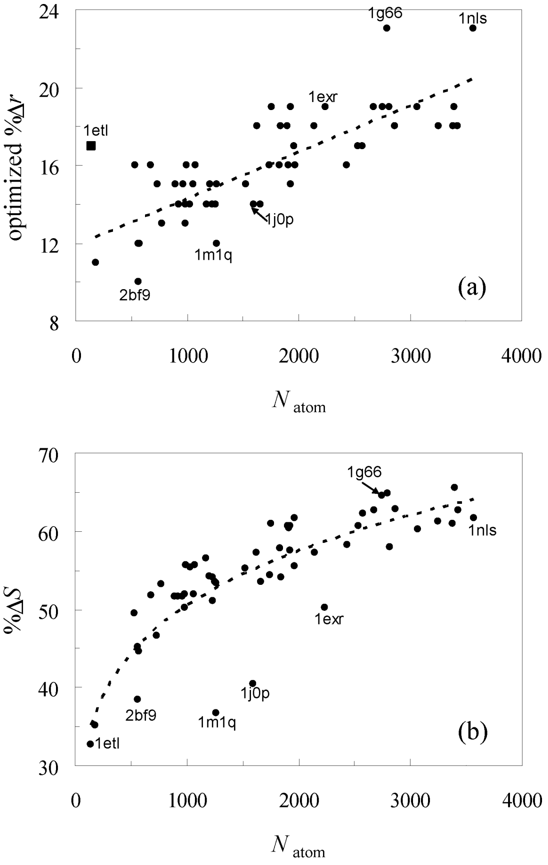


Figure 4. (a) The percent increases in atomic radii, $\%\Delta r$, for optimal agreement between $\Delta G_{\rm solv}{}^{\rm vdW}$ and $\Delta G_{\rm solv}{}^{\rm MS}$ on the 55 test proteins. (b) Percentage of difference in vdW surface area and MS area, $100(S^{\rm vdW} - S^{\rm MS})/S^{\rm vdW}$, against the number of atoms.

We examined the outliers in the correlation between optimized $\%\Delta r$ and $N_{\rm atom}$. Some of the low-lying proteins, such as 2bf9, 1m1q, and 1j0p, in the optimized $\%\Delta r$ vs $N_{\rm atom}$ plot were found to correspond to well-exposed structures (Figure 5a). For these proteins, the difference between the two types of solute surfaces are relatively small, and hence relatively small increases in atomic radii are required to bring vdW-based results for $\Delta G_{\rm solv}$ into agreement with the MS-based target. One way of quantifying the differences between the two types of solute surfaces is by calculating the corresponding surfaces areas. Figure 4b displays $\%\Delta S$, the relative differences in MS area and vdW surface area, against $N_{\rm atom}$. It can be seen that the low-lying proteins, 2bf9, 1m1q, and 1j0p, in the optimized $\%\Delta r$ vs $N_{\rm atom}$ plot are also below the general trend in the $\%\Delta S$ vs $N_{\rm atom}$ plot. However, the correspondence between the two plots is far from being perfect. In particular, a low-lying protein, 1exr, in the $\%\Delta S$ vs $N_{\rm atom}$ plot actually occupies a position above the correlation trend line in the optimized $\%\Delta r$ vs $N_{\rm atom}$ plot, and a high-lying protein, 1etl, in the optimized $\%\Delta r$ vs $N_{\rm atom}$ plot does not take such a position in the $\%\Delta S$ vs $N_{\rm atom}$ plot.

We suspected that the high-lying proteins in the optimized $\%\Delta r$ vs $N_{\rm atom}$ plot correspond to structures with deep channels outside the vdW surface, which become enclosed in the MS and hence are treated as part of the solute dielectric in the MS-based PB calculations. This suspicion did not find

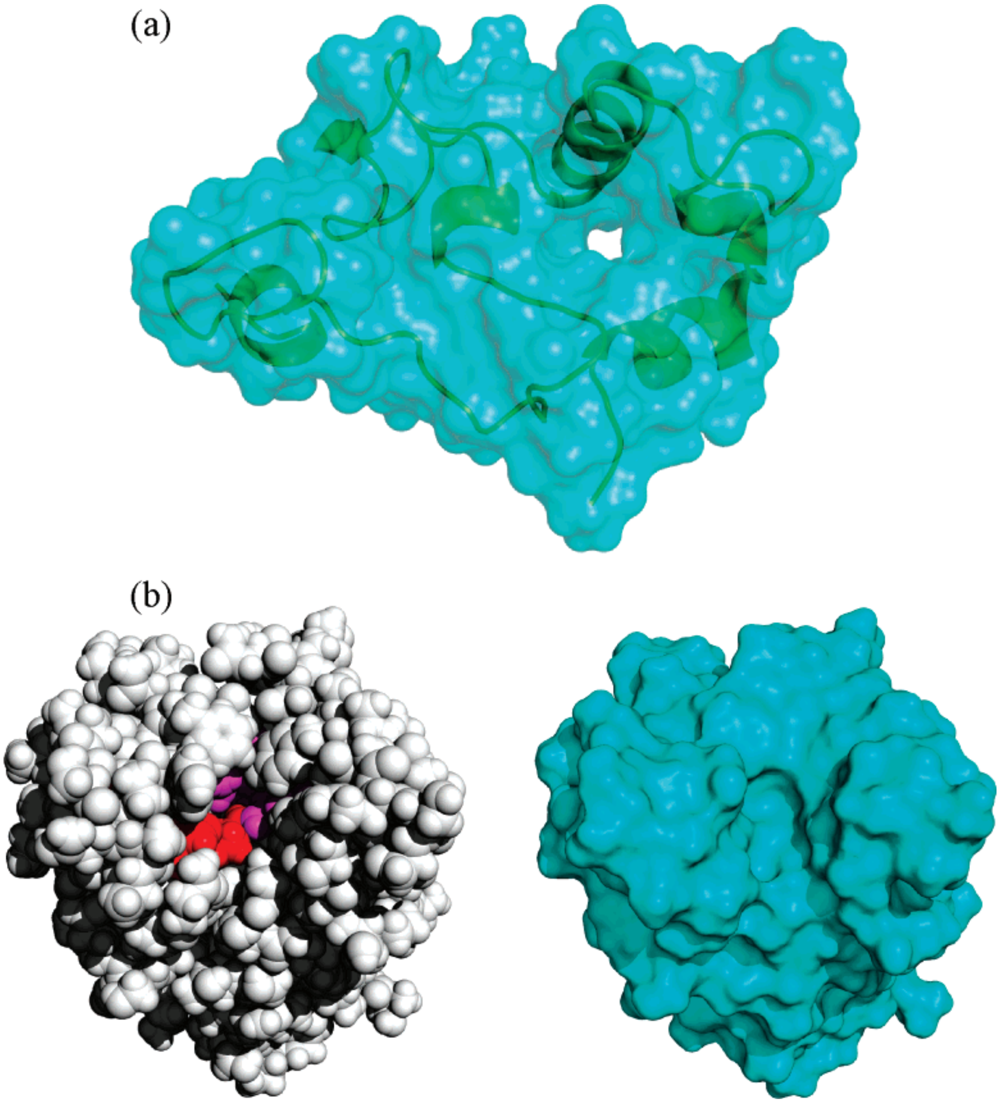


Figure 5. Comparison of van der Waals and molecular surfaces. (a) A well-exposed protein, 1m1q, which has the shape of a thin disk. The green ribbon representation of the protein is enclosed by the molecular surface in cyan; a hole appears near the center of the disk shape. (b) A protein, 1g66, with a deep channel. In the left panel, the van der Waals surface is presented, and residues lining the wall of the channel are displayed in red (for the catalytic triad) or purple. In the right panel, the molecular surface is presented. The active site now appears as an indent, but there is no channel penetrating into the center of the protein.

support in 1etl, which is the smallest (with $N_{\text{atom}} = 145$) of the 55 test proteins but required a relatively large 17% increase in atomic radii to achieve a match between vdW-based and MS-based results for ΔG_{solv}. However, a deep channel in the structure of another high-lying protein, acetylxylan esterase with PDB code 1g66, was identified (Figure 5b). Part of the wall of this channel is lined by the catalytic triad; hence this channel is important for access by solvent as well as the substrate. The channel is inaccessible by the 1.4-Å spherical probe used to defined the MS. This example illustrates the artificial nature of using a spherical probe on a static structure to define the boundary between the solute and solvent. Proteins are dynamic, allowing for transient access of water molecules, as seen in NMR experiments[52] and molecular dynamics simulations.[53] The transient excursions of water molecules into channels and interior positions are accounted for to some extent by choosing the vdW surface as the solute−solvent boundary, which perhaps partly explains the better performance of this choice in reproducing experimental results for electrostatic contributions to protein folding and binding.[8,10,13,20,29,30,32]

We modeled the trend in the $\%\Delta S$ vs N_{atom} plot shown in Figure 4b as a power law

$$\%\Delta S_{\text{pred}} = \alpha N_{\text{atom}}^{\ \nu} \qquad (1)$$

This function, with $\alpha = 13.8\%$ and $\nu = 0.19$, fitted the data with $R^2 = 0.68$. Given that the deviations from the trend of eq 1 could explain some of the outliers in Figure 4a, we included the ratio, $(\%\Delta S)/(\%\Delta S_{\text{pred}})$, as an independent variable along with N_{atom} in a multilinear regression to model the variations of optimized $\%\Delta r$ among the 55 test proteins. The inclusion of the new variable led to a modest increase in R^2, from 0.58 to 0.65. As Figure 6 shows, there are substantial deviations between actual optimized $\%\Delta r$ values and those predicted from multilinear regression, especially for 1etl, 2bf9, 1g66, and 1nls. The significant variations in optimized $\%\Delta r$ became apparent after we tested vdW-based

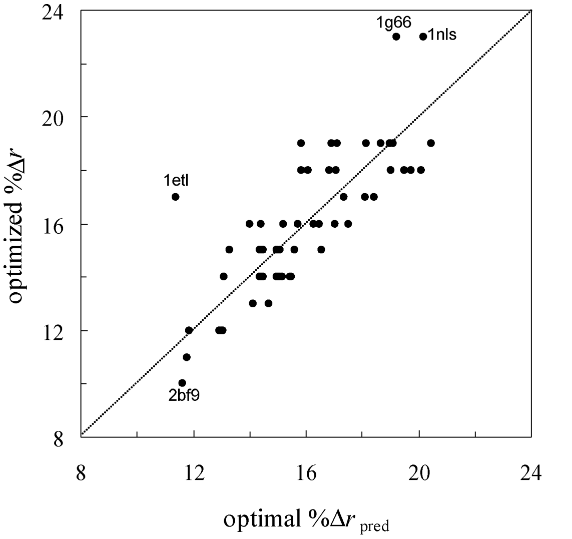


Figure 6. Comparison of actual optimized %Δ*r* values against those predicted from a multilinear relation.

and MS-based PB results on the large, diverse collection of 55 proteins. So our study raises caution against using only a small number of test proteins to parametrize the PB model and draw conclusions.

Based on the efforts reported here, it seems unlikely that a simple way to predict optimized %Δ*r* values can be found. The chance of bringing MS-based and vdW-based PB results into good agreement for a diverse set of proteins through radius reparametrization is thus slim. This finding suggests that significant errors are introduced when vdW-based GB methods are parametrized to approximate MS-based PB results. It is interesting to note that, after parametrizing a vdW-based GB method against MS-based PB results for small compounds,[40] the deviations of this GB from the MS-based PB were found to increase with increasing sizes of test compounds.[54]

The overall increase in optimized %Δ*r* with increasing solute size also raises a cautionary note about the use of experimental and explicit-solvent data on small solutes for parametrizing the PB model. Very similar values of atomic radii will be obtained when MS-based and vdW-based PB calculations are benchmarked against the data on small solutes. However, when these radii are then used in respective PB calculations on proteins, the electrostatic solvation energies can differ significantly. Before the issue of the optimal choice for the dielectric boundary is settled, the value of small-solute data seems open to question. This applies not only to MS- and vdW-based PB calculations but also to alternative choices, such as spline-smoothed surfaces, of the dielectric boundary.[5,9,14,19,26,55−58] A fruitful approach to parametrizing the PB model is to use experimental data obtained on proteins.[8,10,13,20,29,30,32]

Acknowledgment. This work was supported in part by NIH grant GM058187.

References

(1) Gilson, M. K.; Sharp, K. A.; Honig, B. Calculating the electrostatic potential of molecules in solution: method and error assessment. *J. Comput. Chem.* **1987**, *9*, 327−335.

(2) Gilson, M. K.; Honig, B. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins* **1988**, *4*, 7−18.

(3) Nicholls, A.; Honig, B. A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. *J. Comput. Chem.* **1991**, *12*, 435−445.

(4) Madura, J. D.; Briggs, J. M.; Wade, R.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; McCammon, J. A. Electrostatic and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. *Comput. Phys. Commun.* **1995**, *91*, 57−95.

(5) Nina, M.; Im, W.; Roux, B. Optimized atomic radii for protein continuum electrostatics solvation forces. *Biophys. Chem.* **1999**, *78*, 89−96.

(6) 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.

(7) Rocchia, W.; Alexov, E.; Honig, B. Extending the applicability of the nonlinear Poisson-Boltzmann equation: multiple dielectric constants and multivalent ions. *J. Phys. Chem. B* **2001**, *105*, 6507−6514.

(8) Vijayakumar, M.; Zhou, H.-X. Salt bridges stabilize the folded structure of barnase. *J. Phys. Chem. B* **2001**, *105*, 7334−7340.

(9) Grant, J. A.; Pickup, B. T.; Nicholls, A. A smooth permittivity function for Poisson-Boltzmann solvation methods. *J. Comput. Chem.* **2001**, *22*, 608−640.

(10) Dong, F.; Zhou, H.-X. Electrostatic contributions to T4 lysozyme stability: solvent-exposed charges versus semi-buried salt bridges. *Biophys. J.* **2002**, *83*, 1341−1347.

(11) Luo, R.; David, L.; Gilson, M. K. Accelerated Poisson-Boltzmann calculations for static and dynamic systems. *J. Comput. Chem.* **2002**, *23*, 1244−1253.

(12) Zhou, H.-X.; Dong, F. Electrostatic contributions to the stability of a thermophilic cold shock protein. *Biophys. J.* **2003**, *84*, 2216−2222.

(13) Dong, F.; Vijayakumar, M.; Zhou, H.-X. Comparison of calculation and experiment implicates significant electrostatic contributions to the binding stability of barnase and barstar. *Biophys. J.* **2003**, *85*, 49−60.

(14) Lu, Q.; Luo, R. A Poisson-Boltzmann dynamics method with nonperiodic boundary condition. *J. Chem. Phys.* **2003**, *119*, 11035−11047.

(15) Prabhu, N. V.; Zhu, P.; Sharp, K. A. Implementation and testing of stable, fast implicit solvation in molecular dynamics using the smooth-permittivity finite difference Poisson-Boltzmann method. *J. Comput. Chem.* **2004**, *25*, 2049−2064.

(16) Baker, N. A. Improving implicit solvent simulations: a Poisson-centric view. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137−143.

(17) Huang, X.; Dong, F.; Zhou, H.-X. Electrostatic recognition and induced fit in the *κ*-PVIIA toxin binding to Shaker potassium channel. *J. Am. Chem. Soc.* **2005**, *127*, 6836−6849.

(18) Swanson, J. M. J.; Mongan, J.; McCammon, J. A. Limitations of atom-centered dielectric functions in implicit solvation models. *J. Phys. Chem. B* **2005**, *109*, 14769−14772.

(19) Swanson, J. M. J.; Adcock, S. A.; McCammon, J. A. Optimized radii for Poisson-Boltzmann calculations with the AMBER force field. *J. Chem. Theory Comput.* **2005**, *1*, 484–493.

(20) Dong, F.; Zhou, H.-X. Electrostatic contribution to the binding stability of protein-protein complexes. *Proteins* **2006**, *65*, 87–102.

(21) Tjong, H.; Zhou, H.-X. The dependence of electrostatic solvation energy on dielectric constants in Poisson-Boltzmann calculations. *J. Chem. Phys.* **2006**, *125*, 206101.

(22) Zhang, Q.; Schlick, T. Stereochemistry and position-dependent effects of carcinogens on TATA/TBP binding. *Biophys. J.* **2006**, *90*, 1865–1877.

(23) Tan, C.; Yang, L.; Luo, R. How well does Poisson-Boltzmann implicit solvent agree with explicit solvent? A quantitative analysis. *J. Phys. Chem. B* **2006**, *110*, 18680–18687.

(24) Lwin, T. Z.; Zhou, R.; Luo, R. Is Poisson-Boltzmann theory insufficient for protein folding simulations? *J. Chem. Phys.* **2006**, *124*, 034902.

(25) Schnieders, M. J.; Baker, N. A.; Ren, P.; Ponder, J. W. Polarizable atomic multipole solutes in a Poisson-Boltzmann continuum. *J. Chem. Phys.* **2007**, *126*, 124114.

(26) Swanson, J. M. J.; Wagoner, J. A.; Baker, N. A.; McCammon, J. A. Optimizing the Poisson dielectric boundary with explicit solvent forces and energies: Lessons learned with atom-centered dielectric functions. *J. Chem. Theory Comput.* **2007**, *3*, 170–183.

(27) Qin, S. B.; Zhou, H.-X. Do electrostatic interactions destabilize protein-nucleic acid binding? *Biopolymers* **2007**, *86*, 112–118.

(28) Alsallaq, R.; Zhou, H.-X. Prediction of protein-protein association rates from a transition-state theory. *Structure* **2007**, *15*, 215–224.

(29) Alsallaq, R.; Zhou, H.-X. Electrostatic rate enhancement and transient complex of protein-protein association. *Proteins* **2008**, *71*, 320–335.

(30) Qin, S. B.; Zhou, H.-X. Prediction of salt and mutational effects on the association rate of U1A protein and U1 small nuclear RNA stem/loop II. *J. Phys. Chem. B* **2008**, in press.

(31) Richards, F. M. Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151–176.

(32) Qin, S. B.; Zhou, H.-X. Do electrostatic interactions destabilize protein-nucleic acid binding? *Biopolymers* **2007**, *86*, 112–118.

(33) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **1994**, *98*, 1978–1988.

(34) Nina, M.; Beglov, D.; Roux, B. Atomic Born radii for continuum electrostatic calculations based on molecular dynamics free energy simulations. *J. Phys. Chem. B* **1997**, *101*, 5239–5248.

(35) Still, A.; Tempczyk, W. C.; Hawley, R. C.; Hendrikson, R. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(36) Ghosh, A.; Rapp, C. S.; Friesner, R. A. Generalized Born model based on a surface integral formulation. *J. Phys. Chem. B* **1998**, *102*, 10983–10990.

(37) Lee, M. S.; Salsbury, F. R., Jr.; Brooks, C. L., III Novel generalized Born methods. *J. Chem. Phys.* **2002**, *116*, 10606–10614.

(38) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., III New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem.* **2003**, *24*, 1348–1356.

(39) Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. Generalized Born model with a simple, robust molecular volume correction. *J. Chem. Theory Comput.* **2007**, *3*, 156–169.

(40) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

(41) Jayaram, B.; Sprous, D.; Beveridge, D. L. Solvation free energy of biomacromolecules: Parameters for a modified generalized Born model consistent with the AMBER force field. *J. Phys. Chem. B* **1998**, *102*, 9571–9576.

(42) Onufriev, A.; Bashford, D.; Case, D. A. Modification of the generalized Born model suitable for macromolecules *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

(43) Tsui, V.; Case, D. A. Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers* **2000**, *56*, 275–291.

(44) Tsui, V.; Case, D. A. Molecular dynamics simulations of nucleic acids with a generalized Born solvation model. *J. Am. Chem. Soc.* **2000**, *122*, 2489–2498.

(45) Onufriev, A.; Case, D. A.; Bashford, D. Effective Born radii in the generalized Born approximation: The importance of being perfect. *J. Comput. Chem.* **2002**, *23*, 1297–1304.

(46) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* **2004**, *55*, 383–394.

(47) Bondi, A. van der Waals volumes and radii. *J. Phys. Chem.* **1964**, *68*, 441–451.

(48) Tjong, H.; Zhou, H.-X. GBr^6: a parameterization-free, accurate, analytical generalized Born method. *J. Phys. Chem. B* **2007**, *111*, 3055–3061.

(49) Tjong, H.; Zhou, H.-X. GBr^{6}NL: a generalized Born method for accurately reproducing solvation energy of the nonlinear Poisson-Boltzmann equation. *J. Chem. Phys.* **2007**, *126*, 195102.

(50) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER 8*; University of California: San Fransisco, 2004.

(51) Madura, J. D.; Briggs, J. M.; Wade, R.; Davis, M. E.; Lutty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; McCammon, J. A. Electrostatic and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. *Comput. Phys. Commun.* **1995**, *91*, 57–95.

(52) Ernst, J. A.; Clubb, R. T.; Zhou, H.-X.; Gronenborn, A. M.; Clore, G. M. Demonstration of positionally disordered water within a protein hydrophobic cavity by NMR. *Science* **1995**, *267*, 1813–1817.

(53) Damjanovic, A.; Garcia-Moreno, B.; Lattman, E. E.; Garcia, A. E. Molecular dynamics study of water penetration in staphylococcal nuclease. *Proteins* **2005**, *60*, 433−449.

(54) Edinger, S. R.; Cortis, C.; Shenkin, P. S.; Friesner, R. A. Solvation free energies of peptides: Comparison of approximate continuum solvation models with accurate solution of the Poisson-Boltzmann equation. *J. Phys. Chem. B* **1997**, *101*, 1190−1197.

(55) Schaefer, M.; Karplus, M. A comprehensive analytical treatment of continuum electrostatics. *J. Phys. Chem.* **1996**, *100*, 1578−1599.

(56) Im, W.; Lee, M. S.; Brooks, C. L., III Generalized born model with a simple smoothing function. *J. Comput. Chem.* **2003**, *24*, 1691−1702.

(57) Yu, Z.; Jacobson, M. P.; Friesner, R. A. What role do surfaces play in GB models? A new-generation of surface-generalized born model based on a novel gaussian surface for biomolecules. *J. Comput. Chem.* **2006**, *27*, 72−89.

(58) Grant, J. A.; Pickup, B. T.; Sykes, M. J.; Kitchen, C. A.; Nicholls, A. The Gaussian generalized Born model: application to small molecules. *Phys. Chem. Chem. Phys.* **2007**, *9*, 4913−4922.

CT700319X

JCTC Journal of Chemical Theory and Computation

Improved Energy Selection of Nativelike Protein Loops from Loop Decoys

Matthew S. Lin[†] and Teresa Head-Gordon*[,†,‡]

UCSF/UCB Joint Graduate Group in Bioengineering, Berkeley, California 94720, and Department of Bioengineering, University of California, Berkeley, California 94720

Received October 29, 2007

Abstract: We demonstrate the performance of a new implicit solvent model on native protein loop prediction from a large set of loop decoys of 4- to 12-residue in length. The physics-based energy function combines a hydrophobic potential of mean force (HPMF) description with a Generalized Born model for polarization of protein charge by the high dielectric solvent, which we combine with AMBER force field for the protein chain. The novelty of our energy function is the stabilizing effect of hydrophobic exposure to aqueous solvent that defines the HPMF hydration physics, which in principle should be an important stabilizing factor for loop conformations of a protein that typically are more solvent exposed. While our results for short loop decoy sets are comparably good to existing energy functions, we find demonstrable superiority for loop lengths of 8-residue and greater, and the quality of our predictions is largely insensitive to the length of the target loop on a filtered set of decoys. Given that the current weakness in loop modeling is the ability to select the most nativelike loop conformers from loop ensembles, this energy function provides a means for greater prediction accuracy in structure prediction of homologous and distantly related proteins, thereby aiding large-scale genomics efforts in comparative modeling.

1. Introduction

The loop regions of a protein are known to be important for its structure as they determine sequence reversals that allows the chain to collapse and fold[1] as well as for functions such as guiding or gating ligand binding, aiding protein complexation, and for enzymatic activity.[2] The ability to predict the nativeness of loops is thus an important goal, especially since functional differences between homologous proteins differ mostly in their loop conformations.[3,4] However, prediction of native loop structure is a far more difficult problem than prediction of native β-strands and α-helices, since the structures are more highly diverse, and the sequence correlations are even weaker, compared to these other standard secondary structure categories.[5,6]

As is true in prediction of overall protein structure,[7] the techniques that have been used in the prediction of loop structure can be divided into knowledge-based[8−12] and physics-based[13−17] approaches. Knowledge-based approaches rely typically on protein structural databases such as the PDB to derive sample loop conformations directly and typically use empirical criteria for native discrimination against misfolds. A library of known loop fragment structures with the same length as the target loop is fit between the backbone atoms of the residues that precede and succeed the target loop sequence (stem residues), which are then screened by a scoring function to generate plausible candidates, which are sometimes further refined with energy minimization. By contrast, physics-based approaches perform loop generation and selection based on first principles concepts. Ab initio chain growth techniques such as inverse kinematics have been widely implemented[18−21] and combined with more exhaustive conformational sampling to generate loop conformations that interpolate between the two stem residues. Typically a physically motivated energy function that aims

* Corresponding author e-mail: tlhead-gordon@lbl.gov. Corresponding author address: Department of Bioengineering, Stanley Hall 274, University of California, Berkeley, Berkeley, CA 94720.

[†] UCSF/UCB Joint Graduate Group in Bioengineering.

[‡] University of California.

10.1021/ct700292u CCC: $40.75
Published on Web 02/21/2008

to discriminate native loops from misfolded decoys is used to rank the degree of nativeness of a given loop configuration.

Strict categorization of any loop optimization method into these two general approaches is somewhat arbitrary.[22] Until recently, the primary difference between the knowledge-based and physics-based methods is the generation of long loop structures (>8 residues) since they appeared less frequently in the PDB than shorter ones.[6] However, high-throughput structural genomics efforts are diminishing this difference, and nativelike loops, within 2.0 Å for long loops, can be found with high probability in structural database.[11,23] Regardless of whether loops are generated by database searches or ab initio techniques, current evidence suggests that sufficient sampling of loop conformations, even long loops of 8 or more residues, does not appear to be a limiting factor.[11,16,23,24] Instead, the energy or statistical functions used for the discrimination of native and nativelike loop structures against misfolded decoys mostly restrict the prediction accuracy.[11,16,25−27] In fact, a very recent study[26] found that four popularly used commercial software packages rarely selected the most nativelike loop conformer of their generated ensemble, revealing that selection of nativelike loop conformers is the *primary* weakness of the current state-of-the-art approaches for loop modeling. In summary, better energy functions are key to both categories of loop structure prediction in order to push homology modeling efforts toward greater prediction accuracy.

In this work, we demonstrate the performance on native protein loop prediction of our recently developed energy function[28] and compare the prediction accuracy with other physics-based and knowledge-based scoring functions. We then further discuss the limitations of using the raw experimental native loop conformation as the gold standard in the testing procedures and place our results in the context of current state of the art in native loop selection available in several commercial packages. We hope that our energy function provides a better means for loop selection and optimization that should aid experimental structural refinement and large-scale genomics efforts in comparative modeling.

2. Materials and Methods

2.1. Energy Function. We have recently developed a new implicit solvent model[28] to describe a given protein and its aqueous solvent free energy surface. The energy function combines the AMBER ff99 protein force field developed by Wang and co-workers[29] (V_{Protein}), the Generalized Born (GB) description of the electrostatic component of solvent free energy (V_{GB}) developed by Onufriev and co-workers,[30] and the newly developed implicit solvent model, hydrophobic potential of mean force (HPMF), to describe the hydrophobic solute−solute interaction induced by water[31,32] (V_{HPMF})

$$V = V_{\mathrm{Protein}} + V_{\mathrm{GB}} + V_{\mathrm{HPMF}} \tag{1}$$

$$V_{\mathrm{HPMF}} = \sum_{i \in SA_i > A_c}^{N_c} \tanh(SA_i) \sum_{j \in SA_j > A_c}^{N_c} \tanh(SA_j) \times \sum_{k=1}^{3} h_k \exp(-[(r_{ij} - c_k) \times w_k]^2) \tag{1a}$$

$$\tanh(SA) = \frac{1}{2}(\tanh[\mathrm{SLOPE} \times (SA - \mathrm{OFFSET})] + 1) \tag{1b}$$

SLOPE and OFFSET are constants set to 1000.0 and 6.0, respectively. We refer the reader to our previous work[28] for the functional form and the parameter details of the model. The novelty of this energy function is the stabilizing effect of hydrophobic exposure to aqueous solvent that defines the HPMF hydration physics and its apparent improvement over solvent accessible surface area models that penalize hydrophobic exposure. When tested on an extensive number of protein decoy sets, which allows us to compare our performance to other scoring functions for native fold, we find that our energy function outperforms other similarly tested energy and statistical functions with both substantial improvements in native ranking and Z-score drops of 0.5−2.5 units.[28]

In this publication, the entire energy function (AMBERff99+GB+HPMF) will be simply abbreviated as HPMF unless indicated otherwise.

2.2. Decoy Set Selection. Here, we demonstrate the new energy function's performance on native protein loop prediction. We have restricted our comparison to the loop decoys generated by Jacobson et al.,[16] which we and others[12] have found to be more difficult decoys than the RAPPER[14,15] decoy sets. We then compare our results to the physics-based energy function, OPLS/SGB-NP,[16,33,34] and the statistical scoring function, DFIRE,[12] which have been tested on the same loop decoy sets.

Table 1 lists the characteristics of the loop decoys for 4-, 6-, 8-, 10-, 11-, and 12-residues generated by Jacobson et al.[16] As is true in protein structure, some of the loop decoy sets have complications due to exceptional features of the native protein structure that make the testing ambiguous. To combat this problem, Jacobsen et al. provide a filtered list of decoys that (1) eliminates proteins that were crystallized at high or low pH, (2) removes proteins in which target loops have explicit interactions with heteroatoms such as metals or ions, and (3) omits proteins with low-resolution crystal structures in the target loop region which have large measured B-factors.[16] We consider both the filtered and unfiltered decoy sets in this work. Additional problems we encountered were proteins missing from the download source,[35] proteins with segments of missing structure, and nonstandard amino acid in the sequence (see the Supporting Information). Another individual case that we faced was 1DAD_1, which is part of the 11-residue filtered set. It is a dimer protein, and the target loop is at the interface between the two monomers. Since there is only one monomer used in the decoys, and it is ambiguous whether the destabilization caused by the absence of the other monomer would affect the prediction on the target loop, we decided to remove it from the filtered set but keep it in the unfiltered group.

We also consider a set of protein loop decoys recently generated by Rossi et al.[26] In that study, they evaluated the performance of the four commonly used commercial public packages, Prime, Modeller, Sybyl, and ICM, on protein loop modeling; the former two use ab initio loop generation methods coupled with an energy function, and the later ones

Table 1. Jacobson Decoy Set[16] [a]

loop length decoy set	number of protein in unfiltered set	number of protein in filtered set	average number of decoys per protein	decoy RMSD range
4	37	32	300	0.22–1.35
6	141	94	400	0.24–2.98
8	93	60	600	0.32–4.64
10	49	27	900	0.32–5.80
11	31	14	1200	0.27–6.93
12	17	9	1100	0.36–8.96

[a] The RMSD range is the average over the most nativelike decoy as the lower bound and over the least nativelike decoy as the upper bound for each length set.

use knowledge-based techniques to generate loop fragments along with a statistical scoring function. For each tested protein loop, the four modeling packages were instructed to output their top six predictions and to separate the influence of the sampling methods from the scoring functions; two performance measurements were calculated. Top-rank-RMSD is the RMSD of the decoy assigned the lowest energy by the packages' scoring functions, and best-RMSD is the RMSD of the most nativelike decoy among the six predicted models. The latter variable was used to evaluate how well each sampling technique could search around the native basin, and the former one was used to measure if each scoring function could select the most nativelike models. We collected these decoys structures generated by the commercial packages from Rossi and co-workers[36] to test whether our energy function could better detect the most nativelike decoy as the one with the lowest energy compared to the four commercial loop packages.

In the end, we considered over 350 different protein loops with anywhere between 300 and 1200 decoys each; therefore, we believe that we have made a fair comparison to the previous works.[12,16,26,37,38]

2.3. Energy Minimization Procedure. Each loop decoy is locally minimized using eq 1 with the BFGS (Broyden-Fletcher-Goldfarb-Shanno)[39] limited memory quasi-Newton method.[40] During energy minimization, only the atoms at the target loop are allowed to fluctuate, and the rest of the structure is held fixed. After convergence, the backbone heavy atoms (N, Cα, C, O) of the target loop are used to calculate the RMSD between the experimental native and each energy-minimized decoy structure, while two structures are aligned along the nontarget-loop region. We also consider the backbone RMSD values calculated as done in PLOP/Prime based on only three of the backbone heavy atoms (N, Cα, C).[41] All of the RMSD calculations were conducted using MMTSB Tool Set.[42]

2.4. Evaluation Procedure. For the decoys generated by Jacobson et al., we followed the same criteria for success used previously,[12,16] which is, for each protein loop, to take the decoy with the lowest energy ranking and determine its RMSD with respect to the native structure. These RMSD values are then averaged over all proteins in the given loop size set. Note that in some cases the lowest ranked structure is sometimes the native loop in our study, although we follow the procedure of previous studies of not including it in the averages.

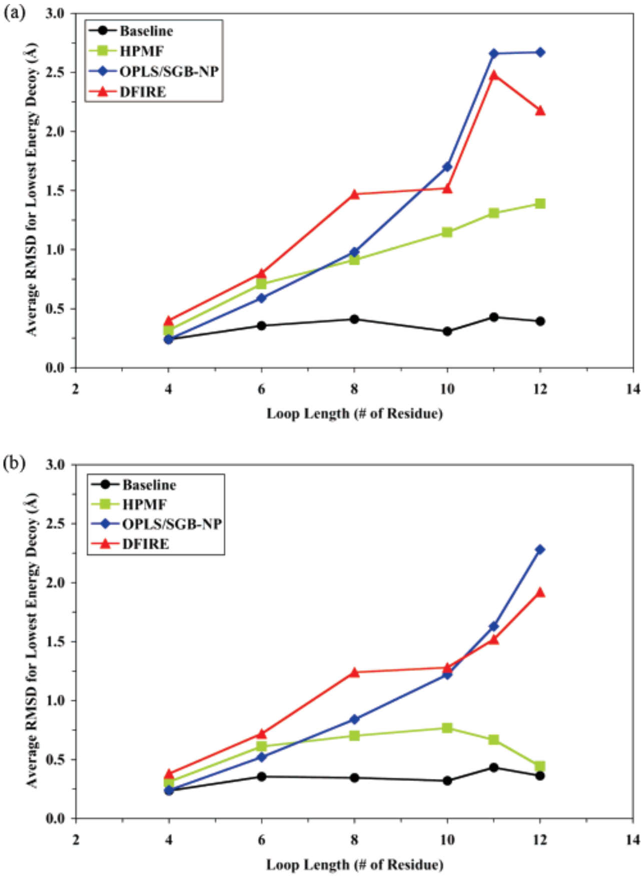


Figure 1. Performance of DFIRE[12] (triangle), OPLS-SGB/NP[16] (diamond), and HPMF[28] (square) energy functions on native loop prediction compared to a baseline of experimental uncertainty (circle): (a) unfiltered decoys and (b) filtered decoys. For comparative purposes we follow the same process for assessing the prediction results as the previous work.[12,16]

For the test on the structures generated by Rossi et al., we adopted the same two variables used in their study to examine our energy function's ability for selecting the most nativelike models. However, we took a slightly different approach when we studied the structures generated by Sybyl and ICM. The structures generated by these two methods sometimes have steric clashes, and those structures were removed if the steric collision could not be resolved using energy relaxation with constrains. Among the remaining structures, we calculated the new average top-rank-RMSD and the new average best-RMSD and used them for the evaluation.

3. Results

3.1. Jacobson Decoy Results. Figure 1 presents our results compared to the OPLS/SGB-NP[16] and DFIRE[12] energy functions on the decoy sets for 4- through 12-residue loops. Figure 1a shows the unfiltered decoy results in which it is evident that the different energy functions perform comparably at short loop lengths, with the OPLS/SGB-NP showing the best performance, 24% and 16% lower than HPMF at 4-residue and 6-residue sets, respectively. For 8-residue and longer loops, the HPMF energy function clearly improves the ability to discriminate against non-nativelike decoys. All

Table 2. Comparison of DFIRE,[12] OPLS-SGB/NP,[16] and HPMF[28] in Regards to Backbone RMSD of the Lowest Energy Decoy against the Experimental Native Structure Averaged over All Proteins in the Loop Decoy Sets of Different Length[a]

				filtered decoys (average lowest RMSD in Å)			
	unfiltered decoys (average lowest RMSD in Å)					HPMF	
loop length decoy set	DFIRE	OPLS SGB-NP	HPMF	DFIRE	OPLS SGB-NP	N,Ca,C,O	N,Ca,C
4	0.40	0.24	0.32 (0.13)	0.38	0.24	0.31 (0.12)	0.25
6	0.80	0.59	0.71 (0.58)	0.72	0.52	0.61 (0.46)	0.52
8	1.47	0.98	0.91 (0.81)	1.24	0.84	0.70 (0.57)	0.62
10	1.52	1.70	1.15 (1.36)	1.28	1.22	0.77 (0.71)	0.68
11	2.48	2.66	1.24 (1.09)	1.52	1.63	0.67 (0.51)	0.59
12	2.18	2.67	1.39 (1.31)	1.92	2.28	0.39 (0.20)	0.36

[a] The results in parentheses show the RMSD over backbone heavy atoms for decoys compared to an energy relaxed native loop structure.

Table 3. Demonstration of the Significant Contribution of HPMF on the Overall Performance of the Energy Function on 11- and 12-Residue Filtered Set

	filtered decoys (average lowest RMSD in Å)	
loop length decoy set	AMBER99+GB	AMBER99+GB+HPMF
11	0.93	0.67
12	0.40	0.39

energy functions show the same behavior of increasing RMSD error with increasing loop size, but for the largest 12-residue loop set the performance of our energy function provides improvements of about 35−50% over the earlier efforts, dropping from an average RMSD deviation from native prediction of ~2.20−2.65 Å to ~1.40 Å. The results are even more dramatic on the filtered set (Figure 1b). Even though the DFIRE and OPLS/SGB-NP energy functions drop about 10−40% in RMSD in the filtered set relative to their performance on the unfiltered set, their trends still show a larger average RMSD as the loop size increases. Our energy function improves from 10 to 70% on the filtered set, and furthermore it shows no strong dependence on the quality of results with the length of target loops. We believe that HPMF outperforms other energy functions at longer loops because of greater solvent exposure. The exact values are tabulated in Table 2. In the same table we also report our results when compared to the minimized native structure and when the backbone RMSD is calculated over only three of the backbone heavy atoms (N, Cα, C) as done in Prime/PLOP.[41]

To demonstrate the significant contribution from HPMF, Table 3 shows the comparison between the results on the 11- and 12-residue filtered loop set of two energy functions, AMBER99+GB and AMBER+GB+HPMF. For the 11-residue loop set, the average lowest RMSD is 0.93 Å and 0.68 Å for AMBER+GB and AMBER+GB+HPMF, respectively. The performance of the energy function including HPMF is 28% better than the other. However, there is no notable difference between the accuracy of the two energy functions for the 12-residue loop set. The subtle difference raises the question as to the sample size of the 12-residue loop set. More 12-residue loop proteins and longer loop sets are needed to further demonstrate the important contribution of HPMF in the overall performance of the energy function.

3.2. Loop Modeling Decoy Results. We tested our energy function on the long loops, 10-, 11-, and 12-residue, reported in Rossi et al.'s work[26] to demonstrate that, for modeling long protein loop, our energy function can consistently select a more nativelike model than the other energy and scoring functions in the popular modeling packages they tested. Similar to the complications that we encountered in the Jacobson decoy set, we needed to remove 1LUC from the 12-residue set and 1CVL and 2ENG from the 11-residue set due to missing fragment structures. We also removed 1FUS from the 11-residue set due to a nonstandard amino acid.

Figure 2 reports the results of our energy function for selection of nativelike loops compared to the predictions of Prime, Modeller, Sybyl and ICM. Because Sybyl and ICM do not filter out structures with steric conflicts, we eliminated those decoys in which energy relaxation (with harmonic constraints on all atom positions) did not resolve the steric collisions to define a sensible energy for comparison. The results show that our energy function can select more nativelike decoys as the lowest energy structures for each length set, such that the average RMSD is 10−40% lower than the predictions made by the standard loop modeling packages (Figure 2a). In addition, the prediction accuracy, which is measured when the most nativelike decoy is ranked as the lowest energy model, is significantly improved by our energy function (Figure 2b).

3.3. Challenges to X-ray Structures. Having established the improved performance of our energy function relative to previous reported physics and statistical potentials, we probe the relevance of these results. X-ray crystallography refinement combines structure factor data with a basic physiochemical model for chain connectivity and excluded volume to derive atom placements in a reported protein three-dimensional structure. The question is how good is this one native structure for assessing loop structure prediction accuracy? This implicit modeling aspect of the experimental data can be tested by optimizing the loop of the native structure, which we do under the usual assumption of a rigid context of the remaining protein. Table 2 shows the results when minimized loops are compared to the minimized native

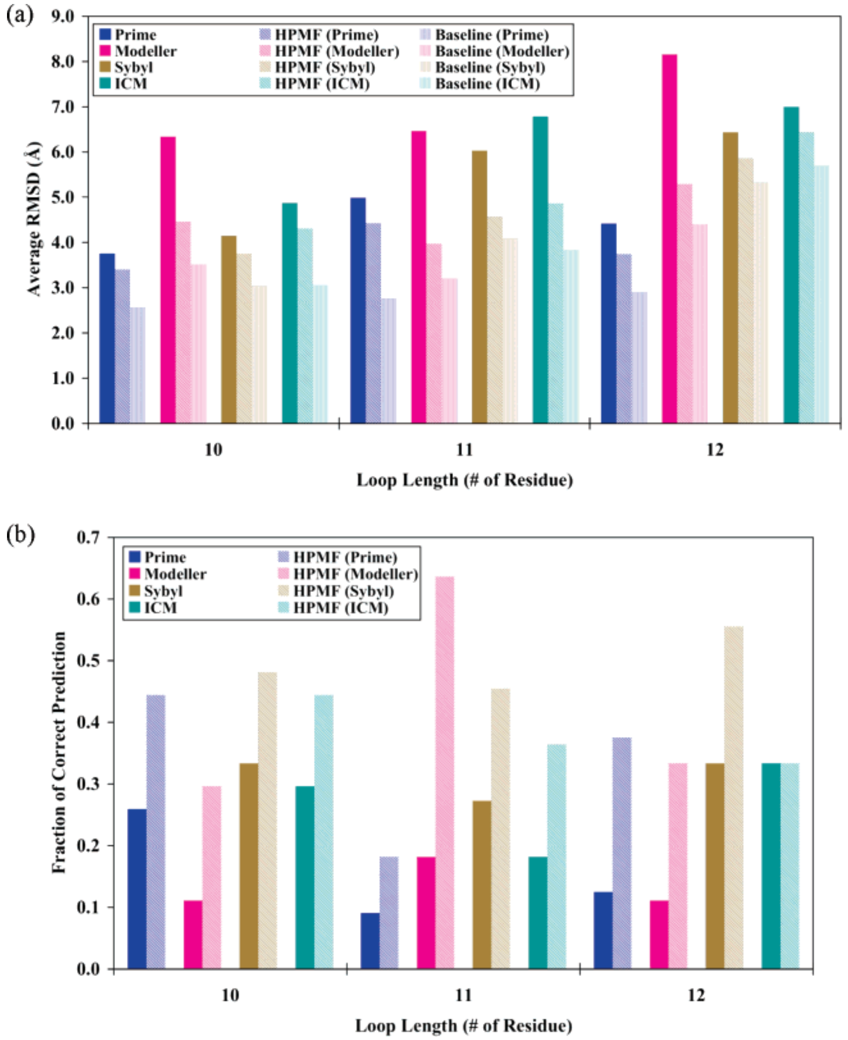


Figure 2. Results of HPMF on the decoy structures (10-, 11-, and 12-residue set) generated by Rossi et al.[26] compared with the modeling packages. The colors blue, pink, brown, and cyan are the decoy models generated by Prime, Modeller, Sybyl, and ICM, respectively. The patterns in the plots, filled, striped, and spotted, represent the average results of each commercial package, HPMF, and the baseline, respectively. (a) It shows the average top-rank RMSD of each energy function. (b) It shows the fraction of correct prediction, which is the percentage of loops that the most nativelike model is ranked the lowest energy.

loop, holding the rest of the protein fixed. Regardless of loop size, and regardless of whether unfiltered or filtered, the average RMSD change between the unoptimized and the optimized native loop structure is 0.4−0.5 Å. Correspondingly, the decoys show an average RMSD change between unoptimized and optimized loop structures of ~0.1 Å. Thus we regard the native state as being uncertain to with ~0.3−0.4 Å due to model effects in deriving the X-ray crystallography structure. We emphasize this point by defining a baseline function as the difference between the RMSD of the unoptimized and the optimized native structure and the RMSD of the unoptimized and the optimized lowest energy decoy, for each decoy set size, as we show in Figure 1. In fact for 4- and 12-residue filtered decoys, the prediction is as good as the baseline function. A similar strategy was developed by Rossi and co-workers[26] in which they determined a permissible range of RMSD values for defining the native loop.

4. Conclusion

In this work we have demonstrated the performance of a new protein and implicit solvent model on native protein loop discrimination from non-native decoys. The novelty of our energy function is the stabilizing effect of hydrophobic exposure to aqueous solvent that defines the HPMF hydration physics, which in principle should be an important stabilizing factor for loop conformations of a protein that typically are more solvent exposed. While our results for short loop decoy sets are comparably good to existing energy functions, we find substantial improvements for loop lengths of 8-residues and greater and find that the quality of our predictions are largely insensitive to the length of the target loop on the filtered decoy sets.

Recent work[37] reported a modification of the OPLS/SGB-NP energy by adding an ad hoc correction to the nonpolar solvation term to model the favorable free energy gained by placing hydrophobic side chains of the target loop in the hydrophobic space created when the target loop is removed, a term that has been successfully used in ligand-protein binding. This is conceptually similar to the more physically grounded hydrophobic interaction model developed in our recent work,[28] and Zhu and co-workers showed demonstrable improvement in loop prediction on a filtered collection of 11- to 13-residue loops over the standard OPLS/SGB-NP energy function.[37] However aspects of the energy model could stand improvement. While they report an average RMSD of 1.00 Å and 1.15 Å for *filtered* 11- and 12-residue loop decoys, our average RMSD on the *unfiltered* decoy set is comparable to these results and clearly superior on similarly filtered decoy sets, reducing the error in native loop selection by over a factor of 2 (Table 2).

More recently, Zhu et al.[38] have improved their energy model by varying the protein internal dielectric constant in the electrostatics energy calculation based on the environment of interacting atoms and by further optimizing the hydrophobic term that they introduced in their previous work.[37] In their latest work,[38] they have achieved the prediction accuracy of 0.39 Å, 0.68 Å, 0.80 Å, and 1.00 Å RMSD for 6-, 8-, 10-, and 13-residue loops, respectively. However the performance criteria was changed from previous studies since they calculate the RMSD of the predictions with respect to the energy-minimized native structures. This concurs with the baseline function used here or the range of permissible native RMSDs used in Rossi et al., that physics-based predictions are better compared to energy relaxed experimental X-ray crystallographic structures.

In the study by Nayeem and co-workers,[43] the authors compared seven homology model building software packages that encompass both knowledge- and physics-based approaches for sequence alignment through to structural loop generation models. One of the tested commercially available software packages, Prime, which uses the ab initio loop generation and physically based energy function[16] was found to be one of the best performers in the study. While Prime showed comparable performance to other software packages for proteins that exhibit sequence identity with known proteins of greater than 50%, Prime showed demonstrable superiority over other approaches as sequence identity decreased.[43] However, a more recent study[26] found that all commercial software packages, including Prime, returned a best energy ranked loop that was rarely the most nativelike of their generated ensemble, revealing a weakness in the energy functions used in selecting nativelike loop conformers. Given the improvement of our energy function over all versions of the OPLS/SGB-NP results,[16,37,38] this result emphasizes that existing commercially available software packages would benefit from improvements in scoring of more nativelike loops by use of the energy function we have developed[28] and which we have tested further for loop prediction in this work.

While our results are “better” than other reported results, we raise questions about the testing procedures that assume the experimental native loop structure is the relevant gold standard for prediction. This is consistent with the recent recognition of the importance of structural ensembles for reporting X-ray crystallographic structures to describe the functional native state, as opposed to the dogma of one native structure.[44] Better measures of structural ensembles consistent with the experimentally derived structure factors could more meaningfully discriminate for or against prediction success, especially for the unfiltered decoys where experimental procedure is clearly a source of uncertainty. A tractable energy function in turn might usefully aid in the experimental refinement procedure for deriving atomic protein models from structure factor data. In a recent study,[45] loop prediction methods such as Prime have been demonstrated to improve the protein structure refinement protocol in NMR experiments. The results of the study evidently indicate that our energy function coupled with a sampling technique not only can be used in theoretical research but also can facilitate experimental studies in solving protein structures.

One of the goals of the structural genomics initiative is to rapidly obtain native structural models from X-ray crystallography and NMR spectroscopy for representative members of protein families.[46] Once the protein structure of a family member is solved, in principle it can serve as a structural template to develop comparative structural models of other family members.[46,47] However, comparative modeling structures that have a *functional* value typically require working with a structural template that exhibits greater than 40−60% sequence identity.[48,49] Below that, function begins to diverge due to structural differences on the protein surface that typically correspond to loop regions.[6,13] We hope that this energy function provides a better means for loop optimization important in recognizing the structural and functional differences between homologous and distantly related proteins to aid comparative modeling efforts.

Acknowledgment. T.H.G. gratefully acknowledges financial support under the NSF-Cyber program. M.S.L. is supported under the DOE/MICS program. We would also like to thank Dr. Karen A. Rossi for providing us the structures generated in her study and Dr. Jonathan E. Kohn’s help on creating the graphic for the issue cover. This research 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. DEAC02-05CH11231.

Supporting Information Available: List of protein loops removed from this work due to missing from the download source, having missing fragments, or having nonstandard amino acids. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Viguera, A. R.; Serrano, L. *Nature Struct. Biol.* **1997**, *4*, 939−946.

(2) Fetrow, J. S. *FASEB J.* **1995**, *9*, 708−717.

(3) Liu, J.; Tan, H.; Rost, B. *J. Mol. Biol.* **2002**, *322*, 53−64.

(4) Blouin, C.; Butt, D.; Roger, A. J. *Protein Sci.* **2004**, *13*, 608−616.

(5) Efimov, A. V. *Curr. Opin. Struct. Biol.* **1993**, *3*, 379−384.

(6) Fiser, A.; Do, R. K.; Sali, A. *Protein Sci.* **2000**, *9*, 1753−1773.

(7) Tramontano, A. *FEBS J.* **2007**, *274*, 1651−1654.

(8) Burke, D. F.; Deane, C. M. *Protein Eng.* **2001**, *14*, 473−478.

(9) Espadaler, J.; Fernandez-Fuentes, N.; Hermoso, A.; Querol, E.; Aviles, F. X.; Sternberg, M. J.; Oliva, B. *Nucleic Acids Res.* **2004**, *32*, D185−188.

(10) Espadaler, J.; Querol, E.; Aviles, F. X.; Oliva, B. *Bioinformatics (Oxford, England*) **2006**, *22*, 2237−2243.

(11) Fernandez-Fuentes, N.; Oliva, B.; Fiser, A. *Nucleic Acids Res.* **2006**, *34*, 2085−2097.

(12) Zhang, C.; Liu, S.; Zhou, Y. *Protein Sci.* **2004**, *13*, 391−399.

(13) Fiser, A.; Feig, M.; Brooks, C. L., III; Sali, A. *Acc. Chem. Res.* **2002**, *35*, 413−421.

(14) DePristo, M. A.; de Bakker, P. I.; Lovell, S. C.; Blundell, T. L. *Proteins* **2003**, *51*, 41−55.

(15) de Bakker, P. I.; DePristo, M. A.; Burke, D. F.; Blundell, T. L. *Proteins* **2003**, *51*, 21−40.

(16) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; Shaw, D. E.; Friesner, R. A. *Proteins* **2004**, *55*, 351−367.

(17) Rohl, C. A.; Strauss, C. E.; Chivian, D.; Baker, D. *Proteins* **2004**, *55*, 656−677.

(18) Canutescu, A. A.; Dunbrack, R. L., Jr. *Protein Sci.* **2003**, *12*, 963−972.

(19) Coutsias, E. A.; Seok, C.; Wester, M. J.; Dill, K. A. *Int. J. Quantum Chem.* **2005**, *106*, 176−189.

(20) Shehu, A.; Clementi, C.; Kavraki, L. E. *Proteins* **2006**, *65*, 164−179.

(21) van den Bedem, H.; Lotan, I.; Deacon, A. M.; Latome, J.-C. *Algorithmic Found. Rob.* **2005**, 345−360.

(22) Deane, C. M.; Blundell, T. L. *Protein Sci.* **2001**, *10*, 599−612.

(23) Du, P.; Andrec, M.; Levy, R. M. *Protein Eng.* **2003**, *16*, 407−414.

(24) Monnigmann, M.; Floudas, C. A. *Proteins* **2005**, *61*, 748−762.

(25) Pellequer, J. L.; Chen, S. W. *Biophys. J.* **1997**, *73*, 2359−2375.

(26) Rossi, K. A.; Weigelt, C. A.; Nayeem, A.; Krystek, S. R., Jr. *Protein Sci.* **2007**.

(27) Smith, K. C.; Honig, B. *Proteins* **1994**, *18*, 119−132.

(28) Lin, M. S.; Fawzi, N. L.; Head-Gordon, T. *Structure* **2007**, *15*, 727−740.

(29) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049−1074.

(30) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383−394.

(31) Crivelli, S.; Eskow, E.; Bader, B.; Lamberti, V.; Byrd, R.; Schnabel, R.; Head-Gordon, T. *Biophys. J.* **2002**, *82*, 36−49.

(32) Head-Gordon, T.; Brown, S. *Curr. Opin. Struct. Biol.* **2003**, *13*, 160−167.

(33) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517−529.

(34) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474−6487.

(35) Jacobson, M. P. http://francisco.compbio.ucsf.edu/∼jacobson/decoy.htm (accessed December 3, 2006).

(36) Rossi, K. A. Private communication.

(37) Zhu, K.; Pincus, D. L.; Zhao, S.; Friesner, R. A. *Proteins* **2006**, *65*, 438−452.

(38) Zhu, K.; Shirts, M. R.; Friesner, R. A. *J. Chem. Theory Comput.* **2007**, *3*, 2108−2119.

(39) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: 1992.

(40) Liu, D. C.; Nocedal, J. *Math. Programm.* **1989**, *45*, 503−528.

(41) Jacobson, M. P. http://francisco.compbio.ucsf.edu/∼jacobson/plop_manual/plop_manual.htm (accessed July 28, 2007).

(42) Feig, M.; Karanicolas, J.; Brooks, C. L., III *J. Mol. Graphics Modell.* **2004**, *22*, 377−395.

(43) Nayeem, A.; Sitkoff, D.; Krystek, S., Jr. *Protein Sci.* **2006**, *15*, 808−824.

(44) Furnham, N.; Blundell, T. L.; DePristo, M. A.; Terwilliger, T. C. *Nature Struct. Mol. Biol.* **2006**, *13*, 184−185.

(45) Rapp, C. S.; Strauss, T.; Nederveen, A.; Fuentes, G. *Proteins* **2007**, *69*, 69−74.

(46) Chandonia, J. M.; Brenner, S. E. *Science* **2006**, *311*, 347−351.

(47) Head-Gordon, T.; Wooley, J. C. *IBM Syst. J.* **2001**, *40*, 265−296.

(48) Tian, W.; Skolnick, J. *J. Mol. Biol.* **2003**, *333*, 863−882.

(49) Todd, A. E.; Orengo, C. A.; Thornton, J. M. *J. Mol. Biol.* **2001**, *307*, 1113−1143.

CT700292U

JCTC Journal of Chemical Theory and Computation

Is the Induction Energy Important for Modeling Organic Crystals?

Gareth W. A. Welch,[†] Panagiotis G. Karamertzanis,[†] Alston J. Misquitta,[†,‡]
Anthony J. Stone,[‡] and Sarah L. Price*[,†]

Christopher Ingold Laboratory, Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, U.K., and University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, U.K.

Received October 16, 2007

Abstract: We compare two methods for estimating the induction energy in organic molecular crystals by approximating the charge density polarization in the crystalline state. The first is a distributed atomic polarizability model combined with distributed multipole moments, derived from ab initio monomer properties. The second uses an ab initio calculation of the molecular charge density in a point-charge field. Various parameters of the models, such as the rank of polarizability model, effect of self-consistent iterations, and damping, are investigated. The methods are applied to a range of observed and predicted crystal structures of three particularly challenging molecules, namely oxalyl dihydrazide, 3-azabicyclo[3,3,1]nonane-2,4-dione, and carbamazepine, as well as demonstrating the importance of induction in the naphthalene crystal. The two models agree well considering the different approximations made, and it is shown that the induction energy can be an important discriminator in the relative lattice energies of structures with substantially different hydrogen-bonding motifs.

1. Introduction

The importance of the induction energy for modeling crystal structures has been the subject of much debate.[1−5] In ionic crystals, the empirical shell model has long been used to model the induced dipole from the strong electrostatic fields.[6,7] The enhanced dipole moment of water in the liquid state has led to a plethora of model intermolecular potentials that include a simple polarizability model.[1] Indeed, the development of the tinker force field for biological modeling now includes a dipole polarizability term,[8,9] with the atomic polarizabilities derived by Thole.[10]

Reliable methods of estimating the polarizability models are demanding for two reasons: first they require a large basis set and high quality wavefunctions to obtain converged polarizability tensors, and second, for all but the smallest of molecules, an accurate description of the molecular polarizability can be obtained only with a distributed polarizability model. Until recently, these requirements have posed an almost insurmountable problem for modeling the induced moments in organic molecules. Furthermore, there has been uncertainty as to whether an approximate polarizability model would lead to greater errors than complete neglect, and almost all organic crystal structure modeling has been performed with model intermolecular potentials that do not include an explicit polarizability term.[11−14] Instead, polarizability effects are to some degree approximately absorbed in the empirically fitted repulsion-dispersion parameters. Notable exceptions are the ab initio potentials developed for crystal structure prediction studies of simple alcohols and alkanes,[5,15] glycol and glycerine,[16] and some self-consistent molecular mechanics work on peptides.[8,17]

There is mounting evidence that induction effects are important within crystal structures of even nonpolar molecules. The PIXEL method, evaluating lattice energies by integrating over semiempirical functions of the in vacuo electron densities placed in the crystal lattice, shows that the induction energies are significant.[4,18−20] Furthermore, experimental analysis of the naphthalene crystal shows

* Corresponding author e-mail: s.l.price@ucl.ac.uk.
[†] University College London.
[‡] University Chemical Laboratory.

10.1021/ct700270d CCC: $40.75 © 2008 American Chemical Society
Published on Web 02/19/2008

evidence of significant polarization on the charge density of the naphthalene molecule.[21] However, the quantification of the induced moments from the difference between the molecular charge density in the crystal and in isolation is hampered by the uncertainty in how crystal charge density should be partitioned between the constituent molecules. This highlights a major difficulty with the evaluation of induction energies in molecular crystals: the charge distributions of molecules in van der Waals contact overlap so much that the long-range models of polarizability may not be valid.

The induction energy is best understood from the theory of intermolecular forces. In particular, the recently developed symmetry-adapted perturbation theory using density function theory (SAPT(DFT))[22−24] provides us with a computationally efficient and accurate method for calculating the induction energy of dimers of small polyatomics. The induction energies from SAPT(DFT) include penetration and charge-transfer effects and therefore provide us with an important benchmark against which approximations can be tested.[25,26]

However, being a two-body theory, SAPT(DFT) does not allow us to estimate the induction energy of the condensed phase. One way of doing this is to use polarizability models. Recently, some of us have developed a method for obtaining distributed polarizability models that is well suited for small polyatomic molecules of around 30 atoms. The Williams-Stone-Misquitta (WSM) method[25,26] allows us to obtain distributed polarizabilities from the ab initio properties of isolated molecules that are optimal at a given rank. From comparisons with SAPT(DFT) induction energies of a variety of dimers, ranging from HF to benzene,[26] we know that the damped WSM models are able to describe not only the long-range induction energy but also an induction energy at short-range, even in the most testing area of hydrogen-bonding contacts. These models result in errors of 2−7% of the dimer interaction energy at typical contact distances. The error would be larger if we included hyperpolarizability effects that are not included in the WSM models. However, for condensed phases the errors are smaller than for van der Waals dimers because of the large number of longer range interactions, for which the WSM models are extremely accurate. Therefore, these polarizability models give us a very powerful tool for computing the induction energy of an organic crystal.

Yet another way of approximating the induction energy of the crystalline phase relies on the ab initio evaluation of the molecular charge density, with the field of the surrounding molecules represented by point charges. When done self-consistently, we obtain an electronic response to point charge field model (SCERP) which does include some of the effects of electron penetration, because the point charges are fitted to the electrostatic potential close to the van der Waals[27] surface. But this model is limited by the accuracy that can be attained by the point charge model and the lack of charge-transfer effects. Yet, once again, these are short-range effects and, for reasons explained above, are not expected to make a significant contribution to the interaction energy of the crystal.

The WSM polarizability model has been validated for dimers against SAPT(DFT) energies.[26] SCERP provides an independent test of the polarizability models in the condensed phase. This paper uses these models to estimate the induction effects in a range of molecular crystals. Our purpose is to establish whether the contributions are sufficiently important that we should implement polarizability models in the crystal structure modeling program DMAREL.[28,29]

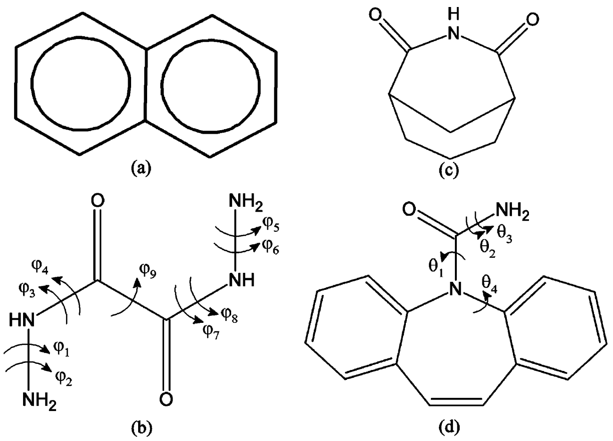


Figure 1. Molecules used in this investigation: (a) naphthalene, (b) oxalyl dihydrazide, (c) 3-azabicyclo[3,3,1]nonane-2,4-dione, and (d) carbamazepine. Double arrows indicate that two atoms independently have a torsion angle defined along the same bond.

The overall aim of the paper is to determine the importance of the induction energy in organic crystal structures, particularly its relevance to the field of organic crystal structure and polymorph prediction,[30] whose promise for aiding the design of new materials and the selection of solid form for pharmaceutical development[31] is severely compromised by uncertainties in the estimation of relative lattice energies. Four contrasting examples are considered for which the molecular structures are shown in Figure 1. The naphthalene (1a) crystal is investigated as a nonpolar system. Charge density studies[21] have shown a change in the electron distribution in the region of the C−H bond involved in a C−H···π interaction in the crystal structure. The other examples are all tests of the differences in induction energy corresponding to different types of hydrogen bonding, as the electrostatic fields involved in hydrogen bonding are among the strongest in crystal structures of neutral organic molecules. The relative induction energies of the 5 different polymorphs of oxalyl dihydrazide[32] (1b) are examined because of the plurality of hydrogen-bonding geometries sampled, including one with a significant intramolecular component. The relative induction energies for sets of experimentally observed and hypothetical crystal structures of 3-azabicyclo[3,3,1]nonane-2,4-dione (1c) and carbamazepine (1d) are computed, to investigate whether modeling induction could improve the prediction of relative lattice energies of crystal structures based on doubly hydrogen-bonded dimers or chain motifs. In both cases, the predictions that the two types of crystal structure were energetically competitive inspired extensive polymorph screening studies to search for the alternative motif.[33,34] For carbamazepine, all known polymorphs are based on a doubly hydrogen-bonded amide dimer (although it does adopt a catemer in a solid solution[35]), whereas the catemer is marginally more stable according to current modeling.[36−38] On the other hand, 3-azabicyclo[3,3,1]nonane-2,4-dione adopts an imide catemer in all its solid forms,[34] although many of the participants in the 2001 international blind test of crystal structure prediction[39] predicted a dimer structure as more stable.

We first define the new computational methods for estimating the induced distributed moments and induction energy contribution to the lattice energy, before applying the two models to this range of issues in organic solid-state chemistry.

2. Method

Using a finite cluster of molecules, sufficiently large to obtain converged electrostatic energies, we estimate the effect of the crystalline environment on the molecular multipole moments calculated using the distributed polarizability and the SCERP models. As will be described below, the modified molecular multipole moments are then used to estimate the induction contribution to the crystal lattice energy. We first describe the cluster model, then the two methods of estimating the induced moments, and finally the method of evaluating the induction energy in the lattice.

2.1. Choice of Crystal Structures and Cluster and Molecular Models. For our calculations we use centrosymmetric crystal structures, from which the clusters are built. Numerical experimentation has shown that a cluster in which a central molecule is surrounded to at least 15 Å in all directions is large enough to converge the electrostatic energy of a molecule in the center, to that of an infinite lattice calculation using DMAREL. This typically means using a cluster of 5 × 5 × 5 unit cells.

The crystal structure used for naphthalene was the 100 K X-ray structure.[21] The molecular structure was optimized in vacuo at the MP2 6-31G** level and then pasted into the experimental structure by minimizing the rms overlap of the carbon atoms. Finally the crystal structure was relaxed using DMAREL with distributed multipoles derived using the same charge density as for the distributed polarizability model.

For oxalyl dihydrazide, the five experimental crystal structures[32] were refined to account for the X-ray determination of the proton positions that are important in the plurality of the hydrogen-bonding in these crystals.[40] This DMAFLEX refinement[36] optimized the lattice energy, including the MP2 6-31G** intramolecular conformational penalty, with respect to the nine torsions shown in Figure 1, and the crystallographic cell parameters and molecular positions. All covalent bond lengths and angles apart from the explicit torsion angles were reoptimized in the ab initio intramolecular calculation at each step. The key difference between the *inter*- and *intra*molecular bonding in the polymorphs (Figure 2) has been preserved, though the model for the ϵ polymorph is more dense that the experimental structure, resulting in one short N···N distance of 2.73 Å. The rms difference[41] between these refined structures and the experimental crystal structures was about 0.2 Å for the α, γ, and ϵ polymorphs and less than 0.6 Å for β and δ, for all non-hydrogen atoms in a 15-molecule cluster (see Table S1). The cluster sizes were varied (9 × 7 × 5 for α and ϵ, 7 × 5 × 7 for γ and δ, and 9 × 5 × 9 for β polymorphs) to give suitable supercells containing between 490 and 980 molecules that conformed to our requirement of 15 Å of material surrounding the polarizable molecule. The intermolecular electrostatic energy for all of these clusters is within 0.5 kJ mol^{-1} of the infinite lattice value (Supporting Information Table S2).

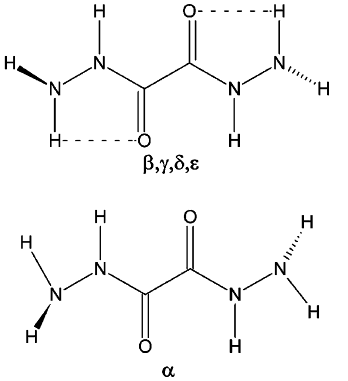


Figure 2. The two major intramolecular conformations of oxalyl dihydrazide. The β, γ, δ, and ϵ polymorphs contain stretched intramolecular hydrogen bonds, indicated by a dashed line. The torsion angles for all five polymorphs are given in Table S1, Supporting Information.

The bicyclic structure of 3-azabicyclo[3,3,1]nonane-2,4-dione[42] makes it essentially rigid; therefore, we used the in vacuo MP2 6-31G** optimized molecular structure. A set of 8 low-energy crystal structures[34] generated using this molecular conformation was considered to represent a range of packing arrangements within 3 kJ mol^{-1} of the global minimum lattice energy. We also examined the minimum obtained with the same computational model, starting from the 297 K experimental crystal structure.[42] The set of structures include both the observed catemer and doubly hydrogen-bonded dimer motifs in a range of space groups. The 5 × 5 × 5 unit cell clusters contained 250, 500, or 1000 molecules.

For carbamazepine,[33] we used DMAFLEX to relax the positions of the amide protons and torsion angles identified in Figure 1, for fifteen low-energy crystal structures[36] obtained from a previous search.[33] These structures covered a wide range of packings including those corresponding to known forms II, III, and IV.

The carbamazepine clusters used in the polarizability calculations consisted of 5 × 5 × 5 unit cells and contained 250 to 1000 molecules.

2.2. Distributed Polarizability Model for Induced Moments. For large molecules, the single-center multipole expansion may not converge, and it is necessary to distribute the polarizabilities[43−45] in an analogous way to the multipole moments. In general, the distributed polarizability α is dependent on the response of a moment at site *a* to a field at another site *a′* in the same molecule. These nonlocal distributed polarizabilities can be obtained very accurately, and in a computationally efficient way, using the methods developed by Misquitta and Stone.[46] However, the summation over two-site terms makes polarizability calculations for large molecules and clusters expensive computationally.

The nonlocal polarizability description may be simplified using a mathematical transformation to remove the nonlocal terms. Such a transformation, put forward by Le Sueur and

Stone,[47] effectively removes any explicit intramolecular polarization terms, resulting in a local distributed polarizability description that depends only on the electrostatic field exerted directly by the surrounding molecules. This is done at the expense of a loss in accuracy[46,47] that increases with molecular size and rank of the polarizabilities. However, the localized polarizability description is still reasonably accurate, and the terms themselves are in good correspondence with what we might expect from physical arguments.

Using these polarizabilities as anchors, the method described by Williams and Stone[48] can now be used to refine the model. The refinement is done by tuning the distributed polarizabilities to reproduce the responses to point charges placed on a grid around the molecule. This final step in the WSM method[25,26] for obtaining a local, distributed polarizability model partially absorbs the effects of higher rank polarizabilities in the lower rank terms. For this work, the distributed polarizabilities are localized to either rank 1 or rank 2 and refined by fitting to responses computed using linear-response DFT at randomly generated points[26] between surfaces at 2 and 4 times van der Waals Bondi radii.[49] We denote the WSM models as being L1 (induced-dipole), L2 (L1 plus induced-quadrupole), and L2/L1 (L1 on hydrogen sites, L2 for all other sites). The fitting procedure includes a penalty function to discourage drifting from the localized values,[26] which has been found to result in accurate models and reduce the occurrence of unphysical values. All distributed properties, both polarizabilities and multipoles, have been derived using CamCASP,[50] from charge densities calculated with DALTON[51] using the asymptotically corrected[52] PBE0[53] functional in combination with the Sadlej basis set.[54]

Having obtained the distributed polarizabilities, the change to the in vacuo multipole moments by polarization effects due to the crystal environment can now be estimated. In our work we obtain distributed multipole moments up to rank 4 using the revised version of Distributed Multipole Analysis.[55] In the condensed phase the charge density distorts due to polarization effects, requiring an alteration in the multipole moments. The change in the multipole moments at a polarizable site *a* in molecule *A*, due to the static field of all other sites *b* in molecules *B* in the surrounding environment, is given by[56]

$$\Delta Q_t^a = -\sum_{B\neq A}\sum_{b\in B}\sum_{uv} \alpha_{tv}^a f_{uv}(\beta R_{ab}) T_{vu}^{ab}(Q_u^b + \Delta Q_u^b) \qquad (1)$$

where α_{tv}^a is the polarizability tensor for site *a*, quantifying the susceptibility of the multipole moment *t* on site *a* to be induced by the field arising from the static multipole moments Q_u^b and induced multipole moments ΔQ_u^b on all sites of other molecules. The subscripts *t*, *u*, and *v*, refer to the component of the multipole moments and run as *00,10,11c,11s...* T_{vu}^{ab} is the interaction tensor containing the distance and orientational relation between sites *a* and *b* and their multipole components *v* and *u*, and $f_{uv}(\beta R_{ab})$ is a Tang-Toennies damping function that is assumed to depend only on distance, a damping parameter β, and the rank of multipoles represented by *v* and *u*.

A damping function is used in an attempt to compensate for the divergence of the multipole expansion at small intersite distances. Little is known about damping functions for induction;[57] see ref 25 for a recent discussion. We use Tang-Toennies damping, which has been used to damp multipolar expansions of the dispersion energy. Examples show that it does not correct fully for the limitations of the multipolar model,[26] but no better form has been proposed. The Tang-Toennies damping function has the form[58]

$$f_{uv}(\beta R_{ab}) = 1 - \left(\sum_{k=0}^{n} \frac{(\beta R_{ab})^k}{k!}\right)\exp(-\beta R_{ab}) \qquad (2)$$

where *n* is the sum of the ranks of multipoles *u* and *v*, and has been effective in reducing the singular behavior of the induction energy when intersite distances are particularly short.[26] The damping expression is used in the calculation of the induced moments, using $\beta = 2\sqrt{2I_X}$, where I_X is the first vertical ionization potential in atomic units.[25] The values of β are as follows: oxalyl dihydrazide α 1.625; β 1.667; δ 1.650; ϵ 1.649; γ 1.657; naphthalene 1.547; carbamazepine 1.510; and 3-azabicyclo[3,3,1]nonane-2,4-dione 1.674.

Thus, the distributed polarizability model estimates the induced moments in the crystal using eq 1, implemented in ORIENT,[59] for a molecule, *A*, at the center of the cluster. At the first iteration with zero induced moments, the induction energy is $E_{ind,d-class}^{(2)}$ and corresponds, within the approximations implicit in the truncation of the multipole and polarization expansions, to the second-order energy in the Rayleigh−Schrödinger theory[60] for a 2-body system at large intermolecular separations

$$E_{ind,pol}^{(2)}(X) = \sum_{r\neq 0} \frac{|\langle \Phi_0^X|\hat{V}|\Phi_r^X\rangle|^2}{E_0^X - E_r^X} \qquad (3)$$

where Φ_r^X and E_r^X are the eigenstates and eigenvalues of the monomer Hamiltonian of molecule *X*, and $\hat{V}$ is the intermolecular electrostatic potential operator arising from the rest of the system. The suffixes *pol* and *d-class* indicate, respectively, the root of the induction energy term in perturbation theory and a damped classical polarizability model.

The coupled eqs 1 for the ΔQ are usually solved by iteration. After one iteration the energy becomes $E_{ind,d-class}^{(2-3)}$, where the change in energy corresponds to a third-order term in perturbation theory. After iteration to self-consistency, the induced multipoles correspond to all orders of the induction energy in the linear-response approximation. The total damped classical induction energy

$$E_{ind,d-class}^{(2-\infty)} = \sum_{k=2}^{\infty} E_{ind,d-class}^{(k)} \qquad (4)$$

was used to examine the effect of polarization on the relative stability of the crystal structures. Figure 3 demonstrates that this iteration procedure makes a significant difference to the calculated induction energy, stabilizing the crystal. The rank 1 model converges rapidly, but higher-ranking polarizabilities do require damping. In practice, Figure 3 shows that the infinite summation in (3) can be truncated, depending on the model, to 5−8 iterations, which is sufficient to achieve convergence of 0.5 kJ mol^{-1}.

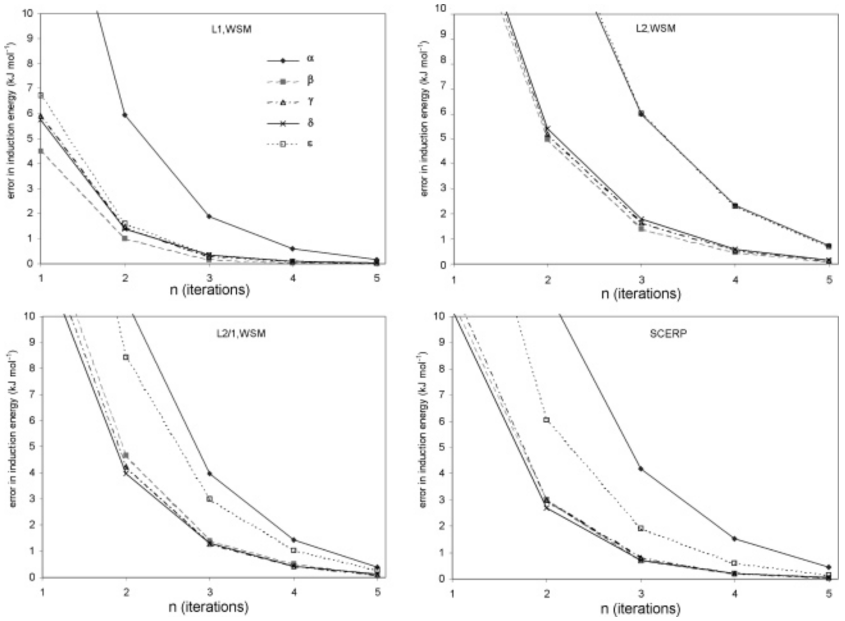


Figure 3. Convergence of $E_{ind,d-class}$ for polymorphs of oxalyl dihydrazide for several induction energy models. The plot shows the error, $E_{ind,d-class}^{(2-n+1)} - E_{ind,d-class}^{(2-\infty)}$, in the induction for different truncations of the infinite sum.

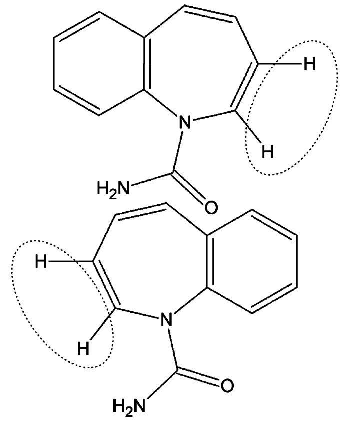


Figure 4. Fragments of carbamazepine used to calculate its atomic polarizabilities. The transferability of the polarizabilities calculated for these molecules to carbamazepine is described in the Supporting Information. The circled atoms are added in place of the 6-membered ring.

The carbamazepine molecule was too large for the WSM polarizability analysis due to computational limitations, and we adopted a different scheme for this molecule. The distributed multipoles were calculated in Gaussian03 using the nonasymptotically corrected PBE0 functional with the Sadlej basis set. The difference between the corrected and uncorrected functionals is insignificant for the calculation of electrostatic energies using distributed multipole moments, but the correction is essential for accurate polarizabilities. The polarizabilities were constructed from two molecular fragments, indicated in Figure 4. The structure of the fragments were held rigid at the MP2 6-31** in vacuo optimized geometry of carbamazepine, except the positions of the hydrogen atoms added in place of the 6-membered ring, which were optimized at the same level. Although polarizability is a molecular property, influenced by all sites, it has been necessary to make the approximation of transferability (see the Supporting Information) for the polarizabilities calculated for these smaller molecules to the larger DMAFLEX minimized structures.

2.3. Self-Consistent Electronic Response to Point Charge Field Model (SCERP). We also present an alternative method of evaluating the effect of induction on the charge distribution directly using the Gaussian03 ab initio package.[61] The CHELPG potential derived charges,[27] which are fitted to a grid of points between the van der Waals atomic radii and 2.8 Å from the nuclei, were obtained for the isolated molecule from an aug-cc-pVTZ charge density with the PBE0 functional. These charges were placed on all the atomic sites of the same clusters as described in paragraph 2.1, except the central molecule which is described using aug-cc-pVTZ basis functions. A DFT calculation using the PBE0 functional is conducted for this molecule within the cluster of charges. The polarized charge density was analyzed by GDMA2.2[55] to obtain $Q_u^b + \Delta Q_u^b$, and hence the induced multipole moments (up to hexadecapole) were obtained by subtraction of the multipoles obtained from the in vacuo calculation.

The potential derived charges of the polarized charge distribution were then used in a further cluster calculation, and the process was repeated until the calculated induction energy had converged as for the distributed polarizability model. Figure 3 shows that, as with the WSM model, iteration is required to capture a significant part of the energy and that around half a dozen iterations are sufficient for convergence within 0.5 kJ mol^{-1}.

This method is more computationally expensive than using the multipole expansion and cannot be used for lattice energy minimization but can be used for testing aspects of the polarizability model. The resources required are almost independent of the number of charges used, and so very large

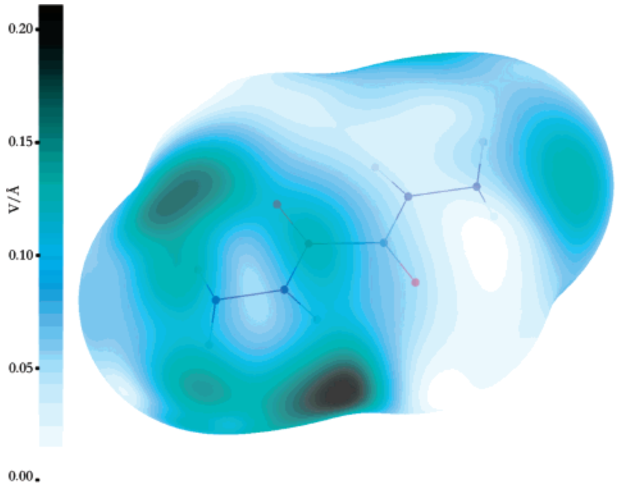


Figure 5. Electrostatic field difference around α oxalyl dihydrazide. The plot shows the norm of the difference between the electrostatic field vectors calculated from distributed multipole moments and from point charges. The surface is the van der Waals surface scaled by 1.8, which is accessible by the hydrogen-bonding protons. The maximum field difference displayed is 0.226 V/Å.

clusters could be used to check convergence with cluster size. The electron density is simply calculated in the fields of the surrounding background charges. Although the use of point charges to model electrostatic field is relatively crude, they are used here to induce a second-order response of the molecular charge density and within the self-consistent nature of the process. Penetration effects from the overlap of the charge distributions in the cluster are absent, except insofar as they are included in the fitting of the potential derived charges to points so close to the molecule.

If the polar hydrogen sites are considered to have a van der Waals radius of zero,[62] the region of interaction with surrounding nuclei in hydrogen bonding arrangement may be approximated by the van der Waals surface scaled by 1.8. In Figure 5 we present a comparison of the electrostatic field at this surface when calculated using multipole moments or point charges, in terms of the norm of the difference vectors at 19 814 points on the surface for α oxalyl dihydrazide. The mean difference is 0.08 V/Å (standard deviation 0.04 V/Å) which is less than 9% of the largest field, 0.92 V/Å, with the multipole moments. The highly localized nature of the error in the electrostatic field can be plainly seen in Figure 5, as dark blemishes around the hydrogen sites. For both the α and the ϵ polymorphs these regions coincide with the shortest hydrogen bonds seen in any of the crystal structures in this work, hence we anticipate the largest errors in our calculations to be for these crystals.

2.4. Calculation of the Induction Contribution to the Lattice Energy. We can evaluate the induction energy for a given crystal structure by the following method using the induced multipole moments that we have calculated by the methods described above. The following easily implemented method is not suitable for optimizing a crystal structure but allows a quick assessment of the importance of induction energy in organic molecular crystals. The classical polarization model for the induction energy is[25,56]

$$E_{ind,d-class}(A) = \frac{1}{2}\sum_{a\in A}\sum_{B\neq A}\sum_{b\in B}\sum_{tu}\Delta Q_t^a f_{(tu)}(\beta R_{ab}) T_{tu}^{ab}\, Q_u^b \quad (5)$$

where the omission of the superscript implies that ΔQ_t^a are converged induced moments. If the damping function is set to unity, then this equation is almost identical to the expression for the electrostatic energy

$$E_{electrostatic}(A) = \frac{1}{2}\sum_{a\in A}\sum_{B\neq A}\sum_{b\in B}\sum_{tu} Q_t^a\, T_{tu}^{ab}\, Q_u^b \quad (6)$$

and this can be exploited to estimate the induction energy of the crystal using the routines already implemented in DMAREL[29] that evaluate this function and perform the lattice summations.

Equation 5 has only one molecule bearing just the induced moments interacting with the electrostatic field of the rest of the crystal and hence cannot be directly calculated by DMAREL, assumes that all symmetry related sites bear equal (or inverted) multipole moments. However, substituting $(Q_t + \Delta Q_t/2)$ into eq 6 gives

$$\frac{1}{2}\sum_{a\in A}\sum_{B\neq A}\sum_{b\in B}\sum_{tu}\left(Q_t^a + \frac{\Delta Q_t^a}{2}\right)T_{tu}^{ab}\left(Q_u^b + \frac{\Delta Q_u^b}{2}\right)$$

$$= \frac{1}{2}\sum_{a\in A}\sum_{B\neq A}\sum_{b\in B}\sum_{tu}\left(Q_t^a\, T_{tu}^{ab}\, Q_u^b + \frac{\Delta Q_t^a\, T_{tu}^{ab}\, Q_u^b}{2} + \frac{\Delta Q_u^b\, T_{tu}^{ab}\, Q_t^a}{2} + \frac{\Delta Q_t^a\, T_{tu}^{ab}\, \Delta Q_u^b}{4}\right)$$

$$= E_{\mathrm{electrostatic}}(A) + E_{\mathrm{ind},d-class}(A) + \Delta E_{\mathrm{error}}(A) \quad (7)$$

Thus the induction energy can be calculated from three evaluations of the “electrostatic” contribution to the lattice energy, one where all molecules have the distributed multipole moments $(Q_t + \Delta Q_t/2)$ to get $E_{electrostatic} + E_{ind,d-class} + \Delta E_{error}$, a second with distributed multipole moments $\Delta Q_t/2$ to give ΔE_{error}, and a third using only Q_t to give $E_{electrostatic}$. All three evaluations use Ewald summation for the charge−charge, charge-dipole, and dipole−dipole terms and sum all the other contributions in direct space for all molecules whose center of mass is within 15 Å. Since there is no facility to include damping of the electrostatic interactions in DMAREL, the necessary damping of the induction energy (5) is included for each iteration of the interaction of induced and static multipole moments in the cluster but is not applied in the final lattice energy calculation.

3. Results

3.1. Oxalyl Dihydrazide: The Effects of Rank, Refinement, and Damping. First we compare the energies calculated using the self-consistent electronic response to potential derived charges (SCERP) with WSM models (Figure 6). We find using the SCERP model for $E_{ind,d-class}$ that the α structure is stabilized the most, followed by ϵ. The β, γ, and δ structures are stabilized less than the ϵ form but by similar magnitudes to one another. Each of the WSM models follow the SCERP results except the L2 models,

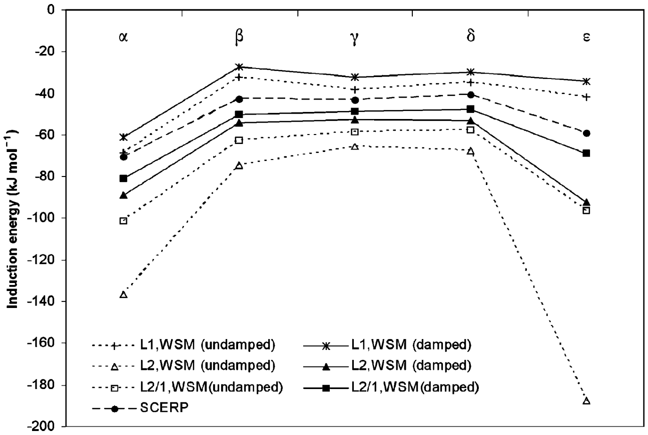


Figure 6. The induction energy of oxalyl dihydrazide for various polarizability models.

which find the relative polarization of the structures to be $\epsilon > \alpha \gg \beta, \gamma, \delta$. This deviation can be explained by examining the crystal structures.

In these five polymorphs there are very short intermolecular contacts. The shortest is an N···H−N contact in ϵ which is shorter than the experimental value and indeed unphysically short. At such short distances the induction energy is very sensitive to distance, and in this case is almost certainly too large. The sensitivity can be reduced by using only rank 1 polarizabilities on hydrogen atoms, which does not lead to significant loss of accuracy overall.

Experimentally, it has been difficult to fully characterize the relative stability of these polymorphs of oxalyl dihydrazide, due to a self-reaction that takes place prior to melting.[32] However, lattice-energy methods that only model the intermolecular repulsion, dispersion, and electrostatic forces, including the conformational energy differences from ab initio gas-phase calculations, predict that the lattice energy of the α form is approximately −110 kJ mol^{-1}, whereas the other four forms range from −130 to −138 kJ mol^{-1} (Supporting Information Table S3). Such a large energy difference is considered to be outside the range of possible polymorphism.[63] By including a correction for the induction energy of the lattice, the predicted lattice energy of the α form becomes comparable with that of the β, γ, and δ polymorphs. It seems apparent that it is important to model charge density polarization for polymorphs that exhibit different intra- and intermolecular hydrogen bonding. This issue is being explored further using electronic structure calculations on oxalyl dihydrazide and other polymorphic systems.[40]

Our results relating to oxalyl dihydrazide strongly suggest that an iterated, damped polarizability model, based on the L1 or mixed L2/L1 models, agrees reasonably well with the self-consistent electronic response to point charges method.

3.2. Naphthalene. Induction is important not only for hydrogen-bonded systems. The crystal structure of naphthalene has been previously analyzed for experimental evidence of induced changes in the charge density.[21] Our SCERP point charge model predicts an induction energy of −1.9 kJ mol^{-1} for the 100 K experimental crystal structure, using the molecular geometry optimized in vacuo. Although small in absolute terms, this is 31% of the electrostatic energy. A damped WSM2/1 polarizability model estimates the induction energy to be 25% of the electrostatic energy. In comparison, the SCERP induction energy for oxalyl dihydrazide polymorphs is 18−38% of the electrostatic energy. Thus, in relative terms, even the charge density of naphthalene is significantly affected by the surrounding molecules in the lattice. By analyzing the change in electrostatic energy due to the induced moments interacting with a unit charge probe, we may indirectly observe the change in charge distribution caused by the crystalline environment. Figure 7 plots the change in the electrostatic energy using the SCERP induced moments, on the van der Waals plus 1.1 Å surface that is sampled by the atomic sites of the surrounding molecules. The anisotropic nature of the induction is clear. The increased electrostatic potential around the C(4)−H bond, in contrast to the C(2)−H bond, shows that the close contact with the π-electrons of the surrounding molecules in the crystal has significantly polarized this bond, as observed in the experimental charge density.[21]

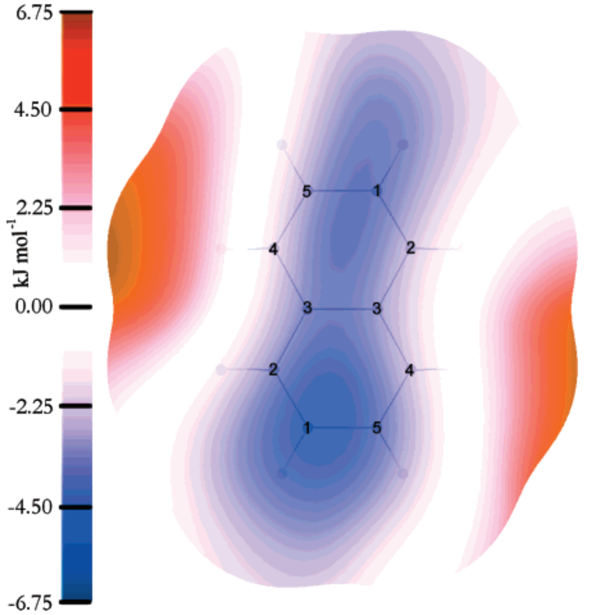


Figure 7. Induced electrostatic energy surface for naphthalene. The energy is calculated from the SCERP model, for the van der Waals + 1.1 Å surface that is accessible by short-contact nuclei. The atom numbering system reflects the symmetry of contacts within the crystal structure, not of the isolated molecule. The energy is calculated using a unit charge probe and ranges from −5.23 kJ mol^{-1} to +6.82 kJ mol^{-1}.

3.3. 3-Azabicyclo[3,3,1]nonane-2,4-dione. 3-Azabicyclo[3,3,1]nonane-2,4-dione presents several challenges in terms of our polarizability calculations: the size of the molecule, in terms of basis functions required and associated computational limits, as well as the volume of space to be sampled for the point-to-point polarizabilities and the C_s symmetry in the molecule. Despite this, and the fact that symmetry of the molecules is not explicitly enforced by CamCASP at any stage, after refinement and localization the resulting polarizabilities are reassuringly symmetric.

We find the induction energy for 3-azabicyclo[3,3,1]nonane-2,4-dione to be 33−36% of the electrostatic energy,

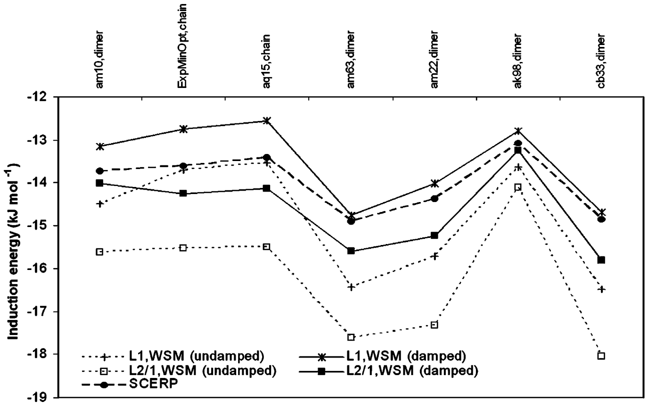


Figure 8. Induction energies for 3-azabicyclo[3,3,1]nonane-2,4-dione. The crystal structures are ordered left-to-right by decreasing lattice stability, as calculated from the distributed static multipole + empirical repulsion-dispersion potential.

with good agreement between SCERP and the damped WSM models (Figure 8). For the crystal structures considered, the induction energy varies by less than 3 kJ mol^{-1}, but this is significant relative to the difference in lattice energies of these structures calculated using a repulsion-dispersion model potential,[64] which range from −95.08 to −97.64 kJ mol^{-1}. Hence, more realistic modeling of the intermolecular interactions to include the induction energy would certainly rerank the structures. However, the observed hydrogen-bonded chain motif is not favored relative to many of the competitive dimer structures,[39] and there is no clear-cut correlation with the hydrogen-bonding motif. Hence neglect of the induction energy does not appear to be the only problem in modeling the relative stability of crystal structures of 3-azabicyclo-[3,3,1]nonane-2,4-dione.[34]

3.4. Carbamazepine. For carbamazepine, we contrast the SCERP with polarizabilities derived from fragment molecules (Figure 4). Despite the additional assumptions, there is still reasonably good agreement in the relative induction energies between SCERP and the damped L1 polarizability model, accounting for an increase in stability of 10.5−18.2 kJ mol^{-1} in the lattice energy. Both models find that the dimer-based structures, and particularly the experimental forms III and IV, are stabilized more by induction than the chain-based structures, and all hydrogen-bonded structures are stabilized more than the structure (ab41) with no hydrogen-bonding (Figure 9). This is significant, as the published crystal structure predictions[33] for carbamazepine found that a structure with a hydrogen-bonded chain motif was more stable than the experimentally known dimer based structures. Improving the modeling of the electrostatic energies by using distributed multipoles from the better charge distribution used in the current work also alters the relative stabilities (Supporting Information), favoring the most stable observed polymorph form III. Hence, more accurate modeling of the electrostatics and adding the induction clearly gives a significant energy lowering to the most stable dimer based structures, which is in accord with experiment.

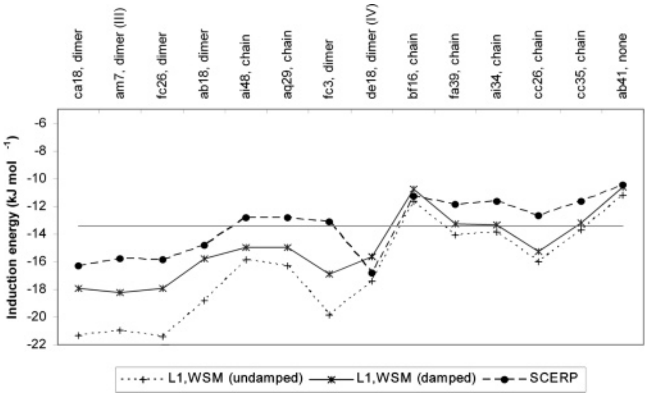


Figure 9. Induction energies for crystal structures of carbamazepine. The structures are ordered left-to-right by decreasing lattice stability, as calculated from the distributed multipoles described, plus an empirical[64] repulsion-dispersion potential. The lattice-energy range for the structures shown is 16 kJ mol^{-1}. The horizontal line indicates the average induction energy with the SCERP model to illustrate the discrimination of structural motifs by the polarizability model.

4. Discussion

4.1. How Important Is the Induction Energy for Organic Crystals? We have used two very different models for estimating the induced moments in organic crystals: an ab initio response to an applied field due to point charges representing the crystal environment, and the use of distributed polarizabilities in the field arising from a distributed multipole representation of the surrounding molecules. The induction energy contribution to the lattice energy, evaluated from these induced moments, is significant. Over this diverse range of crystal structures, the models agree that the induction energy is often between 20 and 40% of the electrostatic contribution to the lattice energy. This order of magnitude is consistent with estimates of the induction energy relative to the electrostatic energy for small polyatomic molecules,[15,65−67] using equally rigorous or better models for the induction energy, although often the polarization is not iterated to self-consistency. It is also comparable with the less rigorous modeling of the induction contribution to lattice energies of neutral organic molecules derived by the pixel method,[3,4,20] where experimental atomic polarizabilities are evenly distributed over the atomic charge density.

More importantly, for different known and predicted crystal structures that otherwise have very similar lattice energies, the two models agree on the relative magnitude of the induction energy. In the case of oxalyl dihydrazide, inclusion of the intermolecular induction is essential for the calculated relative lattice energies to be consistent with the experimental observation of the polymorphs. This is an extreme case, as the intermolecular induction for the α polymorph compensates for the intramolecular hydrogen bonding in the other conformational polymorphs. In the case of carbamazepine, the induction energy favors the observed doubly hydrogen-bonded dimer based structures over the hypothetical catemer based structures. The differences in the induction energies of the low-energy computed structures of 3-azabicyclo[3,3,1]nonane-2,4-dione cannot be so simply ascribed to the hydrogen-bonding motif, but this reflects the relative weakness of the hydrogen bonds for this imide, which forms a plastic phase.[35] In each of these comparisons of known and hypothetical crystal structures, the differences in induction energies are small, only a few kilojoules per mole, but this is sufficient to provide a significant reordering of the relative stability of structures that are virtually equienergetic according to models which do not explicitly model the induction.

To correctly model crystal lattice energy, intermolecular potentials require a reparametrization of the entire repulsion-dispersion potential: adding the induction energy to lattice energies calculated using an empirically fitted potential involves a high degree of double counting. This is sufficient to lead to structures which are too dense if we attempt to minimize crystal structures with an induction term in addition to potentials which have been empirically fitted without the explicit inclusion of induction. It is also important that the model for the induction can be readily implemented in a program that minimizes lattice energies of organic crystal structure.

4.2. Practical Consideration for Using Polarizability Models in the Organic Solid State. A local polarizability model can be implemented in lattice energy minimization packages that use distributed multipole moments. It appears to be feasible to calculate WSM polarizabilities from a reasonable quality ab initio charge density for quite large molecules, with 3-azabicyclo[3,3,1]nonane-2,4-dione probably being the limit with current resources. This is an acceptable limitation, given that the transferable polarizability model calculated from fragments of carbamazepine gave reasonable results compared with the SCERP calculations that used the complete molecule. Thus, it seems that transferable polarizability models could be derived for use in modeling larger molecules.

The induction energy does depend on the order of the polarizabilities included. We have noted some anomalous behavior where rank 2 polarizabilities are used on hydrogen, particularly when involved in short contacts within the crystal structure (most notably on oxalyl dihydrazide ϵ). Given the small amount of charge density associated with polar hydrogen atoms, it seems reasonable that polarizabilities for these sites should be limited to rank 1 for applications to dense systems. The differences between L2 and L1 WSM models for the other atoms are comparable to those between them and the SCERP model.

The error in modeling charge overlap effects in particularly short hydrogen-bonding geometries probably explains the larger variance with polarizability model observed in our oxalyl dihydrazide induction energies, relative to those for the 3-azabicyclo[3,3,1]nonane-2,4-dione crystals which do not have such short contacts. The WSM polarizability model does not account for any density overlap effects, and we have shown that damping is required in order to avoid unreasonable energies for the shorter intermolecular contacts found in hydrogen-bonded crystal structures. We have already shown[26] the agreement between SAPT(DFT) induction energies and WSM models to be very good, and so the WSM polarizability method of modeling the induction energy has a firm foundation. This investigation has shown that damped polarizability models are also suitable for modeling the induction energy in large clusters that represent crystals, with many-body effects, qualitatively different field anisotropy and short contacts.

We consistently find that the ab initio SCERP model falls midway between the L1 and L2/L1,WSM models and that the relative ordering of the energies is consistent. The SCERP model has also approximated the electrostatic field around the molecules (Figure 5), which does lead to significant errors in the hydrogen-bonding region. Thus, we conclude that we cannot at present model the induction energy more accurately than the range indicated by the differences between the SCERP and the L1 and L2/L1 WSM models. It is, however, clear (Figure 3) that the induced moments will need iterating to self-consistency.

5. Conclusions

We have presented two distinct methodologies for approximating the effect of the different crystal environment on the charge distribution of four organic molecules, one using a self-consistent, ab initio derived induced multipoles approach and the other an ab initio derived, distributed, and localized polarizability (WSM) model. We have shown that these models can reproduce experimentally observed changes in the charge density when comparing gas phase and crystalline molecules in the case of naphthalene. We have also shown that the induction energy contribution to the lattice energy of organic molecules is significant and that properly describing the induction energy in lattice energy calculations may improve the relative ranking of the structures to be more in line with experimental observation. The WSM polarizability model and damping scheme can be extended from small polyatomics to crystal structure modeling, on the basis of its rigorous testing for smaller systems, and agreement with an alternative model in crystals (SCERP). A self-consistent WSM polarizability model would be a worthwhile addition to our ability to model molecular crystals. The considerable programming involved is in progress, alongside efforts to improve the theoretical basis of models for all terms in the intermolecular energy.

Acknowledgment. This research was funded by EPSRC EP/C539109/1. The computed crystal structures used are held on STFC Data Portal. G.W.A.W. thanks Dr. Maurice Leslie for useful discussions; A.J.M. thanks Girton College, Cambridge for a Research Fellowship.

Supporting Information Available: S1, structural information for polymorphs of oxalyl dihydrazide; S2, comments of transferability of polarizabilities; and S3, sensitivity of the relative lattice energies of carbamazepine to the electrostatic model. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Halgren, T. A.; Damm, W. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236−242.

(2) Fowler, P. W.; Stone, A. J. *J. Phys. Chem.* **1987**, *91*, 509−511.

(3) Gavezzotti, A. *J. Phys. Chem. B* **2002**, *106*, 4145−4154.

(4) Gavezzotti, A. *J. Phys. Chem. B* **2003**, *107*, 2344−2353.

(5) Mooij, W. T. M.; van Eijck, B. P.; Kroon, J. *J. Phys. Chem. A* **1999**, *103*, 9883−9890.

(6) Lindan, P. J. D.; Gillan, M. J. *J. Phys.: Condens. Matter* **1993**, *5*, 1019−1030.

(7) Catlow, C. R. A.; Norgett, M. J. *J. Phys. C Solid State* **1973**, *6*, 1325−1339.

(8) Ren, P. Y.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497−1506.

(9) Ren, P. Y.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933−5947.

(10) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341−350.

(11) Cox, S. R.; Hsu, L. Y.; Williams, D. E. *Acta Crystallogr., Sect A* **1981**, *37*, 293−301.

(12) Williams, D. E. *Acta Crystallogr., Sect. A*: *Found. Crystallogr*. **1984**, *40*, C95.

(13) Price, S. L. *CrystEngComm* **2004**, *6*, 344−353.

(14) Price, S. L.; Price, L. S. Modelling Intermolecular Forces for Organic Crystal Structure Prediction. In *Intermolecular Forces and Clusters I*; Wales, D. J., Ed.; Springer-Verlag: Berlin, Heidelberg, Germany, 2005; pp 81−123.

(15) Mooij, W. T. M.; van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Eijck, B. P. *J. Phys. Chem. A* **1999**, *103*, 9872−9882.

(16) Mooij, W. T. M.; van Eijck, B. P.; Kroon, J. *J. Am. Chem. Soc.* **2000**, *122*, 3500−3505.

(17) Gascon, J. A.; Leung, S. S. F.; Batista, E. R.; Batista, V. S. *J. Chem. Theory Comput.* **2006**, *2*, 175−186.

(18) Gavezzotti, A. *J. Chem. Theory Comput.* **2005**, *1*, 834−840.

(19) Gavezzotti, A. *Struct. Chem.* **2005**, *16*, 177−185.

(20) Gavezzotti, A. *Z. Kristallogr.* **2005**, *220*, 499−510.

(21) Oddershede, J.; Larsen, S. *J. Phys. Chem. A* **2004**, *108*, 1057−1063.

(22) Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. *Phys. Rev. Lett.* **2003**, *91*, art. no.-033201.

(23) Misquitta, A. J.; Szalewicz, K. *J. Chem. Phys.* **2005**, *122*, art-214103.

(24) Misquitta, A. J.; Podeszwa, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **2005**, *123*, 214103.

(25) Misquitta, A. J.; Stone, A. J. *J. Chem. Theory Comput.* **2008**, *4* (1), 7−18.

(26) Misquitta, A. J.; Stone, A. J.; Price, S. L. *J. Chem. Theory Comput.* **2008**, *4* (1), 19−32.

(27) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361−373.

(28) Willock, D. J.; Price, S. L.; Leslie, M.; Catlow, C. R. A. *J. Comput. Chem.* **1995**, *16*, 628−647.

(29) *DMAREL, version 4.1.1*; Price, S. L.; Willock, D. J.; Leslie, M.; Day, G. M. 2004.

(30) Gavezzotti, A. *CrystEngComm* **2002**, *4*, 343−347.

(31) Price, S. L. *Adv. Drug Delivery Rev.* **2004**, *56*, 301−319.

(32) Ahn, S. Y.; Guo, F.; Kariuki, B. M.; Harris, K. D. M. *J. Am. Chem. Soc.* **2006**, *128*, 8441−8452.

(33) Florence, A. J.; Johnston, A.; Price, S. L.; Nowell, H.; Kennedy, A. R.; Shankland, N. *J. Pharm. Sci.* **2006**, *95*, 1918−1930.

(34) Hulme, A. T.; Johnston, A.; Florence, A. J.; Fernandes, P.; Shankland, K.; Bedford, C. T.; Welch, G. W. A.; Sadiq, G.; Haynes, D. A.; Motherwell, W. D. S.; Tocher, D. A.; Price, S. L. *J. Am. Chem. Soc.* **2007**, *129*, 3649−3657.

(35) Florence, A. J.; Leech, C. K.; Shankland, N.; Shankland, K.; Johnston, A. *CrystEngComm* **2006**, *8*, 746−747.

(36) Karamertzanis, P. G.; Price, S. L. *J. Chem. Theory Comput.* **2006**, *2*, 1184−1199.

(37) Cabeza, A. J. C.; Day, G. M.; Motherwell, W. D. S.; Jones, W. *Cryst. Growth Des.* **2006**, *6*, 1858−1866.

(38) Cabeza, A. J. C.; Day, G. M.; Motherwell, W. D. S.; Jones, W. *Cryst. Growth Des.* **2007**, *7*, 100−107.

(39) Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Dzyabchenko, A.; Erk, P.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Lommerse, J. P. M.; Mooij, W. T. M.; Price, S. L.; Scheraga, H.; Schweizer, B.; Schmidt, M. U.; van Eijck, B. P.; Verwer, P.; Williams, D. E. *Acta Crystallogr., Sect. B*: *Struct. Sci.* **2002**, *58*, 647−661.

(40) Karamertzanis, P. G.; Day, G. M.; Welch, G. W. A.; Kendrick, J.; Leusen, F. J. J.; Neumann, M. A.; Price, S. L. *J. Chem. Phys.* **2008**, submitted.

(41) Chisholm, J. A.; Motherwell, S. *J. Appl. Crystallogr.* **2005**, *38*, 228−231.

(42) Howie, R. A.; Skakle, J. M. S. *Acta Crystallogr., Sect. E* **2001**, *57*, o822−o824.

(43) Angyan, J. G.; Jansen, G.; Loos, M.; Hattig, C.; Hess, B. A. *Chem. Phys. Lett.* **1994**, *219*, 267−273.

(44) Stone, A. J. *Mol. Phys.* **1985**, *56*, 1065−1082.

(45) Le Sueur, C. R.; Stone, A. J. *Mol. Phys.* **1993**, *78*, 1267−1291.

(46) Misquitta, A. J.; Stone, A. J. *J. Chem. Phys.* **2006**, *124*, 024111.

(47) Le Sueur, C. R.; Stone, A. J. *Mol. Phys.* **1994**, *83*, 293−307.

(48) Williams, G. J.; Stone, A. J. *J. Chem. Phys.* **2003**, *119*, 4620−4628.

(49) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441−451.

(50) Misquitta, A. J.; Stone, A. J. *CamCASP: A program for studying intermolecular interactions and for the calculation of molecular properties in distributed form, version 5*; University of Cambridge: 2007. See: http://www-stone.ch.cam.ac.uk/programs.html#CamCASP (accessed Feb 2008).

(51) *DALTON, a molecular structure program, release 2.0*; 2005. See: http://www.kjemi.uio.no/software/dalton/dalton.html (accessed June 21, 2007).

(52) Allen, M. J.; Tozer, D. J. *J. Chem. Phys.* **2000**, *113*, 5185−5192.

(53) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158−6170.

(54) Sadlej, A. J. *Collect. Czech. Chem. C* **1988**, *53*, 1995−2016.

(55) Stone, A. J. *J. Chem. Theory Comput.* **2005**, *1*, 1128−1132.

(56) Stone, A. J. *The Theory of Intermolecular Forces,* 1st ed.; Clarendon Press: Oxford, 1996.

(57) Stone, A. J.; Misquitta, A. J. *Int. Rev. Phys. Chem.* **2007**, *26*, 193−222.

(58) Tang, K. T.; Toennies, J. P. *J. Chem. Phys.* **1984**, *80*, 3726−3741.

(59) Stone, A. J.; Dullweber, A.; Engkvist, O.; Fraschini, E.; Hodges, M. P.; Meredith, A. W.; Nutt, D. R.; Popelier, P. L. A.; Wales, D. J. *Orient: a program for studying interactions between molecules, version 4.6;* University of Cambridge: 2006. See: http//www-stone.ch.cam.ac.uk/programs.html#Orient (accessed Feb 2008).

(60) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887−1930.

(61) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J.; 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, 2003.

(62) Buckingham, A. D.; Fowler, P. W. *Can. J. Chem.* **1985**, *63*, 2018−2025.

(63) Bernstein, J. *Polymorphism in Molecular Crystals;* Clarendon Press: Oxford, 2002.

(64) Coombes, D. S.; Price, S. L.; Willock, D. J.; Leslie, M. *J. Phys. Chem.* **1996**, *100*, 7352−7360.

(65) Chipot, C.; Luque, F. J. *Chem. Phys. Lett.* **2000**, *332*, 190−198.

(66) Chipot, C.; Angyan, J. G. *New J. Chem.* **2005**, *29*, 411−420.

(67) Jansen, G.; Hattig, C.; Hess, B. A.; Angyan, J. G. *Mol. Phys.* **1996**, *88*, 69−92.

CT700270D

JCTC Journal of Chemical Theory and Computation

Structure, Binding Energies, and IR-Spectral Fingerprinting of Formic Acid Dimers

İlhan Yavuz

Physics Department, Marmara University, Göztepe Kampus,
Kadiköy 34772, Istanbul, Turkey

Carl Trindle*

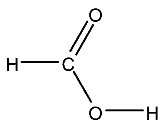
Chemistry Department, The University of Virginia, Charlottesville, Virginia 22904

Received June 27, 2007

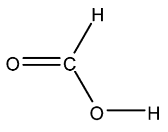
Abstract: We describe equilibrium structures for a variety of species likely to be formed as intermediate species in the dimerization of formic acid to produce the stable C_{2h}-symmetric doubly H-bonded dimer and perhaps produced as the vapor is irradiated. For several low-lying species the rearrangement pathways to the stable form are characterized at the MP2/6-311+G-(d,p) level of theory, with optimized structures and vibrations computed with full counterpoise corrections for basis set superposition error. Estimates of vibrational frequencies with corrections for anharmonicity suggest that infrared transitions (CO stretches and OH out-of-plane motions) could signal the presence of species less stable than the C_{2h} dimer, observable in irradiation studies of formic acid vapor.

Introduction

Formic acid **I** exists primarily as the trans form (the H−C−O−H angle = 180) in the gas phase. The MP2/6-311+G-(d,p) energy difference of 4.65 kcal/mol (4.40 kcal/mol after zero point energy correction[1,2]) suggests that the trans[3] form is about 1000× more abundant than the cis form at room temperature.



I-trans



I-cis

Formic acid forms clusters in the gas phase. The structure and spectra of the C_{2h}-symmetric dimer of formic acid **II** (below) has been thoroughly studied, both by computational modeling[4] and by spectroscopic methods.[5] Considerable emphasis has been placed on the proton exchange and the importance of tunneling in the process.[6]

Shipman et al.[7] have investigated the response of formic acid vapor to IR irradiation in the broad absorption associated with the OH stretch. The breadth has been rationalized by anharmonic coupling to lower-frequency modes, Fermi resonance with combinations of such modes, and (in the case of the symmetric dimer) Davydov coupling between the degenerate OH stretches. Low-temperature studies and temperature-dependent FTIR investigations have provided the basis for the study of intramolecular vibrational relaxation. The studies of Shipman et al. characterize vibrational relaxation of formic acid vapor near room temperature, subjected to an ultrashort (ca. 100 fs) pulse in the OH stretching region. Their observations suggest H-bond breaking with a characteristic time of about 20 ps and perhaps the existence of detectable amounts of a dimer other than the most stable C_{2h} species. The feature associated with the possible new structure, which the authors term the “acyclic” dimer, is broad, centered at about 3230 cm^{-1}. This may be compared with the cyclic dimer’s OH stretch at 3107 cm^{-1} and suggests weaker H-bonding in what might be a short-lived species. The broad feature evolves over a 100 to 200 ps duration. The authors consider the possibility that colli-

* Corresponding author e-mail: cot@virginia.edu.

10.1021/ct700161a CCC: $40.75
Published on Web 02/06/2008

Chart 1. Structures of Species Discussed in the Text

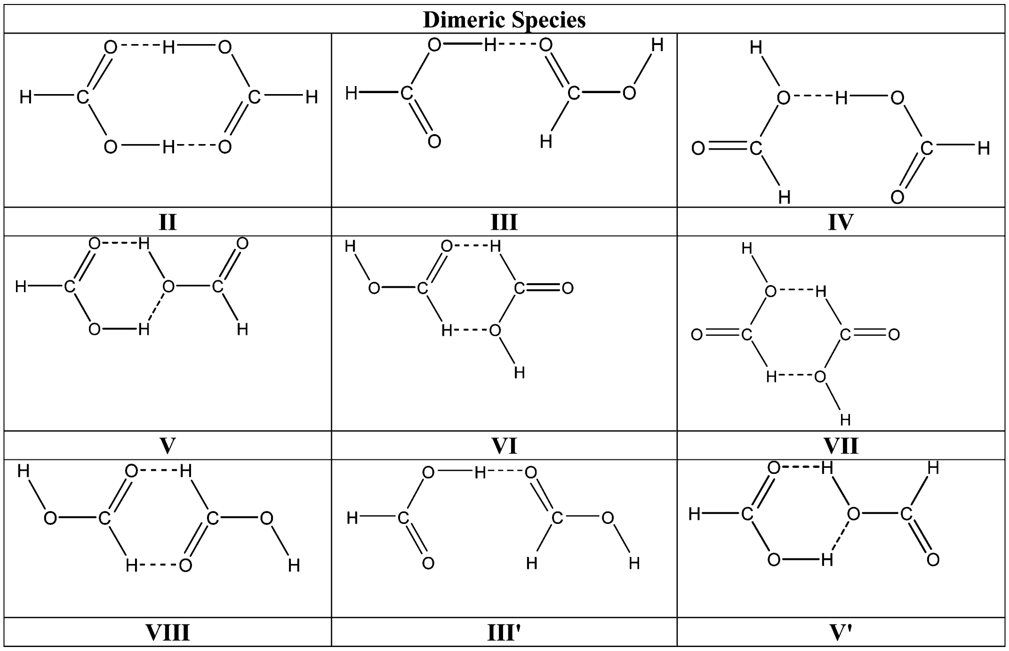
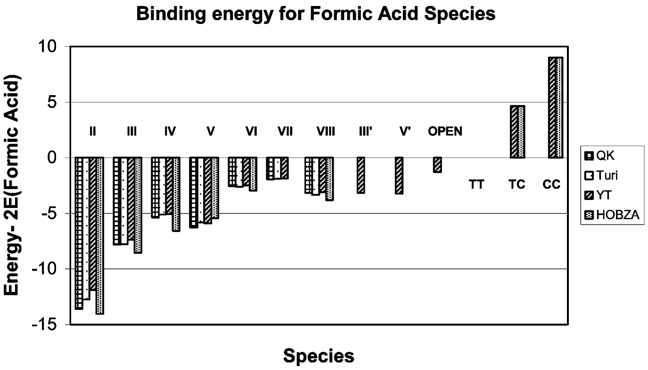


Chart 2. Energies (kcal/mol) Relative to Two Trans Formic Acid Molecules[a]



[a] Turi values from ref 8; QK values from Qian and Krim, ref 9; CVH values from Chocholoušová, Vacek, and Hobza, ref 10; YT=this work (Yavuz and Trindle). Turi and QK have single-point CP corrections; CVH and YT used full counterpoise corrections in optimization. ZPE corrections are not included. TT refers to two isolated formic acid molecules in trans configuration; TC and CC have one and two cis species respectively.

sional cooling of the acyclic dimer(s) may account for the longer-time behavior but concluded that the growth in free OH absorption could not be rationalized in this way. Their preferred account is a dissociation of an acyclic dimer to monomers in the 100 to 200 ps time frame. Direct dissociation requires 14.8 kcal/mol (or about 5000 cm^{-1}) according to photoacoustic measurement,[8] so the 3000 cm^{-1} provided by IR irradiation must be augmented, perhaps by collision.

The purpose of this investigation is to re-examine the energy demands for the formation of acyclic dimers and the further production of monomers for such intermediate species. We intend to identify species within the energy reach of the irradiation and to characterize their IR absorption spectra to provide a basis for more direct identification of the intermediate species. To this end we conduct computations employing correlation-corrected model chemistries (MP2 in extended basis sets), corrected by counterpoise compensation for Basis Set Superposition Errors and including estimates of anharmonicity and mode coupling.

Modeling such small interactions as hydrogen bonds requires accurate methods. This includes a suitably large and flexible basis set, recognition of correlation corrections to energies and structures, and allowance for basis set superposition error (BSSE).[9,10] According to Tzusuki et al.,[3] the extrapolated basis set limit for the counterpoise (CP)-corrected binding energy in CCSD(T) for the C_{2h}-symmetric dimer of formic acid is 13.93 kcal/mol. Their estimate of the MP2 limit for the binding energy is 13.79 kcal/mol. We infer that these values are not ZPE-corrected. The landmark paper of Turi[11] characterized this species and defined a standard notation for other equilibrium structures **II**−**VIII** of the dimers of *trans*-formic acid, using MP2 with basis sets up to D95++(d,p) and single-point counterpoise estimates of the basis set superposition error.

Among the dimers of trans formic acid we find conventional linear =O...HO− and >O...HO− hydrogen bonds (**II** in the first case, **III** and **IV** in the second case), bent -OH...O< H-bonds which are probably slightly weaker (**V** and **V′**), shared H bonds, and much weaker CH...O< and CH...O= interactions (**III**, **III′**, **IV**, **VI**, **VII**, **VIII**). The binding energy can be approximately represented by the energies of the various types of H-bonds: the weak interaction CH...O on average is about 1 kcal/mol; OH...O is about 5−6 kcal/mol; and the distorted bond OH...O is about 3 kcal/mol. Qian and Krimm[12] revisited these systems in their project to construct a suitable potential for molecular mechanics simulations. Their MP2/6-311++G(d,p) binding energies, counterpoise-corrected for BSSE at the equilibrium geometries, are in general agreement with Turi's results. Chocholoušová, Vacek, and Hobza[13] (CVH) have evaluated energies and structures of several of these species, with the MP2/aug-cc-pVDZ model chemistry including both

Table 1. New Equilibrium Structures and Energies Relative to Twice **I-trans** without and with Zero Point Energy Corrections (kcal/mol)

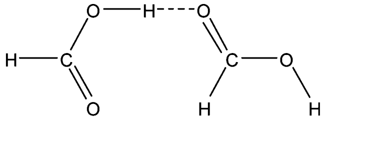
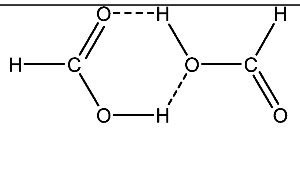
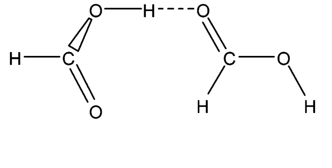
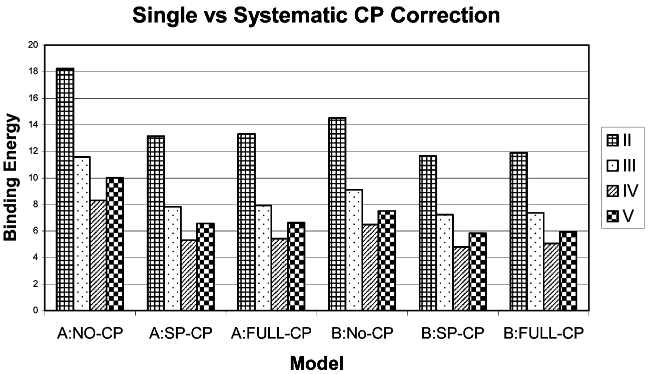
Code	III'	V'	Open
Form			
E	-3.15	-3.24	-1.38
E(ZPE)	-1.75	-2.10	-0.51

Chart 3. Counterpoise Corrections of Two Kinds, Single Point and Full[a]



[a] Binding energy estimates for species **II**, **III**, **IV**, and **V** for various counterpoise corrections. Region A: left to right, MP2/6-31G(d) values with no CP correction, single-point CP, or optimized on a fully CP-corrected surface. Region B: left to right, analogous MP2/6-311+G(d,p) values. All energies are in kcal/mol.

counterpoise-corrected optimization and single-point correction of the energy after conventional optimization. We have returned to these systems as the first step in a study of species that might be generated when formic acid vapor is irradiated by 3000 cm^{-1} photons.

All our calculations employ Gaussian 03(W).[14] In our best calculations our model chemistry is MP2/6-311+G(d,p) with full counterpoise corrections[7] imposed throughout the optimization and frequency calculations. By “full counterpoise corrections” we mean that the method of Simon, Duran, and Dannenberg is employed.[7] At every point in an optimization the energy and gradient E and ∇E are computed for the dimer FA−FA* (where FA and FA* are both formic acid molecules but may be differently disposed in space) in the joint basis BFA∪BFA*; for each monomer in the joint basis; and for each monomer in its own basis. Then the energy is written

$$E = E(\mathrm{FA} + \mathrm{FA}^*;\mathrm{BFA} \cup \mathrm{BFA}^*) + E(\mathrm{FA};\mathrm{BFA} \cup \mathrm{BFA}^*) + E(\mathrm{FA}^*;\mathrm{BFA} \cup \mathrm{BFA}^*) - E(\mathrm{FA}^*;\mathrm{BFA}^*) - E(\mathrm{FA};\mathrm{BFA})$$

The counterpoise-corrected gradient and second derivative tensor require the derivatives of all four correction terms as well as the leading term. Simon, Duran, and Dannenberg report several cases, notably HF in water, where full CP produces significantly different structures and vibrational frequencies compared with single point CP correction.

Chart 2 displays various estimates of the energies of the nine species identified by Turi,[8] relative to the dissociation products, two separated trans formic acid molecules. These values do not include zero-point energy corrections. We have interchanged species **IV** and **V** in CVH Table 1 (III and IV in their numbering), after reproducing their reported numbers for these species. It appears that values in the first column of CVH in Table 1 refer to binding energies obtained by conventional optimization followed by single-point counterpoise correction. Table S2 (Supporting Information) includes more detail of the results of Turi, Qian and Krimm, and Chocholoušová, Vacek, and Hobza as well as our own, including zero point energies and CP corrections. The key difference between our results and these values is the relative stability we find for species **V** which displays a −OH bond participating both as an acceptor and a donor in the six-atom ring stabilized by H bonds.

Brinkmann, Tschumper, Yan, and Schaeffer[15] have studied species **I**, **II**, and **III**, using a variety of basis sets and both MP2 and DFT correlation-corrected model chemistries. They estimate the binding energy of **II** to be about 15.9 kcal/mol (with MP2 with their TZ2P+dif basis) and of **III** to be about 9.5 kcal/mol. The counterpoise correction is about 2.4 kcal/mol for **II** and 0.6 kcal/mol for **III**; this would shift the binding energies of **II** and **III** to 13.5 and 8.9 kcal/mol respectively. These values are apparently not corrected for zero-point vibrational energy.

Our values seem to be consistent with Turi’s values, but our binding energy values are smaller for species **II** and **III**. This may arise either from details of the counterpoise corrections or effects of the difference in basis used; we locate minimum-energy structures and vibrational frequencies on the counterpoise-corrected potential, while Qian and Krimm and also Turi evaluate the CP correction using structures obtained by direct calculations. Chocholoušová, Vacek, and Hobza report results of both single-point CP corrections and counterpoise-corrected optimization. In general one expects smaller binding energies and lower interfragment frequencies for the counterpoise-corrected optimization compared to standard gradient optimization.

Equilibrium structures found on the CP-corrected surfaces may be quite different in geometry and as well as energy relative to the analogous structures located on the uncorrected surfaces. One might wonder whether the relative energies found by a single CP correction are useful approximations to relative energies of structures found on the CP-corrected surface. Chart 3 bears on this question.

Table 2. Interatomic Distances for H-Bonding (Å)

structural feature	I	II	III	IV
−OH...O= (Turi)		1.702	1.788	1.899
−OH...O= (Y−T)		1.816	1.895	2.000
−OH...O= (CVH)		1.68	1.77	1.89
−CH...O= (Turi)			2.387	2.374
−CH...O= (Y−T)			2.493	2.538
−CH...O= (CVH)			2.34	2.39

structural feature	V		VI	VII	VIII
−OH...O< −OH...O= (Turi)	1.963; 1.975	−CH...O< −CH...O= (Turi)	2.535; 2.422	2.527	2.447
−OH...O<−OH...O= (Y−T)	2.078; 2.125	−CH...O< −CH...O= (Y−T)	2.641; 2.535	2.619	2.569
−OH...O< −OH...O= (CVH)	1.91; 1.98	−CH...O< −CH...O= (CVH)	2.50; 2.39	No data	2.43

The single point CP correction seems quite effective for all calculations with or without correlation both in small and larger basis sets, matching the net correction found on the CP surface very closely. Detailed data illustrating this point are in Table S2, Supporting Information. CVH report values of binding energies obtained by full CP corrections which are larger than the values obtained by single-point CP corrections. We confirm this observation.

New Equilibrium Dimeric Forms

We investigated some species not described previously. These include relative minima of the dimer which incorporate the formic acid fragment in a less stable cis (HCOH angle ca. 0°) form. These structures, shown in Table 1, are less stable than the analogs formed with all-trans formic acid, by about the 4−5 kcal/mol energy difference between cis and trans formaldehyde in the gas phase. One interesting form of this type is **III′** which has just been detected experimentally by Marushkevich et al.[16] who induced a trans-to-cis transition in **III** by infrared irradiation. **III′** is still more stable than all the trans−trans species bound only by CH.... O attractions. So is a cis−trans version **V′** of the low-energy species **V** with -H...OH...O< bonds. We also found a system with one conventional −OH...O= bond linking a **I**-cis acceptor with a **I**-trans donor acid, forming an open structure not otherwise stabilized, with the planes of the acids nearly orthogonal.

Geometric Parameters: Influence of the CP Corrections. Table S3 (Supporting Information) shows computed bond lengths for formic acid monomer **I** and Turi's set of dimers **II**−**VIII**. Turi's structures differ from ours in subtle ways in intramonomer C−H, C−O, and C=O distances, which we think are attributable to the differences in basis sets. Both sets of calculations faithfully represent the changes in lengths of OH bonds participating in H-bonding.

As Table 2 shows we predict larger −OH...O= and −OH...O−H bond lengths than Turi or CVH report, which we attribute to our systematic optimization of structures in the CP-corrected regime (in the first case) and perhaps to our superior basis set (in the second case). Turi explored the CP-corrected surface for species **II** and found that the −OH...O= distance increased by about 0.05 Å; our method produces an extension of about 0.1 Å. It appears that Turi made CP energy corrections for all other species at the minima of the noncorrected potentials.

Table 3. Selected Interatomic Distances for Cis−Trans Formic Acid Dimers (Å)

species H...A	III′	V′	Open trans-donor
−OH...O= −OH...O<	1.864	2.015	1.897
−CH...O=		2.055	
	2.384		

Geometry of New Structures. The intramolecular structural parameters of the new structures resulting from the association of one trans and one cis formic acid are unsurprising. Values are collected in the Supporting Information. The intermolecular distances shown in Table 3 suggest that the double role played by the OH group in **V′** (as in **V**) weakens the net attraction in these species. Conventional H-bonding is strongest in **II**, to judge from the short bond =O...HO− distance.

Reaction Paths

Considering the association path forming **II** from monomeric formic acid, BTYS[12] challenged the common notion that **II** forms by synchronous formation of its two hydrogen bonds. These investigators pointed out the statistical advantage of a two-step assembly of the stable species. On that basis one would expect that the association would proceed in at least two steps, as a first strong OH...O H-bond is formed, with the trans-formic acid fragments otherwise almost arbitrarily oriented. If two trans-formic acid monomers interact in this way, then the open form reverts without any apparent activation barrier to species **II** or **III**. When a single H-bond is formed between cis and trans formic acid, a metastable species is formed. It is stable with respect to two trans formic acid species by a mere 0.5 kcal/mol (ZPE corrected) but stable with respect to cis and trans formic acid molecules by about 5 kcal/mol. This open species can rearrange to **III** by cis−trans reversion of one formic acid fragment.

BTYS[12] located a local equilibrium structure which they termed "acyclic" and which appears to be species **III,** lying according to their calculations ca. 6 kcal/mol above **II** and isolated from **II** by an activation barrier they estimated to be ca. 3 kcal/mol. That is, the barrier from **II** to **III** is about 9 kcal/mol. We have re-examined the reaction path between **III** and the C_{2h} species **II**. Our estimate of interconversion barriers of 6.79 kcal/mol (**II** → **III**) and 2.25 kcal/mol (**III** → **II**) (Table 4) are lower than the values quoted by BTYS,[10]

Table 4. Energy (kcal/mol unless Otherwise Designated) and Selected Interatomic Distances (Å) of Transition States for Interconversions Leading to and from **II**[a]

species	**V→II**	**III→II**	**Open→II**
structure			
−OH...O= (Y−T)	1.9744, 4.4892	1.9634, 3.9118	2.0074, 2.1179
−OH...O= (BTYS)		1.847	
barrier from **II** (w. ZPE)	7.41 (6.37)	6.80 (5.76)	15.61 (13.85)
reverse barrier (w. ZPE)	1.41 (1.12)	2.26 (1.77)	5.08 (4.39)
E(hartrees)	−378.7348638	−378.7324325	−378.7183938
ZPE	43.856	43.759	43.033
barrier from **II** −SPCP (ZPE)	7.56 (6.64)	6.73 (5.74)	15.54 (13.93)
reverse barrier −SPCP (ZPE)	1.74 (1.55)	2.31 (1.84)	5.02 (4.36)

[a] The path linking the stable species **II** to the **Open** form is quite different from those leading from **III**, **IV**, or **V**. The **Open** form is a complex of a trans H-bond donor with a cis H-bond acceptor, so passage from the stable dimer **II** requires both partial H-bond breaking and trans−cis isomerization. This is why the activation energy on the order of the sum of the energy of a hydrogen bond added to the activation energy is required for trans−cis isomerization.

which once again we attribute to our use of the CP-corrected potential surface. Zero Point Energy differences reduce the barriers still further, to 5.76 and 1.77 kcal/mol respectively.

We have also established a path connecting the species **V** with **II.** Reversion of **V** to **II** is opposed by only about 1 kcal/mol. The energy and structure of the **V→II** and **III→II** transition states are very similar, though they are distinct at the convergence tolerance imposed in the transition state search. It would seem reasonable to consider the paths to traverse a rather flat upland or mesa rather than the more familiar picture of a high-curvature mountain pass. Our efforts to find a transition state for the passage of **IV** to **II** or **III** have not been successful.

Vibrational Spectra of Formic Acid Species

Irradiation of formic acid vapor with 3000 cm^{-1} light can produce any species which has an activation barrier less than about 8 kcal/mol from species **II**. The possibility that either **III** or **V** is accessible by IR excitation of **II** suggests that we should try to distinguish these species by their vibrational spectra. The absorption in the region of the OH stretch is sufficiently noisy that we should look elsewhere. In following sections we will establish reliability of our computations of harmonic and anharmonic vibrational frequencies and explore the possibility of fingerprinting the low-lying species.

Vibrations I: The Cis and Trans Monomers. First we establish the accuracy of the level of calculation we have chosen. Table 5 contains values for trans-formic acid (trans-**I**) frequencies with and without Barone anharmonicity estimates[17] as incorporated in Gaussian03. The anharmonic frequencies obtained by Barone's method agree remarkably well with most FTIR values. While OH and CH stretches are overestimated by 40 and 80 wave numbers respectively, the other modes all agree within 10 wave numbers.

We can distinguish the *cis* from *trans* computed spectra in the fingerprint region 600 to 670 cm^{-1} by the doublet at about 620 and 640 cm^{-1} computed for the more stable trans form and shown in Chart 4, the schematic representation of the experimental and computed spectra. The less stable cis form displays only a single absorption, well to the blue of this feature. The strong absorption of the OH out-of-plane motion lies very low (ca. 500 cm^{-1}). No such simple discriminating feature of the spectrum is available in the region from 1000 to 1300 cm^{-1} although the computed frequencies for the trans isomer are consistently closer to the observed absorptions.

Table 5. Fundamental Vibrational Frequencies for Cis and Trans Formic (cm^{-1})[a]

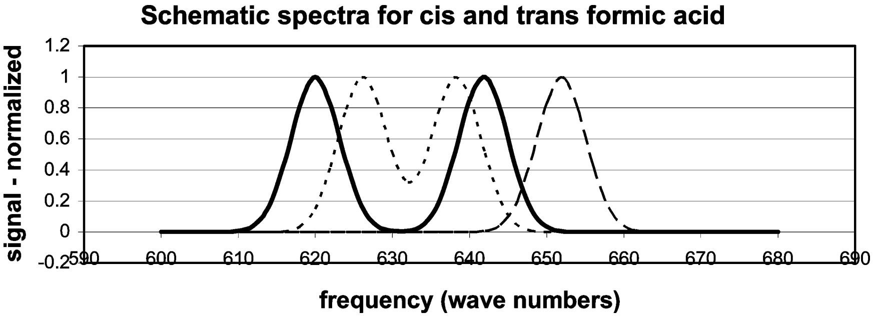
	cis		trans		
MODE	HARM	ANHARM	HARM	ANHARM	EXP[18]
1 OH str	3862.5	3674.6	3797.0	3607.4	3568.0
2 CH str	3035.4	2894.3	3132.0	3020.9	2938.7
3 C=O str	1845.4	1810.81	1807.4	1774.2	1775.5
4 OCH bend	1447.2	1412.1	1428.8	1401.2	1393.5
5 COH bend	1293.5	1248.5	1311.7	1222.4	1216.6
6 CO str	1117.4	1089.9	1143.2	1109.8	1105.5
7 OCO bend	**658.5**	**651.9**	**633.3**	**626.2**	**620.0**
8 CH oop	1042.5	1020.2	1064.2	1041.3	1033.3
9 OH oop	**511.2**	**480.2**	**670.8**	**638.3**	**641.8**

[a] Frequencies in wave numbers for formic acid computed in MP2/6-311+G(d,p) with Barone anharmonicity corrections. Modes 1−7 are in-plane motions, and modes 8 and 9 are out-of-plane motions ("oop"). Intense transitions are in bold face. The primary distinction between cis and trans isomers' spectra is in the splitting of the intense transitions 7 and 9.

Vibrations II: Calibration with the Water Dimer. Water dimer is one of the most thoroughly studied hydrogen-bonded systems, so we wished to use the system as a reference point and to establish expectations for the methods we use. Table 6 shows separately the effect of CP corrections and Barone anharmonicity estimates (as implemented in Gaussian03) on the values of vibrational frequencies. The reported harmonic values listed in the table are not scaled, but if the MP2 factor 0.946 is applied to the frequencies of the harmonic modes, the mean absolute deviation (MAD) is substantially improved from 100 to 43 wave numbers (no CP) or from 77 to 26 wave numbers (CP corrected).

We take note of the well tested vibrational self-consistent method of Chaban, Jung, and Gerber,[23] implemented in GAMESS software as VSCF. This method does extremely well for the intramolecular modes. The soft intermolecular modes, which have large displacements and serious intermode coupling are not realistically described in VSCF, and even the pairwise coupling introduced in a refined version called cc-VSCF does not yield satisfactory results for these

Chart 4. Distinguishing Feature of Trans vs Cis Formic Acid[a]



[a] Solid=experimental spectrum of formic acid showing a doublet centered at 630 cm^{-1} and split by about 30 cm^{-1}; dashed line=computed spectrum of the cis species showing an intense transition about 650 cm^{-1}; dotted line=computed spectrum for the trans species, showing a doublet centered at about 630 cm^{-1} and split by about 10 cm^{-1}. Computed values are the result of Barone's anharmonicity method in the MP2/6-311+G(d,p) model.

Table 6. Barone Anharmonicity Estimates of Frequencies for the Water Dimer, with and without Counterpoise Corrections (cm^{-1})

MP2/6-311+G(d,p)		MP2/6-311+G(d,p) CP				
harmonic	anharmonic	harmonic	anharmonic	**VSCF**	**Exp**	**mode**
133	75	114	56	545	88[a]	A″ PA-rotn
172	138	129	86	414	103[a]	A′ PA-rotn
178	131	155	119	259	108[a]	A″ PD rotn
204	**142**	**156**	**123**	451	**143[b]**	**A′ Diss**
382	286	307	268	550	311[c]	A′ H ‖
665	526	569	459	769	523[c]	A″ H ⊥
1640	1593	1636	1591	1565	1599[c]	A′ PA bend
1664	1604	1655	1603	1612	1616[c]	A′ PD bend
3808	3649	3826	3670	3560	3601[d]	A′ PD sstr
3875	3699	3880	3706	3689	3660[d]	A′PA sstr
3975	3798	3975	3797	3733	3735[d]	A′ PD astr
3989	3805	3995	3821	3763	3745[d]	A″ PA astr
101 (43 [e])	22	77 (26[e])	32	21[f], 285[g]		MAD

[a] Brayly et al.[19] supersonic molecular beam expansion. [b] Keutch et al.[20] supersonic molecular beam expansion. [c] Wuelfert et al.[21] CARS. [d] Huang and Miller[22] molecular beam depletion spectroscopy. [e] Scaled by 0.946 MP2 factor. [f] Intramolecular modes. [g] Intermolecular modes. Total VSCF MAD 153 cm^{-1}, cc-VSCF reduces the MAD values to 21 (intramolecular), 210 (intermolecular), and 115 (overall). PA=proton acceptor; PD=proton donor: sstr=symmetric stretch, astr=asymmetric stretch, H ‖ refers to motion of a H involved in OH...O< H-bonding moving parallel to the O−O axis; H ⊥ refers to motion perpendicular to that axis; diss refers to the dimer's dissociative motion.

difficult modes. This may be attributed to the extensive coupling among low frequency intermolecular modes as well as their severe anharmonicity.

Anharmonicity corrections by Barone's method make significant improvements, especially in the low-frequency intermolecular modes. The MAD is reduced from 101 to 22 cm^{-1} (with no CP) or from 77 to 32 cm^{-1} (with CP). As we noted for monomeric formic acid and water the Barone method does best for motions which are not simple O−H or C−H bond stretches. CP corrections shift most intermolecular frequencies to the red. Scaling CP corrected frequencies or making anharmonicity corrections to vibrations computed without CP shifts produce the most accurate estimates of the cluster's vibrational frequencies. Combining CP and anharmonic corrections produces intermolecular mode frequencies overcorrected to the red and has little effect on the intramolecular modes.

Characterization of the Most Stable Formic Acid Dimer II. The C_{2h}-symmetric dimer has been studied experimentally by Bertie and Michaelian[24] who reported Raman spectra, by Marechal who described gas-phase FTIR data,[25] by Halupka and Sander who reported the IR absorption of matrix isolated species,[26] and by others who investigated specific regions of the spectrum. B3LYP/6-31G(d) calculations followed by Pulay's mode-specific scaling reproduces the experimental data with admirable accuracy (MAD = 15 cm^{-1}).[27] Our calculations include estimates of harmonic frequencies with and without systematic counterpoise corrections and also frequencies corrected for anharmonicity by the Barone's method (Table 7).

All values reported in Table 7 apart from the Pulay entries are unscaled. However if the harmonic values are scaled by the factor 0.946 suitable for MP2 values, then the MAD is reduced from 79 to 60 cm^{-1} for the frequencies obtained on the CP-corrected surface and from 66 to 40 cm^{-1} for the frequencies obtained without CP. The results suggest that one obtains the most reliable representation of the experimental spectrum with unscaled Barone anharmonic values obtained without counterpoise corrections. With a MAD of 21 cm^{-1}, these results come close to the Pulay-scaled values.

Scott and Radom's definition of scaling factors for specific model chemistries and different factors for high frequencies and low frequencies produces a MAD of 42 cm^{-1} for

Table 7. Frequencies for Formic Acid C_{2h}-Symmetric Dimer (cm^{-1})[e]

	MP2/6-311+G(d,p) - CP				MP2/6-311+G(d,p)	
A_g	Harmonic	Anharm	EXP[a-c]	Pulay Scaling[d]	harmonic	Anharm
1(1)	3457.3	3239.5	2949[c]	2968	3330.5	3079.2
2(1)	3144.1	2981.7	3035[c]	2946	3145.0	2970.2
3(1)	1748.9	1707.9	1670[c]	1671	1731.7	1691.6
4(1)	1462.6	1417.2	1415[c]	1409	1481.6	1419.6
5(1)	1400.4	1358.2	1375[c]	1366	1414.8	1337.4
6(1)	1235.8	1200.4	1214[c]	1284	1252.6	1217.3
7(1)	675.4	667.5	677[c]	663	683.9	674.6
8(1)	172.2	160.4		172	188.4	177.2
9(1)	149.3	138.7	137[c]	148	162.4	147.4
A_u						
10(1)	1097.6	1072.1	1050[c]	1061	1105.3	1078.7
11(1)	**917.1**	**867.3**	**942[a], 908[b], 917[c]**	**917**	**939.1**	**934.6**
12(1)	156.1	147.4	163[c]	159	162.4	157.6
13(1)	69.4	67.2	68[c]	66	60.7	66.6
B_g						
14(1)	1086.3	1062.0	1060[c]	1049	1084.4	1064.9
15(1)	885.9	821.3		843	922.1	873.2
16(1)	234.3	217.4	230[c]	237	229.3	230.7
B_u						
17(1)	**3517.1**	**3305.2**	**2992[a], 3000[b], 3000[c]**	**3072**	**3414.3**	**3173.1**
18(1)	**3141.5**	**2978.8**	**2950[a], 2944[b], 2957[c]**	**2957**	**3141.6**	**2969.8**
19(1)	**1795.5**	**1761.2**	**1728[a], 1740[b], 1754[c]**	**1752**	**1788.7**	**1751.8**
20(1)	1454.4	1411.6	1450[c]	1454	1460.3	1413.2
21(1)	1391.1	1347.8	1373[a], 1364[b], 1365[c]	1365	1404.8	1363.7
22(1)	**1241.7**	**1201.7**	**1226[a], 1215[b], 1218[c]**	**1218**	**1258.0**	**1226.9**
23(1)	689.5	678.9	712[a], 699[b], 697[c]	708	703.4	691.5
24(1)	213.0	198.0	248[c]	248	243.1	230.4
MAD	79	45		15	66	21
(scaled)	(60)				(40)	

[a] Ar matrix, Halupka and Sander, ref 13. [b] Gas-phase FTIR, Marechal, ref 12. [c] Raman from Bertie and Michaelian, ref 11. [d] Computations from Fernandez, Gomez Marigliano, and Varetti, ref 14. [e] Very intense transitions in IR are in bold. A_g and B_g absorptions are observed only in the Raman.

frequencies found without systematic CP correction. Omitting two outliers, the OH stretches, reduces MAD to 24 cm^{-1}.

A referee pointed out to us that experimental data and theoretical estimates for the vibrational frequencies of the C_{2h}-symmetric dimer of acetic acid were available[28,29] so we conducted MP2/6-311+G(d,p) calculations of the optimum structure with and without systematic CP corrections and the vibrational frequencies of the species including Barone estimates of anharmonicity corrections. Details are to be found in the Supporting Information. CP corrections reduce the estimated strength of intermolecular H-bonds and reduce the frequencies of relative motion of the monomers. For CP calculations the MAD in cm^{-1} for predicted frequencies relative to experimental values is 52 (unscaled) or 40 (scaled by the MP2 value, 0.953). For the structure optimized without CP corrections, the MAD in Barone estimates is 25 cm^{-1}. This value omits one serious outlier. Simple scaling of the harmonic values yields a comparable MAD.

Vibrational Fingerprinting of Dimeric Formic Acid Species. The C_{2h}-symmetric form **II** must be the dominant dimeric species in the gas phase, but irradiation may produce one or more of the low-lying isomers **III**, **IV**, and **V**. The presence of even rather small amounts of a less stable isomer can be verified by sufficiently sensitive vibrational spectrometry, so it is of interest to see if the less stable isomers have characteristic absorptions which will aid in their detection. Tables S4 and S5 (Supporting Information) contain computed frequencies for low-energy species **III, IV,** and **V**, the harmonic values corrected for anharmonicity and obtained with and without CP corrections.

The C_{2h} symmetry of some of formic acid's dimeric forms makes some transitions symmetry-forbidden. These weakly absorbing transitions are not suitable for fingerprinting. Frequencies of these transitions are parenthesized in Table 8. The lowest energy dimeric species (**II**) has a simple spectrum owing to its C_{2h} symmetry. Where **II** has a single intense absorption for the OH out-of-plane motion at about 920 cm^{-1}, the three energy-accessible isomers **III**, **IV**, and **V** have distinctive OH out-of-plane frequency doublets, shifted to the red relative to the frequency of the analogous motion in **II**. The mean red shifts and splitting in the doublets may be useful as a fingerprint. Those motions which seem most promising as fingerprints are highlighted in Table 8. **IV** could be recognized by its enormous red shift, as shown in Table 9. **V** would have a smaller red shift and a small splitting. **III** is recognizable by its substantial red shift and the large splitting. Compared with **II,** the CO stretching doublets for less stable dimers **III**, **IV**, and **V** also have substantially lower frequency. These do not seem to be so helpful as the OH out-of-plane absorptions since they are

Table 8. Harmonic (A) and Anharmonic (B) Frequencies for Low Energy Species[a]

system	species **II**	species **III**	species **IV**	species **V**
		(A)		
frequencies				
OCO bend	689, (675)	670, 645	663, 632	671, 645
COH oop	**917, (884)**	**854, 692**	**792, 651**	**801, 723**
OCH oop	1097, (1086)	1097, 1079	1075, 1073	1076, 1064
CO str	*1241, (1235)*	*1205, 1170*	*1191, 1107*	*1188, 1143*
COH bend	1390, (1400)	1375, 1333	1366, 1288	1358, 1299

system	species **II**	species **III**	species **IV**	species **V**
		(B)		
frequencies				
OCO	679, (668)	660, 640	643, 621	678, 649
COH oop	**935, (821)**	**806, 639**	**692, 579**	**771, 664**
OCH oop	1072, (1062)	1063, 1051	1048, 1043	1052, 1041
CO str	*1227, (1200)*	*1168, 1135*	*1160, 1065*	*1157, 1106*
COH bend	1358, (1348)	1349, 1292	1328, 1240	1317, 1256

[a] Bold and italicized frequencies are candidates for fingerprints distinguishing Species **III**, **IV**, and **V** from **II**.

Table 9. OOH Out-of-Plane Shifts and Doublet Splitting (cm^{-1})

species	**III**	**IV**	**V**
mean red shift (harmonic)	144	195	155
splitting (harmonic)	158	163	78
mean red shift (anharmonic)	213	299	151
splitting (anharmonic)	167	141	88

Table 10: Selected Harmonic Frequencies and Anharmonic Frequencies in cm^{-1} for Cis−Trans (**III′**) and Trans−Trans (**III**) Formic Acid Complexes[a]

		Harmonic Frequencies		
	species **III′**		species **III**	
modes	Y-T (full CP)	M et al	Y-T (full CP)	M et al
OCO	682, 674		670, 645	
COH out of plane	**899, 595**	946, 573	**854, 692**	936, 699
OCH out of plane	1071, 1069		1097, 1079	
CO stretch	*1204, 1139*	1213, 1153	*1205, 1170*	1205, 1156
COH in plane	1373, 1304		1333, 1375	

		Anharmonic Frequencies		
	species **III′**		species **III**	
modes	Y-T (no CP)	M et al	Y-T (no CP)	M et al
OCO	673, 668		660, 640	
COH out of plane	**874, 525**	946, 573	**806, 639**	936, 699
OCH out of plane	1060, 1052		1063, 1051	
CO stretch	*1184, 1128*	1213, 1153	*1168, 1135*	1205, 1156
COH in plane	1394, 1264		1349, 1292	

[a] Bold and italicized frequencies are candidates for fingerprints distinguishing species **III′** from **III**.

not so easily distinguishable. The weakly absorbing COH in-plane bends are still less useful.

Recently Marushkevich et al.[30] have produced a dimer incorporating both a cis and a trans formic acid, which seems to be the species **III′**. This species lies about 8.8 kcal/mol above the most stable dimer **II**. Frequencies computed by these investigators and our comparable values computed with full CP are recorded in Table 10. There is a potential fingerprint in the OH out-of-plane motion and (less marked) in the CO doublet. The OH out-of-plane motion distinguishes the species, the splitting in the doublet being much larger in **III′** than **III**. The same pattern is displayed in the COH in-plane bends and the CO (single bond) stretches, but to a lesser degree.

Conclusions

We have revisited the dimers of trans formic acid defined by Turi (**II−VIII** in his designation). We used a consistent model chemistry, MP2/6-311+G(d,p) including zero-point-energy corrections and employing full counterpoise (CP) corrections in the optimizations and vibrational frequency calculations. Optimization with systematic CP corrections produces structures with longer and weaker hydrogen bonds than optimization without CP corrections. It is interesting to note that single point CP corrections produce binding energies almost identical with the binding energies obtained by systematic CP corrections.

In addition we have characterized several dimers containing one cis formic acid monomer and one trans monomer. These are called **III′** (recently obtained experimentally) and **V′**. We also found a bound **Open** form with one conventional OH...O= bond but no secondary CH...O interaction. These species are all bound relative to two trans formic acid monomers. Any of the species **II-VII**, **III′**, and **V′** and **Open** may play a role in the association of formic acid monomers. **III**, **IV**, and **V** may appear in experiments in which formic acid vapor is irradiated with infrared photons. Our estimates of the activation barrier for the transformations **II→V** and **II→III** show that these conversions are energetically possible under such circumstances. We can say nothing definite about the passage of species **IV** to **II** or **III**.

The purpose of this study was to identify candidates for the "acyclic" form of the formic acid dimer inferred by Shipman et al. from pulse IR studies of formic acid vapor and to estimate possible characteristic infrared absorptions which could allow experimental identification of the intermediate. According to our estimates of harmonic and anharmonic frequencies, the OH out-of-plane motion and the CO stretches may serve to identify which of the accessible species **III, IV,** and **V** are produced in irradiation experiments.

Acknowledgment. We wish to thank the Body Foundation for equipment and encouragement; the Turkpetrol Foundation for costs of travel; the chemistry department and the office of the provost at the University of Virginia for financial and material assistance; Professor Zikri Altun for practical and personal support, and Professor Brooks Pate and Dr. Stephen Shipman for invaluable guidance. A portion of the data on water dimer (Table 6) was drawn from the M.S. Thesis of John Craig (Virginia Commonwealth University Department of Chemistry 2007).

Supporting Information Available: Numerous tables of structures and frequencies and Gaussian log files for all systems. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) All ZPE corrections obtained by unscaled harmonic frequencies. Use of the ZPE scaling suggested by Scott and Radom (ref 2) or use of anharmonic frequencies has little impact on the ΔZPE values.

(2) Scott, A. P.; Radom, L. *J. Phys. Chem*. **1996**, *100*, 16502.

(3) The convention for *trans* and *cis* nomenclature agrees with Brinkmann et. al. [ref 14] and most other investigators.

(4) Tsuzuki, S.; Uchimaru, T.; Matsumura, K.; Mikami, M.; Tanabe, K. *J. Chem. Phys*. **1999**, *110*, 11906.

(5) Matylitsky, V. V.; Riehn, C.; Gelin, M. F.; Brutschy, B. *J. Chem. Phys.* **2003**, *119*, 10553.

(6) Luckhaus, D. *J. Phys. Chem. A* **2006**, *110*, 3151.

(7) Shipman, S. T.; Douglass, P. C.; Yoo, H. S.; Hinkle, C. E.; Mierzejewski, E. L.; Pate, B. H. *Phys. Chem. Chem. Phys*. **2007**, *9*, 4572.

(8) Winkler, A.; Mehl, J. B.; Hess, P. *J. Chem. Phys.* **1993**, *100*, 2717.

(9) Boys, S. F.; Bernardi, F. *Mol. Phys*. **1970**, *19*, 553.

(10) Simon, S.; Duran, M.; Dannenberg, J. J. *J. Chem. Phys.* **1996**, *105*, 11024.

(11) Turi, L. *J. Phys. Chem*. **1996**, *100*, 11285.

(12) Qian, W.; Krimm, S. *J. Phys. Chem. A* **2001**, *105*, 5046.

(13) Chocholoušová, J.; Vacek, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2002**, *4*, 2119.

(14) 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.

(15) Brinkmann, N. R.; Tschumper, G. S.; Yan, G.; Schaefer, H. F. *J. Phys. Chem. A* **2003**, *107*, 10208.

(16) Marushkevich, K.; Khriachtechev, l.; Lundell, J.; Räsänen, M. *J. Am. Chem. Soc*. **2006**, *128*, 12060.

(17) Barone, V. *J. Chem. Phys*. **2005** *122*, 014108. Barone, V. *J. Chem. Phys.* **2004**, *120*, 3059.

(18) Luiz, G. M. R. S.; Scalabrin, A.; Pereira, D. *Infrared Phys. Technol.* **1997**, *38*, 45

(19) Braly, L. B.; Liu, K.; Brown, M. G.; Keutsch, F. N.; Fellers, R. S.; Saykally, R. J. *J. Chem. Phys*. **2000**, *112*, 10314.

(20) Keutsch, F. N.; Braly, L. B.; Brown, M. G.; Harker, H. A.; Petersen, P. B.; Leforestier, C.; Saykally, R. J. *J. Chem. Phys.* **2003**, *119*, 8927.

(21) Wuelfert, S.; Herren, D.; Leutwyler, S. *J. Chem. Phys*. **1987**, *86*, 3751.

(22) Huang, Z. S.; Miller, R. E. *J. Chem. Phys.* **1989**, *91*, 6613.

(23) Chaban, G. M.; Jung, J. O.; Gerber, R. B. *J. Chem. Phys.* **1999**, *111*, 1823.

(24) Bertie, J. E.; Michaelian, K. H. *J. Chem. Phys*. **1982**, *76* 886

(25) Marechal, Y. *J. Chem. Phys.* **1987**, *87*, 6344.

(26) Halupka, M.; Sander, M. *Spectrochim. Acta, Part A* **1998**, *54*, 495.

(27) Fernandez, L. E.; Gomez Marigliano, A. C.; Varetti, E. L. *Vibr. Spectrosc*. **2005** *37* 179

(28) Turi, L.; Danneberg, J. J. *J. Phys. Chem.* **1993**, 97 12197

(29) Dreyer, J. *J. Chem. Phys.* **2005**, *122*, 184306.

(30) Marushkevich, K.; Khriachtechev, l.; Lundell, J.; Räsänen, M. *J. Am. Chem. Soc*. **2006** *128.*

CT700161A

JCTC Journal of Chemical Theory and Computation

Computational Study on Cesium Azide Trapped in a Cyclopeptidic Tubular Structure

Nerina Armata,[†] John M. Dyke,[‡] Francesco Ferrante,[†] and Gianfranco La Manna*,[†]

Dipartimento di Chimica Fisica "F. Accascina'', Università degli Studi di Palermo, Viale delle Scienze, Parco d'Orleans II - 90128 Palermo, Italy, and School of Chemistry, University of Southampton, Southampton SO17 1BJ, United Kingdom

Received November 13, 2007

Abstract: The structures and the electronic properties of host−guest complexes formed by a cyclopeptidic tubular aggregate and the species CsN_3, $Cs_2(N_3)_2$, and Cs_2N_6 have been investigated by means of density functional theory. Taking advantage of the azide property to act as a bridge ligand between two or more metal cations, it may be possible to trap N_3^- ions inside a confined space. This could be important for the preparation of polynitrogen molecules N_n. Results show that there are significant attractive interactions between the azide ion and the cavity walls, which make the ion stay inside the inner empty space of the cyclopeptidic aggregate. The confinement of the species $Cs_2(N_3)_2$ forces the azide moieties to get closer together. Further, the Cs_2N_6 molecule shows a remarkable interaction with the tubular host, which may indicate a stabilization of N_6.

1. Introduction

Azide ions are versatile ligands, which can coordinate with metal cations, either via their ends or on their sides, like, e.g., copper(II), in $[Cu_2(\text{tetramethylethylenediammine})(N_3)(OH)](ClO_4)_2$,[1] and nickel(II), in $Ni(N_3)_2$(2,2-dimethylpropane-1,3-diamine),[2] as well as Cs and Zn in $Cs_2Zn(N_3)_4$.[3] The property of the N_3^- ion to act as a bridging ligand has been recently discovered to occur in crystals of crown ether complexes such as those formed by $[Cs([18]\text{-crown-6})(N_3)]_2$, where two azide ions form a bridge between two cesium cations each coordinated by crown ethers.[4] It would be useful to take advantage of this property in order to trap azide ions inside a confined space, which would prevent the crystallization of the metal azide, giving rise to possible high-energy releasing molecules.[5] We could speculate that this confined space would force two or more N_3^- ions close one to each other, promoting the formation of polynitrogen N_n clusters.

Our research group is interested in the computational study of the structural and electronic properties of open-ended organic tubular aggregates,[6] which can act as hosts by encapsulating small molecules or ions. In this paper we report a computational investigation of the structures and electronic properties of the host−guest complexes formed by a cyclopeptidic tubular structure and the following guest species: one cesium azide, two cesium azides, and the Cs_2N_6 molecule. There is an analogy between the system built up by two cesium azide molecules capped at each end by two crown ethers, and the system that would be formed by two cesium azide molecules inside the cavity of a tubular aggregate. In both cases the cesium ions can be coordinated by a number of oxygen atoms (ether-like in the first system, carbonylic in the second). However, inside the tubular aggregate the N_3^- ions would be in a confined region, and here could be activated with respect to the formation of a polynitrogen compound like N_6.[7,8]

The host species considered here is a covalent captured dimeric aggregate of the octacyclopeptide from D,L-alternated α-aminoacids, *cyclo*[(L-Ala-D-^{Me}N-Ala-L-Hag-D-^{Me}N-Ala-)$_2$], where D-^{Me}N-Ala is a N-methylated D-alanine residue and L-Hag indicates a residue of L-homoallylglycine. Two molecules of such a macrocycle are able to stack one on top of the other giving rise to an open-ended hollow tubular structure by the formation of eight interunit hydrogen bonds, which involve the carbonyl oxygens of one macrocycle and

* Corresponding author e-mail: lamanna@unipa.it.

[†] Università degli Studi di Palermo.

[‡] University of Southampton.

10.1021/ct700307r CCC: $40.75
Published on Web 02/16/2008

the complementary amidic hydrogens of the other macrocycle. The selective N-methylation of the cyclopeptide rules out the formation of aggregates larger than the dimer. In this system the homoallylglycine residues of the stacked units can react to give a covalent bridge, thus stabilizing the aggregate. The diameter of the tubular cavity is 9 Å ca., and one can estimate a cavity volume of 350 Å^3 if an height of 5.5 Å is considered. The space enclosed by the cavity is thus enough to host two cesium azide molecules. This system, which is very similar to that synthesized and characterized experimentally by Ghadiri et al.,[9] starting from the cyclopeptidic structural unit *cyclo*[(L-Phe-D-^{Me}N-Ala-L-Hag-D-^{Me}N-Ala-)$_2$], was chosen as a model of octacyclopeptidic tubular aggregates.

2. Computational Methods

The geometry optimization of all species discussed in this paper was performed by using the DFT generalized gradient functional BP86[10] along with the Resolution of Identity approximation (RI)[11] in its multipole accelerated variant.[12] In the RI-DFT approximation the electron density is expanded in a set of auxiliary basis functions centered on the nuclei. This procedure results in a reduction of the number of Coulomb integrals and remarkably smaller computational timing. The split valence plus polarization SV(P) basis set[13] was used for light atoms; a *p* function with an exponent of 0.8 has been added to the hydrogen centers, with a resultant contraction scheme of (7s4p1d/4s1p)/[3s2p1d/2s1p]. With regards to the cesium, a pseudopotential containing a polarization d function in the valence basis set was shown to give very satisfactory results on some organocesium compounds.[14] Since we are interested in a qualitative description of the structures of the considered systems, the Stuttgart ’97 relativistic small core pseudopotential[15] with no polarization d function was used for cesium atoms: it describes 46 core electrons, while a basis set with the contraction scheme (7s6p)/[5s3p] is associated with the 9 valence electrons.[16] The harmonic approximation was used to obtain the vibrational frequencies. The calculations were performed by using the TURBOMOLE v5.7 package.[17]

Single point calculations on the optimized geometries were performed by using two different functionals. For comparison with our previous studies on cyclopeptides, the B3LYP hybrid functional[18] was used to calculate the relative energies of different conformers of the structural cyclopeptidic unit and those of the host molecule. On the other hand, in order to obtain a better description of noncovalent contributions, which can play an important role in the host−guest interactions, the interaction energies between the guests and the host molecule were calculated by using the hybrid meta functional MPWB1K.[19] This functional allowed estimates of the van der Waals interactions in biologically relevant systems which were much better than those obtained with the B3LYP functional,[20] including a better description of the dependence on the reduced density gradient in molecular regions that are important for weak interactions. In both series of calculations, the cc-pvdz basis set[21] was used for light atoms and the 3-21G basis set[22] was used for cesium. The basis set superposition error (BSSE) in the host−guest interaction energies was estimated by means of the counterpoise procedure,[23] taking into account also the fragments’ relaxation energies. These calculations were performed by using the Gaussian 03 package.[24]

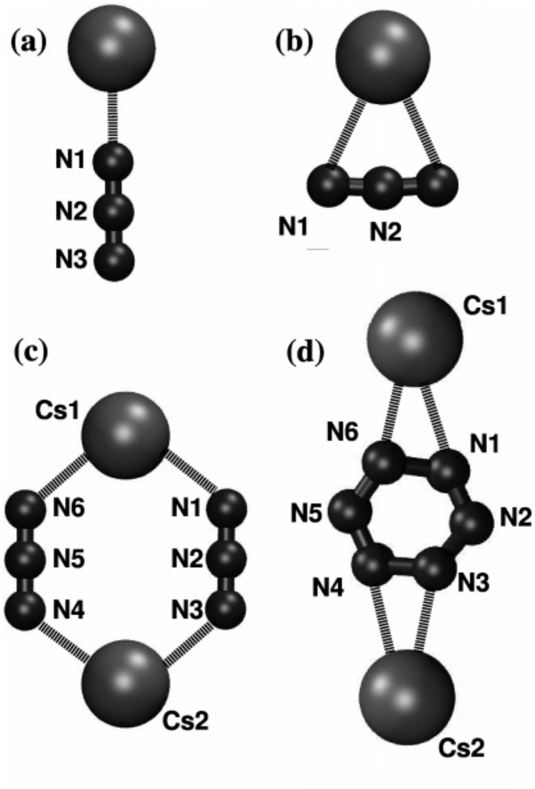


Figure 1. Optimized geometries of the isolated guest molecules: (a,b) linear and side-on CsN_3, (c) $Cs_2(N_3)_2$, and (d) Cs_2N_6. The adopted atoms labeling is reported.

3. Results and Discussion

3.1. The Guest Species. The optimized geometries of the guest molecules are depicted in Figure 1. Experimental characterizations and computational studies have shown[25] that the geometry of an alkaline metal azide in the gas phase or in a nitrogen matrix depends on the size of the alkaline cation. Smaller cations (Li, Na) give rise to linear molecules, while in the case of larger cations (Rb, Cs) a side-on geometry is the most stable, because both the N terminal atoms of N_3^- show a strong interaction with the central atom. However, the energy differences between the linear and side-on geometries, shown in Figure 1a,b, are not large enough to ultimately discriminate one form with respect to the other.

In the CsN_3 case the side-on geometry is 4 kJ mol^{-1} more stable than the linear one according to the computational approaches used in the present study, a value that is intermediate between the B3LYP (1 kJ mol^{-1}) and MP2 (11 kJ mol^{-1}) values reported by Dyke et al.[25] Therefore both the linear and side-on geometries of CsN_3 have been considered as possible guests for the tubular system.

The molecule $Cs_2(N_3)_2$ (Figure 1c) is built up with two azide ions connected by two cesium ions; it has D_{2h} symmetry. To our knowledge, this species has never been characterized as an isolated system, but a similar structure

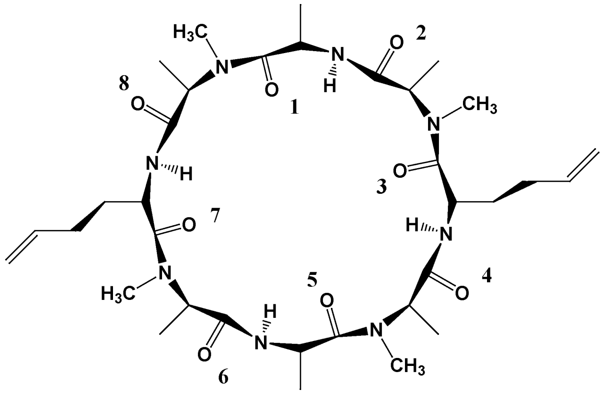


Figure 2. The octacyclopeptide *cyclo*[(L-Ala-D-^{Me}N-Ala-L-Hag-D-^{Me}N-Ala-)$_2$], the structural unit of the host molecule.

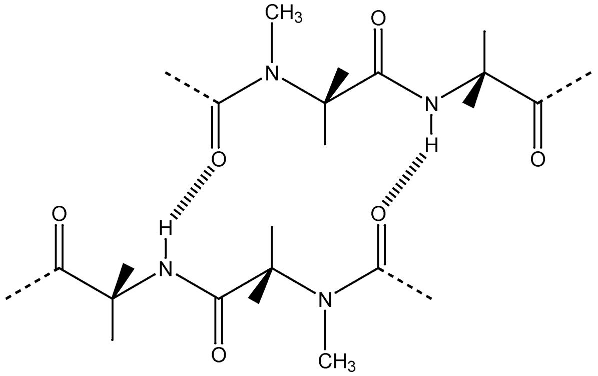


Figure 3. Linearized picture of the hydrogen bonds between two cyclopeptidic structural units stacked in an antiparallel fashion.

is adopted when two CsN_3 moieties are coordinated in the crown ether complex $[Cs([18]\text{-crown-}6)(N_3)]_2$.[4]

In this work, the Cs_2N_6 molecule, containing the hexanitrogen cluster (Figure 1d), has been subjected to a number of optimizations starting from initial geometries where the conformation of the cyclic N_6 and/or the relative positions of the two cesium ions were different. The results showed that a minimum where the two Cs atoms are perpendicular to the N_6 ring does not exist, at least at this level of calculation. On the other hand, when the cesium ions lie in the plane crossing the N_6 ring, different conformations (planar, boat, chair) of the ring converge to the same boatlike structure, which was checked to be a minimum by the absence of imaginary frequencies. This structure has a C_2 symmetry axis passing through the center of the N_6 ring. The geometry of the N_6 ring in Cs_2N_6 corresponds to that of the isolated N_6 with D_2 symmetry, which is a minimum according to refs 7 and 8 and to the optimization performed at the level of theory used here.

3.2. The Cyclopeptide. As a preliminary step, the geometry of the structural unit of the host molecule, *cyclo*[(L-Ala-D-^{Me}N-Ala-L-Hag-D-^{Me}N-Ala-)$_2$], (Figure 2) was optimized. Previous investigations[26] revealed that in the case of octacyclopeptides derived from D,L-alternated α-aminoacids only one conformation is able to stack through hydrogen bond formation (Figure 3). The structure of this conformation has been used to build the initial guess of the cyclopeptidic backbone. The side chains of homoallylglycine (Hag), however, can adopt different conformations. Since the stacking of the present cyclopeptide can occur only on the side without N-methylation, the only relevant conformations of the Hag side chain are those involving the dihedral angle $C\beta{-}C\gamma{-}C\delta{-}C\omega$, that is those corresponding to the rotation around the $C\gamma{-}C\delta$ bond. The energy of the molecule has been calculated (B3LYP) by performing a scan of the above dihedral angle for only one of the Hag side chains (step=24°). An absolute minimum has been obtained at the value of $-120°$, with two relative minima at 0° and 120°, respectively. The value 0° for the dihedral angle has been discarded, because the Hag side chain is directed toward the N-methylated region so that this conformation could not give rise to covalent capture. The $-120°$ and 120° values for the dihedral angle have been assigned to both the Hag residues, and the three conformers so obtained ($-120,-120$; -120,-120; 120,120) were fully optimized. The values $-120°$ and 120° correspond to different conformers, because of the C_4 symmetry of the cyclopeptide backbone. The most stable conformation is that having the dihedral angle $C\beta{-}C\gamma{-}C\delta{-}C\omega$ of $-120°$ on both Hag side chains.

3.3. The Host Molecule. Under appropriate conditions, the cyclopeptide described above gives rise to a dimeric aggregate, and the presence of the homoallylglycine residues allows a covalent capture of the two structural units. In such an aggregate eight hydrogen bonds are formed, but the covalent capture reaction can occur only if the two Hag residues belonging to the different structural units are spatially superimposed one on top of the other. So, in order to arrange the initial geometry of the aggregate, two cyclopeptidic units were stacked in their most stable conformation with regards to the allylic moiety, at the correct distance for the formation of the hydrogen bonds. According to the results of the optimization of this system, the most relevant geometric distortions caused by assembling with only hydrogen-bond formation are those which allow the atoms responsible for these bonds to lie at about 90° with respect to the backbone plane. These distortions cause small variations of the values of bond angles and dihedral angles, whereas the bond lengths of the backbone are almost unaffected.

The cyclopeptidic aggregate with covalent capture (Figure 4) was built starting from the optimized geometry of the same aggregate without covalent capture and considering the product of the reaction between the side chains of the Hag residues. Ghadiri et al.[9] characterized three different conformations for this kind of system. Therefore, the geometry optimization has been performed for the cis−cis, trans−cis, and trans−trans stereoisomers around the double bonds in both sides of the aggregate. The most important difference that can be seen from the structural analysis of these isomers is the difference in lengths of the eight hydrogen bonds, whose values are reported in Table 1. In particular, taking the hydrogen bond lengths in the aggregate with no covalent capture as reference values, a reasonable shortening (0.1 Å) of the four hydrogen bonds close to the covalent capture region is observed in the cis−cis isomer. In the trans−cis form there is a similar shortening for the hydrogen bonds close to the region where the capture is of cis-type, while

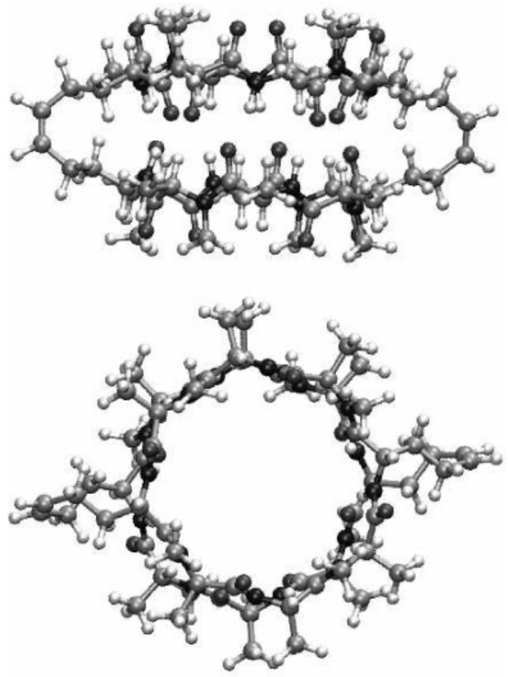


Figure 4. The optimized geometry of the host molecule with covalent bridge, from two different perspectives.

Table 1. Hydrogen Bond Lengths (Å) in the Host without Covalent Capture and in the Three Conformations of the Host with Covalent Capture (CC)

H bond	host	host CC cis−cis	host CC cis−trans	host CC trans−trans
1	1.913	1.824	1.832	2.001
2	1.912	1.918	1.930	1.868
3	1.913	1.921	1.859	1.884
4	1.913	1.815	1.972	1.987
5	1.910	1.821	2.003	1.963
6	1.911	1.909	1.862	1.854
7	1.912	1.932	1.918	1.858
8	1.914	1.821	1.831	1.992

the other two hydrogen bonds, in the trans capture zone, are longer by 0.1 Å. Finally, in the trans−trans isomer all four H-bonds in the covalent capture zone are longer than those placed where the capture is not present. Total energy values of trans−cis and trans−trans isomers are 9.5 and 14.5 kJ mol^{-1}, respectively, larger than the total energy of the cis−cis form (B3LYP results). This can be explained by considering that a covalent bridge in the trans conformation gives rise to a ring strain which disfavors the formation of hydrogen bonds. As a result, the cis−cis isomer of the covalent-captured cyclopeptidic aggregate was used here as a host molecule.

3.4. CsN_3@host. For this system it is possible to propose two starting geometries: one where the cesium atom is coordinated to the carbonyl oxygens not involved in H-bond formation and the azide ion is outside the cavity of the host, and another where the cesium is coordinated the same way but the azide is placed inside the cavity (Figure 5). Both systems were subjected to geometry optimization.

3.4.1. Linear Form of CsN_3. A geometric distortion of the host which enables the approach of the oxygen atoms to cesium was observed, as a result of the variations of the dihedral angles around the peptidic bonds in the two structural unit backbones. This distortion leads the cesium ion to be placed at 3.3−3.5 Å from three of the four oxygen atoms. Values for the Cs−O distances, shown in Table 2 along with the Cs−O distances in the other host−guest systems, are on average 0.4 Å longer than those evaluated by X-ray diffraction experiments in the [Cs([18]-crown-6)-$(N_3)]_2$ complex and are in good agreement with those obtained by calculations at the same level of accuracy on that system.[4]

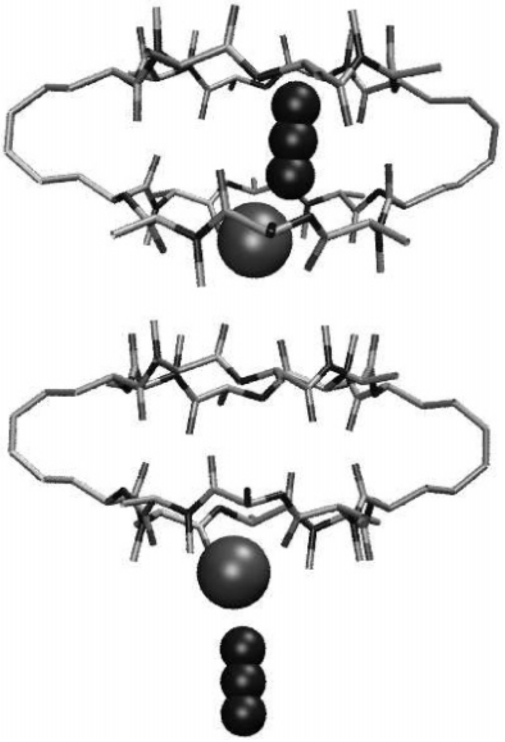


Figure 5. The optimized geometries of the CsN_3@host (a) and CsN_3_out_host (b) complexes. The CsN_3 guest is emphasized.

Table 2. Cesium−Oxygen Distances (Å) in the Host−Guest Complexes

	CsN_3@	CsN_3out	$Cs_2(N_3)_2$@	Cs_2N_6@(I)	Cs_2N_6@(II)
Cs1···O2a[a]	4.589/ 4.213[c]	5.504	4.058	4.592	4.605
Cs1···O4a	3.361/ 3.347	3.329	3.507	3.456	3.388
Cs1···O6a	3.524/ 3.443	3.501	3.713	3.413	3.477
Cs1···O8a	3.532/ 3.405	3.409	3.571	3.441	3.432
Cs2···O2b[b]			4.222	4.602	4.603
Cs2···O4b			3.687	3.476	3.502
Cs2···O6b			3.604	3.390	3.477
Cs2···O8b			3.541	3.432	3.441

[a] Refers to the terminal carbonylic oxygens of the first structural unit. [b] Refers to the terminal carbonylic oxygens of the second structural unit. [c] CsN_3 with end-on geometry on the left, with side-on geometry on the right.

When the cesium azide is inside the cavity it loses its linear geometry, with the Cs−N−N angle equal to 150.7°. The distortion from linearity is attributable to the interaction between the azide ion and the internal walls of the cavity,

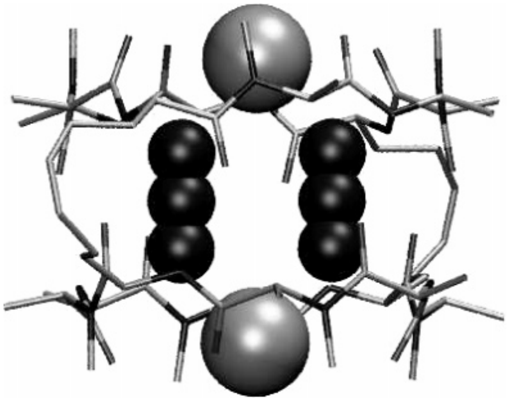


Figure 6. The optimized geometry of the $Cs_2(N_3)_2$@host complex. The $Cs_2(N_3)_2$ guest is emphasized.

to which the azide group gets close. Further, when the azide is outside of the cavity there is only a negligible loss of linearity, with a Cs−N−N angle of 172.5°. In both systems, the Cs−N bond length is increased with respect to isolated CsN_3 (by 0.2 Å when N_3 is inside the cavity, by 0.1 Å when it is outside), whereas there are no significant variations of the N−N bond lengths. The interaction energy values (MPWB1K results), corrected for BSSE, are -77.5 kJ mol^{-1} when the azide is inside the cavity, -69.2 kJ mol^{-1} when it is outside.

3.4.2. Side-On Geometry of CsN_3. The geometry of the system formed by CsN_3 having side-on geometry inside the tubular aggregate was also optimized. An inspection of Table 2 reveals a slightly larger distortion of the cyclopeptidic backbone, driven by the dragging of oxygen atoms toward the cesium, which is more positively charged than in the previous linear case. The N_3 moves toward the cavity walls, placing its two N terminal atoms at 3.02 and 3.28 Å from two carbonyl carbon atoms, respectively. The increase of the Cs−N bond length is about 0.2 Å on either side, whereas the N−N bond lengths are essentially unchanged. The interaction energy is -72.8 kJ mol^{-1}.

3.5. Cs_2 $(N_3)_2$@host. The next step was to build the system formed by the cyclopeptidic structure as the host and two CsN_3 molecules inside it. The initial geometry of the $Cs_2(N_3)_2$ moiety was that obtained from the optimization of the isolated molecule. Two cesium ions were placed along the longitudinal axis of the aggregate, each one in the perpendicular plane passing through the terminal oxygen atoms. Once the full optimization was performed, the structure of the host molecule shows the same geometrical distortions found in CsN_3@host, namely the oxygen atoms get closer to the cesium ion, on both terminal sides (Figure 6). The two azide ions are placed almost exactly at the center of the cavity.

The most relevant issue is that, inside the cavity, the two azide ions are closer to each other (by 1.1 Å) than they are in the isolated $Cs_2(N_3)_2$ species (see the comparison in Table 3). This phenomenon must be due to the increase of the Cs−Cs distance (>1 Å), caused by Cs−O interactions, along with the influence of the confined space in which the two N_3^- ions are trapped. Since changes are not observed in the Cs−N distances or in the N−N bond lengths, the increase of the Cs−Cs distance causes a decrease of the N−Cs−N angle, so that the two azides get closer. It is noteworthy that the species $Cs_2(N_3)_2$ maintains its D_{2h} symmetry inside the cavity, probably because of the local C_4 symmetry of the cavity itself. The interaction distances between the terminal N atoms of the azide groups and the carbonyl C atom of the cavity walls are between 2.80 and 3.01 Å, a range that is slightly lower than the one calculated in the CsN_3@host case. The interaction energy, computed by using the MPWB1K functional on the optimized geometries and corrected for BSSE, is -40.3 kJ/mol.

Table 3. Optimized Geometric Parameters (Å) of Isolated CsN_3 and $Cs_2(N_3)_2$ and of Trapped $Cs_2(N_3)_2$ and XRD Values of the $Cs_2(N_3)_2$ Moiety in the [Cs([18]crown-6)(N_3)]$_2$ Complex[4] [c]

	CsN_3	$Cs_2(N_3)_2$	$Cs_2(N_3)_2$@	crown complex
Cs1···Cs2[a]	-	5.910	7.031	4.679 (4.982)
Cs1−N1	2.733/3.077[b]	3.004	3.003	3.244 (3.197)
N1−N2	1.209/1.191	1.194	1.194	1.179 (1.201)
N2−N3	1.185/1.191	1.194	1.194	1.179 (1.189)
N3−Cs2	-	3.004	3.003	-
N1···N6	-	4.801	3.726	-

[a] See Figure 1 for atom labeling. [b] Linear geometry on the left, side-on geometry on the right. [c] Calculated values in parentheses.

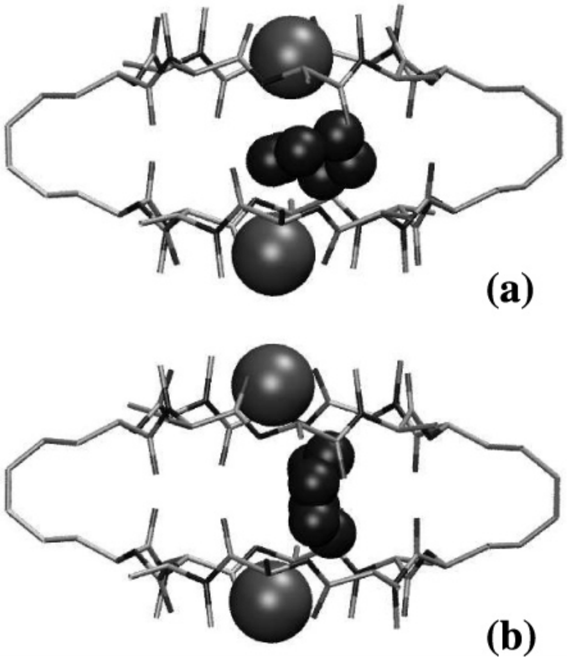


Figure 7. The optimized geometries of the Cs_2N_6@host complexes: **I** form (a) and **II** form (b). The Cs_2N_6 guest is emphasized.

3.6. Cs_2N_6@host. The last system investigated is composed of the N_6 cyclic boatlike polynitrogen species inside the host cavity, with two cesium atoms coordinated to the carbonylic oxygens. Starting from two slightly different geometries, two positions for the trapped N_6 have been obtained. In the first of these, (**I**) (Figure 7a), the plane which crosses the N_6 cycle is parallel to the stacking axis of the cyclopeptidic aggregate; in the second one, (**II**) (Figure 7b), that plane is perpendicular to this stacking axis. In both cases, the polynitrogen moves toward the cavity walls. By comparing the geometry of the isolated N_6 with that adopted inside

Table 4. Optimized Geometric Parameters of Isolated and Trapped Cs_2N_6[b]

	N_6	Cs_2N_6	Cs_2N_6@(I)	Cs_2N_6@(II)
Cs1···Cs2[a]		7.880	6.361	7.063
N1−N2	1.308	1.280	1.412	1.396
N2−N3	1.308	1.347	1.432	1.277
N6−N1	1.370	1.460	1.279	1.418
Cs1−N6		2.911	3.395	3.185
Cs1−N1		3.005	3.457	2.958
Cs1−N5		4.057	3.112	4.345
Cs1−N2		4.270	3.326	3.554
N4−N5			1.404	1.394
N5−N6			1.426	1.277
N3−N4			1.271	1.418
Cs2−N3			3.123	3.170
Cs2−N4			3.123	2.957
Cs2−N5			3.123	3.682
Cs2−N2			3.168	4.372
N1−Cs1−N6		28.5	21.5	26.3
N6−N1−N2	111.3	116.2	117.9	117.2
N1−N2−N3	125.0	119.2	117.2	117.1
N2−N3−N4	111.3	110.3	107.3	108.3
N4−Cs2−N3			17.2	24.4
N3−N4−N5			117.5	117.1
N4−N5−N6			117.8	117.3
N5−N6−N1			116.8	108.4
Cs1−N6−N1−N2		172.7	62.0	−155.1
N6−N1−N2−N3	−22.1	9.3	42.2	44.2
N2−N3−N4−N5	42.9	34.2	2.0	−45.3
Cs2−N3−N4−N5			−88.5	109.9
N3−N4−N5−N6			44.2	44.0
N5−N6−N1−N2			3.6	−45.4

[a] See Figure 1 for atom labeling. [b] The geometric parameters of the isolated N_6 cluster (D_2 symmetry), optimized at the same level of theory used here, are reported for comparison. Lengths in Å, angles in deg.

the cavity (Table 4), a decrease of the Cs−Cs distance is observed, which is more marked in (**I**); further, in the three cases investigated here, isolated Cs_2N_6 and trapped Cs_2N_6 with **I** and **II** geometries, the cyclic polynitrogen has different positions with respect to the axis passing through the two cesium atoms, while maintaining the boatlike structure. In both **I** and **II** systems the Cs−O distances are very similar (Table 1), even if in **I** the cesium atoms lie much more inside the cavity. The interaction energies, corrected for BSSE, are −126.9 kJ/mol^{-1} and −84.6 kJ mol^{-1} for **I** and **II**, respectively. The distances between the atoms involved in the interaction (N of the polynitrogen and carbonylic carbons) are in the range 3.0−3.2 Å in **I** and 3.2−3.5 Å in **II**. Although these distances from the cavity walls are greater than those existing in $Cs_2(N_3)_2$@host, the interaction energies are 2−3 orders of magnitude higher. This seems to indicate an intrinsic stabilization of Cs_2N_6 inside the cavity of the cyclopeptidic aggregate.

In order to evaluate the energy of the fragmentation of the N_6 polynitrogen to three N_2 molecules inside the cavity, the system $Cs_2(N_2)_3$@host, where the Cs atoms are coordinated to the carbonylic oxygen atoms and the three N_2 molecules are inside the cavity of the cyclopeptidic system, has been subjected to geometry optimization and single point energy calculation. The results obtained indicate that the conversion from **I** to $Cs_2(N_2)_3$@host should release 296 kJ mol^{-1} of energy.

4. Conclusion

The present study is intended to suggest a new strategy which gives rise to the confinement of azide ions with the aim of producing polynitrogen compounds, which could be interesting high-energy molecules. In the case investigated, involving the confinement of cesium azide units in a cyclopeptidic tubular structure, the azide ions lie inside the cavity of the host molecule, with the cesium ion coordinated to three carbonylic oxygen atoms of the host. When the $Cs_2(N_3)_2$ species is enclosed in the cavity, the two azide ions are in close proximity, and they may, as a result, be activated to form the polynitrogen species N_6, which could be stabilized in the cyclopeptidic tubular aggregate. Further investigations are currently in progress concerning the interconversion from $Cs_2(N_3)_2$ to Cs_2N_6 inside the host cavity. A possible way to prepare these complexes might involve cocondensation of cesium azide and the cyclopeptide units and the characterization of the products with spectroscopic methods.

Acknowledgment. This work was performed with the contribution of the 60%-funds from the Italian Ministry of Scientific Research.

References

(1) Kahn, O.; Sikorav, S.; Gouteron, J.; Jeannis, S.; Jeannis, Y. *Inorg. Chem.* **1983**, *22*, 2877.

(2) Monfort, M.; Ribas, J.; Solnas, X. *J. Chem. Commun.* **1993**, 350.

(3) Mautner, F. A.; Krischner, H. *Monat. Chem. (Chem. Mon.)* **1990**, *121*, 91.

(4) Brown, M. D.; Dyke, J. M.; Ferrante, F.; Levason, W.; Ogden, J. S.; Webster, M. *Chem. Eur. J.* **2006**, *12*, 2620.

(5) Talavar, M. B.; Sivabalan, R.; Asthana,. N.; Singh, H. *Combust., Explosion Shock Waves* **2005**, *41*, 264 and references therein.

(6) Ferrante, F.; La Manna, G. *J. Phys. Chem. A* **2003**, 107, 91. Ferrante, F.; La Manna, G. *Chem. Phys. Lett.* **2004**, *383*, 376. Ferrante, F.; La Manna, G. *J. Comput. Chem.* **2007**, *28*, 2085.

(7) Glukhovtsev, M. N.; Jiao, H.; Schleyer, P. v. R. *Inorg. Chem.* **1996**, *35*, 7124.

(8) Klapötke, T. M. *J. Mol. Struct. (Theochem)* **2000**, *499*, 99.

(9) Clark, T. D.; Ghadiri, M. R. *J. Am. Chem. Soc.* **1995**, *117*, 12364.

(10) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098−3100. Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822.

(11) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283. Eichkorn, K.; Weigend, F; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1998**, *97*, 112.

(12) Sierka, M.; Hogekamp, A.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *118*, 9136.

(13) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571.

(14) Streitwieser, A.; Liang, J. C.-Y.; Layasree, E. G.; Hasanayn, F. *J. Chem. Theory Comput.* **2007**, *3*, 127.

(15) Leininger, T.; Nicklass, A.; Küchle, W.; Stoll, H.; Dolg, M.; Bergner, A. *Chem. Phys. Lett.* **1996**, *255*, 274.

(16) Institut für Theoretische Chemie, Universität Stuttgart; ECPs and corresponding basis sets. http://www.theochem.uni-stuttgart.de/. Cesium Stuttgart RSC '97 effective core potential is indicated as ECP46MWB.

(17) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kolmell, C. *Chem. Phys. Lett.* **1989**, *162*, 165. Häser, M.; Ahlrichs, R. *J. Comput. Chem.* **1989**, *10*, 104. Von Arnim, M.; Ahlrichs, R. *J. Comput. Chem.* **1998**, *19*, 1746.

(18) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(19) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040. Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664. Zhao, Y.; Trulhar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908.

(20) Zhao, Y.; Trulhar, D. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2701.

(21) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(22) Glendening, E. D.; Feller, D. *J. Phys. Chem.* **1995**, *99*, 3060.

(23) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.

(24) M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, J. A. Pople *Gaussian 03, Revision D.02*; Gaussian, Inc.: Wallingford, CT, 2005.

(25) Ogden, J. S.; Dyke, J. M.; Levason, W.; Ferrante, F.; Gagliardi, L. *Chem. Eur. J.* **2006**, *12*, 3580.

(26) Ghadiri, M. R.; Granja, J. R.; Milligan, R. A.; McRee, D. E.; Khazanovich, N. *Nature* **1993**, *366*, 324. Hartgerink, J. D.; Granja, J. R.; Milligan, R. A.; Ghadiri, M. R. *J. Am. Chem. Soc.* **1996**, *118*, 43.

CT700307R