JCTC Journal of Chemical Theory and Computation

Spin-Unrestricted Calculations of Bare-Edged Nanographenes Using DFT and Many-Body Perturbation Theory

Rodolphe Pollet*,† and Hakim Amara‡

DSM/IRAMIS/SPAM-LFP, CEA, Gif-sur-Yvette, France, and Laboratoire d'Etude des Microstructures, ONERA-CNRS, Châtillon, France

Received April 15, 2009

Abstract: The ability of Density Functional Theory to predict the electronic and magnetic properties of semi-infinite graphene with a single bare edge has been probed. In order to improve the accuracy of spin-unrestricted calculations performed with semilocal density functionals, higher-level methods including double hybrid density functionals and many-body perturbation theory have been applied to the polycyclic aromatic hydrocarbons model systems. We show that the antiferromagnetic or ferromagnetic tendencies of the corresponding electronic ground states strongly depend on the choice of the density functional. In addition the relative stability of the armchair and zigzag edges has been investigated, emphasizing the importance of using methods beyond semilocals density functionals.

Graphene, which is a monolayer of carbon atoms packed into a dense honeycomb crystal structure, belongs to the rich world of carbon nanostructures.[1] It has been the subject of intense scrutinity,[2,3] especially because this kind of structure was previously presumed not to exist in the free state. This exciting object, with unusual electronic and magnetic properties, is considered to be among the most important materials for nanoscale device applications.[4] However, all those remarkable properties can be strongly affected by the presence of edges which are in general a combination of armchair and zigzag regions.[5−7] In particular, theoretical and experimental studies have been devoted to the demonstration that edges of graphene substantially modify their electronic and magnetic properties.[8−10] Therefeore, much effort has been focused on the study of the edges in graphitic nanomaterials, such as graphene and also nanotubes. Indeed, the strong adhesion between the edge of a tube and the metal clusters from which they are produced is a requirement for the nanotube growth.[11]

Graphene exists as flakes also called nanographenes or, for the smallest sizes, polycyclic aromatic hydrocarbons (PAHs). A large spectrum of theoretical methods can therefore be applied according to the size of the system, ranging from intuitive (semiempirical) tight-binding models to highly accurate (ab initio) wave function based methods. Electronic structure calculations of nanomaterials extensively rely on Density Functional Theory (DFT) and especially on semilocal approximations to the exchange-correlation density functional, such as the Local Density Approximation (LDA) or Generalized Gradient Approximation (GGA).[12] The favorable computational cost of these methods together with periodic boundary conditions (PBC) have allowed calculations on infinite or semi-infinite graphene,[13−17] using the pseudopotential plane-waves method. In addition single-hybrid approximations, incorporating a small proportion of exact exchange, have been used for periodic graphenes thanks to a screening procedure.[18−20] Finite nanographenes have also been studied using DFT with a strong focus on hydrogen-terminated PAHs[20−22] (save one exception[23]).

In this Letter, we adopt a new approach, using many-body perturbation theory as well as more elaborate density functionals, to investigate the electronic and magnetic properties of edges present in PAHs. Different PAHs are considered with all the edges hydrogenated but one, thus mimicking semi-infinite graphenes with bare active sites. One can expect that such higher-level methods should indeed be required for an accurate description of edges. This is necessary to understand the key role played by edges in the modification of the properties of PAH as well as their reactivity.

The efficiency of semilocal approximations has been probed by using the Perdew-Burke-Ernzerhof (PBE)[24] GGA density functional in conjunction with triple-ζ plus polarization Gaussian basis sets (TZVPP) and the resolution-of-the-identity (RI) approximation[25] implemented in the TURBOMOLE quantum chemical program package version 5.10.[26] Approximations depending not only on the electron density and on its gradient but also on the Kohn−Sham orbital kinetic energy density, called Meta-Generalized Gradient Approximations (Meta-GGA), have also been considered with the TPSS[27] nonempirical density functional. Wave function based methods constitute a usually more accurate but also time-consuming alternative. For medium-sized systems, second order Møller−Plesset (MP2) perturbation

* Corresponding author e-mail: rodolphe.pollet@cea.fr.
† CEA Saclay.
‡ Onera Châtillon.

10.1021/ct900184d CCC: $40.75
Published on Web 06/01/2009

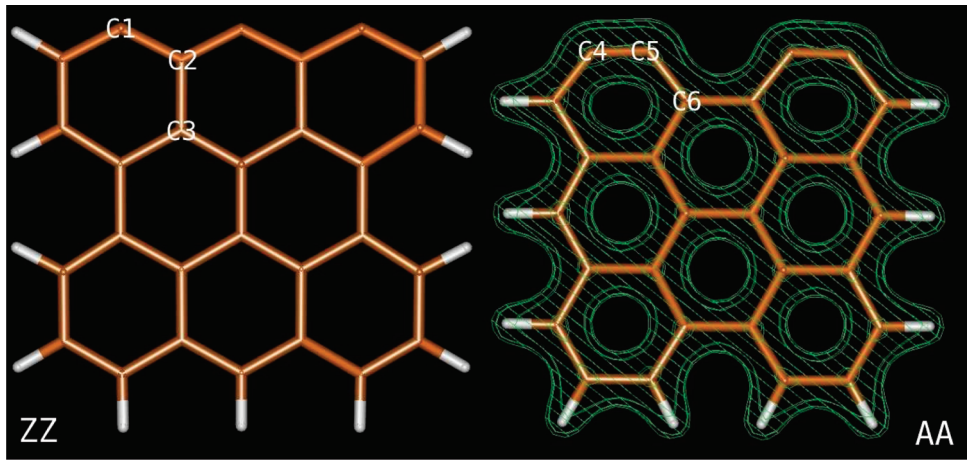


Figure 1. Zigzag (ZZ) and armchair (AA) edges of the bare-edged PAH[3,3] system optimized at the PBE/TZVPP level of theory. For AA, contours of the PBE electronic density are also plotted.

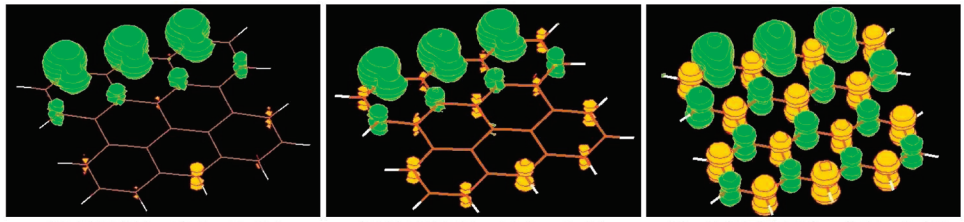


Figure 2. Spin density (green positive, yellow negative) for the zigzag edge of PAH[3,3] obtained with the PBE, TPSS, and B2-PLYP density functionals. The expectation value of the $\mathbf{S^2}$ operator is also reported.

theory (together with the RI approximation) represents an affordable yet powerful method in comparison with DFT, especially in the case of dispersive interactions. Recently a simple, partly empirical, improvement of the method has been proposed by scaling the spin components (SCS) of the pair correlation energies.[28] Like its parent method, the SCS-MP2 method can however fail to describe open-shells species, such as the zigzag (bare) edge of PAHs (*vide infra*). Double-hybrid density functionals (DHDF),[29,30] which lie between pure density and wave function approaches, appear more robust with regard to spin contamination problems occurring in spin-unrestricted calculations.[31] The double hybrid functional tested in this work (B2-PLYP) mixes the B88[32] exchange functional with Hartree−Fock-like exchange and LYP[33] correlation functional with MP2-like correlation energy.

Due to the presence of unpaired electrons along the zigzag (ZZ) edge of graphene, special care must be taken to ensure that first principles calculations correctly reproduce the spin density distribution, i.e., $\rho_s(\mathbf{r}) = \rho_\alpha(\mathbf{r}) - \rho_\beta(\mathbf{r})$, of the electronic ground state.[23] Prior to the energetical comparison between both edges, methods causing a deterioration of $\rho_s(\mathbf{r})$ must therefore be identified and better avoided. This investigation has begun with the bare-edged PAH[3,3] system (28 carbon and 11 hydrogen atoms), where [*x*,*y*] specify the number of adjacent cycles in the horizontal (*x*) and vertical (*y*) directions (see Figure 1). Isosurfaces corresponding to the PBE, TPSS, and B2-PLYP density functionals are reported in Figure 2 together with the expectation value of $\mathbf{S^2}$. Whether the latter value should be used as a diagnostic tool to judge the quality of a spin-unrestricted calculation is still controversial as far as DFT is concerned because only the noninteracting Kohn−Sham wave function is used for the estimation.[34] In this case a clear correlation between the error on $\langle \mathbf{S^2} \rangle$ (equal to 3.75 for a pure quadruplet) and the delocalization of $\rho_s(\mathbf{r})$ can be observed. This effect is particularly significant in the case of the B2-PLYP functional and can be most probably attributed to the large fraction (≈50%) of exact exchange. For comparison, a spin-unrestricted Hartree−Fock (UHF) calculation results in $\langle \mathbf{S^2} \rangle$ as high as 6.32. All density functionals predict a spin density distribution with large positive contributions corresponding to the unpaired electrons of the unsaturated carbons and smaller negative contributions on the opposite hydrogenated edge due to spin polarization effects. These very characteristics have been also observed for the smaller PAH[4,1] system from B3LYP spin-unrestricted calculations.[23] This is also similar to the antiferromagnetic ground state of PAHs fully saturated by hydrogen.[20,22] We have then considered the larger zigzag bare-edged PAH[5,5] (66 carbon and 17 hydrogen atoms). Again the spin density exhibits a larger delocalization when the double hybrid density functional is used (see Figure 3). Interestingly the largest $\langle \mathbf{S^2} \rangle$ value has been obtained for the PBE calculation, whose convergence was particularly difficult to achieve. But the main difference with respect to the smaller PAH[3,3] system is the spin polarization on the opposite, hydrogenated, edge. Contributions to the spin density on this side are indeed positive for all density functionals, showing a tendency toward ferromagnetism. This is not only in contrast to the case of the PAH[3,3] system but also of the fully hydrogenated PAHs, where the energy of the ferromagnetic state always lies higher than for the antiferromagnetic state.[20,22]

We now compare the structures and energetical stability of the ZZ and AA edges. Geometry optimizations have been

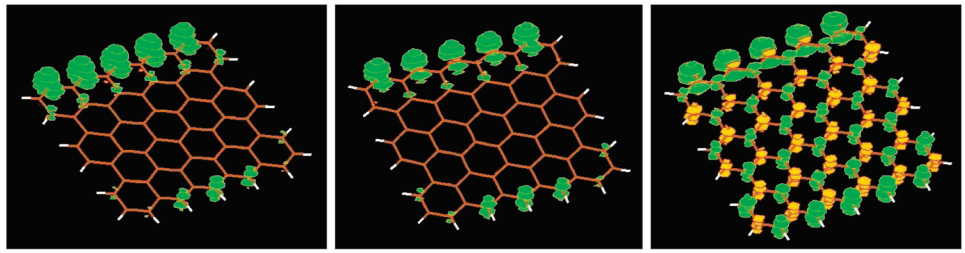


Figure 3. Spin density (green positive, yellow negative) for the zigzag edge of PAH[5,5] obtained with the PBE, TPSS, and B2-PLYP density functionals. The expectation value of the $\mathbf{S}^2$ operator is also reported.

Table 1. Structural Parameters of the Zigzag (ZZ) and Armchair (AA) Edges Obtained from PBE/TZVPP Geometry Optimizations of the PAH[3,3] System (See Labels on Figure *1*)

parameters	PAH[3,3]	∞ nanoribbons
C1−C2 (Å)	1.38	1.37,[35] 1.38,[13] 1.39[15]
C2−C3 (Å)	1.47	1.47[15]
C4−C5 (Å)	1.26	1.23,[13,16,35] 1.24[15]
C5−C6 (Å)	1.38	1.39[15,16]
C4−C5−C6 (°)	127	126[15]

performed at the PBE/TZVPP level of theory. Interestingly the characteristic distances and angle (see Table 1) are in excellent agreement with DFT calculations of infinite nanoribbons. As previously noted,[13,35] the shortening of the C−C bond at the AA edge can be rationalized by the formation of a triple-bond, in agreement with the accumulation of electronic density between both carbon atoms (see Figure 1). The bond length is actually closer to the distance observed in benzyne (1.27 Å)[36] than in acetylene (1.20 Å). The harmonic vibrational frequencies associated with the symmetric and antisymmetric C−C stretch combinations are found at 1921 and 1948 cm^{-1}, respectively. These values are strongly blue-shifted (approximately 400 cm^{-1}) with respect to the other C−C stretches in the 1600−1500 cm^{-1} range: as an example the frequency of the middle C−C bond of the bare edge is found at 1539 cm^{-1}. The present calculations show that evidence for the formation of triple bonds at the armchair edge can be found in the IR spectra of these species. In spite of the weak intensities calculated for these bands, their isolated position in the spectrum should make their experimental observation relatively easy.

Measuring the relative reactivity of both edges should help to identify the best candidate as a precursor to a nanotube growing on a metallic particle. We have calculated the C−H bonds dissociation energy

$$E_{diss} = \frac{1}{n}(E_{bare} + nE_H - E_{H-term}) \qquad (1)$$

where n is the number of unsaturated carbon atoms, E_{bare} is the energy of the bare-edged AA or ZZ system, E_H is the energy of the hydrogen atom, and E_{H-term} is the energy of the fully hydrogenated PAH. Our work has been however not restricted to density functional methods and includes the perturbational approaches detailed previously.[37] Results reported in Table 2 confirm that, except for the MP2 calculation, all methods tested in this work predict that the bare-edged AA structure is more stable than the ZZ one. The SCS procedure succeeds in correcting the MP2 failure and leads to energies in excellent agreement with the double hybrid density functional calculation. Both methods also predict AA and ZZ energies close to the values obtained from DFT calculations of infinite nanoribbons (respectively, 4.36 and 5.36 eV[15]). The PBE density functional significantly underestimates (0.25 eV) the difference between the ZZ and AA dissociation energies, mainly because of too weak C−H bonds at the ZZ edge. We note that the PBE/TZVPP dissociation energy of the ZZ edge is in excellent agreement with the value obtained by May et al.[21] (5.04 eV) on the same system with another GGA density functional. Following Koskinen et al.,[15] the reactivity of both bare-edged structures can also be judged by considering the following reaction

$$\mathrm{P\cdot} + \mathrm{H_2} \xrightarrow{-E'_{diss}} \mathrm{PH} + \mathrm{H\cdot}$$

where P· represents the bare-edged system, PH represents the H-terminated PAH, and E'_{diss} is calculated by subtracting the H_2 dissociation energy to the C−H bonds dissociation energy. The previous reaction can be seen as the first bimolecular reaction step in the mechanism of hydrogenation. According to the values of E'_{diss} reported in Table 3, the bimolecular reaction is endergonic for the AA edge and exergonic for the ZZ edge. The difference is obviously related to the presence of highly reactive dangling bonds in the ZZ case.

In conclusion, magnetic properties resulting from spin-unrestricted DFT calculations are not only influenced by the

Table 2. C−H Bonds Dissociation Energy (eV) for Armchair (AA) and Zigzag (ZZ) Edges of PAH[3,3] Optimized at the PBE/TZVPP Level of Theory[a]

method	AA	ZZ	Δ*E*
PBE/TZVPP	4.37	5.02	0.65
B2-PLYP/TZVPP	4.46	5.36	0.90
MP2/TZVPP	4.34	4.08	−0.27
SCS-MP2/TZVPP	4.44	5.34	0.90

[a] Difference between both values is reported in the last column.

Table 3. Difference between C−H bonds Dissociation Energy and H_2 Dissociation Energy (eV) for Armchair (AA) and Zigzag (ZZ) Edges of PAH[3,3] Optimized at the PBE/TZVPP Level of Theory

method	AA	ZZ
PBE/TZVPP	−0.17	0.48
B2-PLYP/TZVPP	−0.17	0.73
MP2/TZVPP	−0.15	−0.41
SCS-MP2/TZVPP	−0.22	0.68

choice of the density functional, through the delocalization of the spin density, but also by the size of the system, which can alter the relative weight of the antiferromagnetic-like and ferromagnetic-like states. In addition, although all methods tested in this work confirm the stronger stability of the armchair with respect to the zigzag bare edge, semilocal density functionals appear to underestimate this energetical preference. Periodic boundary calculations of nanoribbons using such a method should therefore consider this bias demonstrated here from high level calculations of closely related finite-size systems.

Acknowledgment. The authors thank Eric Gloaguen, Michel Mons, and Cyrille Barreteau (all from CEA/DSM/IRAMIS) for fruitful discussions.

References

(1) Delgado, J. L.; Herranz, M. Á.; Martín, N. *J. Mater. Chem.* **2008**, *18*, 1417.

(2) Novoselov, K. S.; Geim, A. K.; Morozov, S. V.; Jiang, D.; Zhang, Y.; Dubonos, S. V.; Grigorieva, I. V.; Firsov, A. A. *Science* **2004**, *306*, 666.

(3) Zhang, Y.; Tan, Y.-M.; Stormer, H. L.; Kim, P. *Nature* **2005**, *438*, 201.

(4) Castro Net, A. H.; Guinea, F.; Peres, M. R.; Novoselov, K. S. *Rev. Mod. Phys.* **2009**, *81*, 109.

(5) Liu, Z.; Suenaga, K.; Harris, P. J. F.; Ijima, S. *Phys. Rev. Lett.* **2009**, *102*, 015501.

(6) Jia, X.; Hofmann, M.; Meunier, V.; Sumpter, B. G.; Campos-Delgado, J.; Romo-Herrera, J. M.; Son, H.; Hsieh, Y.-P.; Reina, A.; Kong, J.; Terrones, M.; Dresselhaus, M. S. *Science* **2009**, *323*, 1701.

(7) Girit, Ç. Ö.; Meyer, J. C.; Erni, R.; Rossell, M. D.; Kisielowski, C.; Yang, L.; Park, C.-H.; Crommie, M. F.; Cohen, M. L.; Louie, S. G.; Zettl, A. *Science* **2009**, *323*, 1705.

(8) Nakada, K.; Fujita, M.; Dresselhaus, G.; Dresselhaus, M. S. *Phys. Rev. B* **1996**, *54*, 17954.

(9) Son, Y. M.; Cohen, M. L.; Louie, S. G. *Nature* **2006**, *444*, 334.

(10) Enoki, T.; Kobayashi, Y.; Fukui, K. *Int. Rev. Phys. Chem.* **2007**, *26*, 609.

(11) Ding, F.; Larsson, P.; Larsson, J. A.; Ahuja, R.; Duan, H.; Rosén, A.; Bolton, K. *Nano Lett.* **2008**, *8*, 463.

(12) Felice, R. D.; Calzolari, A.; Varsano, D.; Rubio, A. In *Introducing Molecular Electronics*; Cuniberti, G.; Fargas, G.; Richter, K., Eds.; Springer-Verlag: Berlin, 2005; p 77.

(13) Kawai, T.; Miyamoto, Y.; Sugino, O.; Koga, Y. *Phys. Rev. B* **2000**, *62*, R16349.

(14) Gianozzi, P.; Car, R.; Scoles, G. *J. Chem. Phys.* **2003**, *118*, 1003.

(15) Koskinen, P.; Malola, S.; Häkkinen, H. *Phys. Rev. Lett.* **2008**, *101*, 115502.

(16) Okada, S. *Phys. Rev. B* **2008**, *77*, 041408.

(17) Wassmann, T.; Seitsonen, A. P.; Saitta, A. M.; Lazzeri, M.; Mauri, F. *Phys. Rev. Lett.* **2008**, *101*, 096402.

(18) Barone, V.; Hod, O.; Scuseria, G. E. *Nano Lett.* **2006**, *6*, 2748.

(19) Hod, O.; Barone, V.; Peralta, J. E.; Scuseria, G. E. *Nano Lett.* **2007**, *7*, 2295.

(20) Hod, O.; Barone, V.; Scuseria, G. E. *Phys. Rev. B* **2008**, *77*, 035411.

(21) May, K.; Dapprich, S.; Furche, F.; Unterreiner, B. V.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2000**, *2*, 5084.

(22) Jiang, D.; Sumpter, B. G.; Dai, S. *J. Chem. Phys.* **127**, *127*, 124703.

(23) Montoya, A.; Truong, T. N.; Sarofim, A. F. *J. Phys. Chem. A* **2000**, *104*, 6108.

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

(25) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143.

(26) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.

(27) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(28) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095.

(29) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.

(30) Schwabe, T.; Grimme, S. *Acc. Chem. Res.* **2008**, *41*, 569.

(31) Menon, A. S.; Radom, L. *J. Phys. Chem. A* **2008**, *112*, 13225.

(32) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.

(33) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(34) Pople, J. A.; Gill, P. M. W.; Handy, N. C. *Int. J. Quantum Chem.* **1995**, *56*, 303.

(35) Lee, Y. H.; Kim, S. G.; Tománek, D. *Phys. Rev. Lett.* **1997**, *78*, 2393.

(36) Kraka, E.; Cremer, D. *Chem. Phys. Lett.* **1993**, *216*, 333.

(37) In view of the dramatic UHF error on $\langle \mathbf{S}^2 \rangle$, MP2 calculations of the ZZ edge have been performed from restricted open-shell calculations.

CT900184D

JCTC Journal of Chemical Theory and Computation

Spatial Decomposition Analysis of the Thermodynamics of Cyclodextrin Complexation

Takeshi Yamazaki[†] and Andriy Kovalenko*[,†,‡]

National Institute for Nanotechnology, 11421 Saskatchewan Drive, Edmonton, Alberta, T6G 2M9, Canada, and Department of Mechanical Engineering, University of Alberta, Edmonton, Canada

Received February 9, 2009

Abstract: We propose a method of spatial decomposition analysis (SDA) to study the thermodynamics of association in solution, based on three-dimensional molecular theory of solvation. We decompose the solvation thermodynamics quantities into the excluded volume and solvation shell terms and further break them down into partial contributions of the functional groups of the associating species. For illustration, we applied the SDA method to the complexation of β-cyclodextrin and 1-adamantanecarboxylic acid in water. We calculated the changes in the free energy and in the partial molar volume upon the association and decomposed them into the partial contributions of the functional groups to the excluded volume and solvation shell terms. The SDA shows that the adamantyl group of 1-adamantanecarboxylic acid is responsible for the complexation more than its carboxyl group and that the carboxyl has little contribution to the association process. The SDA results are in good agreement with the observation made in a recent molecular dynamics simulation. The SDA method can reveal a microscopic picture for association processes in solution in a number of areas, including protein stability, and might be a useful tool for rational drug design.

1. Introduction

Association of molecules in solution is one of the most fundamental phenomena observed in a wide variety of fields of chemistry, biology, pharmacy, and material science.[1−4] This includes such processes of practical importance as drug−protein binding, micelle formation, and self-assembly of organic nanotubes. Understanding it from a microscopic viewpoint is of substantial value, both for basic science and for control and rational design of important technological processes. In many cases,[3,5,6] the binding affinity of a macromolecule can be subdivided into contributions from its essential fragments. A viable strategy of controlling association is thus to analyze and rationally design the contribution of each fragment, much as in fragment-based drug design.[7,8] The latter has become an important and powerful tool for discovery and optimization of new drug leads. The design and optimization of the ultimate lead compound is carried out by identifying and optimizing individual fragments, followed by synthetic linking or merging them to produce a high-affinity drug lead.

Association in solution is a challenging theoretical problem because the association free energy is typically determined by a subtle balance between the direct interaction potential and the solvent-mediated effective interaction of the associating macromolecules.[9] Association can be assisted not only by solvent thermodynamic driving forces such as hydrophobic attraction between hydrophobic fragments of the associating solutes but also by solvent bridge formation such as water molecules bridging hydrophilic fragments of the solutes with hydrogen bonding.[10] In order to analyze an association process, one needs to employ such theory that properly accounts for the interplay of all these molecular forces of solvation structure and yields their effect on solvation thermodynamics.

In the present article, we propose an approach resolving contributions of functional groups to association in solution

* Corresponding author e-mail: andriy.kovalenko@nrc-cnrc.gc.ca.
[†] National Institute for Nanotechnology.
[‡] University of Alberta.

10.1021/ct9000729 CCC: $40.75
Published on Web 06/01/2009

by using the method of statistical−mechanical, three-dimensional molecular theory of solvation, also known as the 3D reference interaction site model (3D-RISM).[11−14] Starting with atomistic interaction potentials between solution species (force field), 3D-RISM yields the solvation structure in the form of 3D correlation functions of interaction sites of solvent molecules around a solute (including the 3D density distribution function). It then analytically yields the solvation thermodynamics, including the solvation free energy, solvation entropy and enthalpy, and partial molar volume, which are expressed as integrals in terms of the 3D correlation functions. The integrands thus mean the spatial density of the corresponding thermodynamic quantities, resolved in three dimensions of the direct space. This 3D information on solvation structure and thermodynamics provided by 3D-RISM theory is invaluable for microscopic understanding of association in solution. In particular, in the present article we decompose the thermodynamic properties into two terms coming from the excluded volume and solvation shell of the solute macromolecule. We further subdivide them into partial contributions projected onto spatial fragments corresponding to the functional groups of the solute. This reveals how the excluded volume and solvation shell fragments of each functional group contribute to the association thermodynamics. Below we refer to this method as spatial decomposition analysis (SDA).

To validate our approach, we apply it to the complexation of β-cyclodextrin and 1-adamantanecarboxylic acid in water. Harries et al.[15] and Taulier et al.[16] have recently investigated this system by microcalorimetry and volumetric measurements, respectively. Cyclodextrin is a cyclic oligomer of seven units of α-D-glucose. The glucose units are connected through glycosidic α-1,4 bonds, and, as a result, the seven units constitute a doughnut-shaped molecule with a cavity. The presence of this cavity makes cyclodextrin an attractive nanostructured unit, a subject of many studies.[17−20] We obtain the change in the solvation free energy upon the cyclodextrin−adamantane association and analyze the free energy change upon the association by using SDA to reveal which fragment is the most responsible for the association process. Simultaneously, the partial molar volume change is calculated to see whether the theory and the molecular model reasonably describe the system under study. The results demonstrate that SDA can be a helpful tool in elucidating a microscopic picture of molecular association.

2. Methods

2.1. Molecular Model. The molecular geometry for the complex of 1-adamantanecarboxylic acid and β-cyclodextrin was taken from BOGCAB.pdb.[21] These coordinates include two complexes, and we refer to them as complexes A and B, as shown in Figure 1A. On adding the missing hydrogen atoms, as shown in Figure 1B and C, we used the resulting structures, without any further modification, to calculate the internal energy of the solute system and the solvation thermodynamics of the complexation.

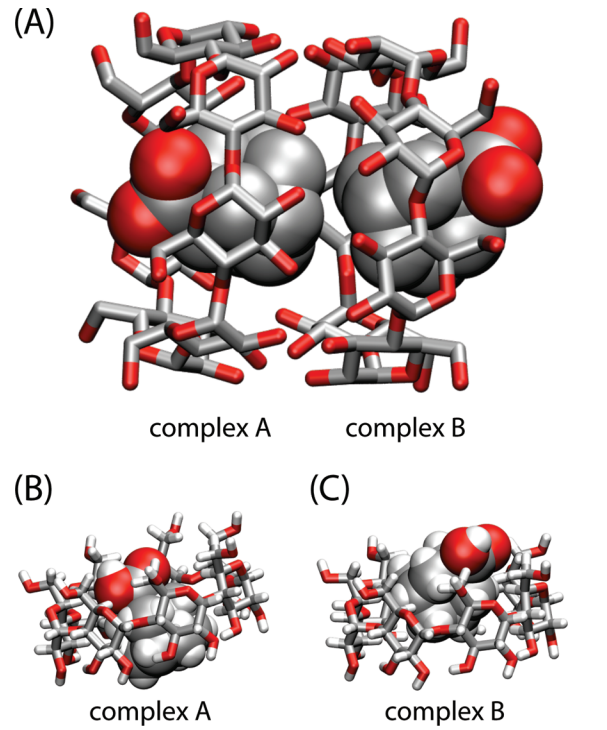


Figure 1. Arrangement of the complex of β-cyclodextrin and 1-adamantanecarboxylic acid. (A) shows the complexes A and B in the arrangement from BOGCAB.pdb.[21] (B) and (C) represent the complexes A and B, respectively, used in the present calculation after adding the hydrogen atoms.

2.2. Three-Dimensional Molecular Theory of Solvation. The 3D-RISM molecular theory of solvation is a powerful tool to study solvation thermodynamics of macromolecules in different environments. In this section, we briefly review the key aspects of the theory which are relevant to the discussion to follow. The 3D-RISM integral equation[11−14]

$$h_\gamma(\mathbf{r}) = \sum_{\gamma'} \int \mathrm{d}\mathbf{r}'\, c_{\gamma'}(\mathbf{r}')\chi_{\gamma'\gamma}(|\mathbf{r} - \mathbf{r}'|) \qquad (1)$$

is coupled with the 3D version of the HNC closure approximation[22]

$$g_\gamma(\mathbf{r}) = \exp\left(-\frac{u_\gamma(\mathbf{r})}{k_\mathrm{B}T} + h_\gamma(\mathbf{r}) - c_\gamma(\mathbf{r})\right) \qquad (2)$$

Here, $h_\gamma(\mathbf{r})$ is the 3D total correlation function related to the 3D distribution function $g_\gamma(\mathbf{r}) = h_\gamma(\mathbf{r}) + 1$, which gives the normalized probability of finding interaction site γ of solvent molecules at position $\mathbf{r}$ around the solute molecule, and $c_\gamma(\mathbf{r})$ is the 3D direct correlation function which has the asymptotics of the solute−solvent site interaction potential: $c_\gamma(\mathbf{r}) \sim -u_\gamma(\mathbf{r})/(k_\mathrm{B}T)$, where $k_\mathrm{B}T$ is the Boltzmann constant times the temperature of the solution. The susceptibility of pure solvent $\chi_{\gamma'\gamma}(r) = \omega_{\gamma'\gamma}(r) + \rho_{\gamma'}h_{\gamma'\gamma}(r)$ splits up into the intramolecular distribution function $\omega_{\gamma'\gamma}(r) = \delta(r - l_{\gamma'\gamma})/(4\pi l_{\gamma'\gamma}^2)$ specifying the geometry of solvent molecules with site separation $l_{\gamma'\gamma}$ and the intermolecular site−site total correlation function $h_{\gamma'\gamma}(r)$ times the solvent site number density $\rho_{\gamma'}$. The radial correlations $h_{\gamma'\gamma}(r)$ of pure solvent are

obtained, in advance of the 3D-RISM calculation, from the dielectrically consistent RISM integral equation theory (DRISM)[23] coupled with the HNC closure. The convolution in the 3D-RISM integral eq 1 is calculated by analytically treating the electrostatic asymptotics of all the correlation functions (both the 3D and radial ones) and applying the fast Fourier transform technique (3D-FFT and 1D-FFT) to the remaining shorter-range part of the correlations.[14,24]

The solvation free energy of the solute is calculated by using the 3D extension[13,14] of the Singer−Chandler formula[25]

$$\mu^{\mathrm{solv}} = k_{\mathrm{B}}T\sum_{\gamma}\rho_{\gamma}\int d\mathbf{r}\,\mathscr{T}(\mathbf{r}) \tag{3a}$$

$$\mathscr{T}(\mathbf{r}) = \frac{1}{2}h_{\gamma}^{2}(\mathbf{r}) - c_{\gamma}(\mathbf{r}) - \frac{1}{2}h_{\gamma}(\mathbf{r})c_{\gamma}(\mathbf{r}) \tag{3b}$$

Under the isochoric condition, the solvation free energy can be decomposed into the solvation energetic and entropic parts

$$\mu^{\mathrm{solv}} = \varepsilon^{\mathrm{solv}} - TS^{\mathrm{solv}} \tag{4}$$

In turn, the solvation energy $\varepsilon^{\mathrm{solv}}$ can be viewed as consisting of two contributions: one arising from creation of a polarized cavity (in pure solvent) and the other corresponding to the energy of embedding the solute molecule into the cavity.[26] By taking the derivative of μ^{solv} with respect to temperature, we can obtain the entropic component $-TS^{\mathrm{solv}}$.[26,27] We calculated the derivative from the equations obtained by analytical variation of the 3D-RISM and DRISM equations, following ref 27.

The partial molar volume $\bar{\mathscr{V}}$ is expressed in terms of the 3D correlation function $h(\mathbf{r})$ between solute and solvent as follows

$$\bar{\mathscr{V}} = k_{\mathrm{B}}T\chi_{T} - \int d\mathbf{r}\,h_{\gamma}(\mathbf{r}) = k_{B}T\chi_{T} + \bar{\mathscr{V}} \tag{5}$$

where χ_T is the isothermal compressibility of pure solvent which can be obtained from DRISM theory for the solvent−solvent correlation functions of pure solvent. In eq 5, $k_{\mathrm{B}}T\chi_T$ is the ideal term of partial molar volume and $\bar{\mathscr{V}}$ is the excess term coming from the solute−solvent interaction. The integration in eqs 3a and 5 is performed with analytically treating the long-range, electrostatic part of the integrands and numerically integrating the remaining short-range terms over the volume of the 3D-FFT supercell. The former appears due to ion−ion correlations in electrolyte solution and becomes significant only in the case of finite but relatively small ionic concentrations with the Debye screening length comparable to the supercell size.

2.3. Spatial Decomposition Analysis of Thermodynamic Properties. Several studies have been reported to investigate the thermodynamic properties in subspaces around a solute molecule.[28−33] For instance, Matubayashi et al.[34,35] have introduced the hydration shell model to analyze thermodynamic properties using the Monte Carlo simulation. We extend their concept to three-dimensional molecular theory of solvation (3D-RISM). First, we subdivide the 3D real space of integration in eqs 3a and 5 into two regions: the excluded volume (EcV) of the solute supramolecule and the space outside it we will refer to as the solvation shell region. We estimate the EcV using a water probe of conventional radius 1.4 Å. In order to obtain the association properties, we need to consider three systems: (i) isolated 1-adamantanecarboxylic acid (AD), (ii) isolated β-cyclodextrin (CD), and (iii) supramolecular complex AD−CD. The excluded volumes of the molecules AD and CD are hereafter referred to as EcV(AD) and EcV(CD). For the AD−CD complex in system iii, we decompose the integration domain into the sum of EcV(AD) and EcV(CD) and the solvation shell region of the complex. The same spatial subdivision of the integration domain is used for the systems (i) AD and (ii) CD as well, in order to trace changes in the local solvent environment in the regions EcV(AD) and EcV(CD) upon the association. Thus, the spaces of EcV(CD) in the system i and the EcV(AD) in the system ii are occupied by solvent water molecules which will be excluded by the association process. Figure 2 schematically illustrates the spatial decomposition. In the present study, we counted the region of overlap between the EcV(AD) and EcV(CD) as belonging to EcV(AD). Next, we perform the integration within each space to obtain its local thermodynamic properties. For example, the solvation free energy inside the EcV(AD) in the system i is calculated as (we drop the superscript "solv" from μ^{solv}, $\varepsilon^{\mathrm{solv}}$, and $-TS^{\mathrm{solv}}$ hereafter)

$$\mu^{\mathrm{sys(i)}}_{\in \mathrm{EcV(AD)}} = k_{\mathrm{B}}T\sum_{\gamma}\rho_{\gamma}\int_{\mathrm{V}\in \mathrm{EcV(AD)}} \mathrm{d}\mathbf{r}\,\mathscr{T}(\mathbf{r}) \tag{6}$$

In this way, the thermodynamic properties obtained by eqs 3a and 5 are broken down into several spaces.

It is worth mentioning about the physical meaning of the solvation free energy in the EcV. In the case of $\mu^{\mathrm{sys(i)}}_{\in \mathrm{EcV(CD)}}$ and $\mu^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(AD)}}$, the local space is occupied by water molecules, and therefore, it is obvious that $\mu_{\in \mathrm{EcV}}$ is the partial contribution of water molecules in that element of space to the total solvation free energy. Then how can we interpret $\mu^{\mathrm{sys(i)}}_{\in \mathrm{EcV(AD)}}$ and $\mu^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(CD)}}$ for the space of EcV holding the solute molecule inside? A general idea can be obtained by representing both solute and solvent molecules simply with hard sphere models and employing the Percus−Yevick closure[22] which is appropriate for hard sphere fluids. We will then get inside the hard sphere cores: $h(\mathbf{r}) = -1$ and $c(\mathbf{r}) = 0$ inside the EcV. Under this condition, the total solvation free energy is equivalent to the total solvation entropy $-TS$. Therefore, $\mu_{\in \mathrm{EcV(AD)}}$ is also equivalent to $-TS_{\in \mathrm{EcV(AD)}}$ which turns out to be equal to the thermal energy times the number of bulk water molecules that can be put inside the empty space of EcV with the factor of 1/2 (because of switching the interaction on). In the present study, we use a more realistic interaction potential (Lennard-Jones plus Coulomb potential between molecular interaction sites) and the HNC closure. Therefore, $\mu_{\in \mathrm{EcV}}$ is strongly perturbed by the direct interaction between solute and solvent molecules. In effect, $\mu^{\mathrm{sys(i)}}_{\in \mathrm{EcV(AD)}}$ calculated as 38 kcal/mol breaks down into $-TS^{\mathrm{sys(i)}}_{\in \mathrm{EcV(AD)}}$ and $\varepsilon^{\mathrm{sys(i)}}_{\in \mathrm{EcV(AD)}}$ as 43 and −5 kcal/mol, respectively (based on the computational condition described in the next section).

2.4. Computational Details. We considered the water solvent with physical density 0.997 g/cm^3 and dielectric constant 78.38 corresponding to the ambient conditions of

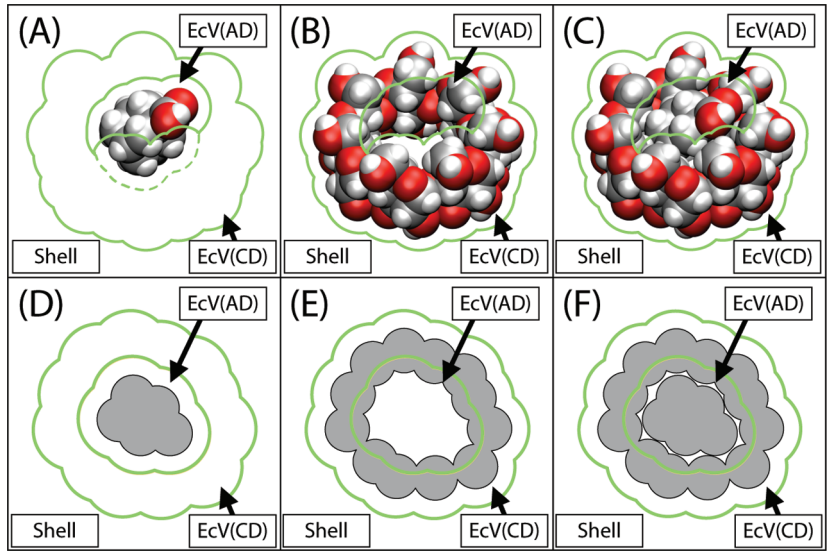


Figure 2. Schematic sketch of the spatial decomposition for systems i−iii (A−C), respectively. The schematic cross section of the space is also shown in (D−F) for systems i−iii, respectively. The entire space is partitioned into three subspaces, EcV(AD), EcV(CD), and Shell, in each system. The green lines indicate the boundary of the subspaces.

Table 1. Solvation Free Energies (in kcal/mol) for the Isolate AD in System i, $\mu^{sys(i)}$, the Isolated CD in System ii, $\mu^{sys(ii)}$, and the Complex in System iii, $\mu^{sys(iii)}$[a]

	$\mu^{sys(i)}$	$\mu^{sys(ii)}$	$\mu^{sys(iii)}$	$\Delta\mu$	E^{int}	ΔA
complex A	22.5	44.6	77.9	10.8	−22.2	−11.4
complex B	24.2	46.0	76.7	6.5	−19.3	−12.8

[a] The association free energy, ΔA, is calculated as the sum of the interaction energy E^{int} between AD and CD and the solvation free energy change upon the association, $\Delta\mu$. The latter is defined as $\mu^{sys(iii)} - \mu^{sys(i)} - \mu^{sys(ii)}$. Data are for complexes A and B (first and second rows, respectively).

Table 2. Partial Molar Volume $\mathcal{V}$ (in cm^3/mol) and Its Change upon the Complexation of AD and CD, versus Experiment[a]

	$\mathcal{V}^{sys(i)}$	$\mathcal{V}^{sys(ii)}$	$\mathcal{V}^{sys(iii)}$	$\Delta\mathcal{V}$
complex A	134.6	693.6	836.0	7.7
complex B	139.7	696.6	825.6	−10.7
exptl	140.5[b]	716 ± 2[c]		∼−5[d]

[a] The data are shown for the isolated AD in system i, $\mathcal{V}^{sys(i)}$, for the isolated CD in system ii, $\mathcal{V}^{sys(ii)}$, for the complex in system iii, $\mathcal{V}^{sys(iii)}$, and for its change, $\Delta\mathcal{V}$, upon the association. $\Delta\mathcal{V}$ is defined as $\mathcal{V}^{sys(iii)} - \mathcal{V}^{sys(i)} - \mathcal{V}^{sys(ii)}$. Theoretical data are for complexes A and B (first and second row, respectively), and experimental results (last row) from references are given in footnotes. [b] Reference 40. [c] Reference 41. [d] Reference 16.

temperature $T = 298.15$ K and pressure = 1 bar. The ideal term in eq 5 was calculated to be 1.5 cm^3/mol by solving the DRISM-HNC equations for pure water. The 3D-RISM-HNC equations were solved on a grid of 128^3 points in a cubic supercell of size 64 Å, large enough to accommodate the complex together with sufficient solvation space around it. We used the OPLS-AA force field[36] for AD and CD and the SPC/E model[37] for water. We have confirmed that the result does not change significantly with the use of of a finer grid of 256^3 points in the same supercell.

3. Results and Discussion

3.1. Free Energy and Partial Molar Volume Changes on Complexation. Table 1 shows the free energy change on the complexation of AD and CD in water for complexes A and B. The association free energy, ΔA, is obtained as the sum of the interaction energy E^{int} between AD and CD, and the solvation free energy change upon the association $\Delta\mu$. The association free energies were calculated to be −11 and −13 kcal/mol for the complexes A and B, respectively, which suggests that AD and CD tend to aggregate in water. Our theoretical prediction is in agreement with the calorimetric measurement showing that the association free energy between AD and CD is −7.7 kcal/mol.[38] A conformational difference between the two complexes is how CD holds AD inside the cavity. As can be seen from Figure 1B and C, the adamantyl group of AD in complex A runs off the edge of the cavity, while that in complex B is inside the cavity. Recent molecular dynamics simulation on a copolymer formed by β-cyclodextrin and adamantane dimers[39] has pointed out, based on the MD snapshots, that the adamantyl group penetrating inside the CD cavities stops well before their centroids pass the mean-square plane made by the glycosidic oxygens of CD. Our result that the association of complex B is more favorable than that of complex A agrees with that MD observation.

Table 2 gives the partial molar volume (PMV) change on the complexation along with the experimental data,[16,40,41] and the present theory reproduces the PMVs reasonably well. Interestingly, the theory predicts that two complexes both of which prefer to aggregate show the opposite trend in the volume change: the complex A has a positive PMV change (+8 cm^3/mol), whereas the complex B has a negative change (−11 cm^3/mol). The volumetric measurements by Taulier et al.[16] have revealed that the complexation of AD and CD has the negative PMV change (∼−5 cm^3/mol). Therefore, it is consistent that the 3D-RISM theory predicts that the association free energy of complex B is more favorable than that of complex A. We will use the complex B as a

Table 3. SDA of the Solvation Free Energy μ and Its Change upon the Association[a]

	$\mu_{\in \mathrm{EcV(AD)}}$	$\mu_{\in \mathrm{EcV(CD)}}$	$\mu_{\in \mathrm{Shell}}$	μ (total)
system i	37.928	−9.1112	−4.6353	24.18
system ii	21.984	68.286	−44.288	45.98
system iii	58.148	66.531	−48.022	76.66
Δ	−1.76	7.36	0.901	6.5

[a] The last column shows the total solvation free energy. Δ in the last row represents the association property defined as system iii − system i − system ii.

Table 4. SDA of the Solvation Entropy $-TS$ and Its Change upon the Association[a]

	$-TS_{\in \mathrm{EcV(AD)}}$	$-TS_{\in \mathrm{EcV(CD)}}$	$-TS_{\in \mathrm{Shell}}$	$-TS$ (total)
system i	43.489	−0.3060	−0.8023	42.38
system ii	35.672	188.306	−2.8900	221.09
system iii	68.340	187.640	−2.6974	253.28
Δ	−10.8	−0.360	0.995	−10.2

[a] The last column shows the total solvation entropy. Δ in the last row represents the association property defined as system iii − system i − system ii.

representative model of the association process between AD and CD for further analysis by using SDA in the following sections.

3.2. Spatial Decomposition Analysis of Solvation Energy and Free Energy Changes. In this subsection, we analyze the association solvation free energy $\Delta\mu$ and the association free energy ΔA. The SDA yields the following decompositions of μ:

$$\mu = \mu_{\in \mathrm{EcV(AD)}} + \mu_{\in \mathrm{EcV(CD)}} + \mu_{\in \mathrm{Shell}} \tag{7}$$

Each term in eq 7 for each of the systems i−iii has been compiled in Table 3. In system i, $\mu^{\mathrm{sys(i)}}_{\in \mathrm{EcV(CD)}}$ and $\mu^{\mathrm{sys(i)}}_{\in \mathrm{Shell}}$ contribute to the stabilization of AD in water, which is the hydration effect on AD. $\mu^{\mathrm{sys(i)}}_{\in \mathrm{EcV(CD)}}$ has a larger negative value of μ than that of $\mu^{\mathrm{sys(i)}}_{\in \mathrm{Shell}}$, because the water molecules in EcV(CD) are closer to AD than those in the Shell. On the other hand, $\mu^{\mathrm{sys(i)}}_{\in \mathrm{EcV(AD)}}$ contributes to the destabilization of AD in water. This is because EcV(AD) holds AD inside and water molecules are excluded from this volume, resulting in loss of the solvation entropy. In system ii, $\mu^{\mathrm{sys(ii)}}_{\in \mathrm{Shell}}$ contributes to the stabilization of CD in water, which is the hydration effect on CD. $\mu^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(CD)}}$ contributes to the destabilization of CD, because EcV(CD) holds CD inside, resulting in loss of the solvation entropy. It is interesting that $\mu^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(AD)}}$ has a positive value of μ, although EcV(AD) includes water molecules inside (the number of water molecules inside this volume is calculated to be about 12). This suggests that water molecules in EcV(AD) unfavorably solvate CD; in other words, EcV(AD) is hydrophobic. Table 4 presenting the decomposition of solvation entropy also confirms that by showing $-TS^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(AD)}}$ has a large positive value. EcV(AD) in system ii covers the cavity space of CD, and there has been a theoretical observation that the cavity is hydrophobic.[42] The SDA conclusion is quite consistent with this observation. In system iii, only $\mu^{\mathrm{sys(iii)}}_{\in \mathrm{Shell}}$ stabilizes the complex in water (hydration effect), and the rest of the terms destabilize the association since they include the cores of the solute molecules.

The association solvation free energy $\Delta\mu$ is decomposed by SDA as

$$\Delta\mu = \Delta\mu_{\in \mathrm{EcV(AD)}} + \Delta\mu_{\in \mathrm{EcV(CD)}} + \Delta\mu_{\in \mathrm{Shell}} \tag{8}$$

where

$$\Delta\mu_{\in \mathrm{EcV(AD)}} = \mu^{\mathrm{sys(iii)}}_{\in \mathrm{EcV(AD)}} - (\mu^{\mathrm{sys(i)}}_{\in \mathrm{EcV(AD)}} + \mu^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(AD)}}) \tag{9}$$

$$= (\mu^{\mathrm{sys(iii)}}_{\in \mathrm{EcV(AD)}} - \mu^{\mathrm{sys(i)}}_{\in \mathrm{EcV(AD)}}) + (-\mu^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(AD)}}) \tag{10}$$

$$\Delta\mu_{\in \mathrm{EcV(CD)}} = \mu^{\mathrm{sys(iii)}}_{\in \mathrm{EcV(CD)}} - (\mu^{\mathrm{sys(i)}}_{\in \mathrm{EcV(CD)}} + \mu^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(CD)}}) \tag{11}$$

$$= (\mu^{\mathrm{sys(iii)}}_{\in \mathrm{EcV(CD)}} - \mu^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(CD)}}) + (-\mu^{\mathrm{sys(i)}}_{\in \mathrm{EcV(CD)}}) \tag{12}$$

$$\Delta\mu_{\in \mathrm{Shell}} = \mu^{\mathrm{sys(iii)}}_{\in \mathrm{Shell}} - (\mu^{\mathrm{sys(i)}}_{\in \mathrm{Shell}} + \mu^{\mathrm{sys(ii)}}_{\in \mathrm{Shell}}) \tag{13}$$

and has been compiled in the last row in Table 3. By rearranging eq 9 to eq 10, and eq 11 to eq 12, we can decompose $\Delta\mu_{\in \mathrm{EcV}}$ into two parts: the change in the volume space which holds solute molecules inside (first term in rhs of eqs 10 and 12), and the change in the hydration (second term in rhs of eqs 10 and 12).

On the association, AD occupies the cavity space of CD and repels water molecules from the cavity. This can be seen as $-\mu^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(AD)}}$ in eq 10. $-\mu^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(AD)}} = -22$ kcal/mol favors the association, because water molecules inside the hydrophobic cavity of CD in system ii favor to be excluded from the cavity. Simultaneously, CD dehydrates AD, which can be seen as $-\mu^{\mathrm{sys(i)}}_{\in \mathrm{EcV(CD)}}$ in eq 12. In contrast to $-\mu^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(AD)}}$, the term $-\mu^{\mathrm{sys(i)}}_{\in \mathrm{EcV(CD)}} = +9$ kcal/mol does not favor the association because water molecules in EcV(CD) in system i interact preferably with AD.

It is a little complicated to interpret the change in the volume space which holds solute molecules inside (first term in rhs of eqs 10 and 12). First, let us discuss the value of $\mu^{\mathrm{sys(iii)}}_{\in \mathrm{EcV(CD)}} - \mu^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(CD)}}$ to see what happens with EcV(CD) when AD occupies the CD cavity. Because AD comes very close to EcV(CD), the interaction potential felt by water molecules in EcV(CD) changes, which changes the excluded volume for water molecules (related to the solvation entropy change) and changes the interaction between CD and water molecules (related to solvation energy change). As a result, $\mu^{\mathrm{sys(iii)}}_{\in \mathrm{EcV(CD)}} - \mu^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(CD)}}$ turns out to be −2 kcal/mol, and its thermodynamic decomposition into the energetic and entropic parts of the change gives $\varepsilon^{\mathrm{sys(iii)}}_{\in \mathrm{EcV(CD)}} - \varepsilon^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(CD)}} = -1$ kcal/mol and $-TS^{\mathrm{sys(iii)}}_{\in \mathrm{EcV(CD)}} - (-TS^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(CD)}}) = -1$ kcal/mol. On the other hand, what happens with EcV(AD) may be simpler than with EcV(CD). The SDA gives us that $\mu^{\mathrm{sys(iii)}}_{\in \mathrm{EcV(AD)}} - \mu^{\mathrm{sys(i)}}_{\in \mathrm{EcV(AD)}} = +20$ kcal/mol, and the thermodynamic decomposition shows that the energetic and entropic parts are $\varepsilon^{\mathrm{sys(iii)}}_{\in \mathrm{EcV(AD)}} - \varepsilon^{\mathrm{sys(i)}}_{\in \mathrm{EcV(AD)}} = -5$ kcal/mol and $-TS^{\mathrm{sys(iii)}}_{\in \mathrm{EcV(AD)}} - (-TS^{\mathrm{sys(i)}}_{\in \mathrm{EcV(AD)}}) = +25$ kcal/mol. This suggests that EcV(AD) in system iii excludes water molecules more than that in system i, and that EcV(AD) in system iii has a stronger repulsive interaction for water molecules. This is related to the fact that EcV(AD) in system iii includes a part of CD and that EcV(AD) in system i does not include it (see Figure 2A, C, D, and F).

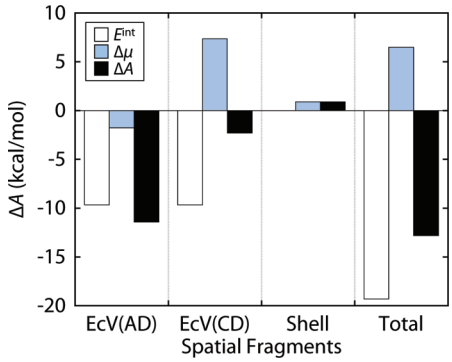


Figure 3. Spatial decomposition of the association free energy ΔA (in kcal/mol). The interaction energy between AD and CD, E^{int}, is divided by 2 and added to $\Delta\mu_{\in EcV(AD)}$ and $\Delta\mu_{\in EcV(CD)}$, respectively. $1/2E^{int}$, $\Delta\mu_{\in EcV(AD)}$, and $\Delta A_{\in EcV(AD)} = 1/2E^{int} + \Delta\mu_{\in EcV(AD)}$ are given in the first column from left, marked as EcV(AD). $1/2E^{int}$, $\Delta\mu_{\in EcV(CD)}$, and $\Delta A_{\in EcV(CD)} = 1/2E^{int} + \Delta\mu_{\in EcV(CD)}$ are given in the second column from left, marked as EcV(CD). $\Delta A_{\in Shell}$ (the third column from left) is defined as being equivalent to $\Delta\mu_{\in Shell}$. The last column marked as total is the sum of these three spatial fragments.

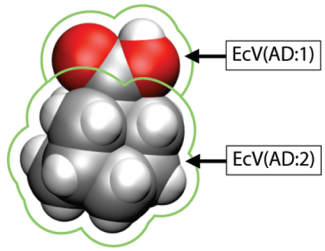


Figure 4. Schematic representation of the spatial decomposition of EcV(AD) into the carboxyl (EcV(AD:1)) and adamantyl (EcV(AD:2)) groups.

In this way, $\Delta\mu_{\in EcV}$ is determined by the balance between the change in the volume space holding solute molecules inside and the change in the hydration. We found that the latter is slightly dominant in the present association process. Therefore, we can assign $\Delta\mu_{\in EcV(AD)}$ and $\Delta\mu_{\in EcV(CD)}$ as the changes in the hydration upon the association. By adding the interaction energy between AD and CD, we finally obtain the SDA of association free energy as shown in Figure 3. In order to calculate $\Delta A_{\in EcV(AD)}$ and $\Delta A_{\in EcV(CD)}$, we divided E^{int} by 2 (according to switching the interaction on) and added them to $\Delta\mu_{\in EcV(AD)}$ and $\Delta\mu_{\in EcV(CD)}$, respectively. We can see from the figure that the dehydration from the cavity, $\Delta\mu_{\in EcV(AD)}$, and the dehydration around AD, $\Delta\mu_{\in EcV(CD)}$, cancel each other. As a result of the balance, $\Delta\mu$ turns out to contribute to destabilization of association. On the other hand, the interaction energy strongly favors the aggregation, and in total, the association in water takes place. Our conclusion is consistent with the experimental observations that the complex is predominantly stabilized by strong host−guest van der Waals interactions.[16,43−45]

By further decomposing EcV(AD) into the carboxyl and adamantyl groups (referred to as EcV(AD:1) and EcV(AD:2), respectively) as shown in Figure 4, we can estimate the contributions of these functional groups to ΔA. The result is compiled in Table 5, along with its thermodynamic decomposition. It is interesting that the SDA shows that $\mu^{sys(i)}_{\in EcV(AD:1)}$ has a negative value, although $\mu^{sys(i)}_{\in EcV(AD)}$ given by the sum of $\mu^{sys(i)}_{\in EcV(AD:1)}$ and $\mu^{sys(i)}_{\in EcV(AD:2)}$ has a positive value, as we have seen above. Its energetic component, $\varepsilon^{sys(i)}_{\in EcV(AD:1)}$, has a large negative value because the carbonyl group is a polar group that favorably interacts with water molecules. Its magnitude is larger than that of $-TS^{sys(i)}_{\in EcV(AD:1)}$, and $\mu^{sys(i)}_{\in EcV(AD:1)}$ turns out to have a favorable contribution to the stabilization of AD in water. The rest of the $\mu_{\in EcV(AD)}$ values in Table 5 are all positive, which can be explained by the EcV holding a solute molecule inside and by the hydrophobicity of the cavity, as discussed above. The SDA predicts the solvation terms as $\Delta\mu_{\in EcV(AD:1)} = +1.1$ kcal/mol and $\Delta\mu_{\in EcV(AD:2)} = -2.9$ kcal/mol, whereas the interaction energies between AD:1 and CD and between AD:2 and CD are calculated as −5.4 and −13.9 kcal/mol, respectively. By adding the interaction energies divided by 2 to the $\Delta\mu$s, we obtain that $\Delta A_{\in EcV(AD:1)} = -1.6$ kcal/mol and $\Delta A_{\in EcV(AD:2)} = -12.6$ kcal/mol, showing that the adamantyl group is largely responsible for the complexation and that the carboxyl group little influences it. From the experimental data, it also has been suggested[39] that the polar residue bonded to the adamantyl group scarcely influences the host−guest interaction. Our analysis is in good agreement with this observation. The present results thus show that SDA is a useful tool for fragment-based analysis to elucidate a molecular picture of association processes in solution.

Table 5. SDA of the Solvation Free Energy $\mu_{\in EcV(AD)}$ and Its Thermodynamic Decomposition into the Solvation Energy E and the Solvation Entropy $-TS$, along with Their Changes upon the Association[a]

	$\mu_{\in EcV(AD:1)}$			$\mu_{\in EcV(AD:2)}$		
	ε	$-TS$	total	ε	$-TS$	total
system i	−12.032	9.9536	−2.079	6.4726	33.535	40.01
system ii	−2.9894	5.0261	2.037	−10.699	30.646	19.95
system iii	−12.263	13.338	1.075	2.0715	55.002	57.07
Δ	2.76	−1.64	1.1	6.30	−9.18	−2.9

[a] Δ in the last row represents the association property defined as system iii − system i − system ii.

3.3. Spatial Decomposition Analysis of Partial Molar Volume Change. SDA is applicable not only to the free energy but also to any thermodynamic property which can be obtained from the distribution functions between solute and solvent. For illustration, we apply the SDA to the excess term of PMV and its change upon the association. In the present case, the ideal term of PMV is relatively small, compared to the whole PMV and its change. Therefore we can consider that the excess term is the primary contributor to the association process. The SDA of PMV is in fact very similar to the SDA of μ, and below we just briefly go through the results to give the relevant interpretations.

The SDA of the excess term in eq 5 yields the following decomposition:

$$\bar{\mathcal{V}} = \bar{\mathcal{V}}_{\in EcV(AD)} + \bar{\mathcal{V}}_{\in EcV(CD)} + \bar{\mathcal{V}}_{\in Shell} \qquad (14)$$

The values of all the components are compiled in Table 6. As we have seen in Table 3, the values $\bar{\mathcal{V}}^{sys(i)}_{\in EcV(CD)}$, $\bar{\mathcal{V}}^{sys(i)}_{\in Shell}$, $\bar{\mathcal{V}}^{sys(ii)}_{\in Shell}$, and $\bar{\mathcal{V}}^{sys(iii)}_{\in Shell}$ represent the hydration effect on the

Table 6. Spatial Decomposition Analysys (SDA) of the Excess Term of the Partial Molar Volume $\bar{\mathcal{V}}$ and Its Change upon the Association[a]

	$\bar{\mathcal{V}}_{\in \mathrm{EcV(AD)}}$	$\bar{\mathcal{V}}_{\in \mathrm{EcV(CD)}}$	$\bar{\mathcal{V}}_{\in \mathrm{Shell}}$	$\bar{\mathcal{V}}$ (total)
system i	209.43	−53.807	−17.367	138.3
system ii	126.87	719.69	−151.44	695.1
system iii	274.79	716.34	−166.97	824.2
Δ	−61.5	50.5	1.84	−9.2

[a] Δ in the last row represents the association property defined as system iii − system i − system ii.

PMV and contribute to the decrease of PMV, which is the so-called electrostriction effect. The term $\bar{\mathcal{V}}^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(AD)}}$ is also the hydration effect on the PMV; however, this term contributes to the increase of PMV. The positive value of PMV is caused by the fact that the distribution of water molecules in EcV(AD) is less than that in the bulk ($g(\mathbf{r}) < 1$), as can be seen from eq 5. This is the manifestation of hydrophobicity inside the CD cavity, which is consistent with the discussion about $\mu^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(AD)}}$ in the previous section. The terms $\bar{\mathcal{V}}^{\mathrm{sys(i)}}_{\in \mathrm{EcV(AD)}}$, $\bar{\mathcal{V}}^{\mathrm{sys(ii)}}_{\in \mathrm{EcV(CD)}}$, $\bar{\mathcal{V}}^{\mathrm{sys(iii)}}_{\in \mathrm{EcV(AD)}}$, and $\bar{\mathcal{V}}^{\mathrm{sys(iii)}}_{\in \mathrm{EcV(CD)}}$ are essentially the geometrical volume of EcV of the molecules, being perturbed by the interaction between AD and CD. From the SDA of the PMV change upon the association, we can see a large negative change of −62 cm^3/mol in EcV(AD) and a large positive change of +51 cm^3/mol in EcV(CD). The former is due to the fact that AD occupies the CD cavity and decreases the cavity volume, accompanied by desolvation of water molecules from the cavity. The latter is due to the desolvation of water molecules around AD. Much as for μ, we found that the PMV change is determined by the balance between the two dehydration events. The magnitude of $\Delta\bar{\mathcal{V}}_{\in \mathrm{EcV(AD)}}$ is larger than that of $\Delta\bar{\mathcal{V}}_{\in \mathrm{EcV(CD)}}$, which suggests that the dehydration from the CD cavity is the predominant factor for the PMV change, and as a result, the PMV change becomes negative, as has been observed in experiment.[16]

4. Concluding Remarks

In the present article, we have presented a method to analyze the association process by decomposing the thermodynamic property into several three-dimensional spaces, based on three-dimensional molecular theory of solvation, also known as 3D-RISM. We refer to this method as spatial decomposition analysis (SDA). In the present SDA, we divided the thermodynamic property into two spaces, the space inside the excluded volume of solute molecules and the outside space. The thermodynamic property projected onto the excluded volume is further divided into a few fragments and is discussed to see how each fragment contributes to the thermodynamic property change upon the association. To demonstrate the SDA, we applied the method to the complexation of β-cyclodextrin (CD) and 1-adamantanecarboxylic acid (AD) in water. By applying the SDA to the association free energy, we found that the complexation is determined by the balance between the interaction between the two molecules and the dehydration contributions from the CD cavity and around AD upon the association. We also found that the adamantyl group of the 1-adamantanecarboxylic acid is largely responsible for the association, whereas the carboxyl group makes a small contribution to the association. In addition, by applying the SDA to the change of the partial molar volume upon the association, we found that the sign of the change is determined by the same balance between the dehydration terms. Our analysis is in good agreement with the observations in the recent study on a copolymer formed by β-cyclodextrin and adamantane dimers by using molecular dynamics simulation.[39] This suggests that the SDA method can be used to elucidate a microscopic picture of a wide range of host−guest association processes in solution.

There are several ways of partitioning the solvation thermodynamics into spatially resolved contributions in the SDA method. For example, the solvation shell region can be further decomposed into the volume parts corresponding to fragments of the solute molecules. This way of partitioning the solvation shell is necessary for associating systems with charged molecules. Although the contribution from the shell region is relatively small in the present case, our preliminary calculations show the possibility of a large contribution to the association process in such systems. SDA would also be particularly useful to study the thermodynamic stability of proteins so as to see which fragment (or residue) of the protein contributes most to its stability. This might render SDA as a useful tool for rational drug design.

Acknowledgment. We are grateful to the National Research Council (NRC) Canada for supporting this research. The computations were supported by the Centre of Excellence in Integrated Nanotools (CEIN) at the University of Alberta. Figure 1,2A−C, and 4 were produced with the visualization program VMD.[46]

Note Added after ASAP Publication. Figures 3 and 4 were presented incorrectly in the version of this paper published ASAP June 1, 2009; the corrected version published ASAP June 3, 2009.

References

(1) Lehn, J.-M. *Science* **1985**, *227*, 849–856.

(2) Lehn, J.-M. *Angew. Chem., Int. Ed. Engl.* **1990**, *29*, 1304–1319.

(3) Schneider, H.-J. *Angew. Chem., Int. Ed. Engl.* **1991**, *30*, 1417–1436.

(4) Houk, K. N.; Leach, A. G.; Kim, S. P.; Zhang, X. *Angew. Chem., Int. Ed.* **2003**, *42*, 4872–4897.

(5) Jencks, W. P. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 4046–4050.

(6) Schneider, H.-J. *Chem. Soc. Rev.* **1994**, *23*, 227–234.

(7) Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. *Nat. Rev. Drug Discov.* **2004**, *3*, 660–672.

(8) Erlanson, D. A.; McDowell, R. S.; O'Brien, T. *J. Med. Chem.* **2004**, *47*, 3463–3482.

(9) Yamazaki, T.; Blinov, N.; Wishart, D.; Kovalenko, A. *Biophys. J.* **2008**, *95*, 4540–4548.

(10) Yamazaki, T.; Fenniri, H.; Kovalenko, A. Submitted for publication.

(11) Beglov, D.; Roux, B. *J. Chem. Phys.* **1995**, *103*, 360–364.

(12) Kovalenko, A.; Hirata, F. *Chem. Phys. Lett.* **1998**, *290*, 237–244.

(13) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **1999**, *110*, 10095–10112.

(14) Kovalenko, A. Three-dimensional RISM theory for molecular liquids and solid-liquid interfaces. In *Molecular theory of solvation*; Hirata, F., Ed.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2003; Vol. 24, pp 169−275.

(15) Harries, D.; Rau, D. C.; Parsegian, V. A. *J. Am. Chem. Soc.* **2005**, *127*, 2184–2190.

(16) Taulier, N.; Chalikian, T. V. *J. Phys. Chem. B* **2006**, *110*, 12222–12224.

(17) Saenger, W. *Angew. Chem., Int. Ed. Engl.* **1980**, *19*, 344–362.

(18) Connors, K. A. *Chem. Rev.* **1997**, *97*, 1325–1357.

(19) Rekharsky, M. V.; Inoue, Y. *Chem. Rev.* **1998**, *98*, 1875–1918.

(20) Davis, M. E.; Brewster, M. E. *Nat. Rev. Drug Discov.* **2004**, *3*, 1023–1035.

(21) Hamilton, J. A.; Sabesan, M. N. *Acta Crystallogr.* **1982**, *B38*, 3063–3069.

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

(23) Perkyns, J. S.; Pettitt, B. M. *J. Chem. Phys.* **1992**, *97*, 7656–7666.

(24) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2000**, *112*, 10391–10402.

(25) Singer, S. J.; Chandler, D. *Mol. Phys.* **1985**, *55*, 621–625.

(26) Pettitt, B. M.; Rossky, P. J. *J. Chem. Phys.* **1982**, *77*, 1451–1457.

(27) Yu, H.-A.; Roux, B.; Karplus, M. *J. Chem. Phys.* **1990**, *92*, 5020–5033.

(28) Mehrotra, P. K.; Beveridge, D. L. *J. Am. Chem. Soc.* **1980**, *102*, 4287–4294.

(29) Mezei, M.; Beveridge, D. L. *Methods Enzymol.* **1986**, *127*, 21–47.

(30) Levitt, M.; Sharon, R. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 7557–7561.

(31) Lounnas, V.; Pettitt, B. M. *Proteins: Struct. Funct. Bioinf.* **1994**, *18*, 133–147.

(32) Ashbaugh, H. S.; Paulaitis, M. E. *J. Phys. Chem.* **1996**, *100*, 1900–1913.

(33) Ashbaugh, H. S.; Paulaitis, M. E. *J. Am. Chem. Soc.* **2001**, *123*, 10721–10728.

(34) Matubayasi, N.; Reed, L. H.; Levy, R. M. *J. Phys. Chem.* **1994**, *98*, 10640–10649.

(35) Matubayasi, N.; Levy, R. M. *J. Phys. Chem.* **1996**, *100*, 2681–2688.

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

(37) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(38) Rüdiger, V.; Eliseev, A.; Simova, S.; Schneider, H.-J.; Blandamer, M. J.; Cullis, P. M.; Meyer, A. J. *J. Chem. Soc., Perkin Trans. 2* **1996**, 2119–2123.

(39) Leggio, C.; Anselmi, M.; Di Nola, A.; Galantini, L.; Jover, A.; Meijide, F.; Pavel, N. V.; Tellini, V. H. S.; Tato, J. V. *Macromolecules* **2007**, *40*, 5899–5906.

(40) Gianni, P.; Lepori, L. J. *Sol. Chem.* **1996**, *25*, 1–42.

(41) Spildo, K.; Høiland, H. J. *Sol. Chem.* **2002**, *31*, 149–164.

(42) Linert, W.; Margl, P.; Renz, F. *Chem. Phys.* **1992**, *161*, 327–338.

(43) Harrison, J. C.; Eftink, M. R. *Biopolymers* **1982**, *21*, 1153–1166.

(44) Cromwell, W. C.; Bystrom, K.; Eftink, M. R. *J. Phys. Chem.* **1985**, *89*, 326–332.

(45) Taulier, N.; Chalikian, T. V. *J. Phys. Chem. B* **2008**, *112*, 9546–9549.

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

CT9000729

JCTC Journal of Chemical Theory and Computation

Fully Numerical All-Electron Solutions of the Optimized Effective Potential Equation for Diatomic Molecules

Adi Makmal,[†] Stephan Kümmel,[‡] and Leeor Kronik*,[†]

Department of Materials and Interfaces, Weizmann Institute of Science, Rehovoth 76100, Israel, and Physikalisches Institut, Universität Bayreuth, D-95440 Bayreuth, Germany

Received November 10, 2008

Abstract: We present an approach for fully numerical, all-electron solutions of the optimized effective potential equation within Kohn−Sham density functional theory for diatomic molecules. The approach is based on a real-space, prolate-spheroidal coordinate grid for solving the all-electron Kohn−Sham equations and an iterative scheme for solving the optimized effective potential equation. The accuracy of this method is demonstrated by comparison with previously reported calculations. New fully numerical benchmark results for selected diatomic molecules are provided.

1. Introduction

The Kohn−Sham formulation of density functional theory (DFT)[1−3] is a widely used approach for calculating the electronic structure of materials from first principles. In Kohn−Sham DFT, the original interacting-electron Schrödinger equation is mapped into an equivalent noninteracting problem. This leads to effective one-particle equations which, in the spin-polarized form[4] are

$$\left(-\frac{\nabla^2}{2} + V_{\text{ion}}(\mathbf{r}) + V_{\text{H}}(\mathbf{r}) + V_{\text{xc},\sigma}(\mathbf{r})\right)\varphi_{i\sigma}(\mathbf{r}) = \varepsilon_{i\sigma}\varphi_{i\sigma}(\mathbf{r}) \tag{1}$$

where σ is a spin index, $\varphi_{i\sigma}$ and $\varepsilon_{i\sigma}$ are the ith Kohn−Sham orbital and energy, respectively, V_{ion} is the ion−electron attraction potential, V_{H} is the Hartree potential, and $V_{\text{xc},\sigma}$ is the exchange-correlation potential that represents all nonclassical electron interactions (hartree atomic units are used throughout unless otherwise stated). The exchange-correlation potential is the functional derivative of the exchange-correlation energy (which is a functional of the charge density) with respect to the (spin-polarized) charge density, $\rho_\sigma(\mathbf{r})$

$$V_{\text{xc},\sigma}(\mathbf{r}) \equiv \frac{\delta}{\delta\rho_\sigma(\mathbf{r})}E_{\text{xc}}[\rho_\uparrow(\mathbf{r}), \rho_\downarrow(\mathbf{r})] \tag{2}$$

Although DFT is exact in principle, the exact form of $E_{\text{xc}}[\rho_\uparrow(\mathbf{r}),\rho_\downarrow(\mathbf{r})]$ is unknown, and in practice approximate forms must be used.

Orbital-dependent functionals are exchange-correlation functionals that use Kohn−Sham orbitals, themselves being functionals of the density, as ingredients in functional construction. Such functionals are currently considered to be one of the most promising avenues in modern DFT, as they hold the promise of overcoming some of the more serious deficiencies of exchange-correlation functionals that are explicit functionals of the density.[5] A major difficulty, however, with implicit density functionals for the exchange-correlation energy is that a direct derivative for determining the exchange-correlation potential is not available. Instead, chain-rule arguments lead to an integro−differential equation, generally known as the optimized effective potential (OEP) equation.[5−12] There are several equivalent formulations for the OEP equation; here we use the (relatively) simple form[11,13,14]

$$S_\sigma(\mathbf{r}) \equiv \sum_{i=1}^{N_\sigma} \psi^*_{i\sigma}(\mathbf{r})\varphi_{i\sigma}(\mathbf{r}) + c.c. = 0 \tag{3}$$

where N_σ is the number of occupied states in the σ spin channel. Here $\psi^*_{i\sigma}(\mathbf{r})$ are called "orbital shifts" and are given by

* Corresponding author phone: +972-8-934-4993; e-mail: leeor.kronik@weizmann.ac.il.

[†] Weizmann Institute of Science.

[‡] Universität Bayreuth.

10.1021/ct800485v CCC: $40.75
Published on Web 06/11/2009

$$\psi^*_{i\sigma}(\mathbf{r}) = -\sum_{j\neq i}^{\infty} \frac{\int \varphi^*_{i\sigma}(\mathbf{r}')[u^{\mathrm{xc}}_{i\sigma}(\mathbf{r}') - V^{\mathrm{OEP}}_{\mathrm{xc},\sigma}(\mathbf{r}')]\varphi_{j\sigma}(\mathbf{r}')\,\mathrm{d}^3r'}{\varepsilon_{i\sigma} - \varepsilon_{j\sigma}}\varphi^*_{j\sigma}(\mathbf{r}), \quad \varepsilon_{i\sigma} \neq \varepsilon_{j\sigma} \tag{4}$$

where

$$u^{\mathrm{xc}}_{i\sigma}(\mathbf{r}) = \frac{1}{\varphi^*_{i\sigma}(\mathbf{r})}\frac{\delta E_{\mathrm{xc}}[\{\varphi\}]}{\delta\varphi_{i\sigma}(\mathbf{r})} \tag{5}$$

Because of the complexity of the OEP equation, many approximate schemes for determining $V^{\mathrm{OEP}}_{\mathrm{xc},\sigma}(\mathbf{r})$ have been suggested.[5] Some of them, e.g., the Krieger, Li, and Iafrate (KLI) approximation,[15] or the common energy denominator approximation (CEDA)[16] (which is equivalent to the localized Hartree−Fock (LHF) approach),[17] often provide an excellent approximation to the correct solution of the OEP equation. Nevertheless, it has been shown that, for various properties, use of potentials other than the full OEP may lead to results that are significantly different quantitatively[18−21] or even qualitatively[22,23] from the full OEP solution. This clearly establishes the need for accurate solutions of the OEP equation.

Typically, the Kohn−Sham equations are solved either by employing pseudopotentials (usually in conjunction with a plane-wave basis or a real-space grid), or by using atomic basis sets. Unfortunately, the use of either approach together with the OEP equation raises new difficulties and controversies:[5] OEP-compatible pseudopotentials require special care in their construction if spurious "potential tails" are to be avoided,[24−27] and even then questions as to the importance of core−valence interaction may be raised.[28] Use of localized basis sets may result in numerical inaccuracies[29] and an ill-defined algebraic problem, an issue whose resolution has recently attracted much discussion.[5,30−36]

In light of the above difficulties, it is highly desirable to obtain benchmark OEP calculations, i.e., "fully numerical" ones, where the only approximation beyond unavoidable roundoff errors is the choice of the exchange-correlation functional. Such calculations could be used for development and testing of new orbital-dependent functionals, as well as for objective testing of various approximate OEP solution schemes. To the best of our knowledge, prior to the present work this was achieved only for single atoms,[7,11,14,37,38] an unsatisfactory state of affairs because chemical bonds cannot be examined. Here, we present a real-space, prolate-spheroidal coordinate[39] based approach for fully numerical, all-electron solutions of the OEP equation for diatomic molecules. In the following, we explain the prolate-spheroidal real-space grid and the main principles of the numerical approach. We then demonstrate the accuracy of the proposed scheme via both OEP and non-OEP calculations for several chemical systems and functionals.

2. Numerical Approach

Following Becke,[39] Laaksonen and co-workers,[40,41] Grabo et al.,[10] and Engel et al.[42,43] we use a real-space grid based on prolate-spheroidal coordinates. These coordinates are useful for describing a system with two atomic centers

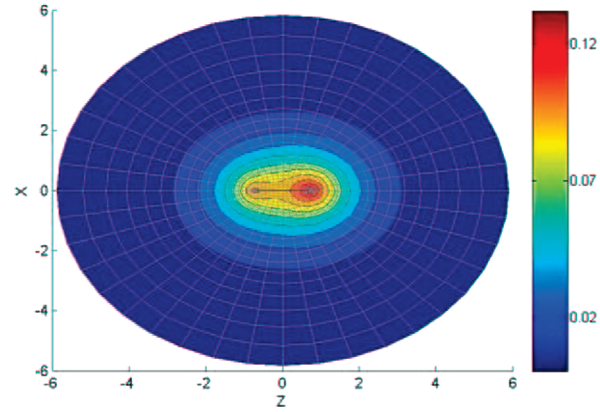


Figure 1. Contour plot of a hypothetical charge density of a heteronuclear dimer, sampled on a prolate-spheroidal coordinate grid. Here, the distance between the two centers is $R = 1.5$ and grid parameters are $N_\mu = 22$, $N_\nu = 20$, and $\mu_{\max} = 2.75$. For clarity, both $\phi = 0$ and $\phi = \pi$ planes are shown, corresponding to positive and negative x values, respectively. During calculation only the plane $\phi = 0$ is explicitly considered.

because the grid is very dense near the two centers, but increasingly coarse with increasing distance from the centers. This property of the grid is most helpful for all-electron calculations because it allows better sampling near the atoms where the ionic potential is singular and the orbitals oscillate rapidly.

For two centers at $A(x = 0, y = 0, z = -R/2)$ and $B(x = 0, y = 0, z = R/2)$, the prolate-spheroidal coordinates $(\mu, \nu, \phi) \in [0, \infty] \times [0, \pi] \times [0, 2\pi]$ are defined by

$$\begin{aligned} x &= \frac{R}{2}\sinh(\mu)\sin(\nu)\cos(\phi), \\ y &= \frac{R}{2}\sinh(\mu)\sin(\nu)\sin(\phi), \\ z &= \frac{R}{2}\cosh(\mu)\cos(\nu) \end{aligned} \tag{6}$$

The geometrical meaning of eq 6 is apparent from the inverse transformation

$$\mu = \cosh^{-1}\left(\frac{r_A + r_B}{R}\right), \quad \nu = \cos^{-1}\left(\frac{r_A - r_B}{R}\right), \quad \phi = \tan^{-1}\left(\frac{y}{x}\right) \tag{7}$$

where $r_A(\mathbf{r})$ and $r_B(\mathbf{r})$ are the Euclidean distances of a general point (x, y, z) from the centers A and B, respectively, and ϕ is the angle of rotation around the interatomic axis, i.e., the z-axis. At any constant-ϕ plane, constant-μ and constant-ν lines correspond to (half) ellipses and hyperbolas, respectively, as shown in Figure 1 for both $\phi = 0$ and $\phi = \pi$.

Due to the cylindrical symmetry of diatomic molecules, the angle ϕ can be treated analytically, and the problem is effectively reduced to a two-dimensional one. Formally, this means that all physical entities (e.g., charge density, potentials, squared absolute wave functions, etc.) are ϕ-independent and that the one-particle wave functions are of the form

$$\varphi(\mu, \nu, \phi) = \varphi(\mu, \nu, 0)\mathrm{e}^{\mathrm{i}m\phi} \qquad m = 0, \ \pm 1, \ \pm 2, \ldots \tag{8}$$

where m is an integer corresponding to the quantum number of angular momentum with respect to the interatomic axis.

Our numerical approach is based on self-consistent solutions of the Kohn−Sham and OEP equations on the uniform, two-dimensional (μ, ν) grid, using the high-order finite difference approach,[44,45] in which several neighbors around each point are used for approximating derivatives. In the context of pseudopotential-based calculations, the high-order finite difference approach has evolved into a powerful software suite, known as PARSEC (pseudopotential algorithm for real-space electronic structure calculations),[46−48] which has found many successful applications to large-scale electronic structure studies in general[48] and to OEP solutions in particular.[20,22,23] This offers a natural starting point for the present numerical approach, which we implemented in a related yet independent code we call DARSEC (diatomic all-electron real-space electronic structure calculations). Because the high-order finite difference approach in general has been discussed in detail elsewhere, here we naturally focus on aspects that are unique to the prolate-spheroidal grid and/or to the solution of the OEP equation.

The main issue that requires careful attention on a two-dimensional prolate-spheroidal grid is the evaluation of the Laplacian operator. Analytically, after elimination of the ϕ coordinate using eq 8, the Laplacian takes the form

$$\Delta_m(\mu,\nu) = \frac{4}{R^2(\xi^2-\eta^2)}\left[\frac{\partial^2}{\partial\mu^2} + \frac{\xi}{\sqrt{\xi^2-1}}\frac{\partial}{\partial\mu} + \frac{\partial^2}{\partial\nu^2} + \frac{\eta}{\sqrt{1-\eta^2}}\frac{\partial}{\partial\nu} - m^2\left(\frac{1}{\xi^2-1} + \frac{1}{1-\eta^2}\right)\right] \quad (9)$$

where $\xi = \cosh(\mu)$ and $\eta = \cos(\nu)$. One complication is that eq 9 implies that a different Laplacian, Δ_m, and ergo a different Hamiltonian, H_m, is to be operated on functions with different $|m|$ values. Fortunately, this is not a prohibitive complication in practice. According to eq 9 the Kohn−Sham Hamiltonian, H_m, is given by

$$H_m = H_0 + m^2 f(\mu,\nu),$$
$$f(\mu,\nu) = \frac{2}{R^2(\cosh^2(\mu)-\cos^2(\nu))}\times \left(\frac{1}{\cosh^2(\mu)-1} + \frac{1}{1-\cos^2(\nu)}\right) > 0 \quad \forall\mu \geq 0, \quad \pi \geq \nu \geq 0 \quad (10)$$

Denoting the lowest energy eigenvalue and orbital, per a given m, by ε°_m and φ°_m, respectively, we obtain from eq 10 and the variational principle that for $|\tilde{m}| > |m|$

$$\varepsilon^\circ_{\tilde{m}} = \langle\varphi^\circ_{\tilde{m}}|H_{\tilde{m}}|\varphi^\circ_{\tilde{m}}\rangle > \langle\varphi^\circ_{\tilde{m}}|H_m|\varphi^\circ_{\tilde{m}}\rangle \geq \min_\varphi\langle\varphi|H_m|\varphi\rangle = \langle\varphi^\circ_m|H_m|\varphi^\circ_m\rangle = \varepsilon^\circ_m \quad (11)$$

Equation 11 shows that the lowest eigenvalue of H_m is lower than the lowest eigenvalue of $H_{\tilde{m}}$. This means that even though there are an infinite number of $|m|$ values, it suffices to consider a finite (and typically small) set of them, starting from $m = 0$ and onward, in order to compute all the filled states with no risk of missing any such states. In each self-consistent cycle, all relevant Hamiltonians H_m are diagonalized and the density constructed from the solutions of all of them is used to update the Hamiltonians.

A second complication associated with the Laplacian of eq 9 is that it is singular along the $\mu = 0$ and $\nu = 0$, π boundaries of the (μ, ν) domain, which together make up the interatomic axis in real space. This can be handled in two ways. One approach, employed by Kobus et al.,[41] is that orbital values on this axis are to be interpolated. In fact such interpolation is necessary only for $m = 0$, because for all $|m| > 0$ all orbitals vanish identically on the interatomic axis. Alternatively, one can shift the grid by half a grid spacing in both the μ and the ν directions, such that no grid points are found on the singular line. We implemented both approaches and did not observe pronounced differences in performance between them.

The third complication associated with the Laplacian of eq 9 is that it is not Hermitian (see also Becke).[39] This makes the employment of algorithms that assume Hermiticity problematic. There are several ways to mitigate this problem considerably and a detailed discussion is provided in Appendix A.

We now briefly review the calculation of the various potential terms in the Kohn−Sham eq 1. The ionic potential is simply taken as the exact one, i.e.,

$$V_{\text{ion}}(\mathbf{r}) = -\frac{Z_A}{r_A(\mathbf{r})} - \frac{Z_B}{r_B(\mathbf{r})} \quad (12)$$

where Z_A and Z_B are the atomic numbers of atoms A and B, respectively, and with $r_A(\mathbf{r})$ and $r_B(\mathbf{r})$ defined after eq 7. Solutions of a single atom are also possible, by setting Z_A (or Z_B) to zero. The Hartree potential, $V_{\text{H}}(\mathbf{r})$, is obtained from a solution of the Poisson equation $\Delta_m V_{\text{H}}(\mathbf{r}) = -4\pi\rho(\mathbf{r})$ (with $m = 0$ due to the cylindrical symmetry of $V_{\text{H}}(\mathbf{r})$) using the conjugate gradient method.[49] Because the grid is typically quite small, boundary conditions are evaluated using direct integration. As for the exchange-correlation potential, in exact-exchange calculations the terms $P_{ij}(\mathbf{r}) = \int(\varphi^*_{i\sigma}(\mathbf{r}')\varphi_{j\sigma}(\mathbf{r}'))/(|\mathbf{r}-\mathbf{r}'|)\,d^3r'$ are needed for constructing $u^{\text{xc}}_{i\sigma}(\mathbf{r})$ of eq 5. They are evaluated by solving Poisson-like equations, with $\rho_{ij}(\mathbf{r}) = \varphi^*_{i\sigma}(\mathbf{r})\varphi_{j\sigma}(\mathbf{r})$ on the right-hand side instead of $\rho(\mathbf{r})$, using the same method. In this case a value of $m = |m_i - m_j|$ has to be used for applying the Laplacian Δ_m. Finally, the OEP equation is solved using the "S-iteration" method.[14,50] Briefly, the exchange-correlation potential is updated iteratively according to

$$V^{\text{OEP}}_{\text{xc},\sigma_{\text{new}}}(\mathbf{r}) = V^{\text{OEP}}_{\text{xc},\sigma_{\text{old}}}(\mathbf{r}) + cS_\sigma(\mathbf{r}) \quad (13)$$

with $S_\sigma(\mathbf{r})$ defined in eq 3 and where c is a real positive parameter that is manually adjusted for optimal convergence. Details regarding the calculation of the orbital shifts defined in eq 4, that are needed to evaluate $S_\sigma(\mathbf{r})$, are given in Appendix B.

Because all potentials are strictly local, they only provide diagonal entries on the (μ, ν) Hamiltonian matrix. Hence, just like in PARSEC, the Hamiltonian is extremely sparse. Here, off-diagonal elements, and very few of them at that, are introduced only by the matrix representation of the Laplacian (the Hamiltonian here is even sparser than that of

PARSEC due to the absence of a nonlocal pseudopotential term). The full Hamiltonian matrix is therefore never computed nor stored. Instead, the operation of the Hamiltonian matrix on an orbital is evaluated by explicitly considering only its diagonal values and the high-order finite difference expansion coefficients, which is very efficient computationally.

A different issue which we found to be of importance is the numerical evaluation of the various integrals arising in the computation. We used the high-order integration scheme recommended by Kobus et al.,[41] which is based on the seven-point Newton−Cotes integration formula, i.e.,

$$\int_{x_1}^{x_7} f(x)\,\mathrm{d}x \cong \sum_{i=1}^{7} c_i f_i h \qquad \text{with} \qquad c_1 = c_7 = \frac{41}{140},$$
$$c_2 = c_6 = \frac{216}{140}, \quad c_3 = c_5 = \frac{27}{140}, \quad c_4 = \frac{272}{140} \quad (14)$$

This integration scheme facilitates convergence with grids that are significantly sparser than those needed using simple summation and allows for huge savings in both computer time and memory. Specifically, we used the expression:

$$\int f(\mu,\nu)\,\mathrm{d}^3 r = \int_0^{2\pi}\mathrm{d}\phi \int_0^{\infty}\mathrm{d}\mu \int_0^{\pi}\mathrm{d}\nu\, \mathscr{J}(\mu,\nu) f(\mu,\nu)$$
$$\cong 2\pi \sum_{i=1}^{N_\mu}\sum_{j=1}^{N_\nu} \tilde{c}_i \tilde{c}_j \mathscr{J}(\mu_i,\nu_j) f(\mu_i,\nu_j) h_\mu h_\nu \quad (15)$$

where $\mathscr{J}(\mu, \nu, \phi)$ is the volume element, given by

$$\mathscr{J}(\mu,\nu,\phi) = \left(\frac{R}{2}\right)^3 (\cosh^2(\mu) - \cos^2(\nu)) \sinh(\mu) \sin(\nu) \quad (16)$$

and $\tilde{c}_i$ are based on the Newton−Cotes coefficients c_i of eq 14.[51] Obviously this means that the number of grid points in the μ and ν directions must be of the form of $6n + 1$, where n is an integer.

We conclude this section with some practical comments on the DARSEC code in which the above concepts are implemented. Like PARSEC, DARSEC is a modern, massively parallel (using the message passing interface protocol) Fortran 90/95 code. Currently, DARSEC uses the ARPACK software package[52] to diagonalize the Kohn−Sham Hamiltonians, but the approach is generic, and other solvers can easily be used (because the Laplacian matrix is real but not symmetic, we use the nonsymmetric, real ARPACK solver). Similarly, below we provide examples based on local density approximation (LDA) and exact-exchange, but the code is designed such that any orbital-dependent functional can be combined with OEP in a straightforward manner. For the systems considered below, DARSEC requires a few minutes to several hours and a modest few hundreds MB of memory for a complete LDA or exchange-only KLI solution with an accuracy of 10^{-5} hartree in total energies and eigenvalues.

As expected, in non-OEP calculations the majority of the computation time is devoted to the iterative diagonalization process. Our empirical experience with ARPACK shows that using a relatively large amount of Arnoldi basis vectors (of the order of thousands) usually decreases the run time dramatically (by up to 2 orders of magnitudes). We attribute this primarily to the above-discussed singularity of the Laplacian operator along the interatomic axis. A numerical convergence of the order of 10^{-5} hartree is usually achieved by less than 15 self-consistent iterations. This number can be reduced significantly, typically to less than five iterations, by using interpolated converged results from a sparser grid as a starting point for the denser one. For OEP calculations, the required run time for achieving comparable accuracies increases, possibly up to several days, for two main reasons: (a) OEP calculations are typically slower to converge and may require tens of self-consistent iterations; (b) the orbital shifts do not always converge smoothly, and different starting vectors for the conjugate gradient computation of the orbital shifts (see Appendix B) may be needed.

Table 1. Total Energy, in hartree, for the CO Molecule, Calculated with Grids of Increasing Density and Size[a]

	CO−xKLI			
$\{N_\mu, N_\nu\}$	μ_{max}	total energy	ΔE	N_A
{61, 43}	3.86	−112.78377		1300
{85, 55}	3.87	−112.78315	0.00062	1300
{109, 73}	3.98	−112.78320	−0.00005	1600
{133, 103}	4.18	−112.783222	−0.00002	1800
{151, 121}	4.39	−112.783223	−0.000001	2300
Engel et al.[b]		−112.78340		

[a] N_μ, N_ν, number of grid points along the μ, ν directions, respectively; μ_{max}, maximum value of μ used in calculation; ΔE, difference between current total energy and the total energy obtained by the previous coarser grid, in hartree; N_A, number of Arnoldi basis vectors used during diagonalization. [b] For comparison the xKLI total energy of CO as calculated by Engel et al. (ref 43) using a similar prolate-spheroidal coordinate grid and bond length of 2.1316 au is also provided.

In Table 1 we provide convergence details for the representative case of the CO molecule with the experimental equilibrium bond length of 2.132 au,[53] studied using the KLI approximation to the exact-exchange functional. We focus on the convergence of the total energy with respect to the sparsity of the grid (number of grid points) and its size (determined by μ_{max}). Clearly convergence to 10^{-5} hartree was achieved for the largest grid (for which the run time was of the order of several hours), and convergence to 10^{-3} hartree was achieved for a very modest grid indeed. It is our general experience, for this and other systems, that the convergence rate of OEP calculations with increasing grid size is similar.

3. Results

In this section we present detailed numerical results obtained using the above-explained approach. Calculations were performed with the LDA[54,55] or with exact-exchange (and no correlation). For the latter functional, the exchange-potential was given either by the KLI approximation (xKLI) or by a full OEP calculation (xOEP). First, we reproduce known results so as to verify our methodology. Then, we provide new fully numerical exact-exchange OEP results for selected diatomic molecules. All calculations were converged to at least 10^{-4} hartree in the total energy (note that the eigenvalues may exhibit larger errors),

Table 2. Numerical Parameters Used in This Work[a]

system	R_{AB} [au]	N_μ	N_ν	μ_{max} (r_{sa} [au])	S	c
			xcLDA:			
H_2	1.446	85	85	4.22 (24.71)		
BH	2.373	85	85	4.32 (44.79)		
Li_2	5.120	85	85	4.32 (96.65)		
			xKLI and xOEP:			
H	0.5	37	25	6.37 (73.31)	1	1.0
H_2^+	2.0	31	31	5.39 (109.35)	1	1.0
He	0.5	61	61	4.43 (28.68)	1	5.0
Li	0.5	61	61	5.43 (28.68)	100	3.0
Be	0.5	73	55	5.16 (21.87)	200	3.5
LiH	3.015	109	91	3.82 (34.38)	100	7.2
BH	2.336	79	79	4.15 (37.00)	200	5.5
Li_2	5.051	109	91	3.44 (39.37)	200	6.5
CO	2.132	109	73	3.98 (28.56)	100	0.7

[a] R_{AB}, interatomic distance; N_μ, N_ν, number of grid points along the μ, ν directions; μ_{max}, maximum value of μ; r_{sa}, length of the semiminor axis; S (OEP calculations only), number of S-iterations used during the convergence process; c (OEP calculations only), typical value of convergence parameter used in the S-iterations, see eq 13.

Table 3. Ground-State Total Energies and HOMO Eigenvalues, in hartree, for H_2, BH, and Li_2, Calculated with LDA

	total energy [hartree]		HOMO [hartree]	
LDA	Grabo et al.[a]	DARSEC	Grabo et al.[a]	DARSEC
H_2	−1.137692	−1.137692(1)	−0.373092	−0.3730920(7)
BH	−24.9770	−24.97695(2)	−0.2041	−0.2040(7)
Li_2	−14.7245	−14.7244(5)	−0.1187	−0.1186(6)

[a] Ref 10.

with the first unconverged digit placed in parentheses. Numerical parameters for all calculations presented below are given in Table 2.

For verifying our solutions to the Kohn−Sham equation, we calculated the electronic structures of H_2, BH, and Li_2 using the LDA. The resulting total energies and highest occupied molecular orbital (HOMO) energies are given in Table 3 and are compared to the all-electron calculations of Grabo et al.,[10] performed on a similar prolate-spheroidal grid. All diatomic calculations were performed using the same bond lengths as used by Grabo et al.[10] The comparison immediately confirms that the results do indeed agree to within the stated accuracy of 0.1 mhartree.

As a first step toward verifying our all-electron xKLI and xOEP, we performed calculations for the one-electron systems of H and H_2^+, for which xKLI and xOEP are identical and should yield the exact results. Indeed we found that the two solutions always agreed on all converged digits. Table 4 compares DARSEC calculations for the first 14 energy levels ($n = 1-3$) of a single H atom with the analytical values, demonstrating excellent accuracy. Table 5 shows a similar comparison, for H_2^+, of DARSEC values with analytical values[56,57] and with numerical Hartree−Fock (HF) values (also exact for single-electron systems),[58] at the equilibrium bond length of 2.0 au. Note that the "analytical" values are not always correct to the last digit reported, because of numerical approximations used in the algebraic solutions,[56] whereas the results of Laaksonen et al.[58] were shown to be correct to at least the ninth digit, via comparison

Table 4. Ground-State Total Energy and Energy Levels, in rydbergs, for the H Atom, Calculated with xKLI and xOEP

H	exact	DARSEC (xKLI, xOEP)	m
E_{tot}	−1	−1.00000000(3)	
1s	−1	−1.00000000(8)	0
2s	−1/4	−0.25000000(6)	0
2p	−1/4	−0.25000000(6)	−1, 0, 1
3s	−1/9	−0.1111111(2)	0
3p	−1/9	−0.11111111(8)	−1, 0, 1
3d	−1/9	−0.11111111(5)	−2, −1, 0, 1, 2

Table 5. Ground-State Total Energy and Energy Levels, in hartree, for H_2^+, at the Equilibrium Bond Length of 2.0 au, Calculated with xKLI and xOEP

H_2^+	analytical results[a]	Hartree−Fock[b]	DARSEC (xKLI, xOEP)
E_{tot}	−0.6026[c]	−0.6026342145[d]	−0.6026342144(7)
$1\sigma_g(1s)$	−1.102625	−1.102634214497	−1.1026342144(7)
$1\sigma_u(2p)$	−0.667535	−0.667534392205	−0.667534392(1)
$1\pi_u(2p)$	−0.428775	−0.428771819894	−0.428771819(9)
$2\sigma_g(2s)$	−0.360865	−0.36086487543	−0.3608648754(0)
$2\sigma_u(3p)$	−0.255415	−0.25541316515	−0.2554131651(7)
$3\sigma_g(3d)$	−0.235775	−0.235777628822	−0.2357776288(4)
$1\pi_g(3d)$	−0.226700	−0.22669962663	−0.2266996266(7)
$1\delta_g(3d)$		−0.21273268176	−0.212732681(8)
$2\pi_u(3p)$		−0.20086482987	−0.200864830(1)
$4\sigma_g$	−0.177680		−0.177681045(7)
$3\sigma_u$	−0.137315		−0.13731292(5)
$5\sigma_g$			−0.130791877(9)
$2\pi_g(4d)$		−0.12671013060	−0.12671013(1)
$4\sigma_u$	−0.126645		−0.12664387(0)
$3\pi_u(4f)$		−0.12619892048	−0.12619892(1)
$1\delta_u$			−0.12496254(3)
$1\phi_u(4f)$		−0.123125506	−0.12312550(0)
$2\delta_g(4d)$		−0.1210194626	−0.12101946(3)
$4\pi_u$			−0.11591529(2)
$6\sigma_g$			−0.1054423(0)
$5\sigma_u$			−0.085938(9)
$7\sigma_g$			−0.08297(4)
$6\sigma_u$			−0.08084(4)
$3\pi_g(5d)$		−0.080833179	−0.08083(5)

[a] Ref 56. [b] Ref 58. [c] Ref 57. [d] Ref 40.

Table 6. Ground-State Total Energies and HOMO Energies, in hartree, for the He, Li, and Be Atoms, Calculated with xKLI and xOEP

	xKLI		xOEP	
atom	Grabo et al.[a]	DARSEC	Grabo et al.[a]	DARSEC
He	−2.8617	−2.8616(8)	−2.8617	−2.8616(8)
Li	−7.4324	−7.43243(5)	−7.4325	−7.4325(0)
Be	−14.5723	−14.57228(3)	−14.5724	−14.5724(3)

[a] Ref 11.

to previous work[59] (except for the $1\phi_u$, $2\delta_g$, and $3\pi_g$ orbitals, for which no such comparison was available).

As a second step in the evaluation of our all-electron xKLI and xOEP schemes, we compare calculations for polyelectron atoms with similar calculations performed on radial grids. Table 6 shows total energies of He, Li, and Be atoms, compared to the data of Grabo et al.[11] Selected eigenvalues for Li[60] and Be are given in Tables 7 and 8, respectively. In Table 7 the Li eigenvalues are compared to those obtained by Engel and Vosko,[37] as well as to independent calculations we performed using a 1D radial code.[14] In Table 8, the Be eigenvalues are compared to those obtained by Kümmel and Perdew.[14] Once again, an accuracy of 0.1 mhartree is achieved throughout. Furthermore, Tables 7 and 8 also

Table 7. Ground-State Energy Levels and $\Delta E_{x\sigma}^{\mathrm{vir}}$, in hartree, for the Li Atom, Calculated with xKLI and xOEP

	xKLI		xOEP		
Li	1D radial grid	DARSEC	1D radial grid	Engel and Vosko[a]	DARSEC
$1s\uparrow$	−2.08145	−2.0814(5)	−2.05453		−2.0545(3)
$2s\uparrow$	−0.19618	−0.1961(9)	−0.19629	−0.1963	−0.1962(8)
$\Delta E_{x\uparrow}^{\mathrm{vir}}$	0.00529	0.0053	1.28×10^{-7}	−0.000014	−0.00004
$1s\downarrow$	−2.46714	−2.4671(5)	−2.46884	−2.4688	−2.4688(5)
$2s\downarrow$		−0.3026(2)			−0.3031(3)
$\Delta E_{x\downarrow}^{\mathrm{vir}}$	10^{-10}	-3×10^{-7}	-2×10^{-9}	0.000004	−0.000009

[a] Ref 37.

Table 8. Ground-State Energy Levels and $\Delta E_x^{\mathrm{vir}}$, in hartree, for the Be Atom, Calculated with xKLI and xOEP

	xKLI		xOEP	
Be	1D radial grid[a]	DARSEC	1D radial grid[a]	DARSEC
1s	−4.1668	−4.1668(3)	−4.1257	−4.1257(1)
2s	−0.3089	−0.3088(5)	−0.3092	−0.3092(3)
$\Delta E_x^{\mathrm{vir}}$	∼1%	0.02 (0.78%)	0.00001,[b] $\sim 10^{-4}$%[a]	0.00001(4×10^{-4}%)

[a] Ref 14. [b] Ref 37.

Table 9. Ground-State Total Energy, Energy Levels, and $\Delta E_x^{\mathrm{vir}}$, in hartree, for LiH, Calculated with xKLI and xOEP

	xKLI		xOEP
LiH	Grabo et al.[a]	DARSEC	DARSEC
E_{tot}	−7.9868	−7.98680(8)	−7.98691(9)
1σ	−2.0786	−2.0786(0)	−2.069(4)
2σ	−0.3011	−0.3010(5)	−0.3015(9)
$\Delta E_x^{\mathrm{vir}}$		−0.02	−0.000098

[a] Ref 11.

Table 10. Ground-State Total Energy, Energy Levels, and $\Delta E_x^{\mathrm{vir}}$, in hartree, for BH, Calculated with xKLI and xOEP

	xKLI		xOEP
BH	Grabo et al.[a]	DARSEC	DARSEC
E_{tot}	−25.1290	−25.12903(0)	−25.12963(6)
1σ	−6.8624	−6.8623(6)	−6.8126(5)
2σ	−0.5856	−0.58561(5)	−0.577(4)
3σ	−0.3462	−0.34621(2)	−0.34729(7)
$\Delta E_x^{\mathrm{vir}}$		0.01	0.0003

[a] Ref 11.

provide the extent to which the exchange virial relation of Levy and Perdew[61]

$$E_{x,\sigma}[\rho(\mathbf{r})] = \int V_{x,\sigma}^{\mathrm{OEP}}(\mathbf{r})[3\rho_\sigma(\mathbf{r}) + \mathbf{r}\cdot\nabla\rho_\sigma(\mathbf{r})]\,\mathrm{d}^3 r \tag{17}$$

is obeyed, i.e., the difference between the left- and right-hand sides of the above equation, which is denoted by $\Delta E_{x\sigma}^{\mathrm{vir}}$. The resulting deviations are similar to those obtained in previous work[37,14] and, as expected, are significantly smaller for OEP than for KLI (except for the singly occupied spin down channel of Li in which the deviations are practically zero for both xKLI and xOEP).

We now turn to exact-exchange calculations of polyelectron diatomic molecules, starting with xKLI. Tables 9−11 provide total energies and energy levels for LiH, BH, and Li_2. These results are again compared to those of Grabo et al.,[11] where the same bond lengths were used. As with the single-atom systems, we find an agreement in the energies to the stated accuracy. Having verified our approach for a wide range of realistic scenarios, we can now turn to providing new fully numerical xOEP results for the total energy and energy levels of LiH, BH, and Li_2, using the same bond lengths that were used for the xKLI calculations. These results are given in Tables 9−11.

Table 11. Ground-State Total Energy, Energy Levels, and $\Delta E_x^{\mathrm{vir}}$, in hartree, for Li_2, Calculated with xKLI and xOEP

	xKLI		xOEP
Li_2	Grabo et al.[a]	DARSEC	DARSEC
E_{tot}	−14.8706	−14.87058(0)	−14.87076(5)
$1\sigma_g$	−2.0276	−2.0275(9)	−2.01(3)
$1\sigma_u$	−2.0272	−2.0272(4)	−2.01(3)
$2\sigma_g$	−0.1813	−0.1812(7)	−0.1818(3)
$\Delta E_x^{\mathrm{vir}}$		−0.04	−0.0000098

[a] Ref 11.

Although xOEP diatomic calculations that are not fully numerical certainly do exist,[5] we are not aware of independent fully numerical xOEP diatomic calculations to compare our results with.[62] In the absence of those, we confirm the correctness of our calculations by verifying that they satisfy several important criteria:[14] First, for the solution to be an OEP solution, it must satisfy eq 3, i.e., the function $S_\sigma(\mathbf{r})$ must vanish. With the xKLI potentials we find $|S_{\max}| \approx 10^{-2}$ a_0^{-3}, where a_0 is the atomic length unit, whereas for xOEP calculations converged to 10^{-4} hartree (in the total energy), $|S_{\max}|$ is typically of the order of 10^{-4} a_0^{-3} or less. Second, we verify that the obtained xOEP total energies are lower than the ones obtained by xKLI, i.e., $E_{\mathrm{tot}}^{\mathrm{xOEP}} \le E_{\mathrm{tot}}^{\mathrm{xKLI}}$. This must be the case as the OEP solution rigorously satisfies a variational principle minimization, whereas the xKLI solution does not.[63] Third, a lower bound for the xOEP total energy is provided by the HF total energy, i.e., $E_{\mathrm{tot}}^{\mathrm{HF}} \le E_{\mathrm{tot}}^{\mathrm{xOEP}}$. This is because the xOEP potential must be a local one, whereas the HF potential has no such constraint.[63] We verified that our xOEP total energies indeed satisfy this relation by comparing them to the following HF total energies, obtained by Laaksonen et al.[40] from fully numerical calculations conducted using a similar prolate-spheroidal grid and the same bond lengths: $E_{\mathrm{LiH}}^{\mathrm{HF}} = -7.9874$, $E_{\mathrm{BH}}^{\mathrm{HF}} = -25.1316$, $E_{\mathrm{Li_2}}^{\mathrm{HF}} = -14.8716$. These HF energies are all lower than the corresponding xOEP total energies. The xOEP total energies of atoms are closer to the corresponding xKLI total energies than to the HF total energies.[11] This behavior was also observed here for the diatomic calculations. Fourth, we calculated the value of $\Delta E_{x\sigma}^{\mathrm{vir}}$ to make sure that the virial exchange relation of eq 17 holds. We find that, whereas it is slightly violated by the KLI approximation, with an error in the order of 0.01 hartree, with OEP calculations converged to 10^{-4} hartree in the total energy, the virial exchange relation is also typically satisfied to $\sim 10^{-4}$ hartree or better. Last, according to available atomic xOEP calculations[11,14,37,50] the xOEP solutions for the highest occupied eigenvalue are lower than those of xKLI. This is reasonable, because OEP leads to stronger binding than KLI.[11] This results in a higher ionization potential, which equals the negative of the highest

occupied eigenvalue in exact Kohn−Sham theory.[64,65] Tables 9−11 confirm that this trend is also found here.

We end this section with our representative case of the CO molecule. Whereas with xKLI we got a total energy of −112.7832 hartree, the xOEP total energy turns out to be −112.785(3) hartrees. Interestingly, a previously reported OEP calculation for CO performed using basis sets[66] yielded a total energy of −112.77652 hartree. This value is higher than the corresponding KLI one, as also pointed out by Engel et al.[43] This underscores the importance of fully numerical all-electron OEP calculations. In calculations that are not fully numerical, even when the computation is numerically converged it may still contain errors that are inherent in the approximations made.

4. Conclusions

We presented an approach for fully numerical, all-electron solutions of the OEP equation within Kohn−Sham DFT for diatomic molecules. The approach is based on a real-space, prolate-spheroidal coordinate grid for solving the all-electron Kohn−Sham equations and an iterative scheme for solving the OEP equation. The accuracy of the approach, as implemented by us in the DARSEC program, was demonstrated by comparison with previously reported results within LDA, xKLI, and xOEP calculations. Finally, new benchmark fully numerical xOEP results for selected diatomic molecules were provided. Because our method is free of any approximation other than the choice of the approximate exchange-correlation energy functional (and unavoidable numerical roundoff errors), we believe that it may serve as a powerful tool for systematic testing and evaluation of orbital-dependent functionals.

Acknowledgment. This work was supported by the Minerva Foundation, the Minerva−Schmidt Center for Supra-Molecular Chemistry, the Lise Meitner−Minerva Center for Computational Quantum Chemistry, and the German−Israeli Foundation. A.M. thanks Dubi Kelmer for many illuminating discussions and Amir Natan for much help.

Appendix A: Non-Hermiticity of the Laplacian

In the prolate-spheroidal coordinate system, the Laplacian operator, Δ_m of eq 9, is not Hermitian (with respect to a standard inner product). This is because it is a weighted sum of second derivatives, which are Hermitian operators, and first derivatives, which are anti-Hermitian operators. This raises additional numerical issues, which are discussed in this Appendix. We show that although the problem may be completely resolved analytically, it reappears upon discretization of the Kohn−Sham equation. The discretization then either approximates the analytical operator extremely well, but is only almost Hermitian (i.e., most, but not all, of its entries, obey Hermiticity), or it yields a strictly Hermitian matrix, but may be an insufficiently accurate approximation to the analytical operator.

For any differential operator that is Hermitian in Cartesian coordinates but is not Hermitian in some other coordinate system, it is always possible to regain Hermiticity by multiplying with the corresponding Jacobian, $\mathcal{J}$. In particular, for the Laplacian operator in prolate-spheroidal coordinates, we have

Table A1. Ground-State Energy Levels, in rydberg, for the H Atom, Calculated with xKLI Using Non-Hermitian (Third Column) and Hermitian (Fourth Column) Hamiltonians[a]

H	exact	$\mathbf{D}_\alpha^m\mathbf{G}_\alpha\mathbf{D}_\alpha^m$	$-(\mathbf{D}_\alpha^m)^{\mathrm{T}}\mathbf{G}_\alpha\mathbf{D}_\alpha^m$
1s	−1	−1.0000000	−1.00065
2s	−1/4	−0.2500000	−0.25053
2p	−1/4	−0.2500000	−0.25007
	−1/4	−0.2500000	−0.24986
	−1/4	−0.2500000	−0.24986

[a] An expansion order of $2L = 12$ neighbors was used.

$$\int f^*(\mathcal{J}\Delta_m g)\,d\mu\,d\nu\,d\phi = \int f^*(\nabla^2 g)\,dx\,dy\,dz = \int (\nabla^2 f)^* g\,dx\,dy\,dz = \int (\mathcal{J}\Delta_m f)^* g\,d\mu\,d\nu\,d\phi \quad \text{(A1)}$$

where the Jacobian, $\mathcal{J}(\mu, \nu, \phi)$, is a real function, given by eq 16. Multiplying both sides of the Kohn−Sham eq 1 by the Jacobian yields

$$\mathcal{J}\left(-\frac{1}{2}\Delta_m + V_{\mathrm{KS}}\right)\varphi_i = \varepsilon_i \mathcal{J}\varphi_i \quad \text{(A2)}$$

where some algebraic manipulation shows that the operator $\mathcal{J}\Delta_m$ is given by

$$\mathcal{J}\Delta_m(\mu,\nu) = \frac{R}{2}\left[\sin(\nu)\left(\frac{\partial}{\partial\mu}\sinh(\mu)\frac{\partial}{\partial\mu}\right) + \sinh(\mu)\left(\frac{\partial}{\partial\nu}\sin(\nu)\frac{\partial}{\partial\nu}\right) - m^2\left(\frac{\sin(\nu)}{\sinh(\mu)} + \frac{\sinh(\mu)}{\sin(\nu)}\right)\right] \quad \text{(A3)}$$

In this formulation, the Kohn−Sham Hamiltonian is rigorously Hermitian, at the cost of having to solve a generalized eigenvalue problem.

We now consider the discrete representation of eq A3. Clearly, the operator $\mathcal{J}\Delta_m$ is Hermitian in its discrete matrix form if and only if the operator $((\partial/\partial\alpha)g_\alpha(\alpha)(\partial/\partial\alpha))$, where $\alpha = \mu, \nu$ and $g_\alpha =$ sinh, sin, respectively, is represented by a Hermitian matrix. We denote the matrix representation for the usual high-order finite difference expansion of a first partial derivative $\partial/\partial\alpha$ by a matrix $\mathbf{D}_\alpha$. The function g_α is represented by the appropriate diagonal matrix $\mathbf{G}_\alpha$. Normally, the complete operator $((\partial/\partial\alpha)g_\alpha(\alpha)(\partial/\partial\alpha))$ can then be represented as either $[\mathbf{D}_\alpha\mathbf{G}_\alpha\mathbf{D}_\alpha]$ or $[-\mathbf{D}_\alpha^{\mathrm{T}}\mathbf{G}_\alpha\mathbf{D}_\alpha]$. The latter expression is manifestly Hermitian, but normally so is the former because $\mathbf{D}_\alpha$ is anti-Hermitian, i.e., $\mathbf{D}_\alpha^{\mathrm{T}} = -\mathbf{D}_\alpha$. Unfortunately, here the matrix representation of $\mathbf{D}_\alpha$ is complicated by the boundary conditions: If a grid point is near the boundary, values of neighbors lying beyond the boundary (i.e., across the $\mu = 0$, $\nu = 0$, or $\nu = \pi$ lines, or, equivalently, across the interatomic axis in real space) need to be taken into account for evaluating the derivative in that point. These are taken as plus or minus the values at the mirror positions inside the boundary, depending on the angular momentum number m, which dictates the parity of the wave function.[41] Formally, the $2L$th order finite difference matrix representation of the first derivative, for a function with angular number m, then assumes the following form (given here for derivation along μ as an example)

$$\mathbf{D}_{\mu}^{m}(i,j)=\begin{cases}0 & i_\nu \neq j_\nu \text{ or } |i_\mu - j_\mu| > L \\ C_{j_\mu - i_\mu} & i_\nu = j_\nu \text{ and } |i_\mu - j_\mu| \leq L \\ & \quad \text{and } i_\mu - (-j_\mu) > L \\ C_{j_\mu - i_\mu} + (-1)^m C_{-j_\mu - i_\mu} & i_\nu = j_\nu \text{ and } |i_\mu - j_\mu| \leq L \\ & \quad \text{and } i_\mu - (-j_\mu) \leq L \end{cases} \tag{A4}$$

Here, i, j are running indices of two arbitrary grid points and i_α, j_α are running indices for the same two points along the one-dimensional direction α.[67] The weight of the kth neighbor for the first derivative in the high-order finite difference scheme is given by C_k.[44] Because $C_{-k} = -C_k$, it follow that $\mathbf{D}_{\alpha}^{m}(i,j) = -\mathbf{D}_{\alpha}^{m}(j,i)$ for most i,j entries. However, for i, j pairs that are near the boundaries, i.e., belonging in the last line of eq A4, $\mathbf{D}_{\alpha}^{m}(i,j) \neq -\mathbf{D}_{\alpha}^{m}(j,i)$, and consequently $\mathbf{D}_{\alpha}^{m}$ is neither Hermitian nor anti-Hermitian. Thus, the representations $[\mathbf{D}_{\alpha}^{m}\mathbf{G}_{\alpha}\mathbf{D}_{\alpha}^{m}]$ and $[-(\mathbf{D}_{\alpha}^{m})^{\mathrm{T}}\mathbf{G}_{\alpha}\mathbf{D}_{\alpha}^{m}]$ are not equivalent, and only the latter must be Hermitian.

Careful consideration of the structure of the matrix $\mathbf{D}_{\alpha}^{m}$ of eq A4 reveals that $[\mathbf{D}_{\alpha}^{m}]^{\mathrm{T}} = -\mathbf{D}_{\alpha}^{m+1}$. Therefore, $[-(\mathbf{D}_{\alpha}^{m})^{\mathrm{T}}\mathbf{G}_{\alpha}\mathbf{D}_{\alpha}^{m}] = [\mathbf{D}_{\alpha}^{m+1}\mathbf{G}_{\alpha}\mathbf{D}_{\alpha}^{m}]$ differs from $[\mathbf{D}_{\alpha}^{m}\mathbf{G}_{\alpha}\mathbf{D}_{\alpha}^{m}]$ solely in the parity of the left-most matrix. Unfortunately, parity considerations show that it is the latter, non-Hermitian form, rather than the former, Hermitian form, which is the correct one. Importantly, this undesirable tradeoff between accuracy and Hermiticity of the representation is a direct consequence of using a high-order expansion. For $L = 1$, i.e., use of immediate neighbors only (and with the interatomic axis explicitly used), the problem does not arise. However, this clearly comes at the cost of a significant reduction in numerical accuracy, per a given grid step.

The numerical consequences of the difference between the two representation schemes are illustrated in Table A1. For a single H atom, it compares the known analytical results to those of xKLI calculations, diagonalized with the above two matrix representations (and all else being equal). Clearly, the "almost Hermitian" representation (third column) yields results that are accurate to all digits shown, whereas the Hermitian representation (fourth column) produces relatively poor results.

A different perspective on the above considerations can be obtained from variational arguments. Equation A2 can also be derived from the variational principle: Let F be a functional given by

$$\mathrm{F} = \underbrace{\frac{1}{2}\int |\nabla\varphi_i(\mathbf{r})|^2 d^3r}_{\mathrm{T}} + \underbrace{\int (V_{KS}(\mathbf{r}) - \varepsilon_i)|\varphi_i(\mathbf{r})|^2 d^3r}_{\mathrm{I}} ,$$

$$\frac{\delta \mathrm{F}}{\delta\varphi_i(\mathbf{r})} = 0 \quad \forall \mathbf{r}, \tag{A5}$$

then setting the variation of F with respect to $\varphi_i(\mathbf{r})$ to zero, where all variation is performed in the prolate-spheroidal coordinate system, yields eq A2.

Becke suggested that a discrete prolate-spheroidal Hermitian representation for the Kohn−Sham equations can be obtained by discretizing the functional of eq A5 and performing the variation on the discrete form.[39] The discrete variation leads to the Hermitian representation, $-[(\mathbf{D}_{\alpha}^{m})^{\mathrm{T}}\mathbf{G}_{\alpha}\mathbf{D}_{\alpha}^{m}]$,[39] obtained above from a different set of considerations. This means that we have obtained an insufficiently accurate representation despite starting from a completely equivalent analytical expression, a result which merits an explanation. A key issue in this respect is that in the functional F of eq A5, the kinetic energy term is expressed as $(1/2)\int|\nabla\varphi_i(\mathbf{r})|^2\ \mathrm{d}^3r$ and not as $-(1/2)\int\varphi_i^*(\mathbf{r})\nabla^2\varphi_i(\mathbf{r})\ \mathrm{d}^3r$. Normally, the two expression are equivalent analytically by virtue of Green's first identity. But numerically, because of the "mirror" or "antimirror" boundary conditions imposed by the prolate-spheroidal coordinate system, Green's identity is no longer obeyed after discretization. To see that, consider that in prolate-spheroidal coordinates, use of the alternative definition for the kinetic energy, $(1/2)\int|\nabla\varphi_i(\mathbf{r})|^2\ \mathrm{d}^3r$, would lead to one-dimensional integrals, and therefore discretization, of the type

$$\int \left(\frac{\partial}{\partial\alpha}\varphi_i(\mu,\nu,\phi)\right)^*\left(\frac{\partial}{\partial\alpha}\varphi_i(\mu,\nu,\phi)\right)g_\alpha(\alpha)\,\mathrm{d}\alpha \Rightarrow$$
$$\langle\mathbf{D}_{\alpha}^{m}\varphi_i|\mathbf{G}_{\alpha}\mathbf{D}_{\alpha}^{m}\varphi_i\rangle = \langle\varphi_i|(\mathbf{D}_{\alpha}^{m})^{\mathrm{T}}\mathbf{G}_{\alpha}\mathbf{D}_{\alpha}^{m}\varphi_i\rangle \quad \alpha = \mu,\nu \tag{A6}$$

However, the original definition for the kinetic energy, $-(1/2)\int\varphi_i^*(\mathbf{r})\nabla^2\varphi_i(\mathbf{r})\ \mathrm{d}^3r$, leads to different one-dimensional integrals, and hence discretization

$$-\int\varphi_i^*(\mu,\nu,\phi)\left(\frac{\partial}{\partial\alpha}g_\alpha(\alpha)\frac{\partial}{\partial\alpha}\right)\varphi_i(\mu,\nu,\phi)\,\mathrm{d}\alpha \Rightarrow$$
$$-\langle\varphi_i|\mathbf{D}_{\alpha}^{m}\mathbf{G}_{\alpha}\mathbf{D}_{\alpha}^{m}\varphi_i\rangle, \qquad \alpha = \mu,\nu \tag{A7}$$

As explained above, the two forms would have been equivalent had $(\mathbf{D}_{\alpha}^{m})^{\mathrm{T}} = -\mathbf{D}_{\alpha}^{m}$, but as this is not the case, only the latter, non-Hermitian form, is accurate.

Appendix B: Computing the Orbital Shifts

The OEP method requires the construction of "orbital shifts", $\psi_{i\sigma}(\mathbf{r})$, defined in eq 4. According to this equation, $\psi_{i\sigma}(\mathbf{r})$ may be interpreted as the negative of the first-order orbital correction that results if a Kohn−Sham orbital, $\varphi_{i\sigma}(\mathbf{r})$, is subjected to the perturbation[15]

$$\Delta v_{i\sigma}(\mathbf{r}) = u_{i\sigma}(\mathbf{r}) - V_{\mathrm{xc},\sigma}^{\mathrm{OEP}}(\mathbf{r}) \tag{B1}$$

The orbital shifts may thus be computed using first-order perturbation theory from

$$(H_{\mathrm{KS}} - \varepsilon_{i\sigma}^{0})\psi_{i\sigma}(\mathbf{r}) = -(\varepsilon_{i\sigma}^{1} - \Delta v_{i\sigma}(\mathbf{r}))\varphi_{i\sigma}(\mathbf{r}) \tag{B2}$$

where $\varepsilon_{i\sigma}^{0}$ is the ith Kohn−Sham eigenvalue and $\varepsilon_{i\sigma}^{1}$ is its first-order correction.[10,50] The solution of eq B2 is obtained numerically by using the conjugate gradient (CG)[49] method. In practice, solving this equation in DARSEC with the prolate-spheroidal coordinates requires several observations:

First, following eq 4 and using the rotational symmetry of the perturbed potential, $\Delta v_{i\sigma}(\mu,\nu,\phi) = \Delta v_{i\sigma}(\mu,\nu,0)$, the orbital shifts can be shown to have the same rotational symmetry as their associated Kohn−Sham orbitals

$$\varphi_{k\sigma}(\mu,\nu,\phi) = \varphi_{k\sigma}(\mu,\nu,0)\mathrm{e}^{\mathrm{i}m_k\phi} \Leftrightarrow \psi_{k\sigma}(\mu,\nu,\phi) = \psi_{k\sigma}(\mu,\nu,0)\mathrm{e}^{\mathrm{i}m_k\phi} \tag{B3}$$

Second, the CG method assumes Hermiticity of the inverted operator, whereas in DARSEC the Kohn−Sham matrix is not Hermitian (see Appendix A). This difficulty is present in all other CG applications in DARSEC, but it is usually solved by multiplying both sides of the equation by the volume element $\mathcal{J}(\mu, \nu, \phi)$ (eq 16), which makes the matrix Hermitian "enough" for the CG method to work properly. In the case of eq B2, however, the non-Hermiticity of the matrix is more severe as it also influences the evaluation of the $\varepsilon^1_{i\sigma}$ term in eq B2. This is because the usual expression, given by $\varepsilon^1_{i\sigma} = \langle\varphi_{i\sigma}|\Delta v_{i\sigma}|\varphi_{i\sigma}\rangle \equiv \overline{\Delta v_{i\sigma}(\mathbf{r})}$ is no longer valid. Instead, first-order perturbation theory for a general (not necessarily Hermitian) operator shows that $\varepsilon^1_{i\sigma}$ generally takes the following form

$$\varepsilon^1_{i\sigma} = \langle\varphi_{i\sigma}|\Delta v_{i\sigma}|\varphi_{i\sigma}\rangle + \langle\psi_{i\sigma}|H_{\mathrm{KS}}|\varphi_{i\sigma}\rangle + \langle\varphi_{i\sigma}|H_{\mathrm{KS}}|\psi_{i\sigma}\rangle \tag{B4}$$

For a Hermitian operator, eq B4 properly reduces to the usual expression due to the orthogonality between $\varphi_{i\sigma}(\mathbf{r})$ and its orbital shift, $\psi_{i\sigma}(\mathbf{r})$.[14] Since the Kohn−Sham Hamiltonian in DARSEC is represented by an almost Hermitian matrix, the resulting orbitals are no longer fully orthogonal and the matrix cannot be applied to the left. As a result, all the terms of eq B4 must be taken explicitly into account when eq B2 is solved. To make this equation compatible with the required CG form—a known matrix **A** on the left-hand side and a known vector **b** on the right-hand side—we rearrange it in the form

$$\underbrace{\left[H_{KS} - \varepsilon^0_{i\sigma} - \varepsilon^0_{i\sigma}|\varphi_{i\sigma}\rangle\langle\varphi_{i\sigma}| - |\varphi_{i\sigma}\rangle\langle\varphi_{i\sigma}|H_{KS}\right]}_{A} \underbrace{|\psi_{i\sigma}\rangle}_{x} = \underbrace{-(\overline{\Delta v_{i\sigma}(\mathbf{r})} - \Delta v_{i\sigma}(\mathbf{r}))|\varphi_{i\sigma}\rangle}_{b} \tag{B5}$$

Because the deviation from Hermiticity of the Hamiltonian matrix is relatively small (see Appendix A), the added terms $\langle\psi_{i\sigma}|H_{\mathrm{KS}}|\varphi_{i\sigma}\rangle$ and $\langle\varphi_{i\sigma}|H_{\mathrm{KS}}|\psi_{i\sigma}\rangle$ are also small, of the order of 10^{-4} hartree. Still, including them in the calculations was found to be crucial for getting converged solutions.

Third, we have found it numerically useful to compute the orbital shifts in two successive steps: At first, we find an approximation to the orbital shifts by solving eq B2 with $\varepsilon^1_{i\sigma} = \overline{\Delta v_{i\sigma}(\mathbf{r})}$, using the symmetric (but inaccurate) Hamiltonian matrix. In the second and final step the CG method is applied once again, but now it solves eq B5 with the highly accurate (almost symmetric) Hamiltonian matrix and with the previously found orbital shift approximations as initial guess vectors.

References

(1) Hohenberg, P.; Kohn, W. *Phys. Rev. B* **1964**, *136*, 864.

(2) Kohn, W.; Sham, L. J. *Phys. Rev. A* **1965**, *140*, 1133.

(3) Dreizler, R. M.; Gross, E. K. U. *Density Functional Theory: An Approach to the Quantum Many-Body Problem*; Springer: Berlin, 1990.

(4) von Barth, U.; Hedin, L. *J. Phys. C* **1972**, *5*, 1629.

(5) Kümmel, S.; Kronik, L. *Rev. Mod. Phys.* **2008**, *80*, 3.

(6) Sharp, R. T.; Horton, G. K. *Phys. Rev.* **1953**, *90*, 317.

(7) Talman, J. D.; Shadwick, W. F. *Phys. Rev. A* **1976**, *14*, 36.

(8) Görling, A.; Levy, M. *Phys. Rev. A* **1994**, *50*, 196.

(9) Görling, A.; Levy, M. *Int. J. Quantum Chem. Symp.* **1995**, *29*, 93.

(10) Grabo, T.; Kreibich, T.; Gross, E. K. U. *Mol. Eng.* **1997**, *7*, 27.

(11) Grabo, T.; Kreibich, T.; Kurth, S.; Gross, E. K. U. In *Strong Coulomb Correlation in Electronic Structure: Beyond the Local Density Approximation*; Anisimov, V. I. , Ed.; Gordon & Breach: Amsterdam, The Netherlands, 2000; pp 203−317.

(12) Görling, A. *J. Chem. Phys.* **2005**, *123*, 062203.

(13) Krieger, J. B.; Li, Y.; Iafrate, G. J. *Phys. Rev. A* **1992**, *46*, 5453.

(14) Kümmel, S.; Perdew, J. P. *Phys. Rev. B* **2003**, *68*, 035103.

(15) Krieger, J. B.; Li, Y.; Iafrate, G. J. *Phys. Rev. A* **1992**, *45*, 101.

(16) Gritsenko, O. V.; Baerends, E. J. *Phys. Rev. A* **2001**, *64*, 042506.

(17) Della Sala, F.; Görling, A. *J. Chem. Phys.* **2001**, *115*, 5718.

(18) Wilson, P. J.; Tozer, D. J. *Chem. Phys. Lett.* **2001**, *337*, 341.

(19) Arbuznikov, A. V.; Kaupp, M. *Chem. Phys. Lett.* **2004**, *386*, 8.

(20) Kümmel, S.; Kronik, L.; Perdew, J. P. *Phys. Rev. Lett.* **2004**, *93*, 213002.

(21) Rinke, P.; Qteish, A.; Neugebauer, J.; Scheffler, M. *Phys. Status Solidi B* **2008**, *245*, 929.

(22) Körzdörfer, T.; Mundt, M.; Kümmel, S. *Phys. Rev. Lett.* **2008**, *100*, 133004.

(23) Körzdörfer, T.; Kümmel, S.; Mundt, M. *J. Chem. Phys.* **2008**, *129*, 014110.

(24) Bylander, D. M.; Kleinman, L. *Phys. Rev. B* **1995**, *52*, 14566.

(25) Städele, M.; Moukara, M.; Majewski, J. A.; Vogl, P.; Görling, A. *Phys. Rev. B* **1999**, *59*, 10031.

(26) Moukara, M.; Städele, M.; Majewski, J. A.; Vogl, P.; Görling, A. *J. Phys.: Condens. Matter* **2000**, *12*, 6783.

(27) Engel, E.; Höck, A.; Schmid, R. N.; Dreizler, R. M.; Chetty, N. *Phys. Rev. B* **2001**, *64*, 125111.

(28) Sharma, S.; Dewhurts, J. K.; Ambrosch-Draxl, C. *Phys. Rev. Lett.* **2005**, *95*, 136402.

(29) Hirata, S.; Ivanov, S.; Grabowski, I.; Bartlett, R. J.; Burke, K.; Talman, J. D. *J. Chem. Phys.* **2001**, *115*, 1635.

(30) Staroverov, V. N.; Scuseria, G. E.; Davidson, E. R. *J. Chem. Phys.* **2006**, *124*, 141103.

(31) Staroverov, V. N.; Scuseria, G. E.; Davidson, E. R. *J. Chem. Phys.* **2006**, *125*, 081104.

(32) Izmaylov, A. F.; Staroverov, V. N.; Scuseria, G. E.; Davidson, E. R.; Stoltz, G.; Cancès, E. *J. Chem. Phys.* **2007**, *126*, 084107.

(33) Hesselmann, A.; Götz, A. W.; Della Sala, F.; Görling, A. *J. Chem. Phys.* **2007**, *127*, 054102.

(34) Heaton-Burgess, T.; Bulat, F. A.; Yang, W. *Phys. Rev. Lett.* **2007**, *98*, 256401.

(35) Görling, A.; Hesselmann, A.; Jones, M.; Levy, M. *J. Chem. Phys.* **2008**, *128*, 104104.

(36) Heaton-Burgess, T.; Yang, W. *J. Chem. Phys.* **2008**, *129*, 194102.

(37) Engel, E.; Vosko, S. H. *Phys. Rev. A* **1993**, *47*, 2800.

(38) Gálvez, F. J.; Buendía, E.; Maldonado, P.; Sarsa, A. J. *Eur. Phys. J. D* **2008**, *50*, 229–235.

(39) Becke, A. D. *J. Chem. Phys.* **1982**, *76*, 6037.

(40) Laaksonen, L.; Pyykkö, P.; Sundholm, D. *Comp. Phys. Rep.* **1986**, *4*, 313.

(41) Kobus, J.; Laaksonen, L.; Sundholm, D. *Comput. Phys. Commun.* **1996**, *98*, 346.

(42) Engel, E.; Höck, A.; Dreizler, R. M. *Phys. Rev. A* **2000**, *61*, 032502.

(43) Engel, E.; Höck, A.; Dreizler, R. M. *Phys. Rev. A* **2000**, *62*, 042502.

(44) Fornberg, B. *Math. Comput.* **1988**, *51*, 699.

(45) Beck, T. L. *Rev. Mod. Phys.* **2000**, *72*, 1041.

(46) (a) Chelikowsky, J. R.; Troullier, N.; Saad, Y. *Phys. Rev. Lett.* **1994**, *72*, 1240. (b) Chelikowsky, J. R.; Troullier, N.; Wu, K.; Saad, Y. *Phys. Rev. B* **1994**, *50*, 11355.

(47) Chelikowsky, J. R. *J. Phys. D* **2000**, *33*, R33.

(48) Kronik, L.; Makmal, A.; Tiago, M. L.; Alemany, M. M. G.; Jain, M.; Huang, X.; Saad, Y.; Chelikowsky, J. R. *Phys. Status Solidi B* **2006**, *243*, 1063.

(49) Reid, J. K. In *Large Sparse Sets of Linear Equations: Proceedings of the Oxford Conference of the Institute of Mathematics and Its Applications*; Reid, J. K. Ed.; Academic Press: London, United Kingdom, 1971; pp 231−254.

(50) Kümmel, S.; Perdew, J. P. *Phys. Rev. Lett.* **2003**, *90*, 043004.

(51) (a) Define $p(i) = ((i-1) \text{ modulo } 6) + 1$. Then $\tilde{c}_i = c_{p(i)}$ except at points which are the upper end of one seven-point integration and the lower end of another, at which $\tilde{c}_i = 2c_{p(i)}$. (b) When the interatomic axis is avoided by shifting the grid points by half-step size (second grid type) we use the "trapezoid rule" to allow for integration up to the interatomic axis. Consequently, the values of $\tilde{c}_{\mu_1}$, $\tilde{c}_{\nu_1}$, and $\tilde{c}_{\nu_{N_\nu}}$ are increased by 1/4 (note that the volume element $\mathcal{J}(\mu, \nu, \phi)$ vanishes at this axis).

(52) Lehoucq, R. B.; Maschhoff, K.; Sorensen, D.; Yang, C. ARPACK-Arnoldi Package. http://www.caam.rice.edu/software/ARPACK/ (accessed Apr 23, 2009).

(53) Carbon monoxide NIST. http://webbook.nist.gov/cgi/cbook.cgi?Formula=CO&NoIon=on&Units=SI&cDI=on (accessed Apr 23, 2009).

(54) Ceperley, D. M.; Alder, B. J. *Phys. Rev. Lett.* **1980**, *45*, 566.

(55) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244.

(56) Bates, D. R.; Ledsham, K.; Stewart, A. L. *Philos. Trans. R. Soc. London, Ser. A* **1953**, *246*, 215.

(57) Wind, H. *J. Chem. Phys.* **1965**, *42*, 2371.

(58) Laaksonen, L.; Pyykkö, P.; Sundholm, D. *Int. J. Quantum Chem.* **1983**, *23*, 309.

(59) Madsen, M. M.; Peek, J. M. *At. Data* **1971**, *2*, 171.

(60) The results for Li, in our work as well as in previous studies, violate the aufbau principle, i.e., the energy levels are not occupied in ascending order. This is clearly seen in Table 7, where the second state of the spin down channel is not occupied even though its eigenvalue is lower than the eigenvalue of the occupied second state of the spin up channel. We interpret this as a failure of the exchange-only approximation and an indicator of the qualitative importance of correlation in the electronic structure of Li.

(61) Levy, M.; Perdew, J. P. *Phys. Rev. A* **1985**, *32*, 2010.

(62) Note that our fully numerical results for the total energy of LiH do agree well with the results obtained using Gaussian basis sets (ref 17).

(63) Krieger, J. B.; Li, Y.; Iafrate, G. J. In *Density Functional Theory*; Gross, E. K. U., Dreizler, R. M. , Eds.; Plenum Press: New York, 1995; p 191.

(64) Levy, M.; Perdew, J. P.; Sahni, V. *Phys. Rev. A* **1984**, *30*, 2745.

(65) Almbladh, C.-O.; von Barth, U. *Phys. Rev. B* **1985**, *31*, 3231.

(66) Ivanov, S.; Hirata, S.; Bartlett, R. J. *Phys. Rev. Lett.* **1999**, *83*, 5455.

(67) (a) The analogous expression for the ν derivative must also consider the boundary at $\nu = \pi$:

$$\mathbf{D}^m_\nu(i,j) = \begin{cases} 0 & i_\mu \neq j_\mu \text{ or } |i_\nu - j_\nu| > L \\ C_{j_\nu - i_\nu} & i_\mu = j_\mu \text{ and } |i_\nu - j_\nu| \leq L \\ & \quad \text{and } (i_\nu - (-j_\nu)) > L \\ & \quad \text{and } 2N_\nu - i_\nu - j_\nu > L) \\ C_{j_\nu - i_\nu} + (-1)^m C_{-j_\nu - i_\nu} & i_\mu = j_\mu \text{ and } |i_\nu - j_\nu| \leq L \\ & \quad \text{and } i_\nu - (-j_\nu) \leq L \\ C_{j_\nu - i_\nu} + (-1)^m C_{2N_\nu - j_\nu - i_\nu + 2} & i_\mu = j_\mu \text{ and } |i_\nu - j_\nu| \leq L \\ & \quad \text{and } 2N_\nu - j_\nu - i_\nu + 2 \leq L \end{cases}$$

(b) This description is suitable for grids that use the interatomic axis. For the second grid type in which the interatomic axis is avoided, entries near the boundaries are given by $C_{j_\alpha - i_\alpha} + (-1)^m C_{-i_\alpha - j_\alpha + 1}$ for $-i_\alpha - j_\alpha + 1 \leq L$ (and for the $\nu = \pi$ boundary: $C_{j_\nu - i_\nu} + (-1)^m C_{2N_\nu - i_\nu - j_\nu + 1}$ for $2N_\nu - i_\nu - j_\nu + 1 \leq L$).

CT800485V

JCTC Journal of Chemical Theory and Computation

Generalization of the New Resonance Theory: Second Quantization Operator, Localization Scheme, and Basis Set

Atsushi Ikeda,[†] Yoshihide Nakao,[†] Hirofumi Sato,*[,†] and Shigeyoshi Sakaki[†,‡]

Department of Molecular Engineering, Kyoto University, Kyoto 615-8510, Japan

Received January 29, 2009

Abstract: We have recently proposed a method to evaluate the weights of resonance structures embedded in a molecular orbital by utilizing singlet-coupling scheme of an electron pair [*J. Phys. Chem. A* **2006**, *110*, 9028]. The method was formulated on the basis of the second quantization, in which a biorthogonal operator related to Mulliken population (MP) was used together with the Boys−Foster (BF) localization scheme. Our method is very easy to use; only a standard localization procedure is required to obtain the resonance weights. In addition, obtained results agreed well with our chemical intuition. In the present Article, the restrictions, namely MP and BF, were removed, and an operator related to Löwdin population (LP) and other various types of localization schemes were employed to examine the generality of the method. We found that computed resonance weights were virtually independent not only on the choice of these combinations but also on basis set. This new finding, the invariant nature in terms of resonance, may suggest that the present approach could be promising for analyzing molecular orbitals.

1. Introduction

The chemical bond is a central concept in chemistry. However, the presently available computational tools are not always related to the concept of the bond. In principle, there are two ways to obtain the electronic wave function: valence bond (VB) method and molecular orbital (MO) method. The former provides an understanding of the chemical bond in a relatively intuitive way, being related to the concepts of covalency, ionicity, and their resonance. Many modern electronic structure theories and their applications are based on the latter method. In this regard, a bridge between the two methods is indispensable. In other words, VB-based characterization of the MO wave function is highly desired to elucidate the nature of chemical bonding. Karafiloglou et al. are working vigorously to address this problem,[1] and several other methods for such purpose have been developed so far, including papers by Shaik et al.,[2] the pioneering work by Hiberty et al.,[3] natural resonance theory (NRT)[4] by Weinhold et al., and the method based on CASSCF-type wave function by Hirao and co-workers.[5] MOVB by Mo and Gao is a direct realization that fits the present purpose.[6] Another type of analysis based on locally defined energy by Nakai et al. can also offer a detailed look inside at the electronic structure of a molecule and its bonds.[7]

Recently, we proposed a new analysis method to evaluate the weights of resonance structures and applied it to several molecular systems.[8] All the results fit in with our chemical intuition. For instance, the method was combined with RISM-SCF,[9] which provides microscopic information on the solvation effect based on statistical mechanics for molecular liquids, and the enhancement of the ionic contribution to electronic structure in solvated molecular system was adequately calculated.[8b,c] Our method is very easy to use: Because of the simple strategy based on the second quantization of singlet-coupling in the target orbital, what we need to obtain the weights is only the density matrix of localized orbitals, and the additional computational cost is negligible. The results also showed excellent agreement with the past reports.[8a]

In the original work,[8] a biorthogonal operator related to Mulliken-population (MP) introduced by Mayer[10,11] was

* Corresponding author e-mail: hirofumi@moleng.kyoto-u.ac.jp.
[†] Kyoto University.
[‡] Fukui Institute for Fundamental Chemistry, Kyoto University, Takano-Nishihiraki-cho 34-4, Sakyo-ku, Kyoto 606-8103, Japan.

10.1021/ct900053r CCC: $40.75
Published on Web 06/01/2009

employed together with the Boys-Foster (BF)[12] localized orbitals. In the present study, we erase these restrictions, and the protocol is extended to a generalized one that includes Löwdin population (LP) related operators[13] with various localization schemes such as Edmiston−Ruedenberg (ER)[14] and Pipek−Mezey (PM)[15] methods. Furthermore, the basis set dependence is examined since it is usually believed to become a serious issue in this type of analysis. The obtained results shows remarkable invariance as explained below, establishing that the present analysis can deliver clear understanding of chemical bondings.

2. Theory

The first order density matrix $(\mathbf{D})_{\nu\mu}$ of wave function $|\Psi\rangle$ is given by eq 1

$$(\mathbf{D})_{\mu\nu} = \langle\Psi|a_\nu^+ a_\mu^-|\Psi\rangle \tag{1}$$

where a_ν^+ and a_μ^- are the creation and annihilation operators related to atomic orbitals (AOs) ν and μ, respectively.[11] In a similar manner, we can define density matrices for an orthonormalized orbital ψ_i ($i = 1, 2, ...$), as follows:

$$(\mathbf{d}^i)_{\mu\nu} = \langle\psi_i|a_\nu^+ a_\mu^-|\psi_i\rangle \tag{2}$$

The matrix holds the idempotency

$$(\mathbf{d}^i)_{\mu\mu} = \frac{1}{2}\sum_\nu (\mathbf{d}^i)_{\mu\nu}(\mathbf{d}^i)_{\nu\mu} \tag{3}$$

and the number of electrons is conserved in each orbital in closed-shell system.

$$2 = \sum_\mu (\mathbf{d}^i)_{\mu\mu} \tag{4}$$

From eqs 3 and 4, a simple equation is obtained:

$$1 = \frac{1}{4}\sum_\mu\sum_\nu (\mathbf{d}^i)_{\mu\nu}(\mathbf{d}^i)_{\nu\mu} = \sum_{\mathrm{M,N}} W^i_{\mathrm{MN}} \text{ where}$$
$$W^i_{\mathrm{MN}} = \frac{1}{4}\sum_{\mu\in\mathrm{M}}\sum_{\nu\in\mathrm{N}} (\mathbf{d}^i)_{\mu\nu}(\mathbf{d}^i)_{\nu\mu} \tag{5}$$

M and N are atomic labels. By introducing spin variables σ_1 and σ_2 ($\sigma_1 \neq \sigma_2$), the quantity is also expressed as the expectation value of an operator

$$\frac{1}{4}(\mathbf{d}^i)_{\mu\nu}(\mathbf{d}^i)_{\nu\mu} = \langle\psi_i|a_\nu^{\sigma_1+} a_\mu^{\sigma_2+} a_\nu^{\sigma_2-} a_\mu^{\sigma_1-}|\psi_i\rangle \tag{6}$$

In the case of $\mu \neq \nu$, both $^1/_4(\mathbf{d}^i)_{\nu\mu}(\mathbf{d}^i)_{\nu\mu}$ and $^1/_4(\mathbf{d}^i)_{\nu\mu}(\mathbf{d}^i)_{\nu\mu}$ represent the weight of the state in which two electrons are singlet-coupled and shared by two AOs, μ and ν. In the case of $\mu = \nu$, $^1/_4(\mathbf{d}^i)_{\mu\mu}(\mathbf{d}^i)_{\mu\mu}$ represents the weight of the state in which two electrons occupy the same AO (μ). Hence, $2W^i_{\mathrm{MN}} = W^i_{\mathrm{MN}} + W^i_{\mathrm{NM}}$ is considered as the weight of the state in which two electrons in ψ_i are shared between M and N atoms, and W^i_{MM} is that of the state in which two electrons in ψ_i are belonging to the atom M.

Let us consider a localized molecular orbital (LMO) ψ_i^{local} ($i = 1, 2, ...$), which has a two-center character between A and B atoms. Equations 5 and 6 are then

$$1 = \sum_{\mathrm{M,N}} W^i_{MN} = W^i_{\mathrm{AA}} + 2W^i_{\mathrm{AB}} + W^i_{\mathrm{BB}} + \left\{\begin{array}{l}\text{all other terms } (W^i_{\mathrm{MN}}) \text{ in which}\\ \text{(M, N) is not (A or B) at the same time}\end{array}\right\}$$
$$= W^i_{\mathrm{AA}} + 2W^i_{\mathrm{AB}} + W^i_{\mathrm{BB}} + \bar{W}^i \tag{7}$$

One can notice that each term corresponds to the weights of ionic and covalent character in the bond between A and B. W^i_{AA} is the weight of the ionic structure ($A^-\ B^+$), W^i_{BB} is that of the ionic structure ($A^+\ B^-$), and $2W^i_{\mathrm{AB}}$ is that of the covalent structure (A−B). $\bar{W}^i$, sum of all the terms in braces corresponding to a many-body term, arises from the fact that LMO often penetrates into other than A and B atoms. As will be shown below, however, $\bar{W}^i$ is actually very small because the concerned two electrons are usually localized in the area between A and B.

Total wave function of a molecule ($|\mathbf{\Psi}\rangle$) is invariant to any unitary transformation among occupied orbitals, and the choice of the orthonormalized orbital is arbitrary. In general, each MO can be localized into either one-center (core orbital, lone-pair orbital, etc.) or two-center (bonding) orbitals. If electrons in different LMOs are independent of each other, the weights of resonance structures of the molecule are simply represented by multiplications of the weights of the two-center bonding orbitals (note that atomic index A and B must be related to the orbital i).

$$1 = \prod_i^{\mathrm{LMOs}} (W^i_{\mathrm{AA}} + 2W^i_{\mathrm{AB}} + W^i_{\mathrm{BB}} + \bar{W}^i) \tag{8}$$

Since the sum of the four terms in parentheses is always 1, normalization of the weights is always guaranteed. It is noteworthy that the contribution from the one-center orbital is regarded as unity because of A = B with negligible $\bar{W}^i$. The alternative view is that the one-center contribution must be simply taken out because it does not participate in the formation of bondings. The separation between one- and two-center orbitals is readily defined, judging from the population assigned to each atom in the localized orbital. As a consequence, the resonance structure of a molecule can be computed by the combinational products of each bonding contribution.

In eq 5, what we need is to compute the density matrix elements related to localized orbital ψ_i^{local}, $(\mathbf{d}^i)_{\nu\mu}$. Now we have two issues that need to be selected in the actual computation of this quantity. One is the choice of the operator described in eq 1, and the other is orbital localization scheme to obtain ψ_i^{local}. For the former choice, both the nonorthogonal and orthogonal AO based operators are examined in this study: nonorthogonal-AO creation operator χ_ν^+ and its biorthogonal-AO annihilation operator ϕ_μ^- related to MP,[10,11] which were introduced in our original work, and the operators, l_ν^+ and l_μ^-, related to LP.[13]

$$\text{Mulliken type: } (\mathbf{d}^i)_{\nu\mu} = (\mathbf{p}^i\mathbf{S})_{\nu\mu}\ (a_\nu^+ = \chi_\nu^+ \text{ and } a_\mu^- = \phi_\mu^-) \tag{9}$$

Löwdin type: $(\mathbf{d}^i)_{\nu\mu} = (\mathbf{S}^{1/2}\mathbf{p}^i\mathbf{S}^{1/2})_{\nu\mu}$ ($a_\nu^+ = l_\nu^+$ and $a_\mu^- = l_\mu^-$) (10)

S is the overlap matrix and $(\mathbf{p}^i)_{\nu\mu}$ is an element constituting **P**-matrix (**P**) for given orbital i.[10,13]

$$(\mathbf{p}^i)_{\mu\nu} = 2(\mathbf{C}^L)_{\mu i}(\mathbf{C}^{L^*})_{\nu i} \quad (11)$$

$(\mathbf{C}^L)_{i\mu}$ is the LCAO coefficient of LMO i. The standard **P**-matrix is computed by summing over all the occupied orbitals, and only the total sum is invariant under the transformation from canonical orbitals (being delocalized) to other localized ones.

$$\begin{aligned}(\mathbf{P})_{\mu\nu} &= \sum_i^{\text{occ}} (\mathbf{p}^i)_{\mu\nu} \\ &= 2\sum_i^{\text{occ}} (\mathbf{C})_{\mu i}(\mathbf{C}^*)_{\nu i} = 2\sum_j^{\text{occ}} (\mathbf{C}^{L1})_{\mu j}(\mathbf{C}^{L1^*})_{\nu j} \\ &= 2\sum_k^{\text{occ}} (\mathbf{C}^{L2})_{\mu k}(\mathbf{C}^{L2})_{\nu i} = \ldots\end{aligned} \quad (12)$$

For the choice of localization as the second option, three major schemes were examined in this article: Boys-Foster (BF), Edmiston-Ruedenberg[14] (ER), and Pipek-Mezey[15] (PM) localization. Hence, the original analysis is generalized in terms of these various combinations of two options. Hereafter, the combination is respectively called BF-, ER- and PM-weights based on the MP (Mullliken) or LP (Löwdin) population related operator. Additionally, the basis set dependence could be crucial from the standpoint of the invariance of the theory. A series of basis sets implemented in the GAMESS program package,[20] DZ, DZP, TZ, TZP, and TZP+,[17] were used for the purpose of a systematic investigation, and further large basis sets were also employed. It is noted that 5d orbitals were used throughout the study since Löwdin population analysis with 6d orbitals is not rotationally invariant.[18,19] All calculations were performed with program code GAMESS[20] modified by us.

3. Results and Discussion

3.1. H_2: A Basal Examination. At first, the weights of resonance structures of H_2 were calculated. This is a basal examination and the so-called minimum requirements in this type of analysis. Because the occupied orbital is unique in this two-electron system, the orbital transformation is not necessary, meaning it is unrelated to the localization scheme. Another special character of this system is that weights do not depend on the choice of basis set at all due to the high symmetry.

The obtained weight of covalent structure H−H was 50%, and that of each ionic structure H^+ H^- and H^- H^+ was, respectively, 25%. These properly exhibit a well-known fact that the electronic structure of H_2 in the Hartree−Fock wave function possesses half-covalent and half-ionic character.

3.2. H_2O and NH_3. Second, the H_2O molecule is examined. After carrying out standard MO computations, the orbitals were localized by BF, ER, and PM procedures

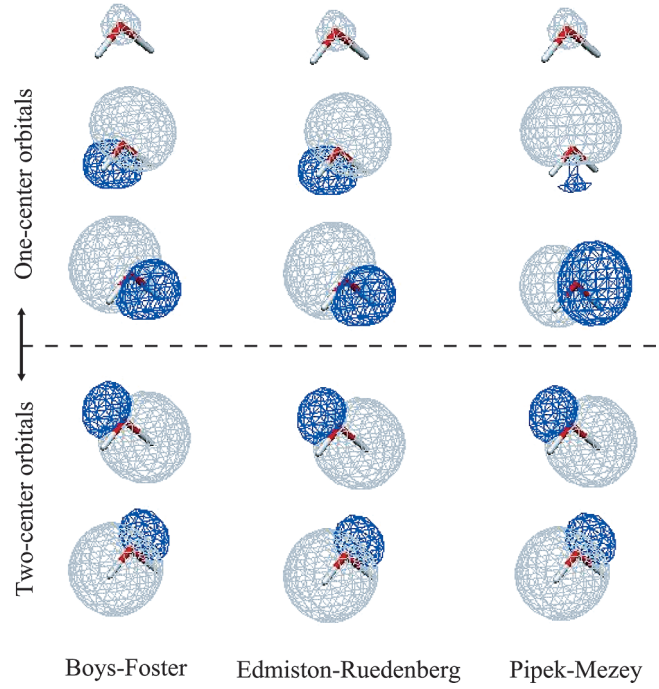


Figure 1. Three one-center and two two-center BF-, ER-, and PM-orbitals of H_2O. TZP basis sets were used.

(Figure 1). Independent of the basis set choice, three one-center ($i = 1,2,3$) and two two-center ($i = 4,5$) orbitals were obtained by each localization method. Regarding one-center orbitals, the core orbital is common to all the procedures. The distinct difference is the PM localization produced one in-plane and one out-of-plane lone pair orbital, whereas BF and ER localizations gave two equivalent lone pair orbitals. But for the two-center orbitals, all the three localization procedures provide very similar orbitals corresponding to the two O−H bonds (OH_1 and OH_2). The weights of resonance structures of H_2O are calculated from $(\mathbf{p}^i)_{\nu\mu}$ of the two two-center orbitals, which participate the bond formation.

$$1 = (W_{OO}^4 + 2W_{OH_1}^4 + W_{H_1H_1}^4 + \bar{W}^4)(W_{OO}^5 + 2W_{OH_2}^5 + W_{H_2H_2}^5 + \bar{W}^5) \quad (13)$$

By opening the brackets, the weights are calculated as shown in Figure 2. The upper panel shows results obtained from the MP operator while the lower panel shows results from the LP one using various basis sets. The six combinations are possible at the choice of the localization (BF, ER, and PM) together with the operator (MP and LP), but the ER-weights were not shown in the figure. This is because they are virtually the same as the BF-weights, and the differences were always less than 0.1% in all the resonance structures. All in all, the most important resonance structure was **2**, in which one O−H bond was ionic and the other is covalent. The next is the totally covalent structure **1**, which is comparable with a totally ionic structure **4**. As illustrated in the figure, BF- and PM-weights are almost the same, meaning that the three localization procedures provide essentially the same results. This may not be surprising because BF, ER, and PM two-center orbitals look very similar, as shown in Figure 1. Essentially, two operators

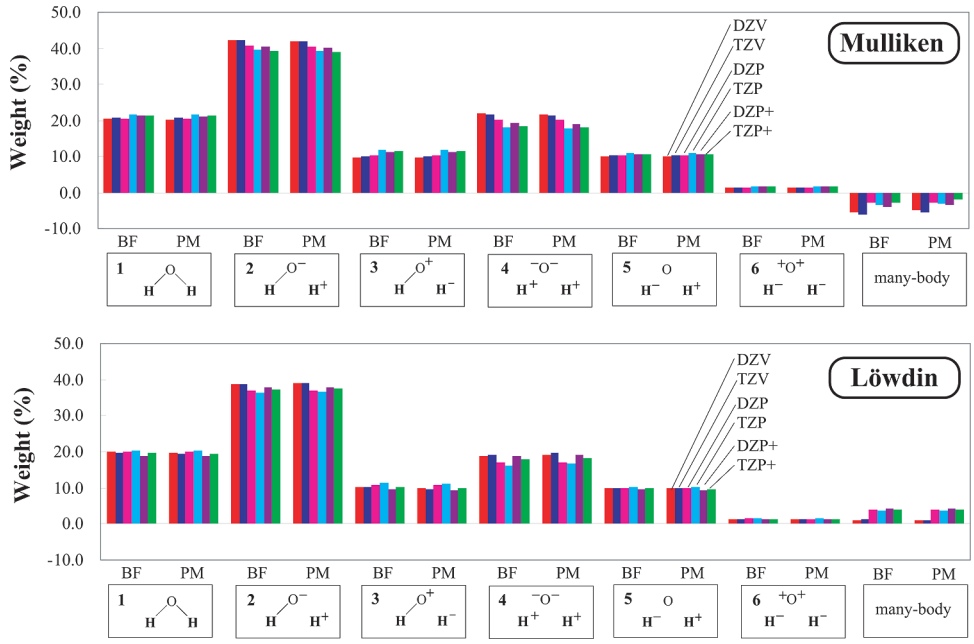


Figure 2. Weights (%) of resonance structures of H_2O evaluated with Mulliken and Löwdin operators.

deliver very similar results, but the weights of **2** and **4** evaluated with the Löwdin type were slightly smaller than those with the Mulliken type, which is consistent with a general trend that polarization is slightly enhanced in MP; for example, populations of the oxygen atom calculated with MP and LP (TZP basis sets) were, respectively, 8.615 and 8.397. Another important difference is found in the many-body term arising from the product of $\bar{W}^i$. The contribution is negligibly small, but the sign is different between two operators. This can be readily understood in terms of a well-known fact that MP analysis often gives negative population due to the nonorthogonality of AOs. However, it must be emphasized that these differences were very small, and two operators provided essentially the same results. Thus, the present procedure seems to be virtually independent from the choice of eq 9 or 10.

All the basis sets provide almost similar results, and the dependence is very small. The differences in the weights are less than 5%. It might sound paradoxical when remembering that the present analysis is related to MP (or LP) analysis, which is usually regarded to exhibit considerable basis-set dependence. Actually, the Mulliken charge of oxygen varies from −0.799 (DZ) to −0.615 (TZP). However, the paradox can be dispelled from the viewpoint of fraction F defined as the ratio of the assigned charge of oxygen to the total electron number in the system: the difference in electron number between the DZ and TZP results, 0.184, corresponds to less than 2% of the total number (0.184/10 × 100). The change in electronic structure often looks remarkable from the point of view of commonly used counting of electron numbers assigned to a specific atom, but it can look changeless when a different viewpoint is introduced. The weights of resonance structure, which characterizes the electronic structure of the whole molecule, is obviously related to the ratio, not to direct number-counting. In fact, the basis set dependence of the weights shows a good correlation with F (Figure 3). In other words,

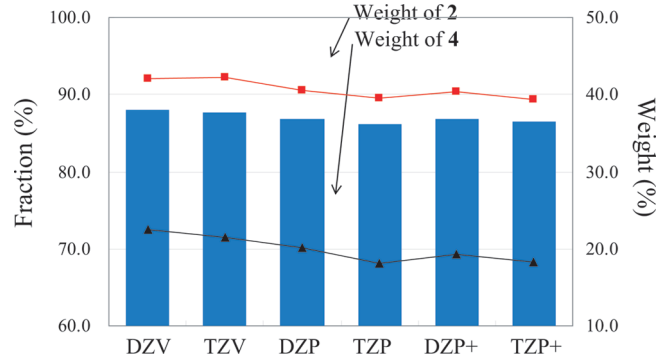


Figure 3. Basis set dependence of F, which is the fraction between number of electrons and Mulliken population of oxygen (bar). The weights of resonance structures **2** and **4** (lines), calculated by BF-localization.

Table 1. Weights of Covalent and Ionic Bond (%) in H_2O by Larger Basis Sets[a]

basis set	1	2	3	4	5	6
cc-pVDZ	24.27	32.97	17.85	11.20	12.13	3.28
cc-pVTZ	22.92	37.82	13.88	15.60	11.46	2.10
cc-pVQZ	21.98	39.64	12.18	17.88	10.98	1.69
aug-cc-pVDZ	25.07	32.69	19.23	10.65	12.53	3.69
aug-cc-pVTZ	22.83	34.86	14.95	13.31	11.41	2.45
aug-cc-pVQZ	21.01	37.48	11.78	16.71	10.51	1.65

[a] BF and MP operator was used. See Figure 2 for the index of the resonance structures.

the viewpoint of resonance structure could offer a robust way to understand the electronic structure of molecules. Table 1 lists the results from much larger basis sets. Again, the obtained wights are virtually independent of the basis set choice. Since the Mulliken population evaluated with these basis sets varies to a considerable degree, the resonance weights also show slight variation. Even so, all the standard deviations of each weight are less than 3.0%.

The next example is NH_3. By the localization, two one-center (the core and lone-pair of nitrogen) and three two-

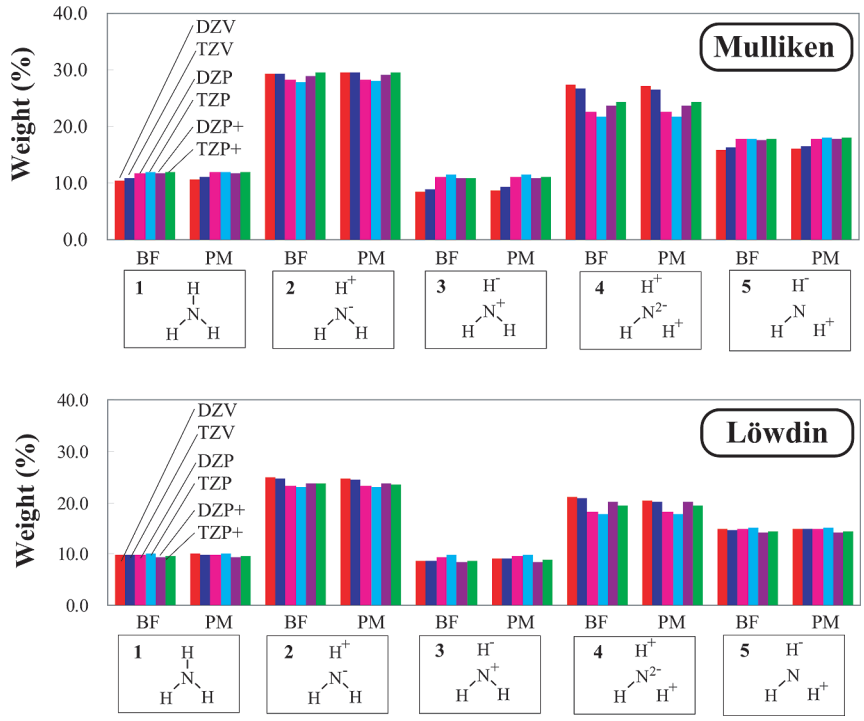


Figure 4. Weights (%) of important resonance structures of NH_3 evaluated with Mulliken and Löwdin operators.

center N−H orbitals were obtained, and wights were calculated from the latter ones. Three orbitals respectively obtained by BF, ER, and PM localizations look similar to the case of H_2O (not shown), and all the weights are also independent to the localization schemes, the operators, as well as the choice of basis sets. In Figure 4, some structures with high weights are selectively shown. The most important structure was **2**, which consisted of two covalent (N−H) and one ionic (N^-H^+) bonds. Next was **4**, which consisted of one covalent (N−H) and two ionic (N^-H^+) bonds. Both **2** and **4** have the character of a negatively charged nitrogen atom, and the weights calculated with Löwdin type were slightly smaller than those with Mulliken type. This feature is again related to the difference in these population analysis.

3.3. H_2CO. Next, the double bond C═O in H_2CO was the focus. In the above-mentioned molecules, bonding orbitals look very similar independent of the localization schemes. The situation is different in the case of a double bond (see Figure 5). BF and ER localization provide two equivalent $\sigma-\pi$ mixed orbitals, exhibiting so-called "banana bond" character (BF1, BF2 and ER1, ER2, respectively), while PM localization provides one σ (PM1) orbital and one π orbital (PM2). In other words, the generality is not obvious compared to the previous cases.

Figure 6 shows calculated BF- and PM-weight from the two orbitals in a similar manner. In all cases, the most important resonance structure is **2**, which comprises one ionic bond and one covalent bond, and the next is **1**, which is a doubly bonded resonance structure. These results accord with common knowledge of polarized character in a C═O bond. One of the most interesting findings is that the weight does not depend upon the choice of operator as well as upon the localization scheme again, even though orbitals BF1 and BF2 looks very defferent from PM1 and PM2. Total weights are

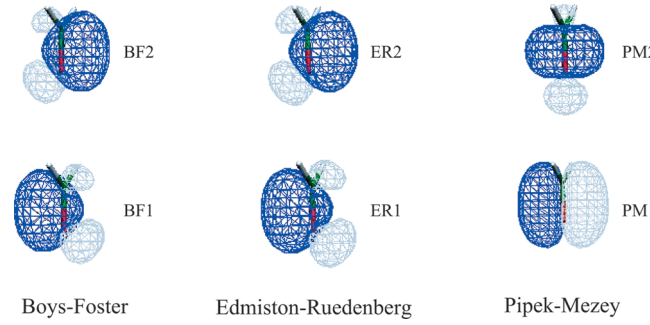


Figure 5. Two two-center BF-, ER-, and PM-orbitals of H_2CO, corresponding to the C═O bond. TZP basis sets were used.

virtually the same among all the combinations, and the difference is less than few percents.

Table 2 compares the localized orbitals in terms of respective weight components defined in eq 5, together with their population calculated by the MP operator. It is unsurprising that BF and ER give virtually the same weights, probably due to their similarity of orbitals. At the same time, PM1 and PM2 are slightly different from those of BF and ER, which are located in the middle of the two PMs. The weights are then calculated from these values, for example,

BF

1: C═O $0.4774 \times 0.4774 = 22.79\%$

2: C^+-O^- $0.4774 \times 0.3834 \times 2 = 36.61\%$

PM

1: C═O $0.4989 \times 0.4505 = 22.48\%$

2: C^+-O^- $0.4989 \times 0.4286 + 0.4505 \times 0.3313 = 36.31\%$

The obtained weights are a little different though they are derived in different ways. It should be noted that all

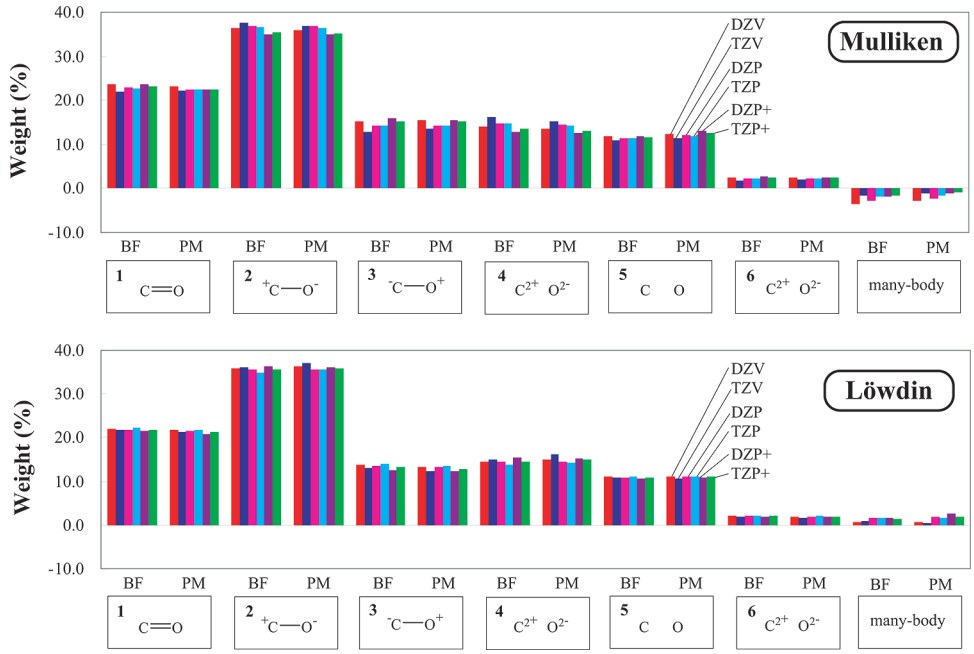


Figure 6. Weights (%) of resonance structures of H_2CO evaluated with Mulliken and Löwdin operators.

Table 2. Weights of Covalent and Ionic Bonds in H_2CO Calculated by All the Localization Schemes[a]

	weights/%				population			
	W_{CC}	$2W_{CO}$	W_{OO}		C	O	H	H
BF1	14.86	47.74	38.34		0.7710	1.2384	−0.0047	−0.0047
BF2	14.86	47.74	38.34		0.7710	1.2384	−0.0047	−0.0047
				subtotal	1.5420	2.4768	−0.0094	−0.0094
ER1	14.91	47.74	38.23		0.7722	1.2365	−0.0044	−0.0044
ER2	14.91	47.74	38.23		0.7722	1.2365	−0.0044	−0.0044
				subtotal	1.5444	2.4730	−0.0088	−0.0088
PM1	18.78	49.89	33.13		0.8668	1.1511	−0.0089	−0.0089
PM2	11.84	45.05	42.86		0.6881	1.3094	0.0013	0.0013
				subtotal	1.5549	2.4605	−0.0077	−0.0077

[a] Computed with the MP operator, and TZP basis sets were used.

the localized orbitals are linked through unitary transformation, but the numbers shown here (weight and its components) are not necessarily the same because the summation in eq 5 is limited over a specific atom (M and/or N) and the transformation is not completed.

Another viewpoint is the population of the localized orbitals. In the table, two orbitals were chosen for each localization scheme, and as can be seen, though the populations assigned to PM1 and PM2 are different from each other, the sum of them is very close to that of BF and of ER. It is noted that the Mulliken population is invariant against unitary transformations of the occupied orbitals, and the gross populations are 5.8310 (carbon), 8.3206 (oxygen), and 0.9242 (hydrogen), respectively. There are two σ-orbitals in carbon and hydrogen, and three one-center (core and lone pair) orbitals in oxygen. If each of them is ideally occupied by exactly two electrons, 1.8310, 2.3206, and −0.0758 are assigned to the population of these two orbitals, which are reasonably close to those of localized orbitals shown in the table. This may suggest that the valence space extracted by all the localization scheme are well separated from the core and lone-pair orbitals, and the obtained valence-space, which is the direct sum of the space spanned by the two orbitals, is very similar each other.

Unfortunately, the formal proof of this invariance seems to be impossible because the agreement is more like qualitative sense. All the numbers look essentially equivalent but are not exactly the same. According to our experience, the invariance about the localization scheme is always found in every case, even in a more complicated compound such as a triple-bond-containing molecule, and the obtained result always matches our chemical intuition for the examined molecules. Hence, the following two facts are worth pointing out. One is that the present analysis is related to MPA that is invariant against unitary transformations. The other point is the present measuring rule, namely, fraction. As mentioned above, an understanding in terms of the ratio seems to be robust enough, and even the basis set dependence of MPA becomes less prominent.

3.4. Some Other Molecules. Finally, two examples are shown. One is a substituent effect to H_2CO, namely XYCO. The same procedures were employed to evaluate the weight using PM localization to select one π and one σ orbital. As shown in Table 3, polarization of the C−O bond properly reproduced, and the contribution from C^+-O^- becomes

Table 3. Selected Weights of Covalent and Ionic Bonding (%) in Substituted H_2CO[a]

X	Y	C=O	$C^+—O^-$	$C^-—O^+$	$C^{2+}—O^{2-}$	C O	$C^{2-}—O^{2+}$
H	H	22.47	36.30	14.36	14.20	11.97	2.22
H	F	22.47	37.66	14.29	14.80	12.72	2.13
F	F	22.25	38.05	14.33	14.77	13.38	2.09
F	Me	22.29	38.24	14.04	15.18	12.94	2.05
H	Me	22.32	37.17	14.05	14.76	12.24	2.11
Me	Me	22.24	37.75	13.89	15.10	12.46	2.05

[a] PM localization and MP operator were used with TZP basis set.

greater by inductive effect compared to the original H_2CO. Interestingly, the contributions from a pure double bond hardly change by the substitution.

The next example is formamide. In relation to the understanding of the nature of an amide bond, its resonance structure was extensively studied by Mo et al.[21] Since the present analysis is built up from separated chemical bonds, treatment of conjugation is not simple. For the sake of simplicity, the following procedure was adopted in the present study. By BF localization, two π orbitals were obtained, and these four electrons are considered to be related to the resonance structure. But the lone-pair electron on the nitrogen atom is also included in this set. Hence, after computing the weights by multiplying the contributions from C−O and C−N bonds, as described above, this contribution (C^+ N^-) was subtracted to obtain the final resonance structure. Using the MP operator with 6-31G(d) basis set, the following weights were obtained: **1**, 28.0%; **2**, 25.6%; **3**, 15.9%; **5**, 7.7%; **6**, 2.5%. (The index was defined by Mo et al. in their work.[21] Since O···N direct interaction is not taken into account, the contribution from **4** does not explicitly appear in the present treatment.) Although this procedure is rather *ad hoc*, the qualitative trend derived shows good agreement with their report.

Nevertheless, the present method is good at analyzing bonds localized at a specific region. The description of conjugated electrons system is relatively poor, and further improvement is highly desired.

4. Conclusions

In the present work, our recently proposed method is generalized to evaluate the weight of resonance structures by taking the various combinations of the operator and the localization scheme. The method is applied to analyze the electronic structure of H_2, H_2O, NH_3, and H_2CO, and the basis set dependency of the method is also examined. Though the chosen operator, namely Mulliken-type or Löwdin-type, is kind of responsible for the weight, it can be concluded that the result is virtually independent from the combination as well as from the choice of basis sets. This suggests that understanding through resonance structure offers a robust and adequate description of molecular electronic structure. Furthermore, our method is very easy to use, and only standard localization procedure is required to obtain the resonance weights. From these results, the method could be a promising tool for analyzing molecular orbitals. On the other hand, it is difficult to apply the method when successful localization is not performed. For example, the electron in the transition state of a reaction inherently spreads over the reaction system, and the one- or two-center orbital picture is no longer valid. Extension to correlated wave function is also interesting. Further work along this line is currently in progress and will be reported elsewhere.

Acknowledgment. We acknowledge financial support by Grant-in-Aid for Scientific Research (C) (20550013) and by Grant-in-Aid for Scientific Research on Priority Areas "Molecular Theory" (461), both from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) Japan.

References

(1) (a) Karafiloglou, P. *J. Comput. Chem.* **2001**, *22*, 306. (b) Karafiloglou, P.; Papanikolaou, P. *Chem. Phys. Lett.* **2007**, *342*, 288. (c) Papanikolaou, P.; Karafiloglou, P. *J. Phys. Chem. A* **2008**, *112*, 8839.

(2) Shaik, S.; Shurki, A. *Angew. Chem., Int. Ed.* **1999**, *38*, 586.

(3) Hiberty, P. C.; Leforestier, C. *J. Am. Chem. Soc.* **1978**, *100*, 2012.

(4) Glendening, E. D.; Weinhold, F. *J. Comput. Chem.* **1998**, *19*, 593.

(5) Hirao, K.; Nakano, H.; Nakayama, K.; Dupuis, M. *J. Chem. Phys.* **1996**, *105*, 9227.

(6) (a) Mo, Y.; Gao, J. *J. Phys. Chem. A* **2000**, *104*, 3012. (b) Mo, Y.; Gao, J. *J. Comput. Chem.* **2000**, *21*, 1458. (c) Mo, Y.; Gao, J. *J. Phys. Chem. B* **2003**, *107*, 1664. (d) Song, L. C.; Gao, J. *J. Phys. Chem. A* **2008**, *112*, 12925.

(7) (a) Nakai, H. *Chem. Phys. Lett.* **2002**, *363*, 73. (b) Kawamura, Y.; Nakai, H. *J. Comput. Chem.* **2004**, *25*, 1882. (c) Yamauchi, Y.; Nakai, H. *J. Chem. Phys.* **2005**, *123*, 034101. (d) Nakai, H.; Kikuchi, Y. *J. Theor. Comput. Chem.* **2005**, *4*, 317. (e) Baba, T.; Takeuchi, M.; Nakai, H. *Chem. Phys. Lett.* **2006**, *424*, 193. (f) Imamura, Y.; Takahashi, A.; Nakai, H. *J. Chem. Phys.* **2007**, *126*, 034103. (g) Nakai, H.; Kurabayashi, Y.; Katouda, M.; Atsumi, T. *Chem. Phys. Lett.* **2007**, *438*, 132. (h) Kobayashi, M.; Imamura, Y.; Nakai, H. *J. Chem. Phys.* **2007**, *127*, 074103. (i) Imamura, Y.; Nakai, H. *J. Comput. Chem.* **2008**, *29*, 1555. (j) Imamura, Y.; Baba, T.; Nakai, H. *Int. J. Quantum Chem.* **2008**, *108*, 1316.

(8) (a) Ikeda, A.; Nakao, Y.; Sato, H.; Sakaki, S. *J. Phys. Chem. A* **2006**, *110*, 9028. (b) Ikeda, A.; Yokogawa, D.; Sato, H.; Sakaki, S. *Chem. Phys. Lett.* **2006**, *424*, 499. (c) Ikeda, A.; Yokogawa, D.; Sato, H.; Sakaki, S. *Int. J. Quantum Chem.* **2007**, *107*, 3132.

(9) (a) Ten-no, S.; Hirata, F.; Kato, S. *J. Chem. Phys.* **1994**, *100*, 7443. (b) Sato, H.; Hirata, F.; Kato, S. *J. Chem. Phys.* **1996**, *105*, 1546. (c) Yokogawa, D.; Sato, H.; Sakaki, S. *J. Chem. Phys.* **2007**, *126*, 244504.

(10) (a) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833. (b) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1841.

(11) (a) Mayer, I. *Chem. Phys. Lett.* **1983**, *97*, 270. (b) Mayer, I. *Int. J. Quantum Chem.* **1983**, *23*, 341.

(12) Boys, S. F. In *Quantum Theory of Atoms, Molecules, and the Solid State*; Lowdin, P. O., Eds.; Academic: New York, 1996; p 253.

(13) Löwdin, P. O. *J. Chem. Phys.* **1950**, *18*, 365.

(14) Edmiston, C.; Ruedenberg, K. *J. Chem. Phys.* **1965**, *43*, S97.

(15) Pipek, J.; Mezey, P. G. *J. Chem. Phys.* **1989**, *90*, 4916.

(16) If the population on are greater than a threshold.

(17) (a) Dunning, T. H. *J. Chem. Phys.* **1971**, *55*, 716. (b) Huzinaga, S. *J. Chem. Phys.* **1965**, *42*, 1293.

(18) Takahashi, O.; Sawahata, H.; Ogawa, Y.; Kikuchi, O. *THEOCHEM* **1997**, *393*, 141.

(19) Mayer, I. *Chem. Phys. Lett.* **2004**, *393*, 209.

(20) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.

(21) Mo, Y.; von Schleyer, P. R.; Wu, W.; Lin, M.; Zhang, Q.; Gao, J. *J. Phys. Chem. A* **2003**, *107*, 10011.

CT900053R

JCTC Journal of Chemical Theory and Computation

Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes

Jan Řezáč,[†,‡] Jindřich Fanfrlík,[†] Dennis Salahub,*[,‡] and Pavel Hobza*[,†,§]

Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Systems, 166 10 Prague 6, Czech Republic, Department of Chemistry, Institute for Biocomplexity and Informatics (IBI) and Institute for Sustainable Energy, Environment and Economy (ISEEE), University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4, and Department of Physical Chemistry, Palacky University, Olomouc, 771 46 Olomouc, Czech Republic

Received February 23, 2009

Abstract: Because of its construction and parametrization for more than 80 elements, the semiempirical quantum chemical PM6 method is superior to other similar methods. Despite its advantages, however, the PM6 method fails for the description of noncovalent interactions, specifically the dispersion energy and H-bonding. Upon inclusion of correction terms for dispersion and H-bonding, the performance of the method was found to be dramatically improved. The former correction included two parameters in the damping function that were parametrized to reproduce the benchmark interaction energies [CCSD(T)/complete basis set (CBS) limit] of the dispersion-bonded complexes from the S22 data set. The latter correction was parametrized on an extended set of H-bonded stabilization energies determined at the MP2/cc-pVTZ level. The resulting PM6-DH method was tested on the S22 data set, for which chemical accuracy (error < 1 kcal/mol) was achieved, and also on the JSCH2005 set, for which significant improvement over the original PM6 method was also obtained. Implementation of analytical gradients allows very efficient geometry optimization, which, for all complexes, provides better agreement with the benchmark data. Excellent results were also achieved for small peptides, and here again, chemical accuracy was obtained (i.e., the error with respect to CCSD(T)/CBS results was smaller than 1 kcal/mol). The performance of the technique was finally demonstrated on extended complexes, namely, the porphine dimer and various graphene models with DNA bases and base pairs, where the PM6-DH stabilization energies agree very well with available benchmark data obtained with DFT-D, SCS-MP2, and MP2.5 methods. The PM6-DH calculations are very efficient and can be routinely applied for systems of up to 1000 atoms. For nonaromatic systems, the use of a linear scaling version of the SCF procedure based on localized orbitals speeds up the method significantly and allows one to investigate systems with several thousand atoms. The method can thus replace force fields, which face basic problems for the description of quantum effects, in many applications.

Introduction

Noncovalent interactions are of fundamental importance for chemistry and molecular biology disciplines. Theoretical description of these interactions is difficult, mainly because they are much weaker than covalent interactions and also because of the key role played by the London dispersion energy. The proper description of these interactions thus requires recovering a large portion of the correlation energy and the use of an extended atomic orbital (AO) basis set. It is now evident that the coupled-cluster method considering the single and double electron excitations iteratively and the triple excitations perturbatively [CCSD(T)] using an extended

* Corresponding authors e-mail: pavel.hobza@uochb.cas.cz (P.H.), dennis.salahub@ucalgary.ca (D.S.).

† Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Systems.

‡ University of Calgary.

§ Palacky University.

10.1021/ct9000922 CCC: $40.75
Published on Web 05/26/2009

basis set, or even performed at the complete basis set (CBS) limit,[1,2] provides accurate energies and other characteristics for different types of noncovalent complexes.[3] The CCSD(T)/CBS method is a genuine ab initio method (i.e., no empirical or experimental characteristics are utilized), but its use for larger systems is limited because of its N^7 scaling with the size of system (N is the number of AOs). All other nonempirical wave function (WF) and density functional theory (DF) quantum mechanical (QM) theories applicable in this field use one or more empirical characteristics mostly parametrized to the benchmark CCSD(T) data. These theories provide static characteristics of noncovalent interactions, and despite better scaling with the size of system, their use is limited to complexes having a maximum of several dozens to several hundreds of atoms. For the description of systems with several thousands of atoms as well as for the understanding of their dynamics, much faster computational procedures should be applied, and molecular mechanics (MM) methods (also called empirical potentials) play an indispensable role here. These methods are efficient enough and provide surprisingly reliable characteristics for various types of noncovalent complexes. The serious drawback of these methods is the fact that they cannot describe quantum effects. The most important among these effects are breaking and formation of a covalent bond, changes in electronic structure of various conformers of complex molecular systems, cooperativity effects, charge transfer, and chemical reactions.

Two different approaches can be utilized to solve these problems. The first represents the use of quantum mechanics coupled with molecular mechanics (QM/MM). This approach, however, is still not efficient enough for extensive sampling of the configuration space of complex molecular systems, and further, it is not free of problems related to the cutting of the chemical bond at the QM/MM boundary. The application of semiempirical QM methods represents another approach. Semiempirical QM methods properly and fully describe all quantum effects mentioned above. Because they were parametrized for covalent bonding, however, their use for noncovalent complexes is not straightforward. Also, in contrast to the nonempirical Hartree−Fock method, which does not recover the correlation energy (and, consequently, also does not recover the London dispersion energy, which forms the dominant part of the intersystem correlation energy), all QM semiempirical methods do account for the dynamic correlation via the scaling of the two-electron integrals, but they do not [without configuration interaction (CI)] include a nondynamic correlation. This situation is identical to the DFT method, where the empirical dispersion correction is applied with great success.

Hydrogen bonding is another important issue in semiempirical QM methods. The original MNDO[4,5] (modified neglect of differential overlap) method was not able to describe H-bonding at all. This serious problem was addressed in later MNDO-based methods (AM1,[6] PM3,[7,8] and others) through the introduction of additional core−core interaction terms and parametrization of the method to reproduce hydrogen bonding. Such a treatment is only empirical, however, and does not solve the problem completely. AM1 was able to describe the interaction, but it yielded incorrect geometries, often featuring bifurcated hydrogen bonds. Another step forward was introduced in the parametrization of the PM3 method, where the geometry of H-bonded complexes was emphasized. Recently, a special parametrization of the PM3 method, PM3(BP), was introduced[9] for application of the method to nucleic acid base pairs. Density functional tight-binding,[10] a semiempirical QM method with parameters adjusted according to density functional considerations, also has problems with hydrogen bonding. These are discussed and addressed in ref 11. An overview of the problem is provided in ref 12. A possible future development toward improving H-bonds within the MNDO framework is outlined in this work, but no further progress has been reported up to now. In general, currently used semiempirical QM methods systematically underestimate the strength of H-bonds by approximately 20−30%. Surprisingly, the problems of semiemprical QM methods with the lack of dispersion energy, which were believed to be more serious, were removed rather easily. Already in 2001,[13] Elstner and one of us (P.H.) modified the semiempirical tight-binding DFT technique by the simple addition of an empirical London dispersion energy, which dramatically improved the performance of the method toward the dispersion-bound noncovalent complexes (e.g., stacked DNA base pairs). Martin and Clark[14] introduced an additional term to treat the dispersion in NDDO-based semiempirical quantum chemical techniques (where NDDO is neglect of diatomic differential overlap). The dispersion energy was calculated using additive "atomic orbital" polarizability tensors. A similar procedure was used later for the modification of AM1, PM3, and OM-x semiempirical methods,[15,16] and in all of these cases as well, much better performance toward dispersion-bound complexes resulted. Unfortunately, for various reasons (parametrization for only a limited number of atoms, strongly overestimated stabilization energies for optimized geometries of H-bonded complexes), these methods are still not accurate enough for most applications in complex molecular systems.

Recently, the new semiempirical method PM6 (parameterized model 6) was introduced,[17] which is superior to other semiempirical QM methods in various aspects. It is an NDDO-based method improved by the adoption of Viotyuk's core−core diatomic interaction term[18] and Thiel's d-orbital approximation.[19−21] These modifications allowed parametrization of 80 elements and also reduced the error for main-group elements.[17,22] The good performance of this method and its applicability to a wide range of problems are the reasons why we selected the PM6 method for further improvement in the direction of noncovalent interactions.

PM6 is available in the MOPAC code[23] from version 2007 and also in the VAMP 10.0 program.[24] The latest version of MOPAC, MOPAC 2009, introduces another interesting feature that makes the PM6 method usable for very large systems: a linear scaling version of the self-consistent field (SCF) procedure using localized orbitals, named MOZYME.[23]

Despite all these advantages, the PM6 method still lacks the ability to accurately describe noncovalent interactions,

specifically the dispersion energy and hydrogen bonding. Even though the method yields surprisingly good geometries of all types of complexes, the interaction energy for dispersion-bound and H-bonded complexes is substantially underestimated. We believe that common NDDO parametrizations do not reproduce hydrogen bonds well. General suggestions as to how to improve semiempirical methods toward reliable description of H-bonds are provided in ref 32. Jug and Geudtner[25] also reported a variant of the SINDO (symmetrically orthogonalized intermediate neglect of differential overlap) method using a p-polarization function on hydrogen to improve the description of hydrogen bonds. However, there is no readily available and applicable method that would give acceptable results for noncovalent interactions. Let us recall once again that, among all widely used ab initio QM procedures (i.e., methods that do not use any empirical or experimental parameters), it is only the CCSD(T)/CBS technique that satisfactorily describes both of these interactions.

In the present study, we introduce an extension of the PM6 method in two directions: (i) including an empirical dispersion energy term that improves the description of complexes controlled by the dispersion energy and (ii) introducing an additional electrostatic term that improves the description of hydrogen-bonded complexes. The resulting method, PM6 with corrections for dispersion and hydrogen bonding, is named PM6-DH. The aim is ambitious: for extended noncovalent complexes to achieve standard ab initio chemical accuracy ($\sim$ 1 kcal/mol). Because of favorable scaling of the code, we would also like to use it in MD simulations. Therefore, in addition to single-point calculations, we also carefully tested the performance of the method when the full gradient optimization was adopted. It should be remembered that just this point was critical for the application of other semiempirical QM methods.

This article also presents benchmarks of the method and a comparison to other methods with a comparable range of applications. The first tests concerned the stacking interactions. Specifically, we investigated the interaction of two porphine molecules, as well as the interaction of various graphene models with nucleic acid bases and base pairs. In the next step, we compared a more complex system that contained both characteristic interaction types, stacking as well as H-bonding. We studied the interaction of DNA with 4′,6-diamidino-2-phenylindole (DAPI), a fluorescent dye that can bind both in the minor groove of the double helix and as an intercalator. Interaction energies obtained with PM6-DH were compared with benchmark calculations in these examples. Finally, as a last example, a DNA tetramer was optimized using the PM6 and PM6-DH methods. The structure of the DNA fragment was determined by both stacking and H-bonding, and this test example was selected to demonstrate the performance of PM6-DH toward the important world of nucleic acids. Another reason supporting this point is the recent finding that the description of the DNA double-helical structure requires inclusion of the dispersion energy.[26]

Methods

Dispersion Correction. The first step in improving the PM6 method was the addition of an empirical dispersion term. This was not a difficult task, because the London dispersion term is well separated from the QM calculation and thus transferable between various methods. Using our experience with empirical dispersion in the Hartree−Fock and density functional theory (DFT) methods,[27−30] including semiempirical tight-binding DFT,[13] we have adopted the formalism described in the work of Jurecka et al.[30]

The correction has the form of a pairwise interatomic force field

$$\sum E_{\mathrm{dis}} = -\sum f_{\mathrm{damp}}(r_{ij}, R_{ij}^{0}) C_{6ij} r_{ij}^{-6} \qquad (1)$$

It consists of the physically sound r^{-6} term, damped at short distances to avoid interfering with the underlying QM potential, that describes the short-range repulsion correctly. Atomic parameters, namely, atomic van der Waals radii (R_{ij}^{0}) and C_6 coefficients, are independent of the QM method and were adopted from the original work. Two parameters in the damping function, the scaling factor for the radii s_r and the exponent α that affects the slope of the damping, were optimized to reproduce interaction energies of dispersion-bonded complexes. A subset of the S22 benchmark data set[31] with hydrogen-bonded complexes removed was used in the fit. Omitting hydrogen-bonded complexes from the training set should yield a better description of the dispersion-bonded complexes; the errors in hydrogen bonds are corrected separately.

The PM6 method with the dispersion correction only is abbreviated as PM6-D in this article.

Hydrogen-Bond Correction. Improving hydrogen-bonding interactions is not as straightforward as correcting dispersion interactions because the electrostatic term, mainly responsible for the description of H-bonding, was included in the original parametrization. Previous attempts to improve the hydrogen-bond description were done at the level of modifying the semiempirical method itself or reparameterizing it.[11,12,32] Because these attempts have not solved the problem, we decided to take another approach. Inspired by the success of the dispersion correction, we aimed to introduce a specific correction that would affect only hydrogen bonds, added on top of an unmodified semiempirical calculation.

Training Set of 104 Hydrogen-Bonded Complexes. The first step in this work was the preparation of an extensive training set of model hydrogen-bonded complexes. We designed the set to cover different types of hydrogen bonds present in biomolecules and organic compounds. Therefore, the training set was composed of 104 hydrogen-bonded complexes formed from the proton donors acetylamine, acetic acid, dimethylamine, phenol, methanol, methylamine, phenylamine, peptide bond, pyrrol, uracil, water molecule, and 1-imino-3-aminocyclohexane, along with the proton acceptors acetic acid, dimethyl amine, phenylamine, phenylm-

ethylamine, furan, methanol, methylamine, propanol, peptide bond, pyrazine, uracil, water molecule, and 1-imino-3-aminocyclohexane. The reference geometries were obtained by RI-MP2/cc-pVTZ gradient optimization, which is known[33] to provide reliable results for various types of noncovalent complexes, including the presently investigated H-bonded ones.

The dissociation curve (scan of the interaction energy as a function of the hydrogen-bond distance) of each of these complexes was calculated using the accurate SCS(MI)-MP2/cc-pVTZ method,[34,35] with the counterpoise correction to eliminate basis set superposition error. This method is parametrized to yield highly accurate interaction energies, while being efficient enough for this task. It should be recalled that the SCS(MI)-MP2 method provides reliable interaction energies not only for stacked complexes (similarly to the original SCS-MP2 technique) but also for H-bonded ones where the SCS-MP2 technique failed. With 14 points per dissociation curve, about 1500 interaction energies had to be calculated to prepare the training set. The calculations were performed using Turbomole 5.9.[36]

The same curves were then calculated using PM6 with the dispersion correction. With this information, we then analyzed the distance-dependent behavior of the error to design the form of the correction.

Selection of Possible H-Bonds. Prior to the calculations, all possible hydrogen bonds were determined from the topology. All possible combinations of hydrogens bonded to an electronegative atom (elements N and O) and electronegative hydrogen acceptors (N, O) were listed. Then, all pairs in the 1−4 configuration (i.e., in the H−X−Y−acceptor pattern) were removed from the list, because they cannot form hydrogen bonds.

The correction was calculated for all of these possible pairs, with the form of the correction ensuring that only the actual hydrogen bonds contributed significantly to the total energy.

Atom Types. An important finding was that there is a significant difference between hydrogen groups with nitrogen as the acceptor and those with oxygen as the acceptor. Further sorting of the error curves showed that there are differences in hydrogen bonds between the elements and that these differences can be correlated with the valence state and environment of the atoms. Later, optimization of different models confirmed that the introduction of atom types is necessary to describe all types of hydrogen bonds accurately.

Rather than using atom types directly, we selected only their combinations that show different behavior and introduced "hydrogen-bond types" to reduce the number of parameters in the model. This step, of course, makes the correction rather empirical and is a source of several limitations, but it is, in our opinion, worth the improved accuracy.

The hydrogen-bond types are as follows:

(1) nitrogen with no hydrogens bonded to it (mostly in aromatic rings) interacting with any hydrogen,

(2) nitrogen with one hydrogen (secondary amines) interacting with any hydrogen,

(3) nitrogen with two or more hydrogens (primary amines, ammonia) interacting with any hydrogen,

(4) oxygen except carbonyl interacting with HN,

(5) carbonyl oxygen interacting with HN,

(6) oxygen interacting with HO hydrogen different from 7 and 8,

(7) oxygen interacting with H in a water molecule, and

(8) oxygen interacting with H in a carboxyl group.

The first limitation that arises from the introduction of hydrogen-bond types is that the method can be applied only to hydrogen bonds for which it was explicitly parametrized. In this work, we attempted to cover all hydrogen bonds in organic compounds involving nitrogen and oxygen. The extension of this set would be straightforward, as there is no reason to believe that the same correction cannot be applied to other elements. We plan to further improve the method, and extending the set of parameters is one of the goals toward this end. Specifically, the addition of parameters for sulfur is necessary for full coverage of the interactions in proteins.

The second limitation directly connected to the hydrogen-bond types in the current implementation is that the bond types are determined only once (at the beginning of the calculation), even for calculations where the geometry can change (i.e., in optimization). This prevents the method from being used to study processes where the valence states of the atoms involved in hydrogen bonds change, because the valence state is what determines the bond type used. The most obvious example of such a reaction is a proton transfer. A simple re-evaluation of the bond types in each step of the calculation would not correct this problem, because the potential energy surface would become discontinuous. On the other hand, this limitation merely prevents the method from being used to study the intermediates of such a reaction. When the reactants and products are studied separately, the different bond types can be readily assigned.

Form of the Correction. Most of the error curves (constructed as energy differences between the benchmark and PM6-D) are similar to the function $-1/r$, which is in agreement with the Coulomb nature of hydrogen bonds. It is obvious that the form of the correction should be similar to the formula for an electrostatic interaction. In some cases, especially in hydrogen bonds of the XH···O type, there is a repulsion at shorter distances, and the correction sometimes becomes positive. An exponential term was added to describe this effect.

Hydrogen bonds differ in their strengths, and so does the correction. Models that do not take this fact into account are not very accurate. To cover these differences, we used partial atomic charges from the PM6 calculations (obtained using Mulliken population analysis) on the hydrogen and on the acceptor. The product of these two charges correlates reasonably well. We tested many different formulas, but the one based on the interaction energy of a point charge and a dipole worked best. Unlike charge−charge interactions, which have an r^{-1} dependence, the distance dependence in the present formula is r^{-2}, and there is an angular part calculated from the angle of the three atoms involved in the hydrogen bond. This angular part is crucial, because the

core−core interaction introduced in semiempirical methods (including PM6) to improve hydrogen bonding is not directional, whereas in reality, H-bonds are very sensitive to the spatial arrangement. Combined together, the final formula for the correction energy (E_{HB}) term in hydrogen bond XH···Y is

$$E_{HB} = c[q(H)\, q(Y)/r^2 \cos(\theta) + c_{rep}A^{-r}] \qquad (2)$$

where r is the H···Y distance and θ denotes the angle XHY. c, c_{rep}, and A are parameters fitted to obtain the best results over the training set.

It must be emphasized that many different models of the correction term were tested, but the one presented here was found to yield the best results.

To prevent problems arising from calculation of pairs in the Y···XH geometry (for example, within a single molecule with more electronegative atoms), an additional rule was added, and pairs with $\theta < 90°$ were not calculated, because they certainly are not hydrogen-bonded.

Optimization Procedure. Different types of hydrogen bonds cannot be parametrized separately, because the contribution of each possible hydrogen bond is calculated in the complex. To gain more control over the fitting procedure, we used gradient optimization of an arbitrary error function, based on numerical gradients of the fitted coefficients. The Broyden−Fletcher−Goldfarb−Shanno (BFGS) algorithm was used to make the optimization efficient. This setup, in contrast to the least-squares method normally used in similar applications, allowed us to experiment with different goals of the optimization. We used two measures of the quality of the fit: the mean absolute error, e_{avg}, and the maximum absolute error, e_{max}. The error function that was minimized by the optimization procedure was a combination of these error measures in variable ratio. We did multiple optimizations with different ratios of e_{avg} and e_{max} to find the best compromise in usability of the fitted method.

Implementation. The basis for our work was the PM6 method implemented in the MOPAC2007 package.[23] The corrections were calculated on top of the finished PM6 calculation, which makes them independent of the PM6 code itself. For the purpose of development of the method, we developed our own code that calls MOPAC for the PM6 calculation and adds the corrections to the result. The latest version of this software has been made available to the public and can be obtained from the authors upon request (http://www.molecular.cz/~rezac/pm6dh.html).

To make the method more useful, we calculated the gradients of both corrections. For the dispersion correction, this was a straightforward task, because the correction was completely separated from the underlying PM6 calculation. The situation was more complicated in the case of the H-bonding correction, which used atomic charges from the PM6 calculation. Determination of the true analytical gradient would require calculation of the derivatives of these charges with respect to nuclear coordinates, which would make the calculation more expensive. In addition, such a calculation is not possible in the current implementation. We made use of the fact that the charges varied only slightly and considered them to be constant to calculate an estimate of the gradient. To justify this approach, we compared this estimate to a full numerical evaluation of the gradients. We found that the differences in the gradient itself and, more importantly, in the optimized energy and geometry were very small and below the convergence limits used in the optimization.

The BFGS optimizer working in Cartesian coordinates is part of the final code.

Validation Sets. The PM6-DH method was tested on multiple sets of benchmark data available from the BEGDB database.[37] All of these data sets feature high-quality geometries and energies extrapolated to the CCSD(T)/CBS level.

First, the results were compared to the S22 data set,[31] which includes model complexes covering hydrogen-bonding and dispersion interactions. In this case, both energies and optimized geometries were utilized. The fact that geometries were tested along with the energies is important because all previous semiempirical QM procedures were tested only for stabilization energies and, on the basis of our experience, their application to geometry optimization is not straightforward. Second, the JSCH2005 database[31] of DNA base pairs in various geometries and some amino acid pairs was used as an example of biologically relevant complexes. Finally, we tested the PM6-DH method in an application different from the calculation of interaction energies (i.e., molecular clusters). The last test was therefore the calculation of the relative stability of conformers of small peptides.[38] In the original work featuring this data set, it was shown that this is a very sensitive measure of the quality of computational methods, and a comparison with benchmark energies is available.

Results

Parameterization of Dispersion Corrections. There were only two parameters to be adjusted in the damping function in the empirical dispersion term. Calculation of the correction was very fast, so we performed a complete 2D scan instead of optimization to obtain the values corresponding to minimal error, expressed as the average absolute value of the difference between PM6-D and the benchmark energy over the training set. This procedure led to $s_r = 1.07$ and $\alpha = 11$ with $e_{avg} = 0.4$ kcal/mol for stacked structures within the S22 benchmark data set.[31]

Parameterization of H-Bond Correction. Optimization of the parameters in the H-bond correction was done in two steps to increase the efficiency of the process. In the first round, only optimal geometries of the complexes in the training set were used in the optimization. The resulting set of coefficients was then used as a starting point for a second round, in which all geometries along the dissociation curve were taken into account.

We tested error functions (the optimized value) with various ratios of the mean and maximum errors, observing the impact on the results. The best error function, which was used to derive the presented parameters, consisted of 90% of the mean error and 10% of the maximum error. Increasing the fraction of the mean error brought no significant

Table 1. Fitted Coefficients in the H-Bond Correction for Hydrogen-Bond Types Discussed in the Text[a]

bond type no.	c	c_{rep}	A
1	14.4209	-1.3273×10^{-2}	7.2847
2	73.3566	-5.3979×10^{-4}	7.0920
3	48.7161	2.9844×10^{-4}	6.4259
4	29.8036	2.1262×10^{-3}	6.9768
5	−6.4578	7.3142×10^{-3}	7.8379
6	23.1582	-4.8015×10^{-5}	6.9382
7	15.3029	2.0789×10^{-3}	7.0365
8	14.8668	-4.6652×10^{-3}	6.9111

[a] Coefficients derived using the following units in eq 2: elementary charge, Å, and kcal/mol.

improvement in e_{avg}, but resulted in an increase in e_{max}. For the opposite change, e_{max} was improved only marginally, whereas e_{avg} became too large. A mean error of 0.8 kcal/mol was achieved for the training set.

Parameters obtained from the optimization are listed in Table 1. Note that the coefficient scaling the repulsion term is negative in some cases. The exponential term is thus no longer repulsive, but it further enhances the attraction, improving the potential of the electrostatic correction where needed. We tried to constrain these coefficients to be nonnegative, but the final results were worse. This issue reflects the major difference between oxygen and nitrogen hydrogen bonds in the PM6 method. (Data demonstrating their different natures are provided in the Supporting Information.) We understand that changing the sign of a term that was originally designed to be repulsive makes our correction further decline from a physically sound formulation, but our goal was to develop an empirical correction, focusing mainly on the final accuracy.

The validation sets yielded a slight increase in the error when the H-bond correction was added to dispersion-bonded complexes that contained heteroatoms that could be involved in hydrogen bonds. We tried to address this problem by a more localized variant of the correction, with a cutoff that would eliminate all contributions that were not real hydrogen bonds. However, the overall results were worse.

Geometry Optimization. After the correction energy term was established, it was tested in geometry optimization to verify its performance. The method was found to yield good geometries, as discussed below, with one exception. In the H-bonded adenine−thymine base pair, we observed a proton transfer along one of the hydrogen bonds, as the gradient of the H-bond correction overpowered the reaction barrier. This happened because the possible hydrogen bonds were determined from the geometry of the initial state, so there was a force driving the hydrogen in one direction but not in the other. The gradient was very steep in this region; here, it became an incorrect extrapolation from the potential at larger distances.

To solve this problem, we had to modify our potential. It had the original form as long as the YH distance (in the Y···HX arrangement) was above a certain limit and remained constant at shorter distances. We set the limit at 1.8 Å, the YH distance in the shortest H-bonds in our training set. This can be easily achieved by using the following rule to determine the value of r in eq 2

$$r = \begin{cases} r & \text{for } r > 1.8 \\ 1.8 & \text{otherwise} \end{cases} \quad (3)$$

The resulting potential has a constant energy and a zero gradient at very short distances, which prevents the anomalous behavior observed before. We tested this modification extensively and found it to be safe. It does not affect calculations in a fixed geometry; it only eliminates the rare problem of geometry optimization described above. This modification was used for all calculations presented here and was implemented in the PM6-DH code.

Tests on Benchmark Data. *S22 Data Set.* The S22 data set is our first choice for evaluating the accuracy of a method in describing noncovalent interactions. It contains complexes featuring hydrogen bonds, dispersion interactions, and their combinations. Biomolecules are represented here by base pairs in both stacked and Watson−Crick (WC) arrangements.

Our first test was an analysis of interaction energies calculated with the PM6, PM6-D, and PM6-DH techniques on the S22 data set. The results are summarized in Table 2, and their differences from the benchmark CCSD(T)/CBS data are plotted in Figure 1. For comparison, we include interaction energies calculated at the much more expensive MP2/cc-pVTZ level (corrected for basis set superposition error).

The performance of the unmodified PM6 is rather poor. The mean absolute error (e_{avg}) is 3.4 kcal/mol, and the maximum absolute error (e_{max}) is 7.5 kcal/mol, occurring in the formic acid dimer. PM6-D brings a great improvement in complexes dominated by dispersion, which reduces e_{avg} to 1.4 kcal/mol, but the maximum error remains high, 6.5 kcal/mol. Finally, by combining the dispersion correction with the correction for hydrogen bonds in PM6-DH, both sources of error are addressed, and the results improve significantly. The mean absolute error is 0.6 kcal/mol, and e_{max} is reduced to 1.8 kcal/mol. These results are even better than in the MP2/cc-pVTZ calculation, for which the mean error is 0.7 kcal/mol and the maximum error is 1.9 kcal/mol.

The errors worsen slightly for the calculation of interaction energies on structures optimized with PM6-DH itself, with $e_{avg} = 0.8$ kcal/mol and $e_{max} = 2.8$ kcal/mol. A possible source of this error is discussed later in this article. However, this behavior is specific only to the strongest hydrogen bonds, and the optimization generally improves the results. Importantly, geometry optimizations performed with other semiempirical QM techniques led to considerably worse results.

It is clear from the results presented that the correction for hydrogen bonding is less perfect than that for dispersion. The complexes with the largest errors are hydrogen-bonded, and the errors can be both positive and negative. The hydrogen-bond correction also introduces some small error into some dispersion-bonded complexes. These are the limitations of the simple form of our correction and its empirical nature. Nevertheless, these errors are very small compared to those obtained with the uncorrected PM6 method, and the overall accuracy is on a par with that of expensive correlated ab initio calculations.

Finally, we tested PM6 and its modifications for the optimization of the S22 complexes. In general, the results

Table 2. Interaction Energy Errors (in kcal/mol) for the S22 Set of Complexes, Calculated as Differences between the Studied Method and the Benchmark CCSD(T)/CBS Results, in kcal/mol[a]

	MP2/cc-pVTZ	PM6	PM6-D	PM6-DH	PM6-DH optimized[b]
ammonia dimer	0.40	0.86	0.33	−0.57	−0.75
water dimer	0.59	1.08	0.70	0.35	0.29
formic acid dimer	1.72	7.47	6.47	1.22	−0.50
formamide dimer	1.71	3.41	2.19	0.57	0.95
uracil dimer (H-bonded)	1.91	7.33	5.55	1.81	1.10
2-pyridoxine···2-aminopyridine	0.80	6.73	4.51	−0.64	−1.79
adenine···thymine (WC)	1.45	7.31	4.90	−1.46	−2.75
methane dimer	0.21	0.47	−0.20	−0.20	−0.20
ethene dimer	0.36	1.11	−0.01	−0.01	−0.02
benzene···methane	0.09	1.03	−0.25	−0.25	−0.38
benzene dimer (stacked)	−1.03	2.86	−0.89	−0.89	−0.86
pyrazine dimer	−1.02	2.61	−0.99	−0.99	−1.32
uracil dimer (stacked)	1.03	5.66	0.53	0.42	0.09
indole···benzene (stacked)	−1.22	5.29	0.02	0.02	−0.77
adenine···thymine (stacked)	−0.07	7.29	−0.04	−0.55	−1.38
ethene···ethine	0.10	0.98	0.42	0.42	0.36
benzene···water	0.35	1.00	−0.13	−0.13	−0.67
benzene···ammonia	0.22	0.82	−0.42	−0.42	−1.47
benzene···HCN	−0.14	2.48	1.26	1.26	1.25
benzene dimer (T-shaped)	−0.27	1.99	−0.10	−0.10	−0.11
indole···benzene (T-shaped)	−0.43	3.33	0.43	0.43	0.51
phenol dimer	0.34	3.67	1.33	0.32	−0.41
e_{avg}[c]	0.7	3.4	1.4	0.6	0.8
e_{max}[c]	1.9	7.5	6.5	1.8	2.8

[a] High-quality geometries from the S22 database used unless otherwise noted. [b] Geometries optimized using the PM6-DH method. [c] Mean and maximum absolute errors used as a measure of the quality of the method.

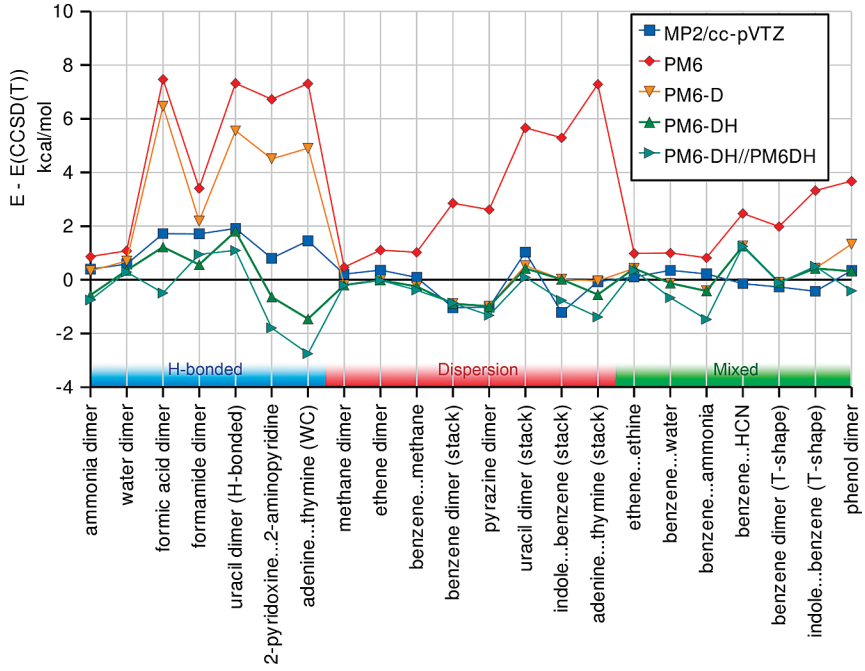


Figure 1. Interaction energy errors in the S22 set of complexes, plotted as the difference between the studied method and benchmark CCSD(T)/CBS results.

were very good even for PM6, and they did not improve when the corrections for dispersion and H-bonding were introduced. The root-mean-square deviation (rmsd) upon optimization, starting from the benchmark geometry, was 0.14 Å for PM6, 0.13 Å for PM6-D, and 0.15 Å for PM6-DH. The hydrogen-bond correction actually made the geometries slightly worse. This effect was more pronounced in complexes featuring very strong H-bonds, such as the formic acid dimer. The problem can be attributed to the form of the correction and the description of hydrogen bonds in the semiempirical method itself. In both cases, the potential responsible for hydrogen bonding acts between the nuclei, but in reality, the hydrogen bond forms between the lone pair on the proton acceptor and an antibonding orbital of the X—H bond of the proton donor. An important feature of H-bonding, namely, its directionality, is due to this point.

Table 3. Analysis of Uncorrected and Corrected PM6 Results for the JSCH2005 Database Featuring Biomolecular Complexes[a]

complex	PM6	PM6-D	PM6-DH
	e_{avg} (kcal/mol)		
all neutral	4.5	2.2	1.6
DNA bases	4.7	2.3	1.6
DNA bases, H-bonds	7.7	5.3	1.9
DNA bases, stacked	3.6	1.2	1.5
amino acids, neutral	3.1	0.9	1.0
amino acids, charged	20.5	17.7	17.5
	e_{avg} (%)		
all neutral	52	25	18
DNA bases	51	25	17
DNA bases, H-bonds	42	29	10
DNA bases, stacked	61	20	26
amino acids, neutral	67	20	22
amino acids, charged	24	21	20
	e_{max} (kcal/mol)		
all neutral	12.1	7.9	6.1
DNA bases	12.1	7.9	6.1
DNA bases, H-bonds	10.7	7.9	4.3
DNA bases, stacked	12.1	6.5	6.1
amino acids, neutral	6.4	3.2	3.3
amino acids, charged	30.1	26.0	25.7

[a] Errors compare PM6-based calculations to benchmark CCSD(T)/CBS results.

Our simple model does not cover this difference while making the interaction stronger, which leads to slightly distorted geometries. This fact should be considered when PM6-DH is used, but the benefits of more accurate interaction energies are much more important. Proper energetics for noncovalent interactions would also improve geometries in more complex structures where intramolecular and/or intermolecular interactions and their balance with other forces are crucial.

JSCH2005. A further step in the evaluation of PM6-DH was to extend the validation set. We used the JSCH2005 data set[31] for this purpose. It consists of DNA base pairs in various geometries and some complexes of amino acids in arrangements found in proteins. In this step, only the interaction and/or relative energies were considered (i.e., no geometry optimization was performed). These complexes were divided into several groups for the analysis. First, we examined separately all complexes of neutral molecules and all charged complexes, because the magnitudes of the interactions are different. In addition, we evaluated the error measures in the groups of hydrogen-bonded base pairs, stacked base pairs, and amino acid complexes. The results, again in terms of average and maximum absolute difference from CCSD(T)/CBS value, are presented in Table 3.

Several important conclusions can be drawn from these results. The first is the progressive improvement of the results when the corrections are added to PM6 calculations, with the PM6-DH method reaching an average error of only 1.6 kcal/mol for the neutral complexes. The error for the charged complexes, the amino acid ion pairs, is much larger, but this is because the interaction energy itself is very large. When the errors are made relative to the magnitude of the interaction, the percentage errors become very similar.

Table 4. Mean and Maximum Absolute Errors (in kcal/mol) of PM6-Based Methods Compared to CCSD(T)/CBS Conformer Energies in a Set of 76 Peptides

	method				
	MP6	PM6-D	PM6-DH	PM6-DH optimized	MP2/cc-pVTZ
e_{avg} (kcal/mol)	1.66	1.76	1.04	0.89	0.92
e_{max} (kcal/mol)	4.69	4.36	5.46	3.31	2.64

Second, these data show that the corrections improve interaction energies in complexes of charged molecules, even though they were parametrized only for neutral ones.

Finally, this large validation set shows that the descriptions of different types of interactions are similar; the method can thus handle both hydrogen bonds and dispersion in a balanced way.

It should be noted that these complexes are the most difficult ones to describe with our method. The hydrogen-bonded base pairs feature strong, cyclic hydrogen bonds for which additional cooperativity of the bonds is important, whereas the correction was parametrized on model complexes featuring single hydrogen bonds. The same applies for stacked bases here, because the molecules feature multiple sites that could be involved in hydrogen bonds. The dispersion energy in nonpolar molecules is easier to describe, and the error can be much smaller, as demonstrated for the S22 set.

Peptides. In our previous work,[38] we thoroughly investigated the performance of a wide range of computational methods for the calculation of conformational energies of small peptides. This turned out to be a very sensitive test case, and not many methods were found to yield satisfactory accuracy. Standard force field methods had problems[38] with atomic charges because these charges depend on the peptide conformation. The stability of the conformers is often determined by intramolecular noncovalent interactions, mainly hydrogen bonds, when present, and the dispersion energy, in peptides containing aromatic amino acid residues. The high sensitivity of the description of noncovalent interactions and the availability of benchmark energies and geometries make this data set very important for the evaluation of the PM6-DH method. The application of semiempirical QM methods

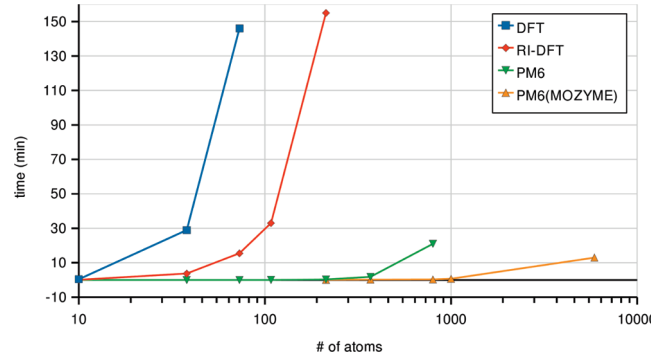


Figure 2. Comparison of DFT, RI-DFT, PM6, and PM6-MOZYME computational resources required for one SCF cycle on a single protein molecule. Note that the horizontal axis is logarithmic to accommodate the studied range of molecule sizes.

Table 5. Stabilization Energies (in kcal/mol) of Stacked (S), Parallel-Stacked (PS), and T-Shaped (T) Structures of the Porphine Dimer Determined by the PM6-DH, DFT-D/TPSS/TZVP, DFT-D/B97/TZV(2df,2pd), SCS-MP2/ TZV(2df,2pd), and MP2.5/CBS[a] Techniques

structure	DFT-D/TPSS/TZVP	DFT-D/B97/TZV	SCS-MP2	PM6	PM6-DH	MP2.5
A(S)	10.6	13.5	17.7	−0.8	16.9	
B(S)	11.2	14.5	19.0	0.4	18.0	17.6
C(S)	16.0	18.3	24.6	0.2	21.1	
D(PS)	16.6	20.0	25.6	0.2	22.0	
E(PS)	16.9	20.5	25.8	1.5	23.3	23.0
F(PS)	17.1	20.6	26.0	0.9	23.2	
G(T)	7.2	8.4	7.3	3.2	8.9	
H(T)	7.3	9.7	9.1	2.3	9.4	

[a] $E^{(2)}$ correction from CBS and scaled $E^{(3)}$ correction from TZV.

in this field is important because empirical potentials, which are usually applied, have failed to describe the variability of atomic charges for different conformers.

On the set of 76 structures of FGG, GFA, GGF, WG and WGG peptides, the relative energies of the conformers (taking the average energy for the same peptide as the zero level) were calculated using PM6, PM6-D, and PM6-DH with the original geometries and also using geometries optimized with PM6-DH. Mean and maximum absolute errors [compared to CCSD(T) relative energies] are listed in Table 4. MP2/cc-pVTZ results from the original work are included for reference.

Achieving an average error of 1.04 kcal/mol with PM6-DH is a very encouraging result. Even more importantly, the errors were reduced by geometry optimization. In the optimized structures, we achieved the limit of chemical accuracy, 1 kcal/mol. Note that these results are comparable to those of the considerably more expensive MP2/cc-pVTZ calculations.

Timing. A comparison of the timing of DFT, the resolution of the identity (RI) approximation to DFT, PM6, and PM6-MOZYME was done on model peptides and proteins. In-silico-built poly(glycine) helices covered the smaller testing systems; experimental protein geometries (PDB numbers 2BEG and 1AFW) served as the largest testing systems. All calculations were run on the same computer featuring a single 2.4 GHz Intel processor. The results are shown in Figure 2. The Meta-GGA functional TPSS and the TZVP basis set were used for both DFT and RI-DFT calculations. Obviously, the current limit of the DFT method is placed at around 100 atoms. The RI approximation[39] to the DFT method allows for the treatment of several hundreds atoms. The NDDO approximation within the PM6 method reduces the computational cost and places the limit of semiempirical methods at around 1000 atoms. The use of the localized molecular orbital method (MOZYME) speeds up the PM6 method significantly and allows for the calculation of systems having several thousand atoms. Following expectations, the localized molecular orbital method was found to be effective only in the case of systems with localized electrons (e.g., amino acids, peptides, proteins). For aromatic systems (such as DNA bases), the acceleration of the calculation resulting from the use of MOZYME option was less dramatic. Notice that the largest models (real proteins) contained aromatic groups, such as phenylalanine. The number of such residues, however, is considerably smaller than the number of nonaromatic ones. In the case of DNA, the ratio of aromatic residues is much higher.

Table 6. Stabilization Energies (in kcal/mol) of Various Graphene Models with Nucleic Acid Bases and Base Pairs Determined by the PM6-DH, DFT-D/B97/TZV(2d, 2p), DFT-D/TPSS/TZVP, and SCS-MP2/ TZV(2df,2pd) Techniques

C		A	T	U	C	G	A···T	A···U	G···C
24	PM6-DH	14.8	15.4	13.6	13.7	18.2	19.7	18.9	20.4
	PM6	2.3	3.9	3.9	3.0	5.3	1.9	1.9	2.4
	DFT-D/TPSS/TZVP	12.1	12.7	11.4	11.6	16.2	15.3	14.7	15.8
	DFT-D/B97/TZV	14.2	15.4	13.2	13.7	18.3	17.7	17.0	18.3
	SCS-MP2	15.1	14.9	13.0	13.4	18.0			
54	PM6-DH	19.0	17.8	15.5	16.9	21.5	29.3	28.2	31.2
	PM6	2.9	3.2	3.2	3.4	4.2	2.9	3.0	3.9
	DFT-D/TPSS/TZVP	15.9	14.8	13.4	15.4	19.8	24.4	23.6	27.4
	DFT-D/B97/TZV	19.3	18.0	15.4	17.8	23.3	29.9	28.5	32.0
	SCS-MP2	20.8	18.4	16.1	18.2	24.1			
96	PM6-DH	19.7	18.0	15.6	17.5	22.4	35.2	33.1	36.5
	PM6	2.9	2.6	2.5	3.4	4.4	4.4	4.5	5.5
	DFT-D/TPSS/TZVP	16.6	15.5	14.0	16.1	20.7	29.8	28.1	32.1
	DFT-D/B97/TZV	20.2	19.0	16.3	18.5	24.2	36.5	34.0	37.9
150	PM6-DH	19.8	18.1	15.8	17.7	22.9	36.7	34.4	37.7
	PM6	2.8	2.4	2.4	3.3	4.2	4.9	4.9	6.0
	DFT-D/TPSS/TZVP	16.7	15.8	14.3	16.4	21.1	30.5	28.8	33.0
	DFT-D/B97/TZV	20.3	19.3	16.5	18.8	24.9	–	34.9	38.7

Application Examples. *Porphine Dimers.* Stabilization energies of various structures of the porphine (the simpliest porphyrine macrocycle) dimer were determined using the PM6-DH technique and these energies were compared to reference data from the work of Muck-Lichtenfeld and Grimme,[40] calculated using DFT-D/B97 and SCS-MP2 [with application of the TZV(2df,2pd) basis set in both cases], as well as our DFT-D TPSS/TZVP and MP2.5/CBS [$E^{(2)}$ correction from CBS and scaled $E^{(3)}$ correction from TZV] calculations.[41] Table 5 shows that the PM6-DH stabilization energies agree very well with the most reliable MP2.5 results, with an average error for two typical structural types of less than 3%. The DFT-D values are systematically (with the exception of structure H) underestimated (by about 13% for the B97 functional, more with TPSS). The SCS-MP2 values are overestimated (by about 11%) for all stacked structures whereas they are underestimated for both T-shaped structures.

Graphene···Nucleic Acid Bases and Base Pairs. PM6-DH stabilization energies of various sheet models of

graphene (having 24, 54, 96, and 150 carbon atoms) with nucleic acid bases as well with H-bonded base pairs were compared with the reference data obtained by the DFT-D/B97/TZV(2d,2p) and SCS-MP2/TZV(2df,2pd) techniques.[42] Our own calculation at the DFT-D/TPSS-TZVP level is included for consistency with the other results presented here. Table 6 shows that PM6-DH results for complexes of graphene with isolated bases agree very well with reference energies for both relative and absolute stabilization energies. A similar conclusion can be made about complexes of graphene with DNA base pairs.

DNA···DAPI. Stabilization energies for the DNA dimer with 4′,6-diamidino-2-phenylindole (DAPI) bound as an intercalator and for the DNA trimer with DAPI bound to the minor groove determined with the PM6-DH method are 58.6 and 88.5 kcal/mol, respectively. Reference stabilization energies obtained using DFT-D/TPSS/TZVP are smaller than the PM6-DH results (40.4 and 67.8 kcal/mol, respectively). Considering, however, the results of the previous applications (i.e., porphine dimers and complexes of graphene with nucleic acid bases), the DFT-D results represent a lower limit of the real (unknown) stabilization energy. The present PM6-DH results can be thus considered as a satisfactory estimate of the respective stabilization energies.

Optimization of the DNA Tetramer. The DNA tetramer was optimized with the PM6 and PM6-DH methods, and the resulting geometries were compared to the DFT-D/TPSS/SVP ones. To demonstrate the flexibility of the method, optimizations were performed in the gas phase as well as in an environment represented by the COSMO implicit solvent model. The structure of the DNA tetramer collapsed during optimization using the PM6 method, and the rmsd for the resulting structure was 2.4 Å (compared to the DFT-D/TPSS/SVP geometry). The D and H corrections to the PM6 method improved its performance. The DNA tetramer retained its characteristic features during optimization with the PM6-DH method, and the rmsd decreased to 1.6 Å. Consideration of an implicit water model in PM6-DH (and also in the reference DFT-D calculations) further decreased the rmsd to an acceptable value of 1.3 Å.

Conclusions

Herein, we present a novel approach to improve semiempirical methods toward a more accurate description of noncovalent interactions acting in molecular clusters as well as in complex molecular systems. London dispersion and hydrogen bonds are treated separately by empirical corrections added to the QM calculation.

(1) Our method (PM6-DH) is based on the recent semiempirical method PM6, which currently represents the state of the art in the development of semiempirical QM methods. The same approach can be used with other QM and semiempirical methods.

(2) We adopted the formalism of a dispersion correction used previously in DFT and reparameterized it for PM6. It systematically improves all dispersion-bonded complexes, yielding an accuracy close to that of high-level correlated QM methods.

(3) The correction of hydrogen bonds has the form of an additional electrostatic term applied to all possible hydrogen bonds in the system. Differences in the nature and magnitude of the error found in PM6 calculations required the introduction of atom types to differentiate several types of hydrogen bonds. The correction is directional and should thus provide a better description of the hydrogen bonds than the standard core−core interaction used in semiempirical methods.

(4) The resulting PM6-DH method was tested on multiple sets of high-quality benchmark data. The results are superior to those obtained with the PM6 method alone. It has to be stressed that the accuracy of the method is close to that of correlated ab initio methods. The long-sought target accuracy for semiempirical QM methods, so-called chemical accuracy (error < 1 kcal/mol), was achieved for the S22 set. On the larger set of biomolecular complexes, the JSCH2005 database, it was shown that the description of different types of interactions is consistent and brings a significant improvement when compared to the uncorrected PM6 method.

(5) Although the method was derived for interaction energies in molecular complexes, the corrections are important also in isolated molecules featuring intramolecular noncovalent interactions. Excellent results were achieved in the description of conformational energies of small peptides, with a mean absolute error [compared to CCSD(T)/CBS results] amounting to 0.9 kcal/mol for a set of 76 structures. This accuracy is comparable to that of MP2/cc-pVTZ calculations.

(6) Implementation of a gradient makes this method useful for the optimization of systems with geometries determined by noncovalent interactions. The analytical calculation of gradients of the H-bond correction is not exact but uses an approximation. Nevertheless, the method performed well in geometry optimization tests, yielding geometries close to those obtained using the best QM methods available. There is a minor problem in the geometries of the strongest H-bonds associated with the form of the correction term, but it is outweighed by the other benefits. Optimization of the structure by PM6-DH leads to improvements in the relative energies of peptide conformers. This is, to the best of our knowledge, a unique feature because geometry optimization with other semiempirical QM methods usually strongly deteriorates the quality of the geometries obtained. Because hydrogen-bond types are determined only once (at the beginning of the calculation), the H-bond correction cannot be used for a continuous description of a reaction that changes valence states of the atoms involved in hydrogen bonds.

(7) The PM6-DH technique was further tested for various extended stacked complexes (porphine dimer, graphene models with DNA bases, and base pairs). It was shown that the method provides excellent stabilization energies that agree very closely with the benchmark values obtained by much more expensive DFT-D, SCS-MP2, or even MP2.5 methods. Finally, for the example of a DNA tetramer, we showed that PM6-DH can be used for geometry optimizations of rather large biomolecules.

(8) The PM6-DH calculations are very efficient and can be routinely applied to systems of up to 1000 atoms. Using the linear scaling MOZYME algorithm available in the PM6 implementation in MOPAC2009, nonaromatic systems with several thousand atoms can be calculated. With this performance, the method alone can replace molecular mechanics in calculations of smaller systems. In contrast to molecular mechanics, the present PM6-DH method fully and properly includes quantum effects. This fact and the efficiency of the method make its application in biological disciplines (concerning static or dynamic descriptions) extremely attractive for use in MD simulations.

Acknowledgment. This work was a part of Research Project Z40550506 of the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, and it was supported by Grants LC512 and MSM6198959216 from the Ministry of Education, Youth and Sports of the Czech Republic. The support of Praemium Academiae, Academy of Sciences of the Czech Republic, awarded to P.H. in 2007 is also acknowledged. D.R.S. is grateful to NSERC-Canada for ongoing Discovery Grant support. We also thank Martin Lepšík and Michal Pitoňák for inspiring discussions. We appreciate the discussion with J. J. P. Stewart, and we are grateful for his assistance and help with the original code. We are also grateful to the reviewers of our article for their insightful comments.

Supporting Information Available: Studying the error in PM6 results for hydrogen bonds and its dependence on the H-bond distance, we found a significant difference between hydrogen bonds with oxygen and nitrogen as the acceptor. This, of course, requires a different form of the correction that, after optimization, differs in the sign of the exponential term. It works as a repulsive term that lowers the magnitude of the correction for hydrogen bonds of oxygen, but it becomes a dominant attractive term for nitrogen. The demonstration of this issue on two typical systems from our training set (methanol···water and pyrazine···water complexes) is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286* (3−4), 243–252.

(2) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Olsen, J. *Chem. Phys. Lett.* **1999**, *302* (5−6), 437–446.

(3) Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2007**, *9* (39), 5291–5303.

(4) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, 4899–4907.

(5) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, 4907–4917.

(6) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107* (13), 3902–3909.

(7) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10* (2), 221–264.

(8) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10* (2), 209–220.

(9) Giese, T. J.; Sherer, E. C.; Cramer, C. J.; York, D. M. J. *Chem. Theory Comput.* **2005**, *1* (6), 1275–1285.

(10) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58* (11), 7260–7268.

(11) Yang, Y.; Yu, H. B.; York, D.; Cui, Q.; Elstner, M. *J. Phys. Chem. A* **2007**, *111* (42), 10861–10873.

(12) Winget, P.; Selcuki, C.; Horn, A. H. C.; Martin, B.; Clark, T. *Theor. Chem. Acc.* **2003**, *110* (4), 254–266.

(13) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114* (12), 5149–5155.

(14) Martin, B.; Clark, T. *Int. J. Quantum Chem.* **2006**, *106* (5), 1208–1216.

(15) Morgado, C. A.; McNamara, J. P.; Hillier, I. H.; Burton, N. A. *J. Chem. Theory Comput.* **2007**, *3* (5), 1656–1664.

(16) Tuttle, T.; Thiel, W. *Phys. Chem. Chem. Phys.* **2008**, *10* (16), 2159–2166.

(17) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13* (12), 1173–1213.

(18) Voityuk, A. A.; Rosch, N. *J. Phys. Chem. A* **2000**, *104*, 4089–4094.

(19) Thiel, W.; Voityuk, A. A. *Theor. Chim. Acta* **1992**, *81* (6), 391–404.

(20) Thiel, W.; Voityuk, A. A. *Theor. Chim. Acta* **1996**, *93* (5), 315.

(21) Thiel, W.; Voityuk, A. A. *J. Phys. Chem.* **1996**, *100* (2), 616–626.

(22) Stewart, J. J. P. *J. Mol. Model.* **2008**, *14* (6), 499–535.

(23) Stewart, J. J. P. *MOPAC2007*; Stewart Computational Chemistry: Colorado Springs, CO, 2007; available at http://OpenMOPAC.net(accessed Nov 10, 2008).

(24) VAMP. In *Materials Studio 4.4*; Accelrys: San Diego, CA, 2008; available at http://accelrys.com/products/materials-studio/modules/VAMP.html(accessed Apr 22, 2009).

(25) Jug, K.; Geudtner, G. *J. Comput. Chem.* **1993**, *14* (6), 639–646.

(26) Černý, J.; Kabeláč, M.; Hobza, P. *J. Am. Chem. Soc.* **2008**, *130* (47), 16055–16059.

(27) Hobza, P.; Mulder, F.; Sandorfy, C. *J. Am. Chem. Soc.* **1981**, *103* (6), 1360–1366.

(28) Hobza, P.; Mulder, F.; Sandorfy, C. *J. Am. Chem. Soc.* **1982**, *104* (4), 925–928.

(29) Hobza, P.; Sandorfy, C. *Can. J. Chem.* **1984**, *62* (3), 606–609.

(30) Jurečka, P.; Černý, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28* (2), 555–569.

(31) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8* (17), 1985–1993.

(32) Clark, T. *J. Mol. Struct. (THEOCHEM)* **2000**, *530* (1−2), 1–10.

(33) Dąbkowska, I.; Jurečka, P.; Hobza, P. *J. Chem. Phys.* **2005**, *122* (20), 204322.

(34) Distasio, R. A.; Head-Gordon, M. *Mol. Phys.* **2007**, *105* (8), 1073–1083.

(35) Grimme, S. *J. Chem. Phys.* **2003**, *118* (20), 9095–9102.

(36) Ahlrichs, R.; Bar, M.; Haser, M.; Horn, H.; Kolmel, C. *Chem. Phys. Lett.* **1989**, *162* (3), 165–169.

(37) Řezáč, J.; Jurečka, P.; Riley, K. E.; Černý, J.; Valdes, H.; Pluháčková, K.; Berka, K.; Řezáč, T.; Pitoňák, M.; Vondrášek, J.; Hobza, P. *Collect. Czech. Chem. Commun.* **2008**, *73* (10), 1261−1270. See also www.BEGDB.com.

(38) Valdes, H.; Pluháčková, K.; Pitoňák, M.; Řezáč, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10* (19), 2747–2757.

(39) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. 1. *Chem. Phy. Lett.* **1993**, *208* (5−6), 359–363.

(40) Muck-Lichtenfeld, C.; Grimme, S. *Mol. Phys.* **2007**, *105* (19−22), 2793–2798.

(41) Pitoňák, M.; Neogrdy, P.; Černý, J.; Grimme, S.; Hobza, P. *ChemPhysChem* **2009**, *10* (1), 282–289.

(42) Antony, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2008**, *10* (19), 2722–2729.

CT9000922

JCTC Journal of Chemical Theory and Computation

Convergence of the CCSD(T) Correction Term for the Stacked Complex Methyl Adenine−Methyl Thymine: Comparison with Lower-Cost Alternatives

M. Pitoňák,[†,§] T. Janowski,[||] P. Neogrády,[§] P. Pulay,[||] and P. Hobza*[,†,‡]

Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, v.v.i., Flemingovo Nám. 2, 166 10 Praha 6, Czech Republic, Department of Physical Chemistry, Palacký University, 771 46 Olomouc, Czech Republic, Department of Physical and Theoretical Chemistry, Faculty of Natural Sciences, Comenius University, Mlynská Dolina, 842 15 Bratislava 4, Slovak Republic, and Department of Chemistry and Biochemistry, Fulbright College of Arts and Sciences, University of Arkansas, Fayetteville, Arkansas 72701

Received March 17, 2009

Abstract: We have performed large-scale calculations for the interaction energy of the stacked methyl adenine−methyl thymine complex at the CCSD(T)/aug-ccpVXZ (X = D,T) levels. The results can serve as benchmarks for the evaluation of two methods, MP2.5, introduced recently, and the widely used ΔCCSD(T) correction defined as the difference between the CCSD(T) and MP2 energies. Our results confirm that the ΔCCSD(T) correction converges much faster toward the complete basis set (CBS) limit than toward the MP2 or CCSD(T) energies. This justifies approximating the CBS energy by adding the ΔCCSD(T) correction calculated with a modest basis set to a large basis MP2 energy. The fast convergence of the ΔCCSD(T) correction is not obvious, as the individual CCSD and (T) contributions converge less rapidly than their sum. The MP2.5 method performs very well for this system, with results very close to CCSD(T). It is conjectured that using a ΔMP2.5 correction, defined analogously to ΔCCSD(T), with large basis sets may yield more reliable nonbonded interaction energies than using ΔCCSD(T) with a smaller basis set. This would result in important computational savings as the MP3 scales computationally much less steep than CCSD(T), although higher than SCS-MP2, a similar approximation.

1. Theoretical Background

The CCSD(T) method in the complete basis set (CBS) limit provides accurate stabilization energies for various structures of molecular complexes. This is a very demanding task, but it must be kept in mind that CCSD(T)/CBS is in fact the only ab initio quantum mechanical method which is consistently capable of delivering benchmark quality results for single-reference systems. Other methods either are too expensive or, if less expensive, fail to provide the right answer (e.g., MP2/CBS strongly overestimates stacking stabilization energies although it describes H-bonding energies reasonably well[1]). The performance of yet another group of less expensive methods is enhanced by incorporating empirical parameters. For example, the SCS-MP2 method[2] and its variants (SOS-MP2,[3] SCS(MI)-MP2,[4] and modified SCS-MP2[5]) provide better results than the MP2 method (the overestimation of the stacking interactions is corrected) but utilize one or two empirical parameters.

CCSD(T)/CBS provides excellent results, but it is very time consuming due to its unfavorable N^7 scaling with the size of the system. Recently, a much more economical but still accurate method, MP2.5 (along with a generalized variant, "scaled MP3"), was proposed.[1] Unlike the SCS-MP2

* Corresponding author. Phone: +420220410311. Fax: +420220410320. E-mail: pavel.hobza@uochb.cas.cz.
† Academy of Sciences of the Czech Republic.
|| University of Arkansas.
§ Comenius University.
‡ Palacký University.

10.1021/ct900126q CCC: $40.75
Published on Web 06/22/2009

method and its variants, MP2.5 has correct asymptotic scaling at large intermolecular distances and uses only a single parameter. The MP2.5 method, more expensive than MP2, computational scaling is ($O(N^6)$ vs $O(N^5)$) but still can be applied to significantly larger systems than CCSD(T). MP2.5 was shown to be superior when compared to SCS-MP2 based methods.[1]

An alternative, and more economical, approach uses modified density functional theory (DFT). DFT, extended by empirical dispersion correction by Grimme[6] or Jurečka et al.,[7] provides good stabilization energies and geometries for molecular complexes and clusters, but two parameters per atom are included in the empirical dispersion energy formula (Grimme's formulation includes an extra global scaling parameter as well). Promising new exchange−correlation functionals of Zhao and Truhlar (the M05 and M06 families[8]) use a large number of fitted parameters. Preliminary results obtained in the Hobza laboratory show good performance of the M06 family. They perform well for both noncovalent interactions and also for IR and visible spectra of isolated molecules. Parameters in most of the procedures mentioned above must, however, be fitted against benchmark data, resulting mostly from the CBS extrapolation of CCSD(T) energies.

The estimated CCSD(T)/CBS energy is defined as

$$\mathrm{CCSD}(T)_{\mathrm{CBS}} \approx \mathrm{MP2}_{\mathrm{CBS}} + \Delta\mathrm{CCSD}(T)_{\text{small basis set}} \quad (1)$$

where the first term is the MP2 interaction energy, covering most of the correlation effects. This term is ideally calculated at a level close to that of the basis set limit. The second term, called the CCSD(T) correction term, is determined as a difference between CCSD(T) and MP2 interaction energies calculated using a smaller, computationally tractable basis set:

$$\Delta\mathrm{CCSD}(T)_{\text{small basis set}} = \mathrm{CCSD}(T)_{\text{small basis set}} - \mathrm{MP2}_{\text{small basis set}} \quad (2)$$

It describes correlation effects such as pair couplings that are omitted in the MP2 method.

Fairly reliable interaction energies close to CBS can be obtained by two- (or more) point extrapolation by schemes proposed, for instance, by Helgaker et al:[9]

$$E_X^{\mathrm{HF}} = E_{\mathrm{CBS}}^{\mathrm{HF}} + A.\exp(-\alpha X) \quad (3)$$

$$E_X^{\mathrm{corr.}} = E_{\mathrm{CBS}}^{\mathrm{corr.}} + B.X^{-3} \quad (4)$$

where A, B, and α are fitting parameters, or by Lee et al.[10,11] (using BSSE corrected, ΔE_X^b, and uncorrected, ΔE_X^n, interaction energies):

$$\delta_X = \Delta E_X^b - \Delta E_X^n \quad (5)$$

$$\epsilon_X = \Delta E_X^b + \Delta E_X^n \quad (6)$$

$$\Delta E_{\mathrm{CBS}} = 1/2(\delta_X\epsilon_{X+1} - \delta_{X+1}\epsilon_X)/(\delta_X - \delta_{X+1}) \quad (7)$$

based on systematically improved Dunning's aug-cc-pVXZ basis sets (e.g., aug-cc-pVDZ and aug-cc-pVTZ or, preferably, aug-cc-pVTZ and aug-cc-pVQZ). The Min et al. extrapolation can utilize two or more arbitrary basis sets.[11] These methods were developed in Kim's laboratory, and they will be referenced as Kim's extrapolation methods from now on.

The determination of the ΔCCSD(T) correction term, needed in eq 1, is computationally expensive despite the fact that it is usually calculated in a small or medium size basis set. It is known that this term converges faster than MP2 interaction energy itself. About 300 CCSD(T)/CBS stabilization energies collected in the S22,[12] S26−07,[13] JCSH2005,[12] and BEGDB[14] databases were determined in the above-mentioned way, using the Helgaker's extrapolation. One very important observation follows from these also referred to as the "benchmark" data. The ΔCCSD(T) correction term is almost negligible for H-bonded complexes while being systematically repulsive (up to 3.5 kcal/mol or more) for stacked complexes. Therefore, MP2/CBS results for stacked structures are strongly overbinding, with relative errors of as much as 50%. The question now remains whether the assumption of fast convergence, which is of key importance for the accurate evaluation of the CCSD(T)/CBS interaction energies for extended complexes, is fulfilled not only for model complexes (see, e.g. ref 12) but also for larger real life examples. Evidently, the only way to test this assumption is to carry out a "brute-force" attempt and perform the very time-consuming CCSD(T) calculations in extended basis sets for the systems of interest.

The benzene dimer is the most extensively investigated model of π interaction at high levels of correlation and large basis sets (up to aug-cc-pVQZ).[15−17,19] The total ΔCCSD(T) term for the stacked dimer exhibits fast and monotonic convergence. Interestingly, its two components ΔCCSD and Δ(T) converge individually and less rapidly indicating a cancellation between these terms. Similar calculations on the uracil dimer[19] at the aug-cc-pVTZ level confirm this conclusion. However, the ΔCCSD(T) correction is rather modest in these systems (below 2 kcal/mol), and it is desirable to check it for systems where it is larger.

Among all the stacked DNA base pairs included in the S22 and JCSH2005 data sets, the largest ΔCCSD(T) correction term, about 3.6 kcal/mol, was for the 9-methyl adenine (mA)−1-methyl thymine (mT) stacked complex. This complex is quite large (110 correlated electrons), so the previous calculations were done with the 6-31G*(0.25) basis set only.

In this work, we are trying to approach the CCSD(T)/CBS for the stacked complex of mA−mT by applying the aug-cc-pVDZ and aug-cc-pVTZ basis sets and extrapolating the results to the complete basis set limit, using standard extrapolation techniques. Besides these benchmark calculations, we also tested the performance of the MP2.5 procedure.

2. Methodology

The geometry of the stacked mA−mT complex (see Figure 1) was taken from the S22 data set where it was determined by counterpoise-corrected all-coordinates gradient optimization at the MP2/cc-pVTZ level. It has been shown previously[20] that these geometries are close to the CCSD(T) ones.

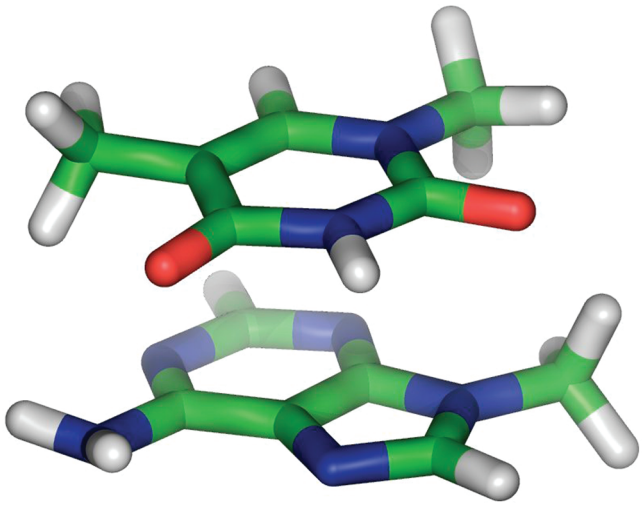


Figure 1. Structure of the methyl adenine−methyl thymine complex.

To investigate the performance of other "small" basis sets which were used for calculation of other extended (stacked) systems, the CCSD(T) calculations were performed also with 6-31G*, 6-31G*(0.25), 6-31G**, 6-31G**(0.25), 6-31G**(0.25,0.15), and 6-31G**(diff aDZ) as well as with 6-31+G*, 6-31+G*(0.25), 6-31+G**, 6-31+G**(0.25), and 6-31+G**(0.25,0.15) (Table 1 provides a complete list; note that all calculations involve spherical harmonic basis sets). Exponents of the *d* (for heavy elements, e.g. C, N, O) and *p* (hydrogen) polarization functions, in the 0.25 and (0.25,0.15) basis sets, were changed from default values to more diffuse ones (0.8→0.25 and 1.1→0.15), following the recommendations of van Lenthe et al.[21] Similarly, in the 6-31G**(diff aDZ) basis set, these exponents of the polarization functions were changed to the most diffuse ones from aug-cc-pVDZ (e.g., 0.8→0.151and 1.1→0.141). All correlated calculations use the frozen core approximation (the 1s orbital of the C, N, and O atoms). The interaction energy was, in all cases, corrected for the basis set superposition error using Boys−Bernardi[22] counterpoise correction.

Calculations involving the largest aug-cc-pVXZ (X = D,T) basis sets were performed using the PQS program package[23] with a recently developed parallel CCSD(T) module[24,25] that utilizes a parallel I/O filesystem.[26] This code was designed for an efficient computation of CCSD(T) energies for very large systems. The linear dependency in the basis set was not removed for this part of the calculations, but a tight integral threshold was applied (10^{-14}) in order to avoid numerical difficulties that may be expected in such an extended system, where the overlap matrix has small eigenvalues.

All other MP3 and CCSD(T) calculations were carried out using the MOLCAS package.[27,28] This program uses Cholesky decomposed (CD) two-electron integrals in the Hartree−Fock module with no local exchange screening.[29] The MP3 and CCSD(T) calculations used a new, highly parallelized closed-shell code[30] and also made use of the CD decomposition of two-electron integrals. For all calculations, a threshold of the order of 10^{-5} was used for the two-electron integrals decomposition; in our experience, this gives an accuracy of higher than 0.01 kcal/mol.

3. Results and Conclusions

The total CCSD(T) stabilization energy (absolute value of the interaction energy) of mA−mT is strongly dependent on the basis set. As shown in the Table 1, going from aug-cc-pVDZ (618 basis functions), where the CCSD(T) calculations are already out of the range of most CCSD(T) codes on standard computer architectures, to aug-cc-pVTZ (1311 basis functions), the stabilization energy increases by as much as ∼1.5 kcal/mol.

A rough measure, of whether a basis set is saturated for calculation of the interaction energy, is a comparison of the BSSE corrected and uncorrected values. In aug-cc-pVTZ, these two values at the CCSD(T) level differ by ∼4.4 kcal/mol. Furthermore, even for such a large basis set as the aug-cc-pVQZ, the corrected and uncorrected stabilization energies at the MP2 level still differ by 2 kcal/mol. It is very important to use diffuse basis functions. For instance, the CCSD(T) stabilization energy obtained using the cc-pVTZ-f basis set (the standard Dunning's cc-pVTZ basis set without f functions) is ∼1.8 kcal/mol lower than at the aug-cc-pVDZ level, i.e., a basis set of the same size. This is also apparent by comparing the results obtained, for instance, with the 6-31+G** and 6-31+G**(0.25,0.15) basis sets. They differ by 3.75 kcal/mol.

Further analysis of results in Tables 1 and 2 clearly indicates where this basis set dependence comes from. The HF results (see Table 1) are repulsive and show almost no basis set dependence, amounting to 7.31 kcal/mol in the least diffuse basis set (6-31G*) and to 7.05 kcal/mol in the most diffuse and extended aug-cc-pVQZ basis set. The effect of the BSSE correction is significant even at this level of theory, being most pronounced in the smallest but very diffuse 6-31G**(diff aDZ) basis set at ∼10 kcal/mol (in absolute value), and leads to artificial stabilization of the complex. Let us skip the MP2 method for a while and focus on the effect of correlation beyond MP2, i.e., ΔCCSD(T) shown in Table 2.

The basis set dependence of the ΔCCSD(T) correction term is significantly smaller than the changes in the total CCSD(T) stabilization energy. The dependence of the ΔCCSD(T) on the diffuseness of basis set is reversed for the 6-31G**, 6-31+G*, and 6-31+G** family of basis sets but increases in going from the aug-cc-pVDZ to the aug-cc-pVTZ basis. Surprisingly, not even the ordering of the BSSE corrected and uncorrected values of ΔCCSD(T) is uniform. This situation is similar to that of the stacked structure of the uracil dimer.[19] The maximum scatter of the ΔCCSD(T) values is much smaller compared to the total CCSD(T) values ∼1.3 vs ∼8.6 kcal/mol or ∼30 vs 66% of the total combination. What differs tremendously is the absolute value of the difference between the BSSE corrected and uncorrected values. The largest ΔCCSD(T) difference is ∼0.8 kcal/mol for the 6-31+G**(0.25,0.15) basis set, while the largest difference for the total CCSD(T) values is more than 27 kcal/mol. For the highest quality ΔCCSD(T) and CCSD(T) values (aug-cc-pVTZ), these differences are

Table 1. Basis Set Dependence of HF, MP2, and CCSD(T) Total Stabilization Energies for the mA−mT Complex[a,b]

basis set	no. of AOs	HF	MP2	CCSD(T)
6-31G*[c]	324	−7.31 (−2.59)	7.65 (17.49)	4.24 (14.15)
6-31G*(0.25)	324	−6.99 (0.66)	12.89 (30.21)	9.32 (26.85)
6-31G**	369	−7.20 (−2.45)	8.10 (17.89)	4.60 (14.41)
6-31G**(0.25)	369	−6.91 (0.69)	13.07 (30.12)	9.48 (26.67)
6-31G**(0.25,0.15)	369	−6.98 (1.69)	13.63 (34.00)	10.28 (30.88)
6-31G**(diff aDZ)	369	−7.22 (2.86)	13.23 (40.56)	10.23 (37.92)
6-31+G*	408	−7.10 (−4.73)	10.61 (21.42)	7.05 (18.48)
6-31+G*(0.25)	408	−6.59 (−1.99)	14.80 (31.35)	11.46 (28.67)
6-31+G**	453	−7.17 (−4.73)	11.09 (22.06)	7.41 (19.05)
6-31+G**(0.25)	453	−6.72 (−2.35)	14.06 (30.15)	10.63 (27.31)
6-31+G**(0.25,0.15)	453	−6.77 (−0.98)	14.36 (34.75)	11.16 (32.33)
cc-pVTZ-f	618	−7.14 (−5.02)	14.20 (21.19)	9.82 (17.04)
aug-cc-pVDZ	618	−7.11 (−4.43)	15.77 (27.00)	11.61 (23.31)
aug-cc-pVTZ	1311	−7.03 (−6.43)	17.31 (21.92)	13.06 (17.44)
aug-cc-pVQZ	2370	−7.05 (−6.89)	17.79 (19.70)	−

[a] Total stabilization energies in kcal/mol; negative values are repulsive. [b] Values in parentheses are stabilization energies without BSSE correction. [c] Spherical harmonics (5d, 7f,...) are used in all basis sets.

Table 2. Basis Set Dependence of ΔMP3 (ΔMP2.5) and ΔCCSD(T) Correction to the Interaction Energy (in kcal/mol)[a]

basis set	ΔMP3	ΔCCSD	ΔCCSD(T)	c_{opt}[b]	ΔMP2.5 error[c]
6-31G*	5.93 (6.77)	5.33 (6.10)	3.41 (3.34)	0.58	0.45
6-31G*(0.25)	7.60 (8.45)	6.30 (7.02)	3.57 (3.36)	0.47	−0.23
6-31G**	6.11 (6.98)	5.49 (6.31)	3.50 (3.48)	0.57	0.44
6-31G**(0.25)	7.67 (8.58)	6.38 (7.17)	3.59 (3.45)	0.47	−0.25
6-31G**(0.25,0.15)	7.67 (8.48)	6.29 (7.10)	3.36 (3.13)	0.44	−0.48
6-31G**(diff aDZ)	7.12 (7.92)	5.90 (6.97)	2.99 (2.64)	0.42	−0.57
6-31+G*	6.95 (7.41)	6.01 (6.71)	3.56 (2.94)	0.51	0.09
6-31+G*(0.25)	7.79 (7.82)	6.34 (6.66)	3.33 (2.68)	0.43	−0.56
6-31+G**	7.15 (7.63)	6.20 (6.94)	3.68 (3.01)	0.52	0.10
6-31+G**(0.25)	7.87 (7.98)	6.44 (6.83)	3.43 (2.84)	0.44	−0.50
6-31+G**(0.25,0.15)	7.81 (7.80)	6.33 (6.71)	3.20 (2.42)	0.41	−0.71
cc-pVTZ-f	8.67 (9.01)	7.34 (7.78)	4.38 (4.15)	0.50	0.05
aug-cc-pVDZ	8.94 (9.43)	7.48 (8.20)	4.16 (3.69)	0.47	−0.31
aug-cc-pVTZ	9.08 (9.46)	7.93 (8.45)	4.25 (4.47)	0.47	−0.29

[a] Values in parentheses are contributions without correction for the BSSE. [b] c_{opt} = ΔCCSD(T)/ΔMP3. [c] ΔMP2.5 error = ΔCCSD(T) − 0.5*ΔMP3.

∼0.2 vs ∼4.4 kcal/mol. The difference between the ΔCCSD(T) calculated in the better performing small basis set, 6-31G**(0.25), and the "reference", aug-cc-pVTZ, is ∼0.7 kcal/mol, while for the better performing "medium" basis set, 6-31+G**, it is ∼0.6 kcal/mol. This represents the limit of accuracy of the "estimated" CCSD(T)/CBS calculated according to the eq 1 using the small basis set for the ΔCCSD(T) correction. Another option is to stay with the "plain" CCSD(T) values, which would, however, lead to much larger errors, ∼2.8 kcal/mol for better performing small basis set 6-31G**(0.25,0.15) and 1.6 kcal/mol for better performing "medium" basis 6-31+G**(0.25,0.15).

The largest basis set dependence is obtained for the MP2 correlation energy. Going from the aug-cc-pVTZ to aug-cc-pVQZ, the stabilization energy still increases by almost 0.5 kcal/mol, while the difference between the BSSE corrected and uncorrected values decreases from ∼4.6 to ∼1.9 kcal/mol. For comparison, going from aug-cc-pVDZ to aug-cc-pVTZ, the ΔCCSD(T) correction term increases only by ∼0.1 kcal/mol and the difference between the BSSE corrected and uncorrected values reduces from ∼−0.5 to ∼0.2 kcal/mol. Considering the slow convergence of MP2 and the rather fast one of ΔCCSD(T), the composite scheme eq 1 should be, in general, successful.

Table 2 includes the ΔCCSD corrections. Comparing them with the ΔCCSD(T) values, it is clear that there is significant cancellation between the ΔCCSD and Δ(T) contributions.

Table 2 also shows the MP3 correction term, ΔMP3. Comparing with ΔCCSD(T), ΔMP3 is too repulsive by 42−59%, see the "c_{opt}" column in the table. Scaling the ΔMP3 correction by 1/2, i.e., doing MP2.5, results are in absolute errors of ∼0.7−∼0.05 kcal/mol relative to the ΔCCSD(T) value, depending on the basis set. The error of MP2.5 is indicated by the difference between the optimum scaling coefficient and 0.5 for the particular structure and basis set. Some trends of "c_{opt}", at least for the nucleic acid base pairs, have already been discussed previously.[1] The optimum scaling coefficient is closer to 0.4 for small and diffuse basis sets and closer to (or larger than) 0.5 for less diffuse basis sets. Interestingly, the basis set dependence of the higher-order correlation terms, e.g., ΔMP3 and ΔCCSD(T), for mA−mT is strong enough to make the MP2.5 method in larger basis sets (aug-cc-pVDZ or aug-cc-pVTZ) more accurate than the estimated CCSD(T)/CBS value, calculated using the exact ΔCCSD(T) correction obtained by small/medium basis sets (e.g., 6-31G*/6-31G**- or 6-31+G*/6-31G+G**-type). The error of MP2.5 is determined by the absolute value of the ΔMP3 term. According to our experience from the S22 test set,[1] which is quite

Table 3. Comparison of the MP3, CCSD, and (T) Wall Clock Timings[a]

basis set	setup	PP/CPP	MP3[b]	CCSD[b]	(T)[b]
6-31G**(0.25,0.15)	6 x A, 4 x B	14/4	0.19	5.93	–
6-31G**(0.25,0.15)	12 x C	24/4	0.05	1.23	5.23
6-31+G**(0.25,0.15)	12 x C	24/4	0.10	2.47	15.75
cc-pVTZ-f	1 x B	1/8	3.78	–	–
cc-pVTZ-f	7 x B	7/8	–	31.50	488.00

[a] The following symbols were used for node architectures: "A" Intel Core2 Quad (4 cores), 2.40 GHz, "B" Intel Xeon E5345 (8 cores), 2.33 GHz, and "C" 2x Quad-Core AMD Opteron 2354, 2.20 GHz. The column setup explains how the machines were utilized. "PP" stands for a number of MPI parallel processes, while "CPP" stands for a number of cores utilized by threaded-BLAS routines per one MPI process. In all the CCSD calculations listed below, 21 iterations were necessary for the convergence. [b] Results are in hours.

diverse in terms of intermolecular interaction types, the optimum value of the scaling coefficient lies between 0.4 and 0.6. This means that the absolute error of the ΔMP2.5 correction is generally within 10% of the ΔMP3 value. This error is comparable with or lower than the basis set effect on ΔCCSD(T).

Timings shown in Table 3 illustrate computational savings of the MP3 method.

MP3 is significantly faster that CCSD despite the same asymptotic scaling of these methods with the system size. The computer time is saved by skipping numerous nonlinear terms from the CCSD equations. However, the major difference is that the CCSD method is iterative, while MP3 is equivalent to a single configuration interaction doubles (CID) iteration. This is especially beneficial in parallel runs. The savings of MP2.5 with respect to CCSD(T) result mainly from its lower asymptotic scaling, $O(N^6)$ vs $O(N^7)$. This makes MP2.5, atlhough it lacks a solid theoretical basis, an efficient alternative for CCSD(T) when the latter exceeds available computational resources. Another problem not yet fully recognized by the computational chemistry community is that the numerical accuracy of the (T) combination obtained using double precision arithmetic may be insufficient due to the very large number arithmetic operations in its evaluation. This is particularly the case if diffuse basis sets are utilized.

For the sake of completeness, we also report a few details involving our calculation for the biggest basis sets. The computation for the dimer in the aug-cc-pVTZ basis set was performed employing a 40 Xeon E5430 with 2.66 GHz nodes, each node with 8 processing cores and 16 GB of memory. The wall clock time for the most expensive part, the (T) correction, was 160 h. Calculation of energies of the monomers in the basis set of the dimer needed for counterpoise correction took less than 120 h on 10 nodes. The basis set dimension for this system is 1311, and the number of correlated occupied orbitals is 55.

To obtain new CCSD(T) benchmark values for the stabilization energy of mA−mT, we performed Helgaker and Kim's type of extrapolations, shown in Table 4.

For both types of extrapolations, two variants are presented. First, rows "aDZ→aTZ" in Table 4 correspond to extrapolations of the total CCSD(T) correlation energies

Table 4. Helgaker's and Kim's Extrapolation of the Total MP2 and CCSD(T) Stabilization Energies and ΔCCSD(T) Correction Term (Repulsive) Calculated from Given Values[e]

extrapolation	MP2	CCSD(T)	ΔCCSD(T)
Helgaker aDZ→aTZ[a]	17.95	13.66	4.29
Helgaker aTZ→aQZ′[b]	17.99	13.70	4.29
Kim aDZ→aTZ[c]	18.38	13.92	4.45
Kim aTZ→aQZ′[d]	18.14	13.74	4.40

[a] HF/aQZ + corr. CCSD(T)/(aDZ→aTZ). [b] HF/aQZ + corr. (MP2/X + ΔCCSD(T)/X − 1)(aTZ→aQZ). [c] Total CCSD(T)/(D→T). [d] (Total MP2/X + ΔCCSD(T)/X − 1)/(aTZ→aQZ). [e] Extrapolation values calculated from the values outlined in footnotes b−e. For details see text; aXZ stands for Dunning's aug-cc-pVXZ (X = D, T, Q) basis sets.

obtained by the aug-cc-pVDZ and aug-cc-pVTZ basis set. In Helgaker's extrapolation, the HF energy was taken from the aug-cc-pVQZ basis set, which was considered converged to the CBS limit. The values in the second row, labeled as "aTZ→aQZ′", were constructed by combining the MP2 correlation (or in case of Kim's extrapolation total) energies with the ΔCCSD(T) obtained in the basis set of one cardinality lower, i.e., MP2/aug-cc-pVQZ + ΔCCSD(T)/aug-cc-pVTZ, and analogously for the aug-cc-pVTZ and aug-cc-pVDZ pair.

Assuming that the "aTZ→aQZ′" extrapolation for both types is closer to the real CBS limit, the total CCSD(T) stabilization energies from the Helgaker and Kim schemes differ by less than 0.05 kcal/mol, both being ∼13.7 kcal/mol, while the ΔCCSD(T) of the both correlation schemes agree within ∼0.1 kcal/mol, both being ∼4.30 kcal/mol.

4. Summary

(1) The ΔCCSD(T) correction term determined from calculations using extended basis sets (aug-cc-pVDZ and aug-cc-pVTZ) is 4.3 kcal/mol, which is ∼0.7 kcal/mol more than that determined using small basis sets (e.g., 3.6 kcal/mol in 6-31G*(0.25)). This means that the CCSD(T)/CBS values in the S22, S26−07, and JCSH2005 data sets (where the CCSD(T) correction term is determined using small or medium basis sets) are reasonably reliable. However, it indicates that the CBS value of ΔCCSD(T) increases by 10−20% (at most), if evaluated using large basis sets.

(2) The ΔCCSD(T) correction converges more rapidly than the ΔCCSD and Δ(T) corrections separately, showing a partial cancellation between these terms.

(3) Helgaker- and Kim-types of extrapolations from extended basis sets (aug-cc-pVXZ, X = D, T, Q) are in good mutual agreement (within 0.1 kcal/mol), which supports the universality and robustness of both schemes.

(4) The MP2.5 method provides excellent values for stabilization energies at a much lower cost and, unlike the SCS-MP2 method[2] and its variants, yields the correct asymptotic behavior for dispersion. Its computational cost, although higher than MP2 based methods, is reasonable, and it is recommended for extended molecular complexes. Based on calculations of other systems, the MP2.5 methods accuracy is expected to be within 10% of the ΔMP3 correction.

Acknowledgment. This work was supported by grants no. LC512 and MSM6198959216 from the Ministry of Education, Youth, and Sports (MŠMT) of the Czech Republic and was part of research project no. Z4 055 0506. It was also supported by the Slovak Research and Development Agency (contract no. APVV-20-018405) and by grant CHE 0515922 of the U.S. National Science Foundation. The authors wish to acknowledge the support of Praemium Academiae of the Academy of Sciences of the Czech Republic, awarded to P.H. in 2007. Generous computation time from the Star of Arkansas supercomputer, which was purchased in part by funds obtained under grant no. MRE 072265 of the U.S. National Science Foundation, is gratefully acknowledged.

References

(1) Pitoňák, M.; Neogrády, P; Černý, J.; Grimme, S.; Hobza, P. *ChemPhysChem* **2009**, *10*, 282.

(2) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095.

(3) Jung, Y. S.; Lochan, R. C.; Dutoi, A. D.; Head-Gordon, M. *J. Chem. Phys.* **2004**, *121*, 9793.

(4) Distasio, R. A.; Head-Gordon, M. *Mol. Phys.* **2007**, *105*, 1073.

(5) Hill, J. G.; Platts, J. A. *J. Chem. Theory Comput.* **2007**, *3*, 80.

(6) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463.

(7) Jurečka, P.; Černý, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555.

(8) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.

(9) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243.

(10) Lee, E. C.; Kim, D.; Jurečka, P.; Tarakeshwar, P.; Hobza, P.; Kim, K. S. *J. Phys. Chem. A* **2007**, *111*, 3446.

(11) Min, S. K.; Lee, E. C.; Lee, H. M.; Kim, D. Y.; Kim, D.; Kim, K. S. *J. Comput. Chem.* **2008**, *29*, 1208.

(12) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.

(13) Riley, K. E.; Hobza, P. *J. Phys. Chem. A* **2007**, *111*, 8257.

(14) Řezáč, J.; Jurečka, P.; Riley, K. E.; Černý, J.; Valdes, H.; Pluháčková, K.; Berka, K.; Řezáč, T.; Pitoňák, M.; Vondrášek, J.; Hobza, P. *Collect. Czech. Chem. Commun.* **2008**, *73*, 1261.

(15) Sinnokrot, M. O.; Sherrill, C. D. *J. Chem. Phys.* **2006**, *110*, 10656–10668.

(16) Hill, J. G.; Platts, J. A.; Werner, H.-J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4072–4078.

(17) Janowski, T.; Pulay, P. *Chem. Phys. Lett.* **2007**, *447*, 27–32.

(18) Pitoňák, M.; Neogrády, P.; Řezáč, J.; Jurečka, P.; Urban, M.; Hobza, P. *J. Chem. Theory Comput.* **2008**, *4*, 1829.

(19) Pitoňák, M.; Riley, K. E.; Neogrády, P.; Hobza, P. *ChemPhysChem* **2008**, *9*, 1636.

(20) Dabkowska, I.; Jurečka, P.; Hobza, P. *J. Chem. Phys.* **2005**, *122*, 204322.

(21) van Lenthe, J. H.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Duijneveldt, F. B. Weakly Bonded Systems. In *Advances in Chemical Physics*; Volume *LXIX*; Prigogine, I., Rice, S. A., Eds.; John Wiley & Sons Ltd.: 1987.

(22) Boys, S. F.; Bernardi, F. *Mol. Phys.* **2002**, *100*, 65.

(23) *PQS version 3.2*; Parallel Quantum Solutions: 2013 Green Acres Road, Fayetteville, Arkansas 72703.

(24) Janowski, T.; Ford, A. R.; Pulay, P. *J. Chem. Theory Comput.* **2007**, *3*, 1368.

(25) Janowski, T.; Pulay, P. *J. Chem. Theory Comput.* **2008**, *4*, 1585.

(26) Ford, A. R.; Janowski, T.; Pulay, P. *J. Comput. Chem.* **2007**, *28*, 1215.

(27) Karlstrom, G.; Lindh, R.; Malmqvist, P.-A.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrády, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222.

(28) Aquilante, F.; De Vico, L.; Ferre, N.; Malmqvist, P.-Å.; Neogrády, P.; Pedersen, T.; Pitoňák, M.; Reiner, M.; Roos, B.; Serrano-Andres, L.; Urban, M.; Veryazov, V.; Lindh, R. J. Comput. Chem. . in press.

(29) Aquilante, F; Pedersen, T. B.; Lindh, R. *J. Chem. Phys.* **2007**, *126*, 194106.

(30) Neogrády, P.; Aquilante, F.; Noga, J.; Pitoňák, M.; Hobza, P.; Urban, M. manuscript in preparation.

CT900126Q

JCTC Journal of Chemical Theory and Computation

Aromaticity of α-Oligothiophenes and Equivalent Oligothienoacenes

Inmaculada García Cuesta,*,† Juan Aragó,† Enrique Ortí,† and Paolo Lazzeretti‡

Instituto de Ciencia Molecular, Universidad de Valencia, P.O. Box 22085, E-46071 Valencia, Spain, and Dipartimento di Chimica, Università degli Studi di Modena e Reggio Emilia, via Campi 183, 41100 Modena, Italy

Received March 17, 2009

Abstract: The aromaticity and the degree of π-electronic delocalization have been theoretically investigated for α,α′-linked oligothiophenes containing three and five rings and for their fused analogs oligothienoacenes. By computing magnetic susceptibilities and ^{1}H NMR shieldings as well as current density maps, it is found that the fused oligomers are more aromatic than the corresponding nonfused partners. The increase of aromaticity with the size of the oligomer—even in the case of quinoidal forms—is also proven. The π-currents induced by an external magnetic field show that oligothienoacenes behave as single cycles since they present an intense diamagnetic current flowing around the whole molecular perimeter. In contrast, nonfused α-oligothiophenes exhibit diamagnetic currents localized over each thiophene ring. For the quinoidal oligomers, local diamagnetic π vortices appear around CC double bonds, indicating that the π electrons are rather localized as in conjugated, nonaromatic polyenes. For quinoidal nonathienoacene, it is however found that the electronic circulation around the ethylenic bonds tends to delocalize all over the carbon skeleton, indicating a more effective π-conjugation and some aromatic character.

1. Introduction

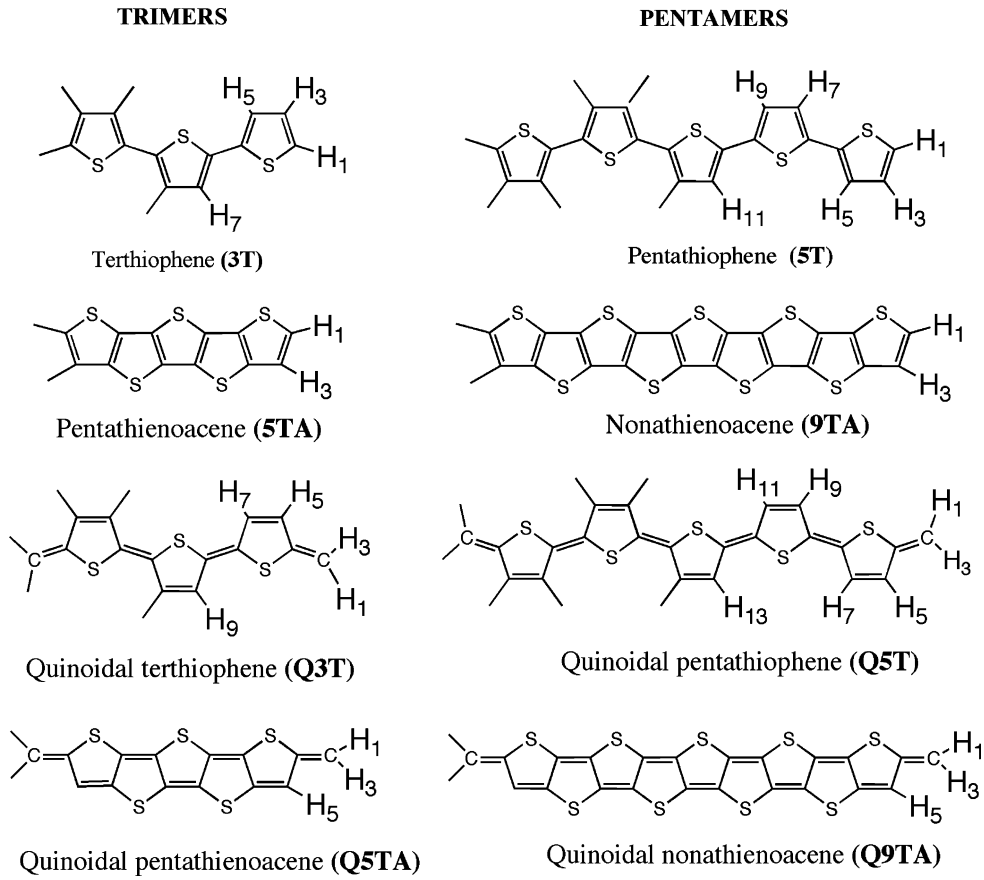
Since Faraday's discovery of benzene almost 200 years ago, the features associated with aromaticity have fascinated chemists.[1] Even at present, when thousands of aromatic compounds are known, research on aromaticity and aromatic compounds is still in progress. In particular, one of the areas in which aromaticity plays an essential role is the field of electroactive organic materials for molecular electronics as, for instance, small-molecule semiconductors and conducting polymers.[2,3] The achievement of the electric and/or optical response always involves structural changes in the molecules constituting the material that imply a gain or a loss in aromaticity. Organic electroactive materials are obtained by chemical or electrochemical synthesis, and their properties can be easily tuned by adding functional groups to their π-conjugated structure.[4−8] It is also important to recall their exceptional stability under different environments. Therefore, it is not surprising that there exists an enormous effort in developing technological applications of these materials such as organic light-emitting diodes (OLEDs), organic field-effect transistors (OFETs), flexible and large area displays, biosensors, lightweight photovoltaic cells, or wearable electronic devices.[9−14]

A vast majority of organic semiconductors and conducting polymers are built from aromatic systems such as benzene or thiophene and their fused derivatives used as monomer units. In these systems, the extended conjugation of the π electrons along the molecular skeleton determines the structural and electronic properties and accounts for their electric and optical behavior. Conducting polymers present two nondegenerate forms in their ground state, the *aromatic* and the *quinoidal* forms.[15−17] Despite the fact that both forms are conjugated, the aromatic structure is usually preferred in the neutral state, and the quinoidal structure is only attained upon charge injection. Electronic delocalization along the π system is not enough to allow the material to become

* Corresponding author e-mail: garciain@uv.es.
† Universidad de Valencia.
‡ Università degli Studi di Modena e Reggio Emilia.

10.1021/ct900127m CCC: $40.75
Published on Web 06/05/2009

Scheme 1. Chemical Structures of the Oligomers Studied



conductor, and doping—oxidation/reduction of the material—is required to achieve the conducting state. Charge injection provokes the progressive quinoidization of the conjugated organic chain. The characteristics of the aromatic form strongly influence the possibilities of the material as conductor not only because they determine the easiness for oxidation or reduction but also because the electronic structure of the aromatic form is partly inherited by the quinoidal structure.[16] In this sense, studying the aromaticity of conjugated oligomers is a relevant task because from this knowledge it is possible to provide useful information for the development of advanced semiconducting materials and at the same time to improve our understanding of aromaticity itself.

Among the different organic semiconducting materials, α-oligothiophenes (thiophene oligomers joined in α,α' positions by single bonds) represent a foremost group of molecules and have been widely employed as the active layers in organic electronic devices.[6,7,11,12,18,19] α-Oligothiophenes show a good environmental stability and are easy to synthesize, and their electronic and solid-state properties can be modulated by introducing either donor or electron-withdrawing groups in their carbon skeleton.[19] A main disadvantage of α-oligothiophenes is that they can deviate from planarity through torsion about the single inter-ring bonds thus decreasing the electron delocalization along the π-conjugated carbon backbone. In this context, oligothienoacenes (linearly fused thiophenes) are emerging as a promising new class of π-conjugated compounds that combine the rigid planarity and extended conjugation of acenes with the chemical stability of oligothiophenes.[20] Oligothienoacenes have been already implemented as the active layers in OFETs exhibiting an excellent field-effect performance with mobilities as high as 0.42 $cm^2\ V^{-1}\ s^{-1}$.[21] The structural and optical properties of penta- and heptathienoacene have been recently analyzed in comparison to those observed in α-oligothiophenes of identical conjugation length.[22]

In this work, we perform a comparative study of the aromaticity of α-oligothiophenes with three (terthiophene, **3T**) and five (pentathiophene, **5T**) rings and of their oligothienoacene analogs of identical conjugation length (pentathienoacene, **5TA**, and nonathienoacene, **9TA**, respectively) in both aromatic and quinoidal forms (see Scheme 1). Our goal is to determine how the aromatic properties change with the following: i) the nonfused/fused character of the conjugated skeleton, ii) the number of monomeric units, and iii) the aromatic/quinoid nature of the carbon backbone. Quinoidal forms are built up by introducing end-capping methylene groups and will be denoted as **Q3T**, **Q5T**, **Q5TA**, and **Q9TA**, respectively. It should be stressed that the main structural difference between α-oligothiophenes and oligothienoacenes is the sulfur atoms that bridge the thiophene rings. For the sake of simplicity, we will refer the studied systems as trimers (**3T** and **5TA**) or pentamers (**5T** and **9TA**).

2. Computational Details

According to the ring-current model (RCM), the exposition of an aromatic cyclic system to an external magnetic field

Table 1. Magnetic Susceptibility Tensors in (cgs) ppm au[a] via the CTOCD-DZ2 Method (Origin in the Center of Mass)[b]

	xx	yy	zz (π)	Av	$\Delta\chi$
terthiophene **(3T)**	−1211.8	−1094.7	−2544.4 (−929.4)	−1616.9	−1391.2
pentathienoacene **(5TA)**	−1475.0	−1373.5	−3400.2 (−1576.0)	−2082.9	−1976.0
quinoidal terthiophene **(Q3T)**	−1440.8	−1199.5	−2116.3 (−376.7)	−1585.5	−796.1
quinoidal pentathienoacene **(Q5TA)**	−1749.8	−1503.4	−2441.8 (−505.5)	−1898.3	−815.2
pentathiophene **(5T)**	−2003.4	−1781.3	−4078.8 (−1450.5)	−2621.1	−2186.5
nonathienoacene **(9TA)**	−2534.5	−2327.2	−5755.3 (−2708.9)	−3539.0	−3324.4
quinoidal pentathiophene **(Q5T)**	−2265.8	−1879.7	−3374.3 (−633.5)	−2506.6	−1301.6
quinoidal nonathienoacene**(Q9TA)**	−2857.8	−2488.5	−4131.6 (−987.3)	−3157.9	−1458.5

[a] The conversion factor from cgs au per molecule to cgs emu per mole is $a_0^3 N_A = 8.9238878 \times 10^{-2}$; further conversion to SI units is obtained by 1 $JT^{-2} = 0.1$ cgs emu. [b] Contributions from π electrons to the zz component are given within parentheses. $\chi_{Av} = (\chi_{xx}+\chi_{yy}+\chi_{zz})/3$. Anisotropy $\Delta\chi = \chi_{zz} - {}^1/_2(\chi_{xx}+\chi_{yy})$.

normal to the molecular plane induces π-electronic ring currents that produce an increase of the modulus of the magnetic susceptibility (χ) mainly due to an enlargement of its perpendicular component (χ_{zz}).[1,23−28] At the same time, the π-currents diminish the perpendicular component of the nuclear magnetic shielding of the proton (σ_{zz}),[1,26] which is normally known as downfield ^{1}H NMR chemical shift. It is important to emphasize that π-ring currents only modify the perpendicular component of the susceptibility and shielding tensors. Indeed, different causes contribute to determine the in-plane components, but none of them is related to the special mobility of the π-electrons and, therefore, to aromaticity. Consequently, criteria for diatropicity and aromaticity can only be based on the out-of-plane component of the magnetic tensors and not on the average (one-third of the trace) values. For planar molecules, the symmetry separation of σ and π orbitals is preserved, and, therefore, it is possible to evaluate only the contribution to those properties from π-electrons.

We have carried out a systematic analysis of the aromaticity of the studied compounds and, with this aim, determined their susceptibilities and NMR shieldings as well as the density maps of the current induced by the magnetic field. We have used the damped variant of a method allowing for continuous transformation of the origin of the current density-diamagnetic zero, CTOCD-DZ2,[29] inside the Coupled Hartree−Fock approach as implemented in the SYSMO suite of programs.[30] The selected basis sets were cc-pCVTZ for carbon[31,32] and sulfur[33] and cc-pVTZ for hydrogen.[31]

Molecular geometries were optimized within the density functional theory (DFT) using the B3LYP functional[34] and the 6-31G** basis set[35] and imposing C_{2h} symmetry constraints. The Gaussian03 program[36] was used to this end. To test the influence of the molecular geometry on the magnetic properties, the geometry of pentathienoacene was also optimized using second-order Møller-Pleset (MP2) perturbation theory. Compared with the B3LYP/6-31G**-optimized geometry, MP2/6-31G** calculations predict slightly shorter single C−C (0.006−0.007 Å), C−S (0.013−0.017 Å), and C−H (0.003 Å) bonds and slightly longer double C═C bonds (0.008−0.009 Å). The variations on the magnetic properties due to these small geometrical changes are calculated to be rather unimportant since the average chemical shieldings vary in 0.07−0.10 ppm and their zz-component in 0.03−0.06 ppm. Changes in magnetic susceptibility are also small: 0.4% for χ_{Av} and 1.2% for χ_{zz}.

3. Results and Discussion

3.1. Magnetic Properties. The two trimers, terthiophene and pentathienoacene, have six carbon−carbon (CC) double bonds in their aromatic forms **3T** and **5TA** and seven CC double bonds in their quinoidal partners **Q3T** and **Q5TA**, but the fused oligomers **(5TA** and **Q5TA)** have two additional sulfur atoms and, therefore, a larger number of π electrons. A similar situation is found for the pentamers, for which the aromatic forms **5T** and **9TA** present one CC double bond less than the quinoidal structures **Q5T** and **Q9TA**—ten vs eleven—, while the four extra sulfur atoms in the fused counterparts provide them with an increased number of π-electrons.

All the considered aromatic systems are diatropic molecules, which means that they are able to sustain intense diamagnetic currents in the presence of an external magnetic field. The diatropic currents are however sensibly less intense for the quinoidal systems. The magnetic susceptibilities calculated for all of them, aromatic and quinoidal, are large and negative (see Table 1). The component normal to the molecular plane, χ_{zz}, is larger than the average in-plane components, making the susceptibility anisotropy, $\Delta\chi$, negative. The structural and electronic differences mentioned above are especially reflected by the degree of anisotropy. $\Delta\chi$ shows significantly larger values for the fused oligomers **5TA** (−1976.0 au) and **9TA** (−3324.4 au) than for their respective nonfused partners **3T** (−1391.2 au) and **5T** (−2186.5 au). Its absolute value increases with the length of the oligomer (57% in passing from **3T** to **5T**, 68% from **5TA** to **9TA**), which is mostly due to the larger contribution of the π electrons to the χ_{zz} component (56 and 72%, respectively). For the aromatic compounds, the value of χ_{zz} is more than twice (2.2 for **3T** and **5T**, 2.4 for **5TA** and **9TA**) the value of ${}^1/_2(\chi_{xx} + \chi_{yy})$, but the ratio between these values is significantly reduced in passing to the quinoidal compounds (1.6 for **Q3T** and **Q5T**, 1.5 for **Q5TA** and **Q9TA**). To compare with, let us recall that in benzene the perpendicular component is almost three times larger than the parallel components.[26]

Obviously, going from aromatic to quinoidal oligomers implies a loss of aromaticity, which can be quantified in terms of the susceptibility anisotropy. There are two important issues to recall. First, the loss of aromaticity is larger the smaller is the oligomer. For the trimers it is of 43% (**3T**) and 59% (**5TA**), while for the pentamers is of 40% (**5T**) and 56% (**9TA**), i.e. some 3% lower, showing an increasing

Table 2. Proton Magnetic Shieldings of Oligothiophenes and Oligothienoacenes in ppm via the CTOCD-DZ2 Method[b]

	xx	yy	zz (π)	Av	δ^1H[a]
Terthiophene **(3T)**					
H1	25.6	25.1	21.3 (−1.67)	24.0	7.5
H3	24.8	27.5	21.4 (−1.71)	24.5	7.0
H5	26.8	26.6	19.3 (−2.61)	24.2	7.3
H7	27.0	26.6	19.3 (−2.45)	24.3	7.2
Pentathienoacene **(5TA)**					
H1	26.0	25.0	20.3 (−2.33)	23.8	7.7
H3	25.7	26.9	19.5 (−3.11)	24.0	7.5
Quinoidal Terthiophene **(Q3T)**					
H1	30.7	24.0	23.7 (−0.26)	26.1	5.4
H3	30.1	23.4	24.2 (−0.08)	25.9	5.6
H5	27.3	24.1	22.6 (−0.62)	24.7	6.8
H7	28.7	24.1	21.1 (−0.67)	24.6	6.9
H9	28.8	24.1	20.9 (−0.72)	24.6	6.9
Quinoidal Pentathienoacene **(Q5TA)**					
H1	30.9	24.0	23.7 (−0.28)	26.2	5.3
H3	30.5	23.4	24.0 (−0.18)	26.0	5.5
H5	28.8	24.8	22.0 (−0.99)	25.2	6.3
Pentathiophene **(5T)**					
H1	25.7	25.0	21.2 (−1.69)	24.0	7.5
H3	24.8	27.5	21.3 (−1.72)	24.5	7.0
H5	26.9	26.5	19.2 (−2.62)	24.2	7.3
H7	27.1	26.6	19.2 (−2.48)	24.3	7.2
H9	27.4	26.6	19.1 (−2.41)	24.4	7.1
H11	27.4	26.6	19.1 (−2.65)	24.4	7.1
Nonathienoacene **(9TA)**					
H1	25.6	25.0	20.2 (−2.38)	23.6	7.9
H3	25.8	26.9	19.3 (−3.15)	24.0	7.5
Quinoidal Pentathiophene **(Q5T)**					
H1	30.7	23.9	23.6 (−0.28)	26.1	5.4
H3	30.2	23.4	24.2 (−0.10)	25.9	5.6
H5	27.5	24.1	22.6 (−0.64)	24.7	6.8
H7	28.7	24.1	21.0 (−0.73)	24.6	6.9
H9	28.9	24.2	20.7 (−0.79)	24.6	6.9
H11	29.0	24.0	20.6 (−0.86)	24.5	7.0
H13	29.0	24.1	20.6 (−0.86)	24.6	6.9
Quinoidal Nonathienoacene **(Q9TA)**					
H1	31.0	23.9	23.6 (−0.33)	26.1	5.4
H3	30.5	23.4	23.9 (−0.23)	25.9	5.6
H5	28.8	24.7	21.7 (−1.08)	25.1	6.4

[a] Chemical shifts are referenced to thiophene values: $\sigma_{Av}(H) = 24.50$, $\delta^1H = 6.96$. [b] Contributions from π electrons to the zz component are given within parentheses. $\sigma_{Av} = (\sigma_{xx}+\sigma_{yy}+\sigma_{zz})/3$.

Table 3. Aromaticity Index HOMA of the Individual Thiophene Rings for Oligothiophenes and Oligothienoacenes

index HOMA	central ring	2nd ring	terminal ring
terthiophene **(3T)**	0.709		0.724
pentathiophene **(5T)**	0.715	0.714	0.727
pentathienoacene **(5TA)**	0.713	0.697	0.727
nonathienoacene **(9TA)**	0.714	0.714	0.730
quinoidal terthiophene **(Q3T)**	0.324		0.228
quinoidal pentathiophene **(Q5T)**	0.496	0.450	0.211
quinoidal pentathienoacene **(Q5TA)**	0.335	0.305	0.137
quinoidal nonathienoacene **(Q9TA)**	0.519	0.478	0.207

aromatic character of the quinoidal form as the size of the oligomer is enlarged. The same numbers illustrate the second important fact: the fused oligomers lose a larger fraction of aromaticity when adopting the quinoidal structure.

The difference in aromaticity between aromatic and quinoidal structures is better observed by comparing the proton magnetic shieldings (Table 2). The values calculated for the ^{1}H NMR shieldings of the **3T, 5TA**, **5T**, and **9TA** oligomers are typical of aromatic compounds. For these oligomers, the deshielding attributable to the π-electrons is in general larger than 2 ppm, while for the quinoidal forms it is smaller than 1 ppm for the thiophene protons and almost negligible for the vinylene protons.

Proton magnetic shieldings can also be used to establish a direct comparison between nonfused and fused oligomers. For the aromatic trimers, protons H_1 of **3T** and **5TA** (see Scheme 1 for atom numbering) are in equivalent environments, and, therefore, the differences in the zz component of the shielding tensor can be precisely associated with the differences in the corresponding ring currents. Accordingly, the values of 21.3 ppm for **3T** and 20.3 ppm for **5TA** show that the ring current is more intense in the terminal ring of the fused oligomer, as the deshielding produced for **5TA** (−2.33 ppm) is larger.

A similar analysis on H_1 protons of the aromatic pentamers **5T** and **9TA** shows that the reduction of the ^{1}H NMR shielding tensor ($\sigma_{zz} = 21.2$ and 20.2 ppm, respectively) is slightly larger than that of the corresponding trimers. This suggests that the intensity of the ring current increases with the size of the oligomer. The trend has been confirmed by extending the study to include also heptathiophene, though using the smaller 6-31G** basis set to calculate the shielding tensor. As shown in Table S1 in the Supporting Information, the computed numerical values are slightly different from those reported in Table 2 using the larger Dunning's basis sets, but the general features are kept. Calculations predict that the proton deshielding due to π electrons increases as the oligomer chain becomes longer. In particular, $\sigma_{zz}(\pi)$ takes values of −2.35, −2.36, and −2.44 ppm for the inner proton of the terminal ring of **3T**, **5T**, and **7T**, respectively. Furthermore, the deshielding slightly decreases in going from the central to the external thiophene rings, with the exception of the terminal rings for which the contribution to deshielding of the π-electrons is the largest.

For the quinoidal trimers **Q3T** and **Q5TA**, it is possible to compare the vinylene protons H_1 close to the sulfur atoms. Again, the deshielding attributed to the π electrons is larger for the fused oligomer (−0.28 ppm) than for the nonfused oligothiophene (−0.26 ppm). The remaining protons of **Q3T** and **Q5TA** present slightly different environments even in analogous positions, but all of them appear more deshielded in **Q5TA** (H_3: −0.18 ppm, H_5: −0.99 ppm) than in **Q3T** (H_3: −0.08 ppm, H_5: −0.62 ppm). In any case, the values of π-deshielding found for σ_{zz} of the vinylene protons are intermediate between those found for a nonaromatic quinoidal phenantrene (−0.10 and −0.16 ppm) and those obtained for a more extended quinoidal system that possesses some aromatic character (−0.40 and −0.46 ppm).[37]

As discussed for the aromatic oligomers, the quinoidal pentamers **Q5T** and **Q9TA** also present larger π-deshieldings than the equivalent trimers, even for the vinylene protons (see Table 2). It is important to note that this effect turns out to be more significant in the environment of the central

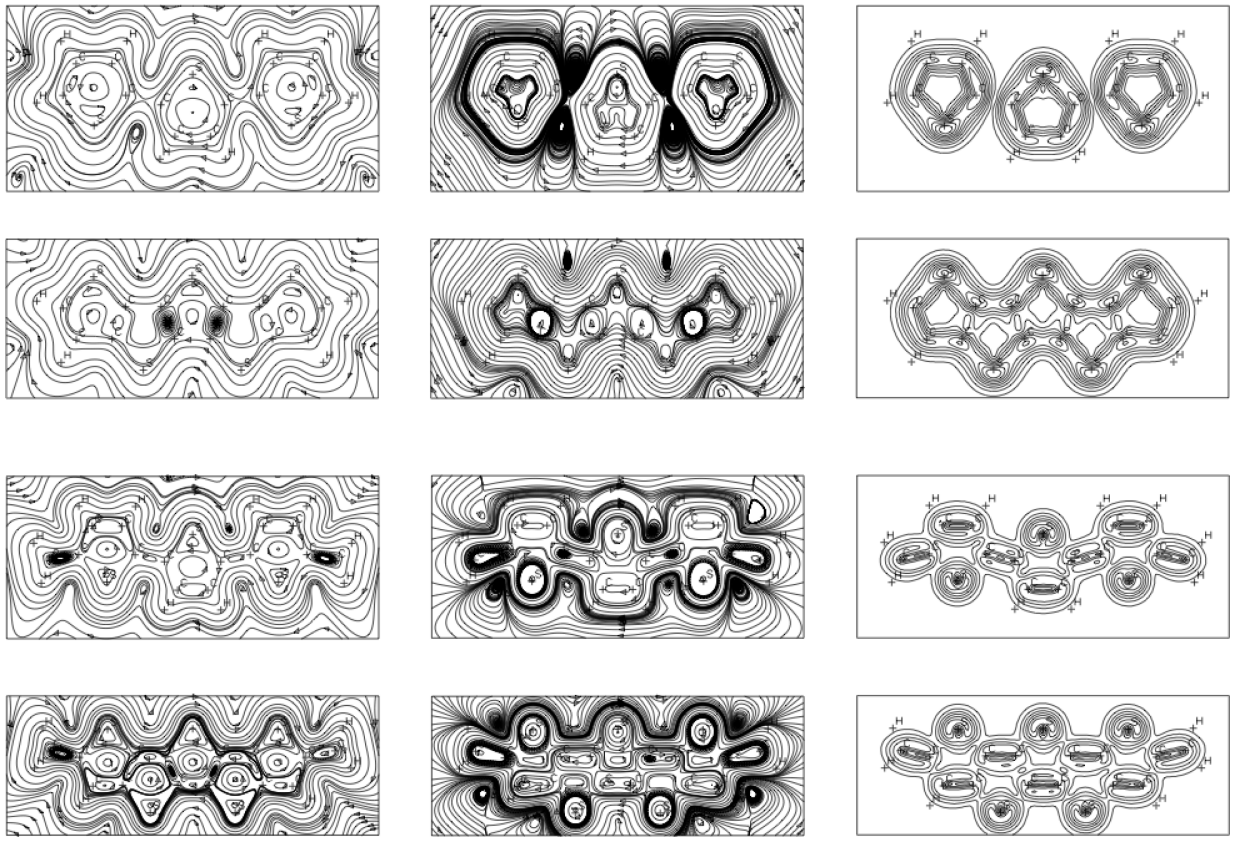


Figure 1. Streamlines of the total current density (left) and π-electron contributions to the current density (center) on a plane at 1.1 bohr above the molecular plane and contour levels for the modulus of the π current (right) for (top to bottom) **3T**, **5TA**, **Q3T**, and **Q5TA**. The maximum modulus (contour step) values are 0.069 (0.007), 0.075 (0.008), 0.033 (0.005) and 0.034 (0.005), respectively, in au.

ring, for which the changes are larger when the oligomer becomes longer. This suggests that the central part of the oligomer becomes more aromatic as the quinoidal chain lengthens. This partial aromatization is in agreement with the reduction in the CC bond length alternation (difference between the length of single and double bonds) calculated for the central ring in passing from **Q3T** (0.070 Å) to **Q5T** (0.047 Å) and with previous geometrical results obtained for quinoidal oligothiophenes end-capped with dicyanomethylene groups.[38−40]

To facilitate comparison to available experimental data, we have also computed ^{1}H chemical shifts taking tetramethylsilane (TMS) as reference. In terthiophenes and similar derivatives it is experimentally observed that for protons in the central ring δ(H) oscillates between 7.0 and 7.7 ppm, while for the central ring δ(H) = 7.0−7.2 ppm.[41] In the case of pentathienoacene, it has been determined that $\delta(H_1) =$ 7.4 ppm and $\delta(H_3) = 7.3$ ppm. In addition, an increase of proton chemical shifts with the number of rings has been found in analogous compounds.[42] A reasonable agreement is encountered between these experimental values and the calculated values reported in Table 2, which illustrates the quality of our theoretical data. Still, let us remark that our analysis is based on the π-contributions to the out-of-plane shielding. Such contribution is not available to experiment, but according to the classical picture is the one directly related to aromaticity.

Other criteria as, for instance, the Harmonic Oscillator Model of Aromaticity, HOMA,[43] (see Table 3) lead to similar conclusions in quantifying the aromaticity of the studied species. The HOMA index takes the value of 0 for a Kekulé structure of a typical aromatic system and the value of 1 for systems with all bond lengths equal to the optimal value as in benzene. We find that the HOMA indices calculated for the thiophene rings in the fused oligomers are larger than those obtained for the nonfused compounds and that they increase with the size of the oligomer. This suggests once more that oligothienoacenes are slightly more aromatic than nonfused oligothiophenes and that the degree of aromaticity increases with the oligomer length. It is also observed that while in the aromatic oligomers the largest indices are encountered for the terminal rings, in the quinoidal structures the highest HOMA index corresponds to the central ring. This correlates with the slight aromatization of the central part of the quinoidal structures inferred from π-deshielding values.

3.2. Current Density Maps. The representations of the current density induced by a uniform magnetic field applied along the positive z axis perpendicular to the molecular plane provide a direct visualization of the phenomenology of molecules in a magnetic field. By applying the RCM, it is possible to quantify the aromaticity of the molecule and the delocalization of the π-electrons through the modulus and the direction of the currents. Two different kinds of current

density maps have been used to this aim. The first type displays the streamlines of the current density with the corresponding modulus represented by contour curves. The second type uses arrows in the direction of the current, their length being locally proportional to the current density in that point. The second type of maps yields a less-detailed description, but it is enough to illustrate the essential features of the flow. The region examined in this study is approximately that of maximum of the π-density, 1.1 bohr above the plane of the molecule.[44] Diamagnetic currents are clockwise, while paramagnetic ones are in the opposite sense.

Figure 1 shows the streamlines of the total and π-current densities as well as the modulus of the π-current density for the four trimers under study. The streamlines for the total current density (left column in Figure 1) present common features in all systems: a diamagnetic flow runs around the molecular periphery and paramagnetic vortices appear on the ring centers, which is typical of planar conjugated cyclic molecules. The π-current density maps (central column in Figure 1) display different characteristics for each molecule. For the aromatic trimers, there exist strong diatropic ring currents, but their topologies are sensibly different depending on the structure of the oligomer. For fused **5TA**, an intense current is delocalized around the whole molecular perimeter, as it were a single cycle, in a way analogous to that found for naphthalene[45] or thieno[3,2-b]thiophene.[44] In contrast, for nonfused **3T**, the aromaticity of the thiophene ring prevails and distinct patterns of noninteracting diamagnetic currents localized on each thiophene ring are observed. For the quinoidal trimers, the localized circulation around the double CC bonds is worth noticing. The different regime of the π-current densities is best shown by the shape of the contour maps in Figure 1.

Figure 1 indicates that the intensities of the current densities are also different. The absolute maximum of the π-current density corresponds in all cases to islands of flow, each centered about a sulfur nucleus. Each circulation has the same diatropic sense as the peripheral ring current but a much smaller radius. The maximum modulus is different for each oligomer: 0.069 (**3T**), 0.075 (**5TA**), 0.033 (**Q3T**), and 0.034 au (**Q5TA**). The π-ring current densities (i.e., the global circulation of current delocalized around the whole molecular perimeter as in benzene) also show different features. In **5TA** the most intense π-ring current reaches a modulus as big as 0.043 au, while in **3T** the maximum values of the ring current intensity are 0.039 and 0.033 au for the terminal and central thiophene rings, respectively.

The intensity and the pattern of the ring currents change drastically for the quinoidal trimers (see Figure 1). A substantial reduction of intensity of the ring currents takes place with respect to the aromatic partners, the maximum values being 0.0080 au for **Q3T** and 0.0081 au for **Q5TA** (c.a. five times smaller than in their aromatic partners). The most distinctive feature observed for the quinoidal trimers is the diamagnetic π-current showing vortices and foci located in the regions of the formal double bonds. This result shows that double CC bonds prevail as the main π-entity in quinoidal oligomers and suggests that π electrons are more localized in these systems than in their aromatic partners.

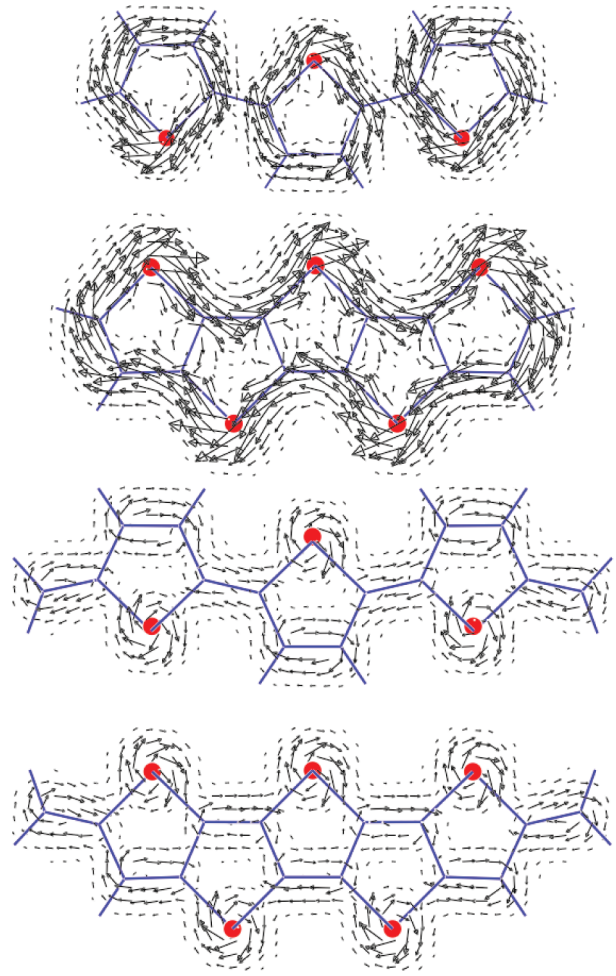


Figure 2. π-electron contribution to the current density on a plane at 1.1 bohr for (top to bottom) **3T**, **5TA**, **Q3T**, and **Q5TA**. The maximum modulus values are 0.069, 0.075, 0.033, and 0.034 au, respectively. The length of the arrows is normalized to the maximum modulus in benzene (0.080 au). Intensities lower than 0.004 au are omitted.

The intensity and direction of the currents can be simultaneously observed in the maps depicted in Figure 2. For the aromatic trimers, the ring currents described above are clearly seen, and it is straightforward to show that fused **5TA** is more aromatic than its nonfused counterpart **3T** since the current density is more intense in the former. The topology of the current density maps for the quinoidal trimers is very different from that for their aromatic partners. First, the π-currents are much less intense as it can be easily checked from inspection of the figure, where the arrows representing the density currents for the quinoidal systems **Q3T** and **Q5TA** are shorter than those for the aromatic systems **3T** and **5TA**. Second, the distribution of the currents is completely changed: the ring currents flowing around the molecular periphery of the quinoidal systems appear in Figure 2 as very short arrows—virtually points—, which describe a weak annular diatropism. At any rate, the currents localized on sulfur atoms and double bonds clearly predominate, and local diamagnetic π vortices, typical of formal C=C double bonds, are observed. Thus, the π electrons are rather localized as in conjugated nonaromatic polyenes.

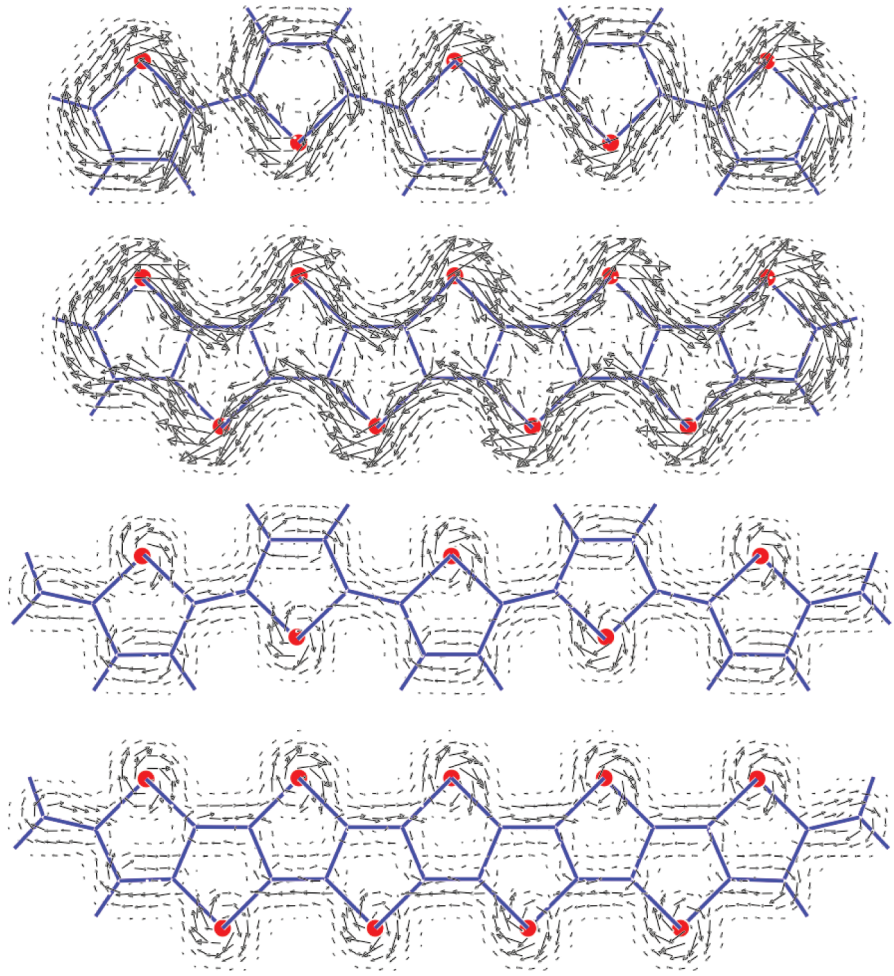


Figure 3. π-electron contribution to the current density on a plane at 1.1 bohr above the molecular plane for (top to bottom) **5T**, **9TA**, **Q5T,** and **Q9TA**. The maximum modulus values are 0.069, 0.075, 0.033, and 0.036 au, respectively. The length of the arrows is normalized to the maximum modulus in benzene (0.080 au). Intensities lower than 0.004 au are omitted.

By close inspection of the two set of maps (Figure 1 and 2) for the quinoidal trimers, one can observe a current streamline flowing over all the double bonds in the thiophene rings, which would imply a small degree of delocalization. Since this current is somewhat more intense in the case of the fused oligomer, the delocalization effects are slightly more sizable for the fused oligomer **Q5TA** than for the nonfused terthiophene **Q3T.** This fact is in agreement with the larger π-contribution to proton deshielding ($\sigma_{zz}(\pi)$) in **Q5TA** than in **Q3T** (see Table 2) mentioned above.

There is no major change in the intensity of the ring currents for the aromatic pentamers **5T** and **9TA** compared with their equivalent trimers. However, significant differences are found for quinoidal oligomers, for which the ring currents have intensities of 0.0083 (**Q5T**) and 0.0088 au (**Q9TA**). The fused **Q9TA** pentamer therefore shows ring currents almost 10% more intense than the corresponding **Q5TA** trimer. The streamlines and modulus maps (Figure S3) present characteristics which are completely similar to those discussed for the corresponding trimers, i.e., peripheral ring currents around the whole aromatic structure for fused **9TA**, in-ring-restricted currents for the aromatic nonfused **5T**, and circulation involving the ethylenic bonds for the quinoidal oligomers **Q5T** and **Q9TA**. The arrows maps displayed in Figure 3 illustrate an important feature that is not detectable in previous maps. For **Q9TA**—and for longer thienoacenes,—it is observed that the outward current in the CC double bonds is much more intense than the return current inside the rings, indicating a larger delocalization of the π-electronic density. In fact, the diamagnetic perimeter circulation grows going from **Q5T** (in which vortices about the double bonds are clearly discernible) to **Q9TA** (in which no closed current loops are found in the region of internal formal double bonds). This effect is only observable for the central rings of fused **Q9TA**. We could not detect it either in the nonfused **Q5T** or the quinoidal trimers, which suggests that the delocalization of the ethylenic bonds requires a minimum size of the oligomer to take place. In **Q9TA** the current density about double bonds cannot form closed loops. On the other hand, it gives rise to continuous π-ring current flowing all over the carbon skeleton visible in the map of **Q9TA** as a peripheral delocalized stream more intense than in the other quinoidal oligomers.

Summarizing, the stronger intensity of the peripheral ring π-current, the absence of closed current loops about the formal double bonds, and the consequent weakening of the return currents indicate that **Q9TA** possesses a higher degree of π-conjugation than other quinoidal oligomers. This is confirmed by the numerical estimates of the π-contribution to the out-of-plane component of proton magnetic shieldings.

4. Conclusions

By means of magnetic criteria, we have evaluated the aromaticity of a series of thiophene-based oligomers with both fused/nonfused and aromatic/quinoidal structures. Our theoretical analysis uses a comparison of results obtained from calculations of magnetic susceptibility, ^{1}H NMR shieldings, and ring currents. All the considered systems, either aromatic or quinoidal, exhibit large and negative values of the magnetic susceptibility and, in particular, of the χ_{zz} component. The deshielding $\sigma_{zz}(\pi)$ contribution to the proton shielding tensor clearly indicates that fused oligothienoacenes are slightly more aromatic than nonfused α-oligothiophenes. It also implies that the degree of aromaticity increases with the oligomer length and going from inner to outer thiophene rings along the chain. π-deshielding values also predict a weak aromatization of the central part of the conjugated chain for quinoidal oligomers increasing with the oligomer size. These trends are also supported by the values obtained for the aromaticity HOMA index. The conversion from aromatic to quinoidal structures obviously implies a loss of aromaticity, which is larger in the case of the thienoacenes.

The π-currents induced by a uniform magnetic field applied perpendicular to the molecular plane display a drastically different topology depending on the structure of the oligomer. Fused oligothienoacenes behave as single cycles and show an intense current flowing around the whole molecular perimeter. In contrast, the aromaticity of the thiophene ring prevails for nonfused α-oligothiophenes which present diamagnetic currents around each thiophene ring. For the quinoidal oligomers, local diamagnetic π vortices, typical of formal CC double bonds, are observed, indicating that the π electrons are rather localized as in conjugated, nonaromatic polyenes. However, the electronic circulation around the ethylenic bonds in the quinoidal nonathienoacene (**Q9TA**) tends to delocalize all over the carbon skeleton giving the molecule some aromatic character. The effect is more pronounced for the central rings and is not observed for the trimer **Q5TA**, thus suggesting that the delocalization of the ethylenic bonds requires a minimum size of the oligomer to be effective. A progressive gain of aromatic character is therefore expected on increasing the length of fused quinoidal systems. This effect has been already predicted for quinoidal benzene-fused systems.[37]

Acknowledgment. This work has been supported by the Spanish Ministry of Education and Science (MEC; projects CSD2007-00010, CTQ2006-14987-C02-02, CTQ2007-67143-C02-01/BQU, and CT2009-08790), the Generalitat Valenciana (ACOMP07/163, GV/2007/093, and GVAINF 2007-051), and European FEDER funds. J.A. acknowledges the MEC for a FPI doctoral grant.

Supporting Information Available: CTOCD-DZ2/6-31G** perpendicular component of ^{1}H NMR shieldings (σ_{zz}) and contributions from π electrons for α,α′-linked oligothiophenes containing three, five, and seven rings, streamlines and modulus of the π current for **5T**, **9TA**, **Q5T**, and **Q9TA**, and a sensibly enlarged version of Figure 1 in order to facilitate its interpretation. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) von Ragué Schleyer, P. *Chem. Rev.* **2001**, *101*, 1115–1117.

(2) For a recent review on organic materials for molecular electronics see the monographic issues: (a) Organic Electronics and Optoelectronics. *Chem. Rev.* **2007**, *107* (4), (b) Organic Electronics. *Chem. Mater.* **2004**, *16*, (23).

(3) (a) *Molecular Nanoelectronics*; Reed, M. A., Lee, T., Eds.; American Scientific Publishers: Stevenson Ranch, CA, 2003. (b) Tour, J. M. *Molecular Electronics: Commercial Insights, Chemistry, Devices, Architecture and Programming*; World Scientific: 2003. (c) Petty, M. *Molecular Electronics: From Principles to Practice*; John Wiley & Sons: 2008.

(4) Roncali, J. *Chem. Rev.* **1997**, *97*, 173–205; **1992**, *92*, 711−738.

(5) Müllen, K.; Wegner, G. *Electronic materials: The oligomer Approach*; Wiley-VCH: 1998.

(6) Fichou, D. *Handbook of oligo- and Polythiophenes*; Wiley-VCH: Weinheim, 1999.

(7) (a) Facchetti, A.; Yoon, M. H.; Stern, C. L.; Hutchinson, G. R.; Ratner, M. A.; Marks, T. J. *J. Am. Chem. Soc.* **2004**, *126*, 13480–13501. (b) Facchetti, A.; Mushrush, M.; Yoon, M. H.; Hutchinson, G. R.; Ratner, M. A.; Marks, T. J. *J. Am. Chem. Soc.* **2004**, *126*, 13859–13874. (c) Yoon, M. H.; di Bennedetto, S. A.; Russell, M. T.; Facchetti, A.; Marks, T. J. *Chem. Mater.* **2007**, *19*, 4864–4881.

(8) Peart, P. A.; Repka, L. M.; Tovar, J. D. *Eur. J. Org. Chem.* **2008**, 2193–2206.

(9) Mitschke, U.; Bäuerle, P. *J. Mater. Chem.* **2000**, *10*, 1471–1507.

(10) Coakley, K. M.; McGehee, M. D. *Chem. Mater.* **2004**, *16*, 4533–4542.

(11) Murphy, A. R.; Fréchet, J. M. J. *Chem. Rev.* **2007**, *107*, 1066–1096.

(12) Dimitrakopoulos, C. D.; Malenfant, P. R. L. *Adv. Mater.* **2002**, *14*, 99–117.

(13) Bao, Z.; Rogers, J. A.; Katz, H. E. *J. Mater. Chem.* **1999**, *9*, 1895–1904.

(14) Katz, H. E. *Chem. Mater.* **2004**, *16*, 4748–4756.

(15) *Handbook of Conducting Polymers*, 2nd ed.; Skotheim, T. A., Elsenbaumer, R. L., Reynolds J. R., Eds.; Marcel Dekker: New York, 1998.

(16) (a) Brédas, J.-L.; Chance, R. R.; Silbey, R. *Phys. Rev. B* **1982**, *26*, 5843. (b) Brédas, J.-L.; Street, G. B. *Acc. Chem. Res.* **1985**, *18*, 309–315.

(17) (a) Kertesz, M.; Lee, Y.-S. *J. Phys. Chem.* **1987**, *91*, 2690–2692. (b) Lee, Y.-S.; Kertesz, M. *J. Chem. Phys.* **1988**, *88*, 2609–2617.

(18) Schulze, K.; Riede, M.; Brier, E.; Reinold, E.; Bäuerle, P.; Leo, K. *J. Appl. Phys.* **2008**, *104*, 074511.

(19) *Handbook of Thiophene-Based Materials: Applications in Organic Electronics and Photonics*; Perepichka, I. F., Perepichka, D. F., Eds.; Wiley: Weinheim, 2009.

(20) (a) Zhang, X.; Côte, A. P.; Matzger, A. J. *J. Am. Chem. Soc.* **2005**, *127*, 10502–10503. (b) Okamoto, T.; Kudoh, K.; Wakamiya, A.; Yamaguchi, S. *Chem.—Eur. J.* **2007**, *13*, 548–556.

(21) (a) Xiao, K.; Liu, Y.; Qi, T.; Zhang, W.; Wang, F.; Gao, J.; Qiu, W.; Ma, Y.; Cui, G.; Chen, S.; Zhan, X.; Yu, G.; Qin,

J.; Hu, W.; Zhu, D. *J. Am. Chem. Soc.* **2005**, *127*, 13281–13286. (b) Gao, P.; Beckmann, D.; Tsao, H. N.; Feng, X.; Enkelmann, V.; Pisula, W.; Müllen, K. *Adv. Mater.* **2009**, *21*, 213–216.

(22) (a) Osuma, R. M.; Zhang, X.; Matzger, A. J.; Hernández, V.; López-Navarrete, J. T. *J. Phys. Chem. A.* **2006**, *110*, 5058–5065. (b) Aragó, J.; Viruela, P. M.; Ortí, E. *J. Mol. Struct., Theochem*, in press.

(23) Pauling, L. *J. Chem. Phys.* **1936**, *4*, 673–677.

(24) Lonsdale, K. *Proc. R. Soc. (London)* **1937**, *A159*, 149–170.

(25) London, F. *J. Phys. Radium* **1937**, *8*, 397. 7ème Série.

(26) Lazzeretti, P. in *Progress in Nuclear Magnetic Resonance Spectroscopy*; Emsley, J. W., Feeney, J., Sutcliffe, L. H., Eds.; Elsevier: Amsterdam, 2000; Vol. 36, p 188.

(27) Gomes, J. A. N. F.; Mallion, R. B. *Chem. Rev.* **2001**, *101*, 1349–1383.

(28) Lazzeretti, P. *Phys. Chem. Chem. Phys.* **2004**, *6*, 217–223.

(29) Zanasi, R. *J. Chem. Phys.* **1996**, *105*, 1460–1469.

(30) Lazzeretti, P.; Malagoli, M.; Zanasi, R. Technical Report on Project "Sistemi informatici e calcolo parallelo"; Research Report 1/67, CNR; 1991.

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

(32) Woon, D.; Dunning, T. H., Jr. *J. Chem. Phys.* **1995**, *103*, 4572–4585.

(33) Peterson, K. A.; Dunning, T. H., Jr. *J. Chem. Phys.* **2002**, *117*, 10548–10560.

(34) Becke, A.D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

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

(36) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *GAUSSIAN03*; Gaussian Inc.: Wallingford, CT, 2004.

(37) Faglioni, F.; Ligabue, A.; Pelloni, S.; Soncini, A.; Viglione, R. G.; Ferraro, M. B.; Zanasi, R.; Lazzeretti, P. *Org. Lett.* **2005**, *7*, 3457–3460.

(38) (a) Casado, J.; Miller, L. L.; Mann, K. R.; Pappenfus, T. M.; Higuchi, H.; Ortí, E.; Milián, B.; Pou-Amérigo, R.; Hernández, V.; López-Navarrete, J. T. *J. Am. Chem. Soc.* **2002**, *124*, 12380–12388. (b) Casado, J.; Pappenfus, T. M.; Mann, K. R.; Ortí, E.; Viruela, P. M.; Milián, B.; Hernández, V.; López Navarrete, J. T. *ChemPhysChem.* **2004**, *5*, 529–539.

(39) Milián, B.; Ortí, E.; Hernández, V.; López Navarrete, J. T.; Otsubo, T. *J. Phys. Chem. B* **2003**, *107*, 12175–12183.

(40) (a) Ortiz, R. P.; Casado, J.; Hernández, V.; López Navarrete, J. T.; Ortí, E.; Viruela, P. M.; Milián, B.; Hotta, S.; Zotti, G.; Zecchin, S.; Vercelli, B. *Adv. Funct. Mater.* **2006**, *16*, 531–536. (b) Ponce Ortiz, R.; Casado, J.; Hernández, V.; López Navarrete, J. T.; Viruela, P. M.; Ortí, E.; Takimiya, K.; Otsubo, T. *Angew. Chem., Int. Ed.* **2007**, *46*, 9057–9061.

(41) (a) Merz, A.; Ellinger, F. *Synthesis* **1991**, 462–464. (b) Henze, O.; Feast, J.; Gardebien, F.; Jonkheijm, P.; Lazzaroni, R.; Leclère, P.; Meijer, E. W.; Schenning, A. P. *J. Am. Chem. Soc.* **2006**, *128*, 5923–5929. (c) Yoon, M.-H.; Facchetti, A.; Stern, C. E.; Marks, T. J. *J. Am. Chem. Soc.* **2006**, *128*, 5792–5801.

(42) Sato, N.; Mazaki, Y.; Kobayashi, K.; Kobayashi, T. *J. Chem. Soc., Perkin Trans. 2* **1992**, 765–770.

(43) Krygowski, T. M.; Cyranski, M. K. *Chem. Rev.* **2001**, *101*, 1385–1419.

(44) Cuesta, I. G.; Soriano, R.; Sánchez de Merás, A.; Lazzeretti, P. *Mol. Phys.* **2005**, *103*, 789–801.

(45) Cuesta, I. G.; Sánchez de Merás, A.; Pelloni, S.; Lazzeretti, P. *J. Comput. Chem.* **2009**, *30*, 551–564.

CT900127M

JCTC Journal of Chemical Theory and Computation

Multicore Parallelization of Kohn−Sham Theory

Christopher J. Woods, Philip Brown, and Frederick R. Manby*

Centre for Computational Chemistry, School of Chemistry, University of Bristol, Bristol BS8 1TS, U.K.

Received March 23, 2009

Abstract: A multicore parallelization of Kohn−Sham theory is described, using standard commodity multisocket and multisocket/multicore shared-memory processors. Near-linear scaling of the parallel parts of the code was observed up to the maximum of sixteen cores that were available for benchmarking, and an order of magnitude reduction in run time was achieved running using sixteen threads on a quad-socket quad-core Xeon system. The speed-ups achieved using multisocket/multicore processors were competitive with those achieved using numerical accelerator cards.

Introduction

Recent advances in method development, computing technology, and algorithm design have allowed electronic structure theory methods to be applied routinely to large biological molecules.[1−3] However, *ab initio* electronic structure methods are not yet widely used in biomolecular simulations for drug design or computational enzymology. The computational expense of *ab initio* calculations of biomolecular systems is such that access to large high performance computing (HPC) facilities is typically required. Drug design or computational enzymology simulations are commonly performed using commodity computing resources (e.g., desktop computers or Beowulf clusters), and so most calculations rely on inexpensive semiempirical methods, such as AM1,[4] PM3,[5] or tight binding.[6]

Recent changes in the way that commodity processors are designed have made it increasingly difficult for programmers to extract the full double-precision computing power that is available. In this paper we describe a parallel implementation of Kohn−Sham density functional theory (DFT)[7,8] optimized for modern commodity processors and fully benchmark this port to assess whether it is practical to use widely available computer hardware to run DFT simulations on biomolecular systems routinely.

Numerical Accelerators and Multicore Processors

Moore's law—devised in 1965—states that the number of transistors that can be placed inexpensively into an integrated circuit doubles every two years.[9] This observation has remained true for over forty years, and for most of this time, as the number of transistors in the processor doubled, so too did the clock speed. As clock speeds increased, each generation of processor became capable of performing more floating point operations per second, and so scientific applications automatically ran more quickly.

Recently, there has been a sea change. While transistor counts are still doubling, problems of managing power consumption and heat dissipation mean that the processor clock speed of commodity processors has remained static since around 2004 or indeed has even been falling. The extra transistors have been used to provide larger on-chip memory caches or to add extra processor units (called cores). Thus commodity dual-core processors (which can run two independent threads of execution simultaneously per processor) became available from 2005 and quad-core from 2007. Oct-core processors are likely to become available in 2009[10] (note that this follows the lead of the development of HPC processors, where dual-core chips, such as the IBM POWER4, have been available since 2001 and multisocket symmetric multiprocessor [SMP] nodes are common). Many existing scientific applications that were developed for commodity processors were designed to use only a single thread of execution. They can thus only use a single core of the processor and cannot automatically take advantage of any additional cores that are available. Thus, the automatic and dramatic reduction in run-times for calculations on commodity hardware is no longer assured.

* Corresponding author e-mail: fred.manby@bris.ac.uk.

10.1021/ct900138j CCC: $40.75
Published on Web 06/24/2009

One solution to this problem is provided by numerical accelerators. In these processor chips the large numbers of available transistors are arranged to create hundreds of minicores, such that large numbers of floating point operations can be performed in parallel. For example, general purpose graphic processing units (GPGPUs) evolved from 3D graphics processors and can perform hundreds of floating point operations simultaneously within each clock cycle. Dedicated numerical accelerators have also been developed, such as the ClearSpeed CSX600 and CSX700 chips, each of which can also perform hundreds of floating point operations in parallel.[11] While these accelerators provide large amounts of computing power, algorithms must be redesigned to fit the paradigm of performing hundreds or thousands of independent parallel computations. This requires significant effort, but the reward can be dramatic reductions in the run time of the calculation.

There has been significant interest and success in porting *ab initio* quantum chemistry programs to numerical accelerators.[12−18] Recently, we demonstrated a port of the DFT code from the Molpro quantum chemistry package[19] that was parallelized for ClearSpeed numerical accelerator cards.[13] In this work the Coulomb problem was reformulated to be dominated by numerical quadrature, and, as a result, good scaling was found over the 2304 processor elements available in a ClearSpeed CATs system. This port was capable of speeding up DFT calculations of medium-sized (30−50 atom) molecules by over an order of magnitude.[13] Calculations on molecules of about this size are needed for QM/MM calculations on biomolecular systems.[1]

The algorithms developed to port the DFT calculation to ClearSpeed are generally applicable to any processing platform that is capable of operating on multiple double-precision values in parallel. Commodity processors are now capable of this, both via vector instruction sets (such as SSE for Intel or AMD processors[20]) and by providing access to multiple cores. Dual-core or quad-core desktop processors are now readily available, and modern Beowulf computer clusters are now being built from large numbers of multi-socket/multicore processors. For example, a dual-socket/quad-core machine, here called a dual-quad, contains two quad-core processors that both share the same memory address space; on such a system eight threads of execution can run efficiently in parallel. If SSE2 is available, then a dual-quad system can perform sixteen double-precision operations at a time. Comomodity oct-core processors are likely to become readily available some time over the next year,[10] and it is not unreasonable to expect that commodity quad-oct (which could operate on 64 doubles at once) or even oct-oct platforms (128 doubles at once) will be marketed.

An advantage of these platforms over accelerators, in addition to their wide availability, is that they can be programmed using long-established and portable languages, such as OpenMP and MPI. This makes maintenance and porting of the code as processors evolve, and the number of cores increase, more straightforward. Indeed, there is a long history of parallelizing quantum chemical programs over HPC multisocket and multicore supercomputers. Through a combination of MPI and OpenMP, efficient scaling over hundreds or thousands of processor cores is readily achievable.[21−28]

One method of taking advantage of multicore/multisocket systems is to completely design the quantum chemical program from the ground up to use multiple threads of execution. For example, this is the approach taken by the PQS quantum chemical program,[29] that was developed since 1998 to use the multiple processors available in commodity Beowulf clusters. The ONETEP[30,31] DFT program also takes this approach, being designed from the start to run in parallel on a range of different architectures, including commodity clusters. An alternative approach, and the one we adopt in this paper, is to identify the computationally demanding bottlenecks of the calculation and to adapt just those to run over multiple cores. Parallelization of the bottlenecks of the calculation can be an effective strategy; for example, Kleinschmidt et al.[32] parallelized the matrix-vector multiplication which was the most time-consuming part of their direct multireference configuration interaction (MRCI) code. They achieved good scaling over commodity multicore processors but noted that the memory bandwidth to the processor became a significant bottleneck, particularly as the number of cores per processor increased.

Theory

DFT has two main bottlenecks when applied to 30−50 atom systems: the evaluation of the Coulomb matrix

$$J_{\alpha\beta} = \sum_{\gamma\delta} \gamma_{\gamma\delta}(\alpha\beta|\gamma\delta) \tag{1}$$

and the numerical quadrature used to evaluate the exchange-correlation contribution to the Fock matrix

$$V^{\mathrm{xc}}_{\alpha\beta} = \int \mathrm{d}\mathbf{r} v^{\mathrm{xc}}(\mathbf{r})\chi_\alpha(\mathbf{r})\chi_\beta(\mathbf{r}) \approx \sum_\lambda w_\lambda v^{\mathrm{xc}}_\lambda \chi_{\alpha\lambda}\chi_{\beta\lambda} \tag{2}$$

Here, and throughout, we use the notation $(\cdot|\cdot)$ to denote a two-electron repulsion integral (ERI), so for example

$$(\alpha\beta|\gamma\delta) = \int \mathrm{d}\mathbf{r}_1 \int \mathrm{d}\mathbf{r}_2 \frac{\chi_\alpha(\mathbf{r}_1)\chi_\beta(\mathbf{r}_1)\chi_\gamma(\mathbf{r}_2)\chi_\delta(\mathbf{r}_2)}{r_{12}} \tag{3}$$

The numerical quadrature runs over the points $\mathbf{r}_\lambda$, with weights w_λ, and $v^{\mathrm{xc}}_\lambda = v^{\mathbf{xc}}(\mathbf{r}_\lambda)$, and $\chi_{\alpha\lambda} = \chi_\alpha(\mathbf{r}_\lambda)$.

It is straightforward to parallelize numerical quadrature by distributing batches of integration points between processing cores. The Coulomb term is more problematic. Direct calculation of the Coulomb contribution requires four-index ERIs $(\alpha\beta|\gamma\delta)$. Density fitting[33,34] can be used to treat the Coulomb contribution using just two- and three-index integrals. The conventional Kohn−Sham density

$$\rho(\mathbf{r}) = \sum_{\alpha\beta} \gamma_{\alpha\beta}\chi_\alpha(\mathbf{r})\chi_\beta(\mathbf{r}) \tag{4}$$

is approximated using an auxiliary basis of fitting functions, Ξ_A

$$\tilde{\rho}(\mathbf{r}) = \sum_B d_B \Xi_B(\mathbf{r}) \quad (5)$$

Here d_B are the density-fitting coefficients, calculated by minimizing the Coulomb self-energy of the fitting residual[34]

$$\Delta = \frac{1}{2}(\rho - \tilde{\rho}|\rho - \tilde{\rho}) \quad (6)$$

The Coulomb Fock contribution becomes

$$J_{\alpha\beta} = (\alpha\beta|\rho) \approx (\alpha\beta|\tilde{\rho}) = \sum_B d_B(\alpha\beta|B) \quad (7)$$

However, this requires the evaluation of a large number of three-index integrals. To achieve the very fine-grained parallelism required for the port to the ClearSpeed accelerator, we introduced a quadrature-based Coulomb method.[13] The details can be found elsewhere,[13] but for convenience the key points of the so-called grid-based density fitting Poisson method (GDFP) for the Coulomb problem are as follows:

1) The density is expanded in a basis that contains a small number of standard, atom-centered basis functions Ξ_A, and many Poisson functions of the form $\nabla^2\Xi_B$.[35,36]

2) Three-index Coulomb integrals involving Poisson functions become short-ranged overlap-like integrals, which are evaluated by quadrature.

3) The remaining, small number of conventional Coulomb integrals are evaluated analytically.

By using quadrature, both the grid-based Coulomb method and the treatment of exchange-correlation fall into the class of embarrassingly parallel problems, and it is expected that a parallel implementation for multisocket/multicore commodity processors will result in perfect linear scaling for these steps.

Bottlenecks

There are three bottlenecks in GDFP Kohn−Sham theory for our target system sizes. These, in order of decreasing time, are

1) The evaluation of the conventional Kohn−Sham density at each grid point, ρ_λ.

2) The quadrature-based evaluation of the exchange-correlation contributions to the Fock matrix.

3) The quadrature-based GDFP evaluation of the Coulomb contribution.

The first bottleneck is the calculation of the density on the grid, ρ_λ, which is evaluated during each Kohn−Sham iteration based on density matrix, $\gamma_{\alpha\beta}$, and the representation of each orbital at each grid point, $\chi_{\alpha\lambda}$

$$\rho_\lambda = \sum_{\alpha\beta} \gamma_{\alpha\beta}\chi_{\alpha\lambda}\chi_{\beta\lambda} \quad (8)$$

The density on the grid is used to evaluate the exchange-correlation potential on each grid point, v_λ^{xc}, which is used to calculate the exchange-correlation contributions to the Fock matrix

$$V_{\alpha\beta}^{\mathrm{xc}} = \sum_\lambda w_\lambda v_\lambda^{\mathrm{xc}} \chi_{\alpha\lambda}\chi_{\beta\lambda} \quad (9)$$

(Similar terms arise for gradient-corrected functionals).

The density on the grid is also used in the GDFP-based Coulomb numerical quadrature. The quadrature is used to evaluate the contribution to the Coulomb Fock matrix that arises from the Poisson part of the fitted density

$$J_{\alpha\beta} \leftarrow \sum_\lambda w_\lambda \chi_{\alpha\lambda}\chi_{\beta\lambda}\left[\sum_B d_B \Xi_{B\lambda}\right] \quad (10)$$

Program Design

The three bottlenecks in the calculation each involve evaluation of values at all of the large number of independent quadrature grid points. It is therefore natural to design the program so that the outer loop is over grid points, and thus parallelization is achieved by using vectorization to process multiple grid points per processor cycle and by dividing batches of grid points between processor cores. Vectorization of double-precision operations on x86 and x86-64 processors such as Intel Xeon or AMD Opteron is achieved by using version 2 of the SSE instruction set (SSE2),[20] which can be generated automatically by a good compiler, or which are available directly to the programmer in the C and C++ languages via intrinsics (in the emmintrin.h header file).

Dividing batches of grid points between processor cores can be achieved using OpenMP. OpenMP provides a set of compiler pragmas in C, C++, or Fortran, which the programmer can use to delineate parallel and serial parts of the code. OpenMP was first published in 1997, at which time commodity multisocket or multicore platforms were not available. Ports of code to OpenMP were thus limited to specialist hardware, such as the Silicon Graphics Origin 2000, on which a parallel port of part of the Gaussian 98 quantum chemical program using OpenMP was demonstrated.[37]

OpenMP is easy to use, but efficient parallelization requires that the programmer ensures that data synchronization between the serial and parallel parts of the code occurs infrequently and that each parallel batch computes as much as possible using thread-local data. Thread-local data are data which exist in memory that is owned and accessed by only a single thread of execution, and so no synchronization is needed between threads when reading or writing that data. This requires tight control of memory read/write operations, and of which data are thread-local and which are global. This is best achieved using scoping, where variables that represent data that are thread-local exist only within the parallel regions of the code.

C++ is a computer language that provides scoped variable declarations (e.g., a variable can exist only within a loop, and is not accessible outside the loop). Because of this, and because C++ provides direct access to SSE2, it was decided to rewrite the computationally expensive kernels of the Molpro DFT implementation in C++.

Vectorization Using SSE2. Vectorization of the DFT calculation was achieved by batching together grid points into vectors. A C++ class, called MultiDouble, was created to represent a vector of doubles within the code. An ancillary

class, called MultiPoint, was then created to represent a vector of grid points. SSE2 intrinsics were then added to the member functions of MultiDouble to allow hardware vectorization to be used where it was available. The SSE2 code was hidden behind conditional statements, so the code is portable to platforms where SSE2 is not available.

Parallelization Using OpenMP. Each of the three bottlenecks is written as a loop over MultiDouble batches of grid points. All application-global read-only data are accessed via read-only pointers, thereby ensuring that synchronization between threads is not necessary to access that data. Each thread, when entering a parallel region, creates a local workspace in which to assemble the intermediates during the calculation. This local workspace is accessed via variables local to the scope of the parallel region and is cleared at the end of each parallel section. Synchronization between threads occurs only when thread-local contributions to the exchange-correlation and Coulomb parts of the Fock matrix are summed.

Screening Orbitals and Caching Data. All three bottlenecks of the calculation involve loops over all pairs of atomic orbitals and over each grid point. For large molecules, screening small contributions of orbitals on remote grid points reduces the scaling of computational work from cubic to linear with molecular size.

The representation of each orbital at each grid point, and the decision of whether or not an orbital should be screened, is a constant throughout the DFT calculation and does not need to be recalculated at each Kohn−Sham iteration. However, the memory required to store the representation of the orbitals on the grid scales with the number of orbitals times the number of grid points and can take several gigabytes for 30−50 atom molecules and commonly used basis sets. This memory requirement is naturally reduced by the application of screening, as only the non-negligible orbitals at each grid point need to be saved. The memory requirement then scales linearly with molecule size for sufficiently large cases.

Screening has been implemented on the basis of MultiDouble batches of grid points. Each MultiDouble batch is assigned a unique identification number, and an array of MultiDouble objects is then used to store the representation of only the non-negligible orbitals for this batch of grid points. This is then cached using a dictionary container, indexed using the unique identification number. A dictionary cache is used to allow the program to gracefully cope with cases where there is insufficient memory to store all of the orbitals on the grid. In these cases, as many MultiDouble batches that can fit into the cache are saved, while the remainder are recomputed for each Kohn−Sham iteration as needed.

Results

Benchmarking System and Platforms. The OpenMP DFT code was benchmarked by calculating the DFT B-LYP single point energy of an analogue of the neuraminidase inhibitor, DANA (Figure 1). This molecule is typical of those proposed during rational drug design, and it consists of 34 atoms, of which 17 are hydrogen.

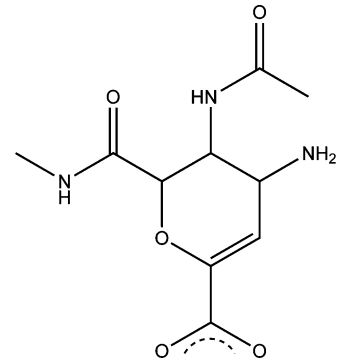


Figure 1. DANA, the test molecule used to benchmark the OpenMP DFT code.

Table 1. Four Multisocket/Multicore Platforms Used To Benchmark the OpenMP DFT Code

vendor	model	speed/ GHz	platform	memory/ GB	cores	peak[a]
AMD	Opteron 2218	2.59	dual−dual	8	4	20.7
Intel	Xeon E5472	2.80	dual-quad	8	8	44.8
AMD	Opteron 8220	2.80	oct-dual	32	16	89.6
Intel	Xeon X7350	2.93	quad−quad	32	16	93.8

[a] Peak double-precision performance in GFLOPS.

The DFT energy was calculated using three different basis sets, 6-31G*, cc-pVDZ, and aug-cc-pVDZ, to investigate how well the code scaled as the computational cost for each grid point increased. Density fitting was performed using a fitting set optimized for the cc-pVTZ atomic orbital basis.[38] For a carbon atom, the set consists of 1s1p1d standard Gaussian functions and 10s7p5d2f1g Poisson functions.

The code was benchmarked on four different platforms (see Table 1), chosen to represent a balance between AMD and Intel processors, and spread the range from two-socket to eight-socket platforms and dual-core to quad-core processors. To ensure that comparisons between the platforms were not biased by different compiler or OpenMP implementations, the C++ DFT code was compiled on all platforms using GCC 4.2,[39] with the libgomp[40] library provided with GCC.

Selecting a Vector Size. First, the sensitivity of performance to the size of the MultiDouble vector was investigated. While SSE2 hardware vectors are currently limited to just two doubles (as 128-bit registers are used), the MultiDouble class provides a software vector, the size of which can be controlled using a compile-time constant. Using a large vector is likely to improve the speed of the code, as it should allow for efficient pipelining. However, large vectors require more memory, which can have a negative impact on cache management and performance. To investigate this, the MultiDouble vector size was varied in powers of two from 8 to 128 doubles, and the impact on the speed of each of the three bottlenecks was measured (Figure 2).

While the exchange-correlation and numerical Coulomb bottlenecks benefit from larger vector sizes, the density bottleneck performs best using a vector size of 16 or 32. It is also clear that while the performance of the three bottlenecks is roughly equal using both the 6-31G* and cc-pVDZ basis sets, the performance of the exchange-correlation

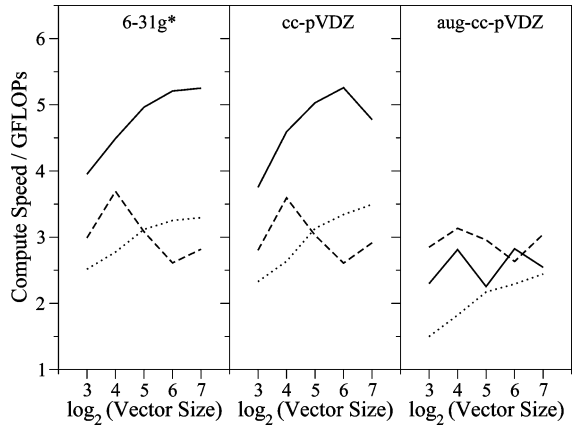


Figure 2. Average speed of calculation for the three main bottlenecks from each Kohn−Sham iteration for different MultiDouble vector sizes, for three different basis sets. Calculated using four threads on a dual−dual Opteron 2218 (solid: exchange-correlation contributions, dashed: density, dotted: numerical Coulomb).

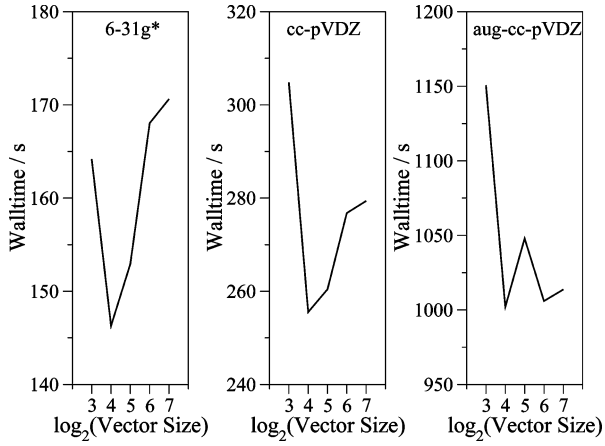


Figure 3. Total run time (wallclock) for different MultiDouble vector sizes, for three different basis sets. Calculated using four threads on a dual−dual Opteron 2218.

and density code is significantly reduced using aug-cc-pVDZ. This suggests that the increased memory-transfer requirements of this larger basis set may be stalling the calculation.

In addition to looking at the speed of the bottlenecks, the total run time of the calculation was also investigated as a function of vector size (Figure 3). From this it is clear that the preference of the density calculation for smaller vectors out-competes the preference for the other bottlenecks for larger vectors, and so the best performance is obtained using a vector size of 16 or 32. This observation was checked and found to be true on all of the other benchmark platforms, and so a vector size of 16 was chosen for all further tests.

Screening and Caching Orbitals on the Grid. The performance gains of orbital caching and screening were assessed. First, the total run time of the DFT calculation for each of the three basis sets was measured on the dual−dual Opteron 2218, both with the cache enabled and with the cache disabled. Using a dictionary-based cache significantly improves the run time, as shown in Table 2. The run time is reduced by 40−50% for the 6-31G* and cc-pVDZ calculations, but there is a comparatively smaller reduction for aug-cc-pVDZ. This is because the time required to perform each iteration (63 s) is much larger than the time required to build the orbitals on the grid (7 s). In addition, while look-up of the orbitals from the cache remains fast (taking just 0.1 s per iteration), the actual transport of the data for the orbitals on the grid from main memory appears to be hitting the bandwidth limit. This is evidenced by a reduction in computational efficiency from 3.3 GFLOPs to 2.8 GLOPs for the density calculation when the cache is used, compared to when the cache is disabled. Despite this, the use of the cache is still beneficial. Building the orbitals on the grid takes between 5−7 s, and, without the cache, this must be repeated for each Kohn−Sham iteration.

Table 2. Required Size of the Orbital Cache in Megabytes for Each of the Three Basis Sets, Together with the Run Time (Wallclock), in s, of the DFT Calculation with Both the Cache Enabled and Cache Disabled[a]

basis set	cache size	cache enabled	cache disabled
6-31G*	1819	146	296
cc-pVDZ	2128	256	409
aug-cc-pVDZ	3060	1002	1066

[a] Calculated using four threads on a dual−dual Opteron 2218.

Table 3. Time Required To Build the Orbital Cache, in s, Compared to the Total Run Time (Wallclock), in s, for the DFT Calculation, for Three Basis Sets[a]

basis set	threshold	% screened	build	run
6-31G*	10^{-15}	49	5.2	146
	0	9	5.9	306
cc-pVDZ	10^{-15}	47	5.5	256
	0	9	6.2	407
aug-cc-pVDZ	10^{-15}	43	7.3	1066
	0	8	7.6	1722

[a] Calculated using four threads on a dual−dual Opteron 2218. Times with screening (threshold = 10^{-15}) and without screening (threshold = 0) are shown, together with the percentage of orbital/grid batches that are less than or equal to the screening threshold.

The cost of evaluating orbitals on the grid can be reduced through a per-orbital/per-grid point screening algorithm. At the time of building the orbitals, a screening test is performed which compares the absolute value of each orbital at each grid point with a screening threshold (10^{-15}). The screening threshold was chosen to ensure that there was no detectable numerical difference between the screened and unscreened energies, but obviously greater savings could be made at the expense of some numerical precision. If the density is less than the screening threshold for each grid point for a MultiDouble batch (16 grid points), then this orbital is discarded for this batch of grid points. This slightly reduces the cost of building the orbitals on the grid, but, more importantly, this test absolutely removes all insignificant orbitals at each batch of grid points. As the subsequent parts of the calculation scale with the square of the number of significant orbitals times the number of grid points, the cost of building all orbitals at all grid points is more than recovered by the savings in the later stages. This is demonstrated in Table 3, which shows that the total time paid to screen and build the orbitals on the grid is significantly less than the time saved when performing screening.

Table 4. Average Speed, in Gigaflops, of the Three Main Bottlenecks over Each Kohn−Sham Iteration for Three Different Basis Sets, Both with Enabled and Disabled Manual SSE2 Code[a]

bottleneck	SSE enabled	6-31G*	cc-pVDZ	aug-cc-pVDZ
exchange-correlation	yes	13.8	14.0	11.0
	no	13.9	14.0	11.2
density	yes	11.7	10.9	9.6
	no	8.6	8.3	7.7
numerical Coulomb	yes	8.7	9.6	8.3
	no	8.3	9.0	7.9

[a] Calculations performed using eight threads of a dual-quad Xeon E5472.

Table 5. Total Run Time (Wallclock), in s, for the DFT Calculation Using Three Different Basis Sets, Both with Enabled and Disabled Manual SSE2 Code[a]

basis set	SSE enabled	1 core	8 cores
6-31G*	yes	308	59
	no	357	62
cc-pVDZ	yes	550	91
	no	639	101
aug-cc-pVDZ	yes	1606	290
	no	1854	314

[a] Calculations performed using either a single thread or eight threads of a dual-quad Xeon E5472.

Benchmarking Manual SSE. All of the benchmark processors support the use of SSE2 vector instructions. Manual SSE2 code can be easily enabled or disabled using a preprocessor flag within the implementation of the MultiDouble vector class. Manual SSE2 code is used extensively in the density calculation, while it is used sparingly when calculating the exchange-correlation contributions (for the reason, see the Appendix). The impact on enabling manual SSE2 on computational efficiency is shown in Table 4.

While manual SSE2 has little effect on the raw speed of the exchange-correlation and numerical Coulomb bottlenecks, the calculation of the density is the major bottleneck that dominates the total run time (approximately 40%). The extra 2−3 GFLOPs that is obtained by enabling manual SSE2 therefore has a large impact on the run time, particular when using a small number of threads (see Table 5). Because the use of manual SSE2 always reduced the run time, it was enabled for all further tests.

Scaling of the Bottlenecks. The previous benchmarks had validated the benefits of using a vector class, caching the representation of the orbitals on the grid, performing grid-based screening, and using manually coded SSE. The final part of the design to test was the use of OpenMP to parallelize the loops over MultiDouble batches of grid points. First, the scaling of the computational speed of each bottleneck was measured as a function of the number of threads. The speed was measured during each Kohn−Sham iteration of a full calculation and then averaged. The dependence of speed on the number of threads is shown in Figures 4, 5, and 6 for exchange-correlation contributions, calculation of the density, and numerical Coulomb respectively.

These results show that, as expected, the use of quadrature has resulted in linear scaling for each of the bottlenecks, even up to 16 threads. However, while this is seen for the 6-31G*

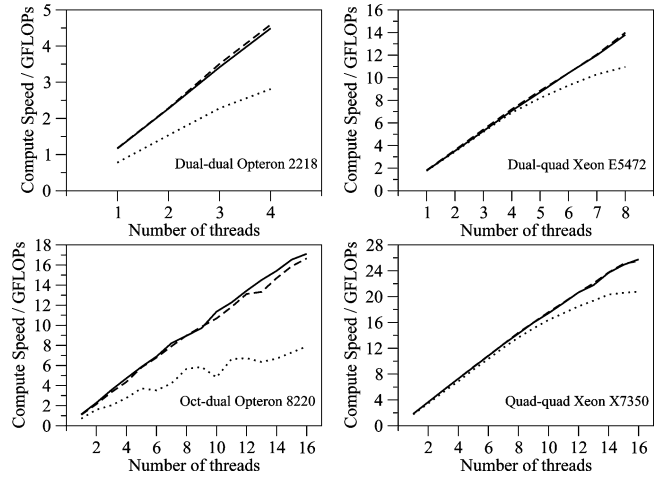


Figure 4. Average speed of the calculation of the exchange-correlation contributions to the Fock matrix, for three different basis sets (solid: 6-31G*, dashed: cc-pVDZ, dotted: aug-cc-pVDZ) as a function of the number of threads of execution on each of the four benchmark platforms.

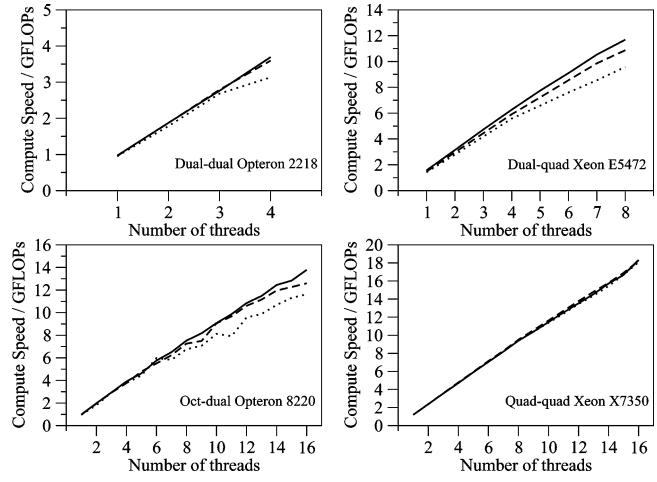


Figure 5. Average speed of the construction of the density, for three different basis sets (solid: 6-31G*, dashed: cc-pVDZ, dotted: aug-cc-pVDZ) as a function of the number of threads of execution on each of the four benchmark platforms.

and cc-pVDZ calculations, the limits of memory bandwidth mean that the scaling of the bottlenecks of the aug-cc-pVDZ calculations quickly tops out. Here, the quad−quad Xeon platform performs best, with linear scaling up to 8 threads for the exchange-correlation and numerical Coulomb bottlenecks and with linear scaling maintained up to all 16 threads when evaluating the density. This contrasts with the dual−dual and oct-dual Opteron platforms, which both suffer from substantially degraded performance of the exchange-correlation and numerical Coulomb bottlenecks (only 8% of the peak performance of the oct-dual Opteron is used for the 16 threads aug-cc-pVDZ exchange-correlation calculation, compared to 22% using 16 threads on the quad−quad Xeon). When making this comparison, it must be remembered that the Opteron processors used in these benchmarks are older designs than the Xeons, and it is likely that the latest AMD processors are likely to perform better in this regard.

In general these results show that the code performs well across each of the benchmark platforms. The evaluation of

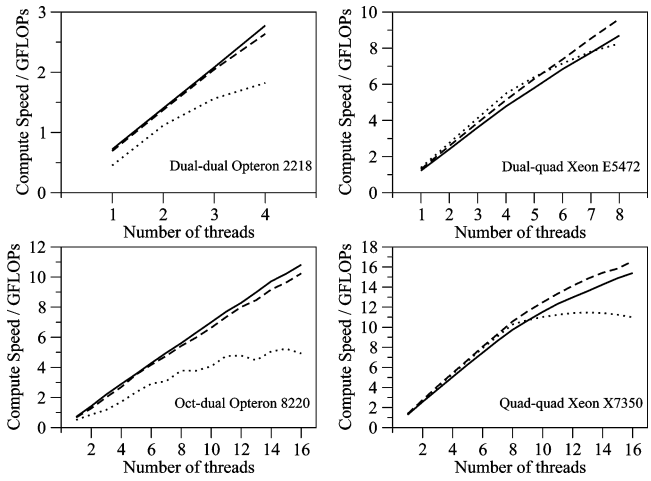


Figure 6. Average speed of the calculation of the Coulomb numerical quadrature, for three different basis sets (solid: 6-31G*, dashed: cc-pVDZ, dotted: aug-cc-pVDZ) as a function of the number of threads of execution on each of the four benchmark platforms.

the exchange-correlation components is particularly efficient on the Xeon platforms, while the dual-quad Xeon E5472 is consistently the most efficient platform tested. There does, however, seem to be some room for improvement in the efficiency of the code used to evaluate the Coulomb numeric quadrature contributions, with this region of the code having consistently the lowest performance of the three bottlenecks and having the poorest scaling and greatest performance degradation when using the large aug-cc-pVDZ basis set. One of the reasons for the comparatively poor performance of this bottleneck is that it is essentially just a product of all pairs of orbitals with all Poisson orbitals, at each grid point (see eq 10). This requires large amounts of data (all orbitals and Poisson orbitals), but very little actual computation (just three multiplies on four values loaded from four different regions of memory, in a naive implementation). It is thus not surprising that this memory-access-heavy but computationally light bottleneck has the poorest scaling and the biggest performance degradation with increasing basis size.

Discussion

Benchmarking of the OpenMP DFT implementation has shown that the code performs well up to the sixteen cores that were available for testing. Ultimately though, the user of the code is not really interested in the performance characteristics of parts of the code. Instead, their primary interest is likely to be the total run time of the calculation. This is shown in Table 6, where the total run time of the DFT calculations on each of the benchmark platforms is reported using both one thread and using the maximum number of threads available on each platform.

The original aim of this work was to reduce the run time to the extent that DFT calculations on commodity processors are sufficiently fast to allow practical applications in QM/MM free energy simulations. This has been achieved, with calculations using 6-31G* or cc-pVDZ basis sets reduced to less than 100 s for the eight and sixteen core machines. Indeed, a run time of just 46 s is required using the 6-31G* basis set on the quad−quad Xeon. QM/MM free energy calculations based on a recently developed Monte Carlo method[41] require of the order of 1000−2000 QM energy evaluations run in sequence to calculate a converged QM/MM relative free energy. This would take approximately 13−26 h using the quad−quad Xeon, using a 6-31G* basis set, or 1−2 days using a cc-pVDZ basis set on the more readily available dual-quad Xeon. This is well within the time scale necessary to make these calculations practical for drug-discovery type protein−ligand relative binding free energy simulations or for use in computational enzymology.

Table 6. Total Run Time (Wallclock), in s, for the DFT Calculation Using Three Different Basis Sets, on the Four Different Benchmark Platforms, As Calculated Using Just One Core, and Using All Available Cores[a]

platform	basis set	1 core	all cores
Opteron	6-31G*	493	146
2218	cc-pVDZ	894	256
dual−dual	aug-cc-pVDZ	3371	1002
Opteron	6-31G*	495	59
8220	cc-pVDZ	916	94
oct-dual	aug-cc-pVDZ	3364	369
Xeon	6-31G*	308	59
E5472	cc-pVDZ	550	91
dual-quad	aug-cc-pVDZ	1606	290
Xeon	6-31G*	345	46
X7350	cc-pVDZ	599	63
quad−quad	aug-cc-pVDZ	1751	187

[a] E.g. all eight cores of the dual-quad Xeon.

Table 7. Comparison of the Run Time (Wallclock), in s, for the Different Versions of Molpro on a Dual−Dual Opteron 2218 with Attached ClearSpeed CATs[a]

version	6-31G*	cc-pVDZ	aug-cc-pVDZ
Fortran	519	801	2070
OpenMP1	493	894	3371
OpenMP4	146	256	1002
ClearSpeed1	283	326	429
ClearSpeed12	39	46	86

[a] The original serial (Fortran) DFT implementation is compared to the OpenMP implementation on one core (OpenMP1), the OpenMP implementation on all four cores (OpenMP4), the ClearSpeed implementation using one ClearSpeed card (ClearSpeed1), and the ClearSpeed implementation using all twelve ClearSpeed cards (ClearSpeed12).

Finally, the performance of this OpenMP DFT code must be compared against the other implementations. Such a comparison is not entirely fair, due to differences in the implementation of screening, the use or not of numerical quadrature for the Coulomb calculation, a different reliance on and efficiency of the BLAS library, etc. To minimize the effect of these difference, this comparison was made using only the dual−dual Opteron (as it was this system that was connected to the ClearSpeed CATs). Table 7 shows the comparison of the total run time for the DFT calculations of the OpenMP code using one thread and four threads against the run time using an unmodified serial Molpro executable and against the ClearSpeed-enabled Molpro using one ClearSpeed dual socket CSX600 card ($2 \times 96 = 192$ cores) and using twelve ClearSpeed dual socket CSX600 cards (2304 cores).

Table 8. Total Run Time of the DFT Calculations, in s, Measured Using Different Numbers of ClearSpeed Cards Available via a ClearSpeed CATs Attached to a Dual−Dual Opteron 2218

number of cards	6-31G*	cc-pVDZ	aug-cc-pVDZ
1	283	326	429
2	150	173	231
3	107	122	177
4	89	102	149
5	49	59	130
6	46	55	122
7	43	51	90
8	41	49	88
9	40	47	87
10	39	46	86
11	39	47	86
12	39	46	86

This comparison raises two interesting points. First, the speed of the OpenMP implementation using one thread is competitive with standard Molpro, at least for the smaller basis sets. For the aug-cc-pVDZ basis set, standard Molpro is significantly faster than the OpenMP implementation, probably because in serial the numerical Coulomb calculation is not competitive with analytic integral evaluation. The second interesting point is that using four threads in the OpenMP code is quicker than using one ClearSpeed card, despite the higher peak double-precision performance (66 GFLOPs for one card versus 41 GFLOPs for the dual−dual Opteron). Again, this is only observed for the smaller basis sets. The ClearSpeed port is most efficient for larger basis sets, hence why the performance degradation is significantly lower when moving up to aug-cc-pVDZ compared to any of the other implementations.

The observation that the four-thread OpenMP implementation is faster than using a ClearSpeed card does raise an interesting question. For smaller systems it may be more effective to use a multiprocessor/multicore platform as opposed to using a single numerical accelerator. This can be quantified, again by using a slightly unfair comparison, by asking how many ClearSpeed accelerator cards are necessary to reduce the run time so that it is less than the minimum run time on each of the multiprocessor/multicore platforms. For example, the run time for sixteen threads for the cc-pVDZ calculation on the quad−quad Xeon was 63 s (Table 6). The run time using four ClearSpeed cards on the dual−dual Opteron was 102 s (see Table 8). Thus, in this case, a user would be better served by using the quad−quad Xeon. However, the run time using five ClearSpeed cards is 59 s, and so a user would be better served by using the accelerators. The number of ClearSpeed cards necessary to outperform each benchmark platform for each basis set has been calculated and is shown in Table 9. Again, it must be stressed that this is not a completely fair comparison, as the serial parts of the calculation are computed using an Opteron 2218, which was the slowest of the four processors used in this study. Despite this, it is clear that it is only the larger basis sets that benefit from the use of small numbers of accelerator cards and that large numbers of accelerator cards must be used in parallel in order to compete against a multiprocessor/multicore system for drug-sized molecules with commonly used basis sets.

Table 9. Number of ClearSpeed Cards That Must Be Used To Exceed the Speed of the OpenMP Implementation with the Maximum Number of Threads Available on Each of the Four Benchmark Platforms[a]

platform	6-31G*	cc-pVDZ	aug-cc-pVDZ
Opteron 2218	3	2	1
Opteron 8220	5	5	2
Xeon E5472	5	5	2
Xeon X7350	7	5	3

[a] The ClearSpeed calculations ran on a dual−dual Opteron 2218 attached to a ClearSpeed CATs, and this comparison does not account for the slower speed of the serial parts of the code on this system compared to the others.

Given that three of the benchmark systems are faster using 8 or 16 cores than using four ClearSpeed cards (with 768 cores) for the 6-31G* and cc-pVDZ basis sets, questions should be asked as to whether or not more effort should be spent adapting algorithms for existing multiprocessor/multicore architectures. These questions are particularly pertinent now, as the standards for writing code for numerical accelerators and the underlying hardware model of these accelerators have yet to mature. On the other hand, multicore platforms are ubiquitous, and the languages and methods used to target them are well-established and performance-portable. Scientific applications are written over time scales of years to decades and contain large amounts of accumulated knowledge. Rewriting them so that they run efficiently in parallel represents a significant undertaking. Targeting multiprocessor/multicore machines leads to code that can be deployed efficiently across a range of platforms and which is easier to maintain and port over the coming years.

Acknowledgment. We thank the EPSRC for funding this work (EP/F010516/1) and for the Universities of Bristol, Cardiff, and Manchester for providing computer resources. In particular, the authors thank Ian Stewart, Christine Kitchen, and Simon Hood for providing access to, and for their help getting us set up on, the benchmark systems.

Appendix

The use of manual SSE2 in the MultiDouble vectors is only possible if the iteration over grid points occurs in an inner loop (as this MultiDouble is being used to vectorize over grid points). The calculation of the density is a sum over all pairs of orbitals, at all grid points (see eq 8), and so it is most efficient to use the iteration over grid points as the inner of the three loops. In contrast, the calculation of the exchange-correlation contributions is a sum over all grid points of products of all pairs of orbitals. In this case, it is most efficient to place the iteration over grid points as the outer loop, as otherwise poor cache performance is encountered during the iteration over all pairs of orbitals. Poor cache performance arises because the representation of orbitals in memory is arranged in contiguous blocks of 16 doubles (one for each grid point represented by the MultiDouble batch). This means that the distance in memory between adjacent orbitals, commonly called the stride, is 16 doubles. Loops

over pairs of orbitals are therefore inefficient, as the stride between each iteration is larger than a cache line, and thus it is likely that different orbitals will exist on different cache lines. In the original implementation, this lead to poor performance. This was resolved by preceding the loop over pairs of orbitals with a loop over the 16 grid points within a MultiDouble batch, that copied the value of the orbital for that grid point into a temporary array of doubles. This packed the orbitals on a single grid point together into a single contiguous block in memory. In effect, this transposed the memory layout from using grid point as the inner index, to using orbital as the inner index. The cost of transposing memory was found to be negligible (formally scaling linearly with the number of orbitals), while the time saved was significant (as the cost of looping over pairs of orbitals scales quadratically with the number of orbitals). However, because the iteration over grid points was now the outer, not the inner, loop, manual SSE2 could no longer be used.

References

(1) Claeyssens, F.; Harvey, J. N.; Manby, F. R.; Mata, R. A.; Mulholland, A. J.; Ranaghan, K. E.; Schütz, M.; Thiel, S.; Thiel, W.; Werner, H. J. *Angew. Chem. Int. Ed.* **2006**, *45*, 6856–6859.

(2) Scuseria, G. *J. Phys. Chem. A* **1999**, *103*, 4782.

(3) Gogenea, V.; Suárez, D.; van der Vaart, A.; Jr., K. W. M. *Curr. Opin. Struct. Biol.* **2001**, *11*, 217.

(4) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

(5) Stewart, J. J. P. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1.

(6) Frauenheim, T.; Seifert, G.; Elstner, M.; Hajnal, Z.; Jungnickel, G.; Porezag, D.; Suhai, S.; Scholz, R. *Phys. Stat. Sol. (B)* **2000**, *217*, 41.

(7) Kohn, W.; Sham, L. J. *Phys. Rev. A* **1965**, *140*, 1133.

(8) Kohn, W.; Becke, A. D.; Parr, R. G. *J. Phys. Chem.* **1996**, *100*, 12974.

(9) Moore, G. E. *Electronics* **1965**, *38*, 8.

(10) Intel announced an oct-core Xeon in February 2009 at the International Solid State Circuits Conference.

(11) Advance e620 Product Brief. http://www.clearspeed.com (accessed March 11, 2008).

(12) Anderson, A. G.; Goddard, W. A.; Schröder, P. *Comput. Phys. Commun.* **2007**, *177*, 298–306.

(13) Brown, P.; Woods, C. J.; McIntosh-Smith, S.; Manby, F. R. *J. Chem. Theory Comput.* **2008**, *4*, 1620–1626.

(14) Yasuda, K. *J. Comput. Chem.* **2008**, *29*, 334–342.

(15) Yasuda, K. *J. Chem. Theory Comput.* **2008**, *4*, 1230–1236.

(16) Vogt, L.; Olivares-Amaya, R.; Kermes, S.; Shao, Y.; Amador-Bedolla, C.; Aspuru-Guzik, A. *J. Phys. Chem. A* **2008**, *112*, 2049–2057.

(17) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2008**, *4*, 222–231.

(18) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2009**, *5*, 1004–1015.

(19) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M. et al. *MOLPRO, version 2008.1, a package of ab initio programs*; 2008. See http://www.molpro.net (accessed Feb 2009).

(20) Intel C++ Compiler for Linux - 9.X: Intrinsics Reference. http://software.intel.com/en-us/articles/intel-c-compiler-for-linux-9x-manuals (accessed May 28, 2009).

(21) Smith, L.; Kent, P. *Concurrency: Pract. Exper.* **2000**, *12*, 1121–1129.

(22) Shellman, S. D.; Lewis, J. P.; Glaesemann, K. R.; Sikorski, K.; Voth, G. A. *J. Comput. Phys.* **2003**, *188*, 1–15.

(23) Medvedev, D. M.; Goldfield, E. M.; Gray, S. K. *Comput. Phys. Commun.* **2005**, *166*, 94–108.

(24) Hutter, J.; Curioni, A. *Parallel Comput.* **2005**, *31*, 1–17.

(25) Almasi, G.; Bhanot, G.; Chen, D.; Eleftheriou, M.; Fitch, B.; Gara, A.; Germain, R.; Gunnels, J.; Gupta, M.; Heidelberg, P.; Pitman, M.; Rayshubskiy, A.; Sexton, J.; Suits, F.; Vranas, P.; Walkup, B.; Ward, C.; Zhestkov, Y.; Curioni, A.; Andreoni, W.; Archer, C.; Moreira, J.; Loft, R.; Tufo, H.; Voran, T.; Riley, K.; Cunha, J. C. *Euro-Par 2005 Parallel Process.* **2005**, *3648*, 560–570.

(26) Bottin, F.; Leroux, S.; Knyazev, A.; Zérah, G. *Comput. Mater. Sci.* **2008**, *42*, 329–336.

(27) Fan, P.-D.; Valiev, M.; Kowalski, K. *Chem. Phys. Lett.* **2008**, *458*, 205–209.

(28) Rayson, M.; Briddon, P. *Comput. Phys. Commun.* **2008**, *178*, 128–134.

(29) Baker, J.; Wolinski, K.; Malagoli, M.; Kinghorn, D.; Wolinski, P.; Magyarfalvi, G.; Saebo, S.; Janowski, T.; Pulay, P. *J. Comput. Chem.* **2009**, *30*, 317–335.

(30) Mostofi, A.; Haynes, P.; Skylaris, C.-K.; Payne, M. *Mol. Simul.* **2007**, *33*, 551–555.

(31) Skylaris, C.-K.; Haynes, P.; Mostofi, A.; Payne, M. *J. Phys.: Condens. Matter* **2008**, *20*, 064209.

(32) Kleinschmidt, M.; Marian, C.; Waletzke, M.; Grimme, S. *J. Chem. Phys.* **2009**, *130*, 044708.

(33) Baerends, E. J.; Ellis, D. E.; Ros, P. *Chem. Phys.* **1973**, *2*, 41.

(34) Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. *J. Chem. Phys.* **1979**, *71*, 3396.

(35) Manby, F. R.; Knowles, P. J.; Lloyd, A. W. *J. Chem. Phys.* **2001**, *115*, 9144.

(36) Manby, F. R.; Knowles, P. J. *Phys. Rev. Lett.* **2001**, *87*, 163001.

(37) Sosa, C. P.; Scalmani, G.; Gomperts, R.; Frisch, M. J. *Parallel Comput.* **2000**, *26*, 843–856.

(38) Polly, R.; Werner, H.-J.; Manby, F. R.; Knowles, P. J. *Mol. Phys.* **2004**, *102*, 2311–2321.

(39) GNU Compiler Collection 4.2.4. http://gcc.gnu.org (accessed March 21, 2009).

(40) libgomp: The GNU implementation of OpenMP. http://gcc.gnu.org/onlinedocs/libgomp (accessed March 21, 2009).

(41) Woods, C. J.; Manby, F. R.; Mulholland, A. J. *J. Chem. Phys.* **2008**, *128*, 014109.

CT900138J

JCTC Journal of Chemical Theory and Computation

Integrated Continuum Dielectric Approaches To Treat Molecular Polarizability and the Condensed Phase: Refractive Index and Implicit Solvation

Jean-François Truchon,[†,‡] Anthony Nicholls,[§] Benoît Roux,[||] Radu I. Iftimie,[†] and Christopher I. Bayly*[,‡]

Département de chimie, Université de Montréal, C.P. 6128 Succursale centre-ville, Montréal, Québec, Canada H3C 3J7, Merck Frosst Canada Ltd., 16711 TransCanada Highway, Kirkland, Québec, Canada H9H 3L1, OpenEye Scientific Software, Inc., Santa Fe, New Mexico 87508, and Institute of Molecular Pediatric Sciences, Gordon Center for Integrative Science, University of Chicago, 929 East 57th Street, Chicago, Illinois 60637

Received January 15, 2009

Abstract: The idea of using a dielectric continuum inside a molecule to accurately model molecular polarizability is extended to include a larger spectrum of bioorganic molecules and the condensed phase. Atomic polarization radii and an internal dielectric (ε_{in}) were fitted to reproduce ab initio B3LYP/aug-cc-pVTZ polarizability tensors taken from a data set of 707 molecules. The average unsigned error on the isotropic polarizability and anisotropy are 2.6% and 5.2%, respectively. It is shown that usual Poisson−Boltzmann contact radii and a low internal dielectric are not appropriate and require major revision. To account for the anisotropy of polarizability, the internal dielectric (ε_{in}) constant needs to be larger than 6.0. Reinterpreting the theoretical link between ε_{in} and the experimental refractive index (n), this study shows, with a set of 23 organic molecules spanning the entire range of n, that even with $\varepsilon_{in} = 24$ the obtained refractive indices can correlate well with experiment (slope of 1.00, intercept of 0.05, and R = 0.95). The novel methodology used here to calculate a macroscopic-like refractive index shows that the application of the EPIC parametrization to condensed phase leads to suitable behavior. Although the primary goal in developing EPIC was to include polarizability in explicit solvent calculations, we also extend the model to work with implicit solvent. This requires the use of a 3-zone smooth dielectric function to transition from the polarization dielectric inside the molecules to the dielectric continuum of the solvent. The parametrization and validation of this model are performed against 485 experimental free energies of hydration. Using 8 solvent cavity atomic radii and a single surface tension an average unsigned error of 1.1 kal/mol and a correlation coefficient of 0.9 are obtained, validating the use of the EPIC model in the condensed phase.

1. Introduction

The newly introduced treatment of electronic polarization by an internal continuum (EPIC) was shown to be accurate in reproducing experimental and density functional (DFT) molecular polarizability tensors with a remarkably small number of adjustable parameters.[1] Moreover, the high accuracy found when computing intermolecular interaction energies, in which the appropriate treatment of electronic polarization is crucial, opens up the possibility of using EPIC to include polarizability in force fields.[2] This led us to propose the use of EPIC to embed polarizability in all-atom explicit-solvent calculations. EPIC uses continuum dielectric

* Corresponding author phone: (514)428-3403; fax: (514)428-4930; e-mail: christopher_bayly@merck.com.
[†] Université de Montréal.
[‡] Merck Frosst Canada Ltd.
[§] OpenEye Scientific Software, Inc.
[||] University of Chicago.

10.1021/ct900029d CCC: $40.75
Published on Web 06/05/2009

electrostatic theory to account for the way electronic density polarizes under the presence of an external electric field that can come from either other molecules in explicit condensed phase calculations or the reaction field in an implicit solvent calculation. In contrast to the point inducible dipoles[3−5] or the Drude's oscillator models[6,7] that use the atomic nuclear positions as polarizable centers, EPIC employs a polarizability density that induces a dipole density, normally referred to as polarization, throughout the molecule volume as a response to the local electric field. In a recent study, Schropp and Tavan[8] proposed that the use of single centers in point inducible dipole polarizable calculations was responsible for the large difference between the best condensed phase atomic polarizability and the best vacuum phase atomic polarizabilities previously noticed.[9,10] Other studies, based on Quantum Mechanical (QM) assessment, suggest that the polarizability in condensed phase should only be slightly reduced.[11] The idea of using a continuum dielectric to account for electronic polarization was first formulated by Sharp et al.[12] but was not further pursued until Tan and Luo[13] optimized the internal dielectric of solutes to produce the electrostatic potential in the context of Poisson−Boltzmann calculations with different implicit solvents. In their two studies,[13,14] they do not attempt to give a detailed molecular polarizability description but rather focus on the shift in dipole moments when a solute is put in different solvent. It is difficult to decouple the solvent polarization from the solute polarization and the cooperative polarization when calculations in implicit solvent are done. The current study uses previously developed techniques[1,2] to separate the charge fitting from the polarizability fitting by optimizing, in the absence of atomic charges, an *electronic volume* to fit quantum mechanics (QM) polarizability tensors for molecules in vacuum, as was done originally with other polarizable models.[3,5,15] Curiously, we found that in order to accurately reproduce the polarizability tensors of even challenging molecules, the atomic radii needed to be much smaller than the van der Waals (vdW) contact radii usually used in implicit solvent calculations (e.g., Bondi radii[16]). At the same time, the internal dielectric needed to be surprisingly high in order to reproduce the anisotropy of the polarizabilities. While that work allowed for a systematic way of adjusting a dielectric function to account for electronic polarization, it raised two issues: the abnormally high internal dielectric of 14 seems questionable and the small radii made implicit solvent calculations impractical. Regarding the first issue, the dielectric inside the molecule is closely related to the refractive index squared ($\varepsilon_\infty = n^2$) of the pure liquid, which adopts values between 1.7 and 2.9 for organic liquids, far below our large values. Regarding the second issue, if such small atomic radii were used to define the molecular cavity in solvent, the free energy of charging would become unrealistically negative e.g. in Poisson−Boltzmann (PB) calculations. In this work, we specifically address both issues and demonstrate the physical soundness of the approach. An important change from our previous work is the use of a smooth dielectric boundary to represent both the solute and the solvent polarization. We present a newly designed dielectric function with 3 zones (3-zone dielectric) that permits the use of EPIC for implicit solvent calculations. We show that describing the dielectric function this way better reflects the underlying physical principles involved in solvation than the usual 2-zone dielectric (i.e., inside and outside the cavity).

Another question that we examine is the ability to optimize the EPIC parameters in a general and robust way with few parameters on a larger variety of chemical functionality than in earlier work. For this purpose, we have formed a large database of QM molecular polarizability tensors for 707 diverse bioorganic molecules (for a total of 4242 polarizability tensor elements) along with their optimized molecular geometries (cf. the Supporting Information). As will be outlined below, this data set contains a large variety of chemical functional groups representing a significant component of bioorganic chemistry. This substantially enlarged parametrization of the polarizable EPIC model is then used for the calculation of refractive indices and hydration free energies. The validity of both the internal dielectric function and the 3-zone dielectric function is assessed with the independent fit of the solvent cavity atomic radii (which define the third zone of the function) on 485 experimental free energies of hydration.

In the remainder of this article, section 2 presents the theoretical basis and methods employed, where we present the 3-zone dielectric function for implicit solvent calculations and we review the polarizability tensor calculation. This is followed by the theoretical background for the calculation of the refractive index. A theoretical layout for free energy of hydration calculations and computational details related to quantum calculations close this section. Section 3 describes the chemical data sets used in section 4 where the results and their analysis are presented. Section 4 closes with a 3-zone dielectric optimization on experimental hydration free energies, leading into the conclusons.

2. Theory and Methods

2.1. 3-Zone Dielectric in Implicit Solvents. The dielectric function in continuum approaches is fundamental as it is modulating all sources of polarization. In this work, we move away from our previous use of vdW envelope surfaces[17] toward a smooth functional form based on a sum of atomic Gaussians which has previously proven successful[18,19] in PB applications. Although useful, the hard dielectric boundary often leads to numerical problems: iterative convergence failure, slower convergence, strong dependency on orientation and translation, and unstable force evaluations.[18,20] The use of smooth solute/solvent dielectric boundary was shown to improve over the hard boundary on all these aspects. More specifically, the molecular dielectric function used in the present work is given by

$$\varepsilon(\vec{r}) = \varepsilon_{in} - (\varepsilon_{in} - \varepsilon_{ext})\exp(-A \cdot f_{in}(\vec{r})) \qquad (1)$$

where ε_{in} is the dielectric constant inside the molecular volume, and ε_{ext} the dielectric value outside. The dielectric here is expressed as a permittivity relative to the vacuum permittivity. The exponential behaves as a switching function that is turned on or off depending on the value of a molecular

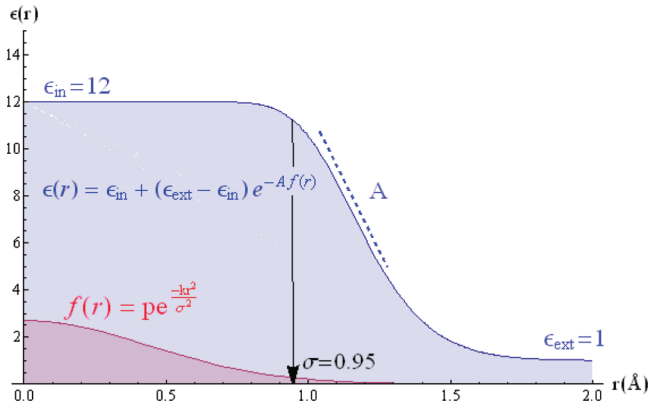


Figure 1. This figure shows the smooth dielectric function used in this work for a single atom with $\sigma = 0.95$ Å, $\varepsilon_{in} = 12$, $\varepsilon_{ext} = 1$, and $A = 10.0$ (Cl of the G1−12 set). Starting from the center of the atom ($r = 0$), the dielectric (blue curve) stays constant until the 'density', expressed with a sum of Gaussians (pink curve), reaches a certain small value that causes the dielectric to smoothly transition to the external dielectric value. The steepness and the position of the switching region depends on the value of the A parameter. The sum-of-Gaussians density expression is explained in eq 2 (see text).

'density' function $f_{in}(\vec{r})$. The A parameter modulates the steepness of the switching function. The details of the dielectric are then incorporated into the 'density' function

$$f_{in}(\vec{r}) = \sum_{i=1}^{Natoms} p \cdot \exp\left(-k\frac{|\vec{r}_i - \vec{r}|^2}{\sigma_i^2}\right) \qquad (2)$$

The summation runs over all atoms, and a 3-dimensional Gaussian defines the radial extent of the atomic volume; σ_i are atomic radii and $\vec{r}_i$ their positions. The σ_i will be the subject of an extensive parametrization in the next sections. The constant k is set to 2.3442 and p to 2.7 following the Grant et al. recommendation.[18] Equation 1 can be conceptually understood in terms of electronic density that would have a constant susceptibility (polarizability density) inside and drops rapidly as the density vanishes as shown in Figure 1.

The main methodological novelty proposed in this work is the 3-zone dielectric for the coupling of EPIC with implicit solvation. When atomic radii are optimized on QM-based molecular polarizability tensors, their resulting small size prohibits their use to define the cavity formed by the solute in implicit solvent calculations. Indeed, it presents a dilemma: on the one hand, accurate solute polarization requires atomic radii far smaller than accepted contact radii. On the other hand, the solvent boundary for implicit solvation requires atomic radii as large or larger than contact radii. The resolution to this dilemma is found in challenging the assumption that the atomic radii for solute polarization and for the solvent boundary should be the same. There is no underlying physical reason why the polarization response of an atom in a molecule would be uniform all the way out to its contact radius; on the contrary our QM model for molecules tells us the electron density (the source of electronic polarization) drops exponentially in moving from an atomic nucleus toward the contact surface of the molecule. We believe that it is more reasonable to think that the radial extent of the electronic polarization can be different from the vdW radius used for the solvent cavity. The idea presented here is that both kinds of smooth surfaces could be simultaneously used: one for solute polarization, formed with the smaller atomic polarization radii, and one for solvent polarization, defined with the solvent cavity atomic radii. In between the two surfaces is a transition region of low dielectric since it describes where the solvent and the solute electrons are both at a minimum. This leads to a 3-zone dielectric function to which we give the form

$$\varepsilon(\vec{r}) = \varepsilon_{in} + (\varepsilon_{trans} - \varepsilon_{in}) \exp[-A \cdot f_{in}(\vec{r})] + (\varepsilon_{solv} - \varepsilon_{trans}) \exp[-B \cdot f_{solv}(\vec{r})] \qquad (3)$$

where ε_{in} is the dielectric constant inside the molecular cavity, ε_{solv} is the bulk solvent dielectric constant (80 for water), and ε_{trans} is the dielectric constant in the zone of transition between the solute and the solvent. For the smooth inner dielectric boundary, A has the same meaning as in eq 1, and $f_{in}(\vec{r})$ is given by eq 2. The additional exponential term, for the outer dielectric boundary (with solvent), is a switching function that turns on when a second Gaussian sum ($f_{solv}(\vec{r})$) becomes sufficiently small. The $f_{solv}(\vec{r})$ term is also given by eq 2 with the difference that the atomic radii are larger as they define the solvent cavity. The B parameter is responsible for the steepness of the cavity boundary, but with a sufficiently large value it has the effect of moving the position of the boundary as if the radii were scaled. The radial behavior of the 3-zone dielectric is illustrated in Figure 2 for a single atom and for the 4-pyridone molecule, both with typical parameters. In eq 3, it is important to set $\varepsilon_{trans} = 1$ when the first zone of the dielectric function is fitted on molecular polarizability since the shape of the dielectric function needs to drops to one in order to present the same ability to polarize. Also, if the atomic partial charges are fitted with DRESP, a change in the first zone boundary would also change the ability of the dielectric to form the full internal polarization taken into account during the charge fitting process.

2.2. Molecular Polarizability Tensor. In this section, we review the methodology previously developed to calculate molecular polarizability tensor with a finite difference Poisson solver,[1] and we summarize how the parameters involved are optimized in this work.

2.2.1. Method. Our formulation of electronic polarization based on continuum electrostatics allows the calculation of induced multipolar moments by considering the bound charge density, which results from the polarizability density of the media (from the *bound* the electrons in our case). A formula to calculate the bound charge density is[21]

$$\frac{\rho^b(\vec{r})}{\varepsilon_0} = -\vec{\nabla}([\varepsilon(\vec{r}) - 1]\vec{E}(\vec{r})) = -\vec{\nabla} \cdot \vec{P}(\vec{r}) \qquad (4)$$

where ρ^b is the bound charge density, and $\vec{E}(\vec{r})$ the total electric field. Physically, ρ^b is a consequence of the formation of dipoles at each point in space due to the electric field (the polarization $\vec{P}(\vec{r})$ or dipole density). The bound charge density can be thought of as an induced charge density from

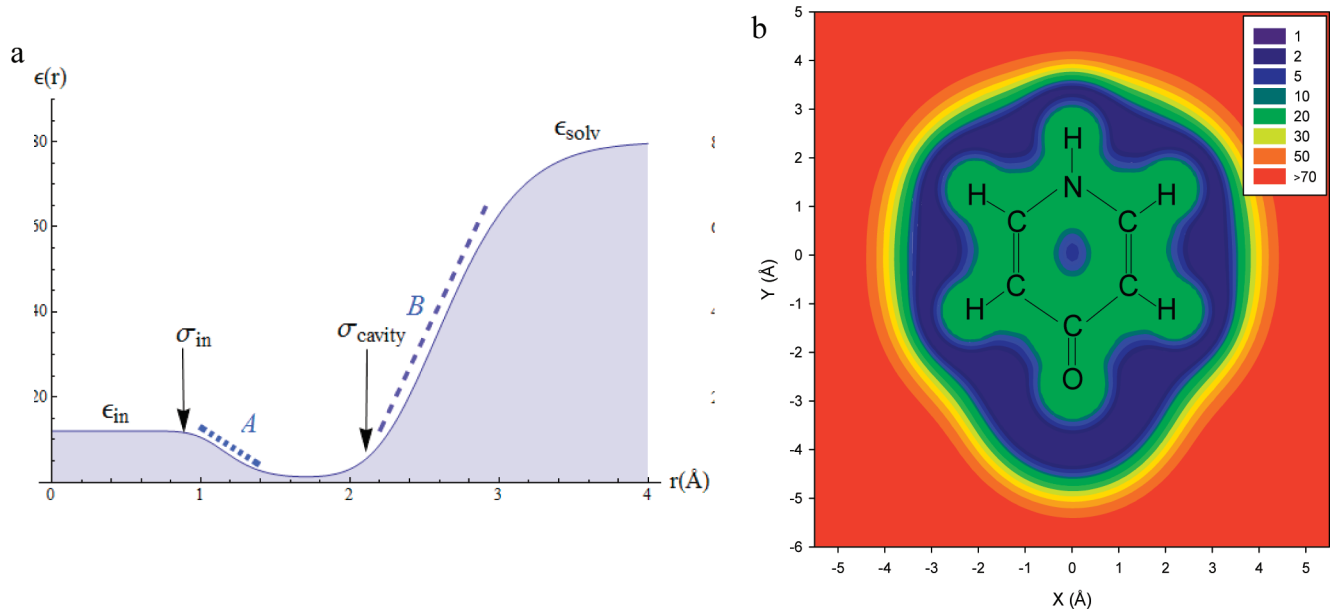


Figure 2. The 3-zone dielectric function allows an accurate description of both the solute polarization and the solvent polarization within the EPIC approach. (a) The radial component of the dielectric for a single atom (G1−12 aromatic carbon) is shown together with the polarization (σ_{in}) and the solvent cavity (σ_{cavity}) atomic radii. Each plateau of the dielectric function defines a zone. The middle intermediate zone corresponds to the solute/solvent contact distance. (b) The resulting dielectric function is also shown in the ring plane of 4-pyridone (b) when applying the G2−12 parameters.

the dielectric polarization that appears where the dielectric varies, as an excess of charge builds due to the head or tail of the dipole density. Although the polarization occurs everywhere the dielectric is greater than one, the bound charge density appears in regions of spaces where $\varepsilon(\vec{r})$ varies, such as the dielectric boundary of a molecule. Equation 4 is useful since it transforms the locally induced dipoles into a scalar value, the bound charge density, which can be used more easily as done below. In eq 4, $\varepsilon(\vec{r}) - 1$ plays the role of a local polarizability density, also called the electric susceptibility, and $\vec{P}(\vec{r}) = (\varepsilon(\vec{r}) - 1)\vec{E}(\vec{r})$ corresponds to the induced dipole density (polarization). The analogy with the point inducible dipole model, a different polarizable model, is obvious since, in that case, the atomic induced dipole is given by $\vec{\mu}(\vec{r}_i) = \alpha_i\vec{E}(\vec{r}_i)$ where $\vec{\mu}(\vec{r}_i)$, α_i, and $\vec{E}(\vec{r}_i)$ are the dipole induced at the atomic position $\vec{r}_i$, the atomic polarizability, and the electric field at $\vec{r}_i$. Here, the polarization is more smoothly distributed over the molecular volume. Equation 4 is intrinsic to the definition of Poisson's equation.

A classical example, for which an analytical solution exists, is the dielectric sphere in vacuum experiencing an external electric field. In this case the mathematics show that bound charges appear on the surface of the sphere with opposite charge sign on both hemispheres, resulting in an induced potential equivalent to an ideal induced dipole moment aligned with the external field located at the center of the sphere. The induced dipole moment is proportional to the external electric field, and the sphere polarizability α_{sphere} is given by the Clausius-Mossoti equation

$$\alpha_{sphere} = \left(\frac{\varepsilon_{sphere} - 1}{\varepsilon_{sphere} + 2}\right) R_{sphere}^3 \tag{5}$$

where R_{sphere} is the sphere radius. For a molecular system, the analytical solution is unknown, and we use a finite difference algorithm to solve Poisson's equation numerically with a uniform electric field in the form of a voltage clamp applied by means of the boundary conditions. More precisely, a uniform electric field in the z direction can be produced with a null potential on one side of the grid boundary and the value $-E_{ext} \times L_z$ on the opposite side, where L_z is the box size in the z direction, and E_{ext} is the magnitude of the applied field. On the four other sides, parallel to the field, the grid boundary potential is simply calculated as a linear interpolation along the z direction: $\varphi(z\text{-}z_0) = -(z\text{-}z_0) \times E_{ext}$. As with the dielectric sphere in vacuum, a molecular dielectric volume responds linearly to the applied field (given an isotropic dielectric function), and the proportionality constant is the molecular polarizability tensor. The field is applied in three orthogonal directions to build the polarizability tensor, which depends on the orientation of the molecule

$$\bar{\alpha} = \begin{bmatrix} \frac{\mu_{x,x}}{E_{ext}} & \frac{\mu_{x,y} + \mu_{y,x}}{2E_{ext}} & \frac{\mu_{x,z} + \mu_{z,x}}{2E_{ext}} \\ & \frac{\mu_{y,y}}{E_{ext}} & \frac{\mu_{y,z} + \mu_{z,y}}{2E_{ext}} \\ & & \frac{\mu_{z,z}}{E_{ext}} \end{bmatrix} \tag{6}$$

where $\mu_{x,y}$ is the x component of the induced dipole moment when an external electric field of magnitude E_{ext} is applied in the y direction. Some experimental values are available for the eigenvalues of this tensor in vacuum ($\varepsilon_{ext} = 1$); also, the polarizability tensor can be calculated using approaches based on quantum mechanics (QM) methods such as density functional theory.

The induced dipole moment is calculated analogously to the sphere dielectric system, integrating the bound charge

density over space. From eq 4 (or simply from Gauss's law), one can show that

$$\rho^b(\vec{r}) = -\rho^f(\vec{r}) + \varepsilon_0 \vec{\nabla} \cdot \vec{E}(\vec{r}) \quad (7)$$

In the present context, there is no free charge density $\rho^f(\vec{r})$ (from atomic partial charges, for instance), and as such the bound charge density, induced only by the external uniform electric field, is given by the divergence of the field. With a finite difference solver, the total charge (bound and free charges) can be calculated by integrating over each differential volume element (grid cube) which leads to bound charges on grid points. This can be done simply by calculating

$$\frac{q_{ijk}}{\varepsilon_0} = \frac{q^b_{ijk} + q^f_{ijk}}{\varepsilon_0} = -\left(\frac{h_y h_z}{h_x}\right)(\varphi_{i+1jk} + \varphi_{i-1jk} - 2\varphi_{ijk}) - \left(\frac{h_x h_z}{h_y}\right)(\varphi_{ij+1k} + \varphi_{ij-1k} - 2\varphi_{ijk}) - \left(\frac{h_x h_y}{h_z}\right)(\varphi_{ijk+1} + \varphi_{ijk-1} - 2\varphi_{ijk}) \quad (8)$$

where q_{ijk}, q^b_{ijk}, and q^f_{ijk} are the total charge, the bound charge, and the free charge inside the volume element associated with the *ijk* grid point, and φ_{ijk} and $\varphi_{ijk\text{-}1}$ are the electrostatic potential at the (*x,y,z*) and (*x,y,z-dz*) grid points, respectively. The grid spacing in *x*, *y*, and *z* are given by h_x, h_x, and h_z. The grid free charge q^f_{ijk} are zero for this calculation, and, in general, it is given by the atomic partial charges as distributed on the grid. Finally, the total dipole moment is given by

$$\vec{\mu} = \sum_{i,j,k}^{Grid} \vec{r}_{ijk} q_{ijk} \quad (9)$$

With the free charges equal to zero (no atomic partial charge), the dipole calculated is then the induced dipole, and the only contributor is the bound charge density. More generally, any molecular electric moment can be calculated with analogs to eq 9. The overall procedure to calculate the polarizability tensor requires three solutions from the numerical solver. The calculation does not involve atomic partial charges (free charges) which allows them to be fit independently (although this must still be done in the context of the molecular dielectric).

2.2.2. Computational Details. The finite difference Poisson calculations were performed with a modified version of the OpenEye Inc. ZapTK.[22] The distance between two grid points was set to 0.35 Å, and the grid boundary was at least 5 Å away from the surface defined by the polarization radii. Atomic charges of $\pm 0.001e$ were assigned randomly on the atoms as the grid energy was used to determine the convergence of the algorithm set to 0.000001 k_BT. The results were not sensitive to these small charges. Atom typing was assigned via SMARTS[23−25] with the OpenEye Inc. OEchem toolkit.[26]

2.2.3. Optimization of the Polarizabilities. The atomic radii were optimized in order to minimize a chi-square function using a Levenberg−Marquardt algorithm as implemented in scipy,[27] a scientific Python library. The error was defined as the difference between the 6 components of the polarizability tensor obtained with B3LYP and EPIC

$$\chi^2 = \sum_i^{molecules} \sum_{k=xx,xy,xz,yy,yz,zz} (\alpha^{EPIC}_{k,i} - \alpha^{QM}_{k,i})^2 \quad (10)$$

where $\alpha_{xy,i}$ is one of the six independent polarizability tensor elements of molecule i either under optimization (EPIC) or from the QM target values. By using the six independent tensor elements, we included both the magnitude and the direction of the polarizability in a natural way.[28] We optimized the cube of the polarization radii because their contribution to the polarizability grows with the atomic volume (cf. eq 5). For analysis purposes, we also defined the average polarizability (eq 11) and the anisotropy of the polarizability tensor (eq 12) below

$$\alpha_{avg} = (\alpha_1 + \alpha_2 + \alpha_3)/3 \quad (11)$$

$$\Delta\alpha = \sqrt{\frac{(\alpha_1 - \alpha_2)^2 + (\alpha_1 - \alpha_3)^2 + (\alpha_2 - \alpha_3)^2}{2}} \quad (12)$$

where $\alpha_1 \leq \alpha_2 \leq \alpha_3$ are the eigenvalues of the polarizability tensor. The polarizability anisotropy is significantly harder to fit than the average polarizability. We defined the error in the average polarizability (eq 13) and anisotropy (eq 14) for a set of molecules as

$$\delta_{avg} = \frac{1}{N}\sum_i^N \frac{|\alpha^{QM}_{i,avg} - \alpha_{i,avg}|}{\alpha^{QM}_{i,avg}} \quad (13)$$

$$\delta_{aniso} = \frac{1}{N}\sum_i^N \frac{|\Delta\alpha^{QM}_i - \Delta\alpha_i|}{\alpha^{QM}_{i,avg}} \quad (14)$$

where *N* is the total number of molecules considered and *QM* corresponds to the target value. Finally, the relative root-mean-square deviation (RRMS) of the tensor was defined as

$$RRMS = \frac{\sum_i^{molecules} \sum_{k=xx,xy,xz,yy,yz,zz} (\alpha^{EPIC}_{k,i} - \alpha^{QM}_{k,i})^2}{\sum_i^{molecules} \sum_{k=xx,xy,xz,yy,yz,zz} (\alpha^{QM}_{k,i})^2} \quad (15)$$

and constituted a single metric for the overall fitness of the optimized polarizability tensors. If the RRMS was calculated for a single molecule, the summations on the molecules in the numerator and the denominator were simply omitted.

2.3. Refractive Index Calculations. *2.3.1. Theory.* The dielectric constant of an isotropic material at the high frequency limit (ε_∞) is related to the material refractive index[29] *n* by

$$n^2 = \varepsilon_\infty \quad (16)$$

where *n* is usually measured with the D line of the sodium spectrum at 589 nm (n_D). The ε_∞ corresponds to the material's

dielectric constant solely due to the electronic polarization since the frequency of the visible light is too high for nuclear relaxation to contribute. Typically, a pure liquid of an organic compound will have a refractive index between 1.3 and 1.7 leading to a ε_∞ between 1.7 and 2.9. Since the work of Debye and Onsager,[12,13,30,31] it has become a dogma that the interior dielectric (ε_{in}) of a solute cavity in implicit solvent models should be close to the experimental ε_∞ in order to capture the dipole moment change due to the cooperative solute−solvent polarization. It is when we seek for accuracy in solute polarization that we found the generally accepted relation $\varepsilon_\infty = \varepsilon_{in}$ to badly fail.[1] A way to reconcile this puzzling finding is by computing the macroscopic refractive index that corresponds to what is measured instead of assuming it is the same as the *internal refractive index* (quoting Onsager[30]). The Clausius-Mossoti equation relates the polarizability of a sphere to its interior dielectric. Since ε_∞ and n are macroscopic intensive quantities, their measurement should not depend on the size of the studied sample, given that it is large enough to exhibit a macroscopic behavior, the worst case being the use of a single molecule. It is not to say that Onsager's uses of the Clausius-Mossoti equation with the radius of a single molecule were not justified. In fact, he was primarily interested in the molecular polarizability (α_{mol}) and used the formula

$$\alpha_{mol} = \left(\frac{n_D^2 - 1}{n_D^2 + 2}\right)\frac{3V}{4\pi N} \tag{17}$$

where V is the volume of the liquid sphere considered, and N the number of molecules it contains. In eq 17, the rightmost factor corresponds to the cube of an effective single spherical molecule radius. It is however understood that the same molecular polarizability is obtained as long as the V/N factor is preserved and is therefore size independent with the key assumption that ε_∞ is filling the space uniformly, i.e. that it is a spatially averaged value. In order to calculate the refractive indices for the general case where the internal dielectric is not uniformly distributed in the liquid, we generated pure liquid configurations from molecular dynamics (MD) simulations at room temperature and cut out spherical clusters (or droplets) from individual snapshots. We maintained the V/N ratio by fixing the density to experiment and calculating the droplet effective ε_{in} with the formula

$$n^2 = \frac{R_{droplet}^3 + 2\alpha_{droplet}}{R_{droplet}^3 - \alpha_{droplet}} \tag{18}$$

where $R_{droplet}$ and $\alpha_{droplet}$ are the droplet radius and polarizability. We assigned the dielectric function on all molecules and applied the procedure outlined above to calculate the droplet polarizability and thereby access the droplet refractive index. Averaging the droplet refractive index over many droplets yields an approximation of the bulk refractive index.

2.3.2. Computational Details. To obtain the liquid phase droplets, molecular dynamic simulations, using the AMBER 8.0 package, were performed on 3375 molecules (15 × 15 × 15) in a cubic box. The NVT ensemble and periodic boundary conditions allowed the density to be fixed to the experimental value. The temperature was set to 20 °C to match the experimental conditions used to report refractive indices and maintained constant with the Berendsen's weak coupling algorithm[32] with the kinetic energy adjusted every 1 ps. The short-range nonbonded interaction cutoff was set to 8.0 Å, and long-range interactions computed with particle mesh Ewald[14,33] using the default Amber 8.0 setup. The molecules were charged with AM1-BCC,[34,35] and the Generalized Amber Force Field (GAFF)[36] was used. The SHAKE procedure[37] was used to fix all bond lengths to hydrogen.

The initial liquid box was generated by positioning the molecules on a cubic lattice, randomly oriented with the Marsaglia[38] quaternions method. The system was first minimized until the root-mean-square (rms) of the gradient is less than 0.1 kcal/mol/Å. This was followed by a 8 ps annealing phase integrated by steps of 1 fs, during which the nonbonding interactions were gradually turned on and the temperature increased from 0 K to 40 K and decreased to 0 K. The system was then heated over 20 ps up to 293.15 K with a 2 fs integration time step. Following a 1 ns equilibration, 50 evenly spaced snapshots were written over a 2 ns production run. Each of the liquid boxes for a given molecule was then wrapped in the primary cell. A sphere with a diameter set to 85% of the box length formed a liquid droplet when picking all molecules with an atom lying inside the sphere. The droplet radius was then determined by considering the position of the outermost non-hydrogen atoms. The precise definition of the radius is not unique, and we have verified, for example, that using the experimental density to calculate the radius of the corresponding ideal sphere gives refractive indices within ±0.01 of those obtained by the chosen algorithm. Also, this model assumes a perfectly spherical object, ignoring the dimples formed because of the finite size of the spheres. The relatively large size of the droplet and the averaging over 50 independent configurations reduced the effect of this approximation.

The solution to Poisson's equation in the presence of the voltage clamp boundary conditions was obtained on a rectangular grid sized to encompass the full droplet plus half its radius on each side of the droplet. The target grid spacing was set to be 0.5 Å. The smooth dielectric functions (eq 1), fitted on the molecular polarizability tensors only, were assigned together with the matching atomic polarization radii, internal dielectric ε_{in}, and A parameter. The external dielectric was always set to the vacuum value $\varepsilon_{ext} = 1$. The convergence criteria for the ZapTk solver was based on the grid energy and set to 0.0001 kT. This convergence criteria required the assignment of atomic charges that we choose to be $\pm 0.001e$ were randomly placed on half the atoms, keeping an overall neutral system. Given the strength of the external field applied, this was not perceptibly affecting the answer.

2.4. Free Energy of Hydration. *2.4.1. Theory.* Implicit solvent models are commonly used to incorporate the effects of solvation in molecular models as a mean field.[39−43] These models considerably reduce the computational burden needed to sample the solvent configurational space when each atom of the solvent are explicitly simulated. An important valida-

Table 1. Reported Optimal Polarization Radii (σ_{in}) and Atom Typing for the Four G1 Sets Defining the Internal Dielectric (Eq *1*)

SMARTS[23−25]	typical functional groups	radius (Å)			
		G1−4[a]	G1−9[a]	G1−12[a]	G1−24[a]
ε_{in}[b]		4	9	12	24
A[b]		10	5	10	4.19
	H				
[H]	all H	0.83	0.65	0.55	0.52
	C				
[CX4]	alkanes	0.78	0.79	0.67	0.62
[c,CX2,CX3]	aromatic, sp, sp^2	1.25	1.02	0.87	0.78
	N				
[n,NX1,NX3,$(Nc),$(NN)]	aromatic, nitriles, sp^3, aniline, hydrazine,	1.09	0.89	0.76	0.69 0.74[c]
[$(N[C,S]=*)]	amides, amidines, sulfonamides	0.89	0.77	0.64	0.58
[$(N=C)]	imine, amidine	1.07	0.93	0.81	0.76
[$([#7]∼[OX1])]	N-oxides, nitro	0.00	0.79	0.68	0.59
	O				
[$([OX2]([H])[#6,#7]),o,$([OD2]([CX4,c])[CX4]), $(O=[c,C,S])]	alcohols, furan, hydroxamic acids, ethers, ketones, aldehydes, amides, sulfones	0.88	0.73	0.63	0.60
[$(OC=[O,N])]	esters, carboxylic acids	0.68	0.55	0.46	0.46
[$([OX1]∼[#7])]	N-oxide, nitro	1.08	0.89	0.77	0.74
	Others				
[S,s]	all sulfur atoms	1.44	1.22	1.06	1.01
[F]		0.77	0.62	0.53	0.51
[Cl]		1.30	1.09	0.95	0.91
[Br]		1.47	1.24	1.07	1.03
	Water				
	special fit				
[$([OX2]([H])[H]]		0.93	0.86	0.76	0.75
[$([H][OX2][H])]		0.64	0.45	0.36	0.31
	Charged Atoms				
[$([#1][#7+]),$([#1][#7][#6]=[#7+][#1]), $([#1][#7][#6]=[#7+]),$([#1][n+]∼c∼n), $([#1]n∼c∼[n+])]	proton in guanidiniums, amidiniums, ammoniums, pyridiniums	0.44	0.43	0.37	0.01
[$([O-]C=O),$(O=C[O-])]	O in carboxylates	1.20	1.02	0.88	0.85
[$([NX4+]),$([#7+]=C−N),$(N−C=[N+]))]	N in ammoniums, guanidiniums, amidiniums,	0.00	0.34	0.39	0.52
[$([n+]∼c∼n),$([n]∼c∼[n+])]	N in imidazoliums	0.00	0.00	0.00	0.42

[a] Model name. [b] Parameter kept fixed during optimization. [c] Nitrile nitrogen radius made different for G1-24.

tion for solvation models comes from the experimental free energy of hydration (ΔG_{hyd}) that consists in the chemical potential difference for the transfer of a solute from vacuum to bulk solvent. The computational evaluation of ΔG_{hyd} is separated into two processes. First, the nonpolar free energy of hydration (ΔG_{np}) comes from the formation of the solute-shaped cavity in the bulk solvent that causes a reorganization of the solvent molecules and nonpolar interactions between the solute and the solvent. Second, the electrostatic free energy of hydration (ΔG_{elec}) results from the electrostatic work necessary to place the solute charge density in the solute cavity, involving interactions between solute and solvent charge densities and their response to one another. This results in the equation

$$\Delta G_{hyd} = \Delta G_{elec} + \Delta G_{np} \tag{19}$$

The long-standing use of implicit solvent to evaluate ΔG_{elec} is based on a high continuum dielectric solvent region that gets polarized by a static solute electric field. While the solute cavity is traditionally formed with a molecular surface with a discrete transition of the dielectric function at the solute−solvent boundary, we chose a smooth boundary transition as explained earlier. The solute cavity volume and shape is determined by atomic radii. For a given set of charges, atomic radii that are too small exaggerate the affinity of the solute for water, while radii that are too large will have the opposite effect. The calculation of ΔG_{elec} is normally done with a nonpolarizable solute, or, if the cavity is assigned a $\varepsilon_{in} > 1$, the very significant screening of the atomic partial charges requires a special treatment that was not done until recently.[2] For nonpolarizable solutes, knowing that water increases the dipole moment of solvated molecules often by as much as 15%, the atomic charges should not be fit on a gas phase QM ESP. For this reason, the charges are often generated from RESP[44] or AM1-BCC[34,35] that are known to be sufficiently overpolarized compared to the gas phase.

In the 3-zone dielectric model that we propose in this article (cf. Figure 2a and eq 3), the first zone should accurately account for the solute polarizability, which allows for the use of vacuum phase atomic charges obtained taking into account the internal dielectric function. The second zone located between the internal dielectric and the solvent is set to vacuum, and the transition to the full implicit solvent model of the third zone needs to be parametrized. Following the suggestion of Grant et al.[18] for their nonpolarizable 2-zone dielectric function, we fixed the B parameter in eq 3

to 11.8, which leaves the solvent cavity atomic radii to be fitted on the experimental free energy of hydration. However, in order to compare the calculated ΔG_{hyd} to experiment, we needed to use existing values or methods for ΔG_{np}. Fortunately, converged molecular dynamics free energy[45,46] calculations based on free energy perturbation (FEP) calculations are available for each compound from our hydration free energy data set. We feel this is the best achievable theoretical estimation of ΔG_{np}, so this is our preferred estimation in current study. However, since this is not very useful for prospective evaluations of ΔG_{hyd}, due to the heavy computational demands for such FEP calculations, we also tested a surface area based model that calculates ΔG_{np} as

$$\Delta G_{np} = \gamma \times S \tag{20}$$

where γ is a surface tension, and S is the surface area of the molecule as defined by a solvent accessible surface[17] created with a 1.4 Å rolling probe and the Bondi radii.[16] This crude approximation has proven useful, and it can be improved upon by atom typing the γ[47] or by using some treatment of the dispersion energy[48−51] instead; however, in this work a single value of γ was fitted for each model.

2.4.2. Computational Details. The atomic partial charges responsible for the permanent electrostatic potential (ESP) were determined by a least-squares-fit on the QM ESP calculated on a face-centered-cubic grid of points. Following Jakalian et al.,[34] the grid spacing was set to 0.5 Å, and the grid points were positioned around the molecule in a volume formed by two vdW surfaces, each built with Bondi radii scaled by a factor of 1.4 and 2.0. The dielectric scales down by a factor of $1/\varepsilon_{in}$ the effect of the charges; this is partly compensated by the bound charges appearing from the internal polarization. Hence, the least-squares-fit requires a Poisson solver in order to capture the overall effect, which depends on the shape of the dielectric boundary. It is noteworthy that the EPIC polarizability model is independent of the charge fitting process; as a result, charges are fitted after the solute dielectric parameters are optimized. The details of the procedure, called DRESP, can be found elsewhere.[2]

A finite difference Poisson solver was written to allow the implementation of the 3-zone dielectric model. Here is a brief description of the algorithms implemented. We use successive over-relaxation (SOR) and a Gauss-Seidel iterative scheme[52,53] where the over-relaxation parameter w is estimated by

$$w = \frac{2}{1 + \sqrt{1 - \lambda_{\max}^2}} \qquad \lambda_{\max} = 1 - \frac{\pi^2}{2(N-1)^2} \tag{21}$$

where N is the number of grid points in one of the dimension of the grid.[52] This crude estimate of the spectral radius of the A matrix in the finite difference form of the Poisson's equation used (see the Appendix of ref 18) was sufficient to reduce by a factor of approximately 30 the number of Gauss-Seidel steps necessary.

The free charges of the system were assigned on the grid with a quadratic inverse interpolation scheme[18] that has the advantage of conserving the dipole moment, has a continuous first derivative, and is more robust to the effects of rotation and translation. The same interpolation rule is used to calculate the potential in between grid points. In our calculations, we use a convergence criteria base on grid energy defined as the sum of the electrostatic potential times the distributed free charges on the grid. This convenient criterion is directly related to the energy in an absolute way and thus ensures that relative energies are also converged. The boundary conditions, in energy calculations, were determined with a Coulomb potential.

The ΔG_{elec} was computed by taking the grid charge energy difference between a solution obtained in vacuum ($\varepsilon_{ext} = 1$) and another solution in water ($\varepsilon_{ext} = 80$) from the resulting Poisson's equation and calculated with

$$\Delta G_{elec} = \frac{1}{2} \sum_{i}^{Atoms} q_i(\varphi(\vec{r}_i)^{water} - \varphi(\vec{r}_i)^{vacuum}) \tag{22}$$

where q_i is the atomic partial charge of atom i, and $\varphi(\vec{r}_i)^{\text{vacuum}}$ is the interpolated electrostatic potential at atom i position $\vec{r}_i$. The grid spacing for the solver was set to 0.35 Å, and the minimum distance between the solute internal radii and the grid boundary was set to 7 Å. In those cases where the solute was nonpolarizable, ε_{in} was set to one. Finally, the parameters (solvent cavity atomic radii and surface tension) were adjusted with the same Levenberg−Marquardt algorithm used for the fit to the polarizability tensor. All parameters were simultaneously optimized.

2.5. Quantum Calculations. The B3LYP exchange-correlation functional[54,55] was used for all DFT quantum calculations of this work within the Gaussian 03 software.[56] All molecular structures of this work were initially relaxed with B3LYP and the 6-31++G(d,p) basis set.[57−59] Property calculations required larger basis sets for accuracy. The electrostatic potential values were obtained with B3LYP and the 6-311++G(3df,3pd) extended triple-ζ basis set.[57−59] The molecular polarizability tensor computations used the aug-cc-pVTZ basis set,[60] as it was shown to lead to accurate results.[61] The implemented method in Gaussian 03 to calculate the molecular polarizability tensor is the Coupled Perturbed Hartree−Fock (CPHF) method.[62] The Hartree−Fock calculations performed to fit water-adapted atomic partial charges were also performed with the Gaussian 03 software with the 6-31G(d,p) basis set.

3. Data Sets

In this work, we made extensive use of three kinds of data: B3LYP/aug-cc-pVTZ polarizability tensors, free energies of hydration, and refractive indices. A total of five data sets were created.

3.1. Polarizability Training Data Set (PTD). A training data set was used to optimize the internal radius in order to match B3LYP polarizability tensors. To this end, we made use of the previously published training data sets[1] and added new molecules for a total of 265 polarizability tensors. In this data set, many neutral functional groups are represented:

alkanes, alkenes, alkynes, halogens (bromo, fluoro, chloro), alcohols, thiols, amines, ethers, thioethers, nitriles, aldehydes, ketones, esters, thioesters, amides, acids, ureas, imines, amidines, sulfones, sulfoxides, sulfonamides, heteroaromatics, hydrazines, hydroxamic acids, N-oxides, pyridones, and peptides. In addition, charged functional groups were also included with the sole purpose of examining charged side chains in amino acids. They were carboxylates, guanidiniums, imidazoliums, and ammoniums. The strength of the polarizability training data set is in the wide coverage of functional groups, but its weakness is the lack of polyfunctional molecules. To get this level of coverage would require calculations on a great many more larger molecules and consequently an enormous amount of computational power. The intention in this paper is to assess whether a small and reasonably general first set of parameters can adequately treat a wide variety of bioorganic small molecules in addition to most biomolecules.

3.2. Polarizability Validation Data Set. The polarization validation data set is composed of the previously published validation sets[1] and 401 molecules from the hydration free energy data set (below) not included in the polarizability training data set. In addition, a few special molecules such as neutral and charged peptides, melamine, sugars, etc. were added, giving a total of 442 molecules.

3.3. Polarizability Data Set. The polarizability data set is the combination of the validation and training data sets, making available all 707 polarizability tensors together with the molecule coordinates (see the Supporting Information).

3.4. Hydration Free Energy Data Set. This data set is built from a compilation of 504 experimental free energies of hydration of neutral molecules recently published with the corresponding ΔG_{np} and ΔG_{chg} from Molecular Dynamics based absolute free energy calculations.[45] We took the published data set, eliminated the iodine- and phosphorus-containing compounds, and formed a data set of 485 molecules on which we could fit the solvent part of the dielectric function (eq 3) and the surface tension coefficient (γ).

3.5. Refractive Indices Data Set. The refractive indices data set contains 23 small organic molecules (cf. Figure 5) that are liquids at 20 °C, for which the density and the refractive indices are taken from the CRC Handbook of Chemistry and Physics.[63] They span a variety of functional groups, and most of the entire spectrum of refractive indices measured for bioorganic molecules.

4. Results and Discussion

4.1. Polarizability Tensor. This work follows the precedent of ref 1 in fitting atomic polarization radii and a single inner dielectric constant to QM molecular polarizability tensors to produce an accurate EPIC model of electronic polarization. This is done independently of the permanent electrostatic potential; all atomic partial charges are therefore set to zero. In this section, we generalize the parametrization to account for most of the biomolecules and a significantly wider spectrum of bioorganic functional groups. In contrast

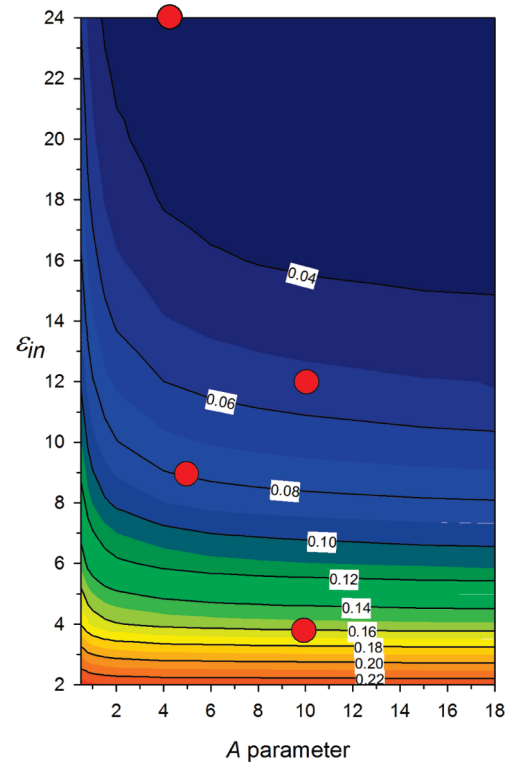


Figure 3. The iso-contour plot of the RRMS error between B3LYP/aug-cc-pVTZ and EPIC polarizability tensors are shown as a function of the ε_{in} and *A* parameters of eq 1. This RRMS surface was generated from a simultaneous fit of the H, alkyl C, aromatic C, and aromatic N atomic polarization radii on training set of 11 aromatic and 14 alkane molecules against their B3LYP polarizabilities. It shows that in order for a single dielectric model to fit the polarizabilities of these two chemical classes to within 10% error, the ε_{in} needs to be sufficiently large (>6). Deviations in the anisotropy of the polarizability are the main source of error for lower values of ε_{in}.

to our previous work, we use a smooth dielectric function as described earlier and a single internal dielectric (ε_{in}) value.

4.1.1. Choice of ε_{in} and A Parameters. It was previously shown that a more accurate polarizable model was obtained when different ε_{in} were fitted for alkanes and aromatics. However, the single-ε_{in} model performed as well as the multi-ε_{in} model and DFT against experimental directional polarizabilities. Furthermore, in another study[2] that examined the local electronic polarization, the same single-ε_{in} model was only slightly worse than the multi-ε_{in} model. In this work we pursue the single-ε_{in} model because it greatly simplified the Poisson solver implementation and the robust parametrization for a wide spectrum of bioorganic chemistry.

Before the global parametrization of polarizability atomic radii, a range-finding study was performed with a smaller training set examining which combination of ε_{in} and *A* (cf. eq 1) is best to use for extending the EPIC parametrization previously initiated.[1] We used a set of 13 alkanes (set g in ref 1) including methane, propane, cyclopropane, butane (cis, trans), hexane (cis, trans), and neopentane, together with a set of 10 heteroaromatic molecules (set a in ref 1). We formed the two-dimensional grid of ε_{in} and *A* pairs and optimized four radii (hydrogen, alkane carbon, aromatic

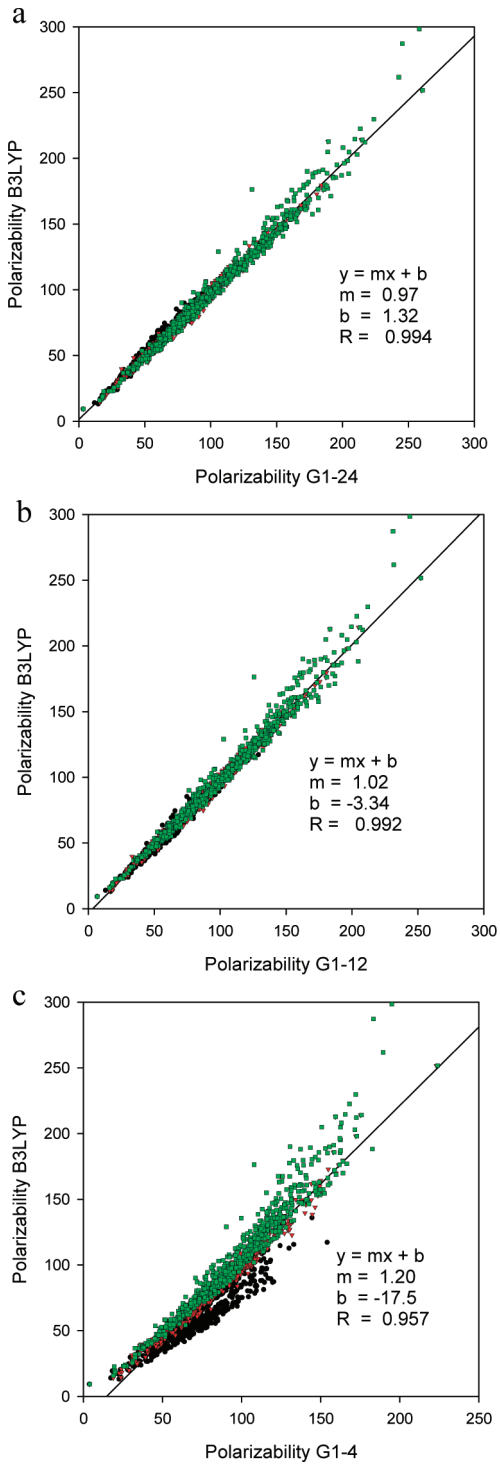


Figure 4. Correlation graph between the B3LYP/aug-cc-pVTZ directional polarizabilities (α_1 black circles, α_2 red triangles, and α_3 green squares in au) for three G1 dielectric parameter sets (cf. Table 1). Each figure shows the data for 707 molecules for a total of 2121 points. From these figures, it is clear that a small number of parameters (optimized on 265 molecules) can generalize well. A large $\varepsilon_{in} = 24$ (a) produces the best fit, a medium range $\varepsilon_{in} = 12$ produces slightly larger discrepancies, and a small $\varepsilon_{in} = 4$ produces significantly larger deviations, in keeping with the results of the range-finding study on the small data set.

carbon, and aromatic nitrogen) for each point of the grid. The polarizability tensor RRMS deviation from QM for this

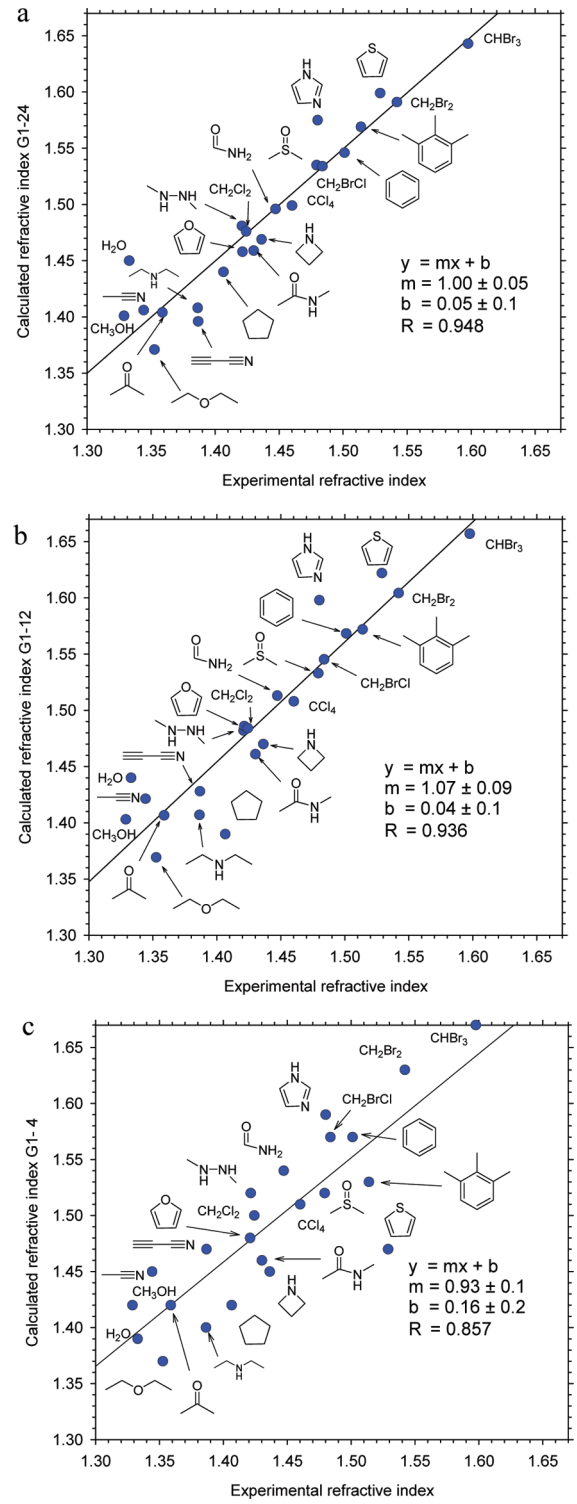


Figure 5. The calculated refractive indices (n) of 23 organic molecules are compared to experiment. Three dielectric parameter sets are used a) G1−24, b) G1−12, and c) G1−4 (Table 1). For each set, the preoptimized radii can be found in Table 1. The reported refractive indices (n) were obtained by polarizing a liquid droplet formed by carving spheres from periodic MD liquid simulation snapshots. The Clausius-Mossoti equation leads to $n^2 = \varepsilon_\infty$ close to experiment, in spite of the large ε_{in}. The predicted values are systematically higher than experiment, which can be explained by potential artifacts or a polarizability shift when passing from vacuum to condensed phase (see text). As with the polarizabilities, the predictions deteriorate with decreasing ε_{in}, in keeping with the results of the range-finding study on the small data set.

data set at each (ε_{in}, A) pair is shown as an iso-contour plot in Figure 3. It is clear that in order to fit a general dielectric function, a sufficiently large ε_{in} is needed. Also, the flatness of the error surface allows for multiple equivalent choices, a potential advantage if other criteria become more stringent in the development of the polarizable model. As shown by red circles in Figure 3, four starting points were selected for further examination: G1−24 ($\varepsilon_{in} = 24$, $A = 4.188$), G1−12 ($\varepsilon_{in} = 12$, $A = 10$), G1−9 ($\varepsilon_{in} = 9$, $A = 5$), and G1−4 ($\varepsilon_{in} = 4$, $A = 10$). In the case of G1−24 only, the A parameter was relaxed to a value of 4.18. The G1−12 seems slightly superior to the G1−9. Finally, while the G1−4 parameter set showed the worst RRMS, it was still a good case for having a small value of ε_{in}, identified by Tan and Luo[13] as being optimal. Each of the G1 ε_{in} and A choices was fixed in the global parametrization of atomic polarization radii described below. Finally, Figure 3 shows that making a poor selection of (ε_{in}, A), in particular having $\varepsilon_{in} < 6$, cannot be redeemed by adjusting either the radii or the A parameter.

4.1.2. The Optimized Polarization Radii. The parametrization of the four G1 sets on the 265 molecules of the polarizability training data set proceeded as described in the Method section. The ε_{in} and A values were fixed, and the atomic polarization radii σ_i were adjusted to optimize the fit to the B3LYP polarizability tensors. The atom typing of the radii was a primary concern, and we aimed at minimizing the number of radii fitted to reduce the fitting complexity, ensuring a better generalization of the chemistry. Each nonsymmetric molecule produced 6 data points from their polarizability tensor; structurally symmetric molecules produced fewer data points. The number of fitted parameters was kept small compared to the number of associated data points to prevent overfitting. The determination of the atom typing was done iteratively by hand: first, the polarizability training data set was designed in terms of chemical functional-group classes. Adjustable parameters were added gradually with new molecules having unmet chemical functionalities. Often, the addition of a new chemical functionality class led to one or two additional parameters. We also merged atom types when the radii values were similar and the fitness metrics (χ^2, δ_{avg}, and δ_{aniso}) were not significantly affected. For example, the alkane H and C radii were the first to be fitted. This was followed by aromatic C, H, and N. It was determined early that a single aromatic and alkyl atom type for C and H could be utilized. Then the alcohol oxygen radius, halogen radii, alkene carbon, and alkyne carbon radii were individually fitted. The final stage was a global simultaneous fit of all radii with the entire polarizability training data set. Because of its special importance as a solvent, water was treated separately with its own special O and H radii.

The resulting polarization radii are given in Table 1 for the four G1 parameter sets. The ordering of atom types in Table 1 is important since the atom typing was done in sequence from top to bottom to deal with the issue of a particular atom falling in more than one category (H for instance). The first observation is that all polarization radii are significantly smaller than vdW contact radii such as Bondi,[16] Pauling,[64] or Parse[40] often used in Poisson−Boltzmann approaches. This finding unveils the two different natures of the physical phenomena described. On the one hand, the polarization radii aim at calibrating how the electrons polarize in reaction to an external field created, for example, by an interacting molecule. On the other hand the vdW radii determine the position of the repulsive molecular wall toward other molecules. It is also expected that the larger the ε_{in}, the smaller will be the radii: to maintain the overall polarization the dielectric must increase as the radii decrease. This is illustrating a general feature of the model that produces larger polarizabilities when either the 'electronic volume', decided by the radii, or the internal dielectric increase. The sort of relationship involved is given in eq 5 above for a hard sphere and elsewhere for a diatomic.[1]

It is also interesting to compare polarization radii between elements and between the different chemical environments. First, it is remarkable that the diverse carbon atom contexts can be covered by only two atom types: sp^3 and others. The smaller sp^3 carbon radius implies that carbon makes a much smaller contribution to the overall polarizability when sp^3 hybridized than when pi electrons are involved, i.e. in the sp or sp^2 hybridization states. This can be rationalized by the presence of π^* molecular orbitals, the different number of connected H atoms, and the difference in the molecule shape and the related anisotropy.

The nitrogen atoms were subdivided into four atom types among which two encompass almost all instances in the data sets. The first of these is a general nitrogen type assigned to amines, nitriles, hydrazines, or anilines for example. The smaller second major nitrogen radius makes amide, amidine, or sulfonamide nitrogen less polarizable. Surprisingly, the more specific nitro and N-oxide nitrogen radius, in the G1−4 set, has a radius of zero. The dielectric on this nitrogen atom is only slightly smaller than ε_{in} because of the large bound oxygen radii and the short N−O bond, typically 1.2 Å, which allows dielectric from the oxygen to spread over onto the nitrogen. It is also interesting to note that in the G1−24 set, there was a gain in accuracy when the nitrile nitrogen had its own radius.

The oxygen atom behavior can mainly be accounted for by two adjustable radii types, which was a significant advantage in the fitting process—the N-oxide and nitro functional groups still being an exception. Another interesting result is the large radius of the sulfur atom that is comparable to the bromine radius. However, it is not to say that the polarizability contribution of sulfur is equivalent. In fact, the bromine bonds are longer and hence offer a larger polarizable volume. This argument is also useful to explain why the fluorine radius is smaller than the hydrogen radius. For example, the model predicts a polarizability for tetrafluoromethane of 18 au compared to 17 au for methane and a polarizability of 76 au for hexafluorobenzene compared to 70 for benzene, all in close agreement with B3LYP. Because of water's special importance as a solvent, both the oxygen radius and the hydrogen radius were optimized to exactly match the B3LYP polarizability tensor. These various behaviors of the atomic polarization radii underscore their

Table 2. Error Obtained with the Optimized Polarization Radii of the G1 Sets When EPIC Molecular Polarizability Tensors Are Compared to B3LYP for Different Molecule Data Sets

model[a]	δ_{avg}[b] (%)	δ_{aniso}[b] (%)	RRMS[b] (%)
Polarizability Training Data Set: 265 Molecules			
G1−4	5.0	20.9	12.7
G1−9	3.2	9.1	6.7
G1−12	2.9	5.3	5.0
G1−24	2.3	5.2	4.4
Polarizability Validation Data Set: 442 Molecules			
G1−4	4.0	18.2	12.3
G1−9	2.7	7.6	6.7
G1−12	2.6	5.1	5.3
G1−24	2.1	5.4	4.6
Polarizability Data Set: 707 Molecules			
G1−4	4.4	19.2	12.4
G1−9	2.9	8.2	6.7
G1−12	2.7	5.2	5.2
G1−24	2.2	5.4	4.6

[a] Model using the parameters given in Table 1. [b] Cf. section 2.2.3.

difference in nature and purpose from vdW contact radii, which is why they must be treated differently.

Finally, charged species pose a special challenge that we decided to address specifically for charged side chains in proteins: Arg, Lys, Asp, Glu, and His. Further generalization of the radii for charged species while retaining the same level of accuracy in the polarization tensor would require a more extensive parametrization. One reason for this is the expected reduction in polarizability on the neighbor atoms through the strong induction caused by the charged site. On the other hand, the electrostatic interactions around charged centers will be dominated by the monopole (i.e., the distribution of the charge itself), so high accuracy in the effects of polarization may become less important than with neutral species.

4.1.3. Polarizability Tensors. The G1 parametrizations clearly showed the capacity of EPIC to produce accurate polarizabilities with a minimum of atom types. The choice of ε_{in} and A combinations made based on the very small range-finding subset showed the same behavior in the polarizability training data set, the polarizability validation data set, and their combination (polarizability data set), made of 265, 442, and 707 molecules, respectively. Table 2, which summarizes the errors, shows the accuracy of the obtained models. The G1−24 data set has an unsigned average error of 2% on the average polarizability (eq 13) and a 5% error on the anisotropy of the tensor (eq 14). With the point inducible dipole polarizable models, such a low level of error was obtained only when anisotropic atomic polarizabilities were fitted,[5,15,65−67] making their generalization very challenging. The other G1 models are worse, and as predicted from the range-finding study results shown in Figure 3, the G1−4 set is inadequate to reproduce the directional difference in the polarizability (the large δ_{aniso} values in Table 2). That the error obtained on both the training data set and the validation data set was similar indicates that our radii are not overfit. Finally, the three directional polarizabilities (eigenvalues of the tensor) obtained for the 707 molecules (2121 data points) are compared to the corresponding B3LYP values in Figure 4 for three representative G1 sets. The excellent correlation is obvious for the G1−24 and G1−12 and deteriorates in the G1−4 EPIC model (the Pearson correlation coefficients are 0.99, 0.99, 0.96 and the slopes 0.97, 1.02, and 1.20, respectively). An apparent outlier is the α_3 (longitudinal polarizability) of (3E)-hexa-1,3,5-triene for which B3LYP gives a value of 176 au compared to the EPIC value of 125 au. For this specific molecule, Sekino et al.[68] showed that B3LYP greatly overestimates the α_3 value of acetylene chains. Their better estimate, based on very accurate CCSD and MP2 QM results, predicts a value of ∼135 au, close to the EPIC value. Another remarkable discrepancy between EPIC and B3LYP is observed in Figure 4 for the α_3 of 1,4-dioxidopyrazine (doubly oxidized nitrogen on pyrazine) that is predicted to be 103 au by the G1−12 model versus 129 au by B3LYP. A similar observation can be made for 4-nitroaniline. Although we have not found better estimates for these molecules, they most certainly constitute a challenge both for classical and *ab initio* polarizability calculations.

4.2. Refractive Indices. In the previous subsection, we have developed dielectric functions that predict remarkably well, relative to QM, the polarizabilities of a single molecule in the gas phase. In this section we present the *macroscopic* refractive index calculations and the corresponding effective high frequency limit dielectric (ε_∞). In a previous publication,[1] we proposed that the vacuum of the intermolecular spacing may be sufficient to reduce the effective macroscopic ε_∞ resulting from the high intramolecular ε_{in} obtained in the optimization to polarizability tensors. Here we use a theoretical approach to verify this hypothesis. Another important point addressed by the refractive index calculation is the transferability of the dielectric function from the gas phase to the condensed phase.

As explained in further detail in the Theory and Method sections, we form liquid droplets containing thousands of molecules from snapshots obtained by MD simulations and calculate the effective ε_∞ by the use of the Clausius-Mossoti equation. The small range spanned by experimental refractive indices makes this test somewhat stringent. Figure 5 shows the correlation between the results obtained with three representative EPIC parametrizations and experiment; G1−9 is omitted here and for the remainder of the article because the results are so similar to those of G1−12. The first observation is the close agreement between the magnitudes of the ε_∞ values. This clearly demonstrates that the effective ε_∞ of the liquid droplets have the appropriate value in spite of the high ε_{in}, confirming our hypothesis. Figure 6 provides a visual explanation for the apparent mismatch between the low effective ε_∞ compared to the high ε_{in}. This figure shows the molecular dielectric inside a CCl_4 droplet when it is sliced through its center. The G1−24, G1−12, and G1−4 models have a quite variable low-dielectric intermolecular space. The coloring scheme of the dielectric function (eq 1) assigns red when $\varepsilon(r) = \varepsilon_{in}$ and dark blue when $\varepsilon(r) = 1$. The low dielectric intermolecular space increases with ε_{in} as the atomic radii decrease. It is striking that these three parametrizations produce the same refractive index and the same molecular polarizability in spite of the very different ε_{in}. If

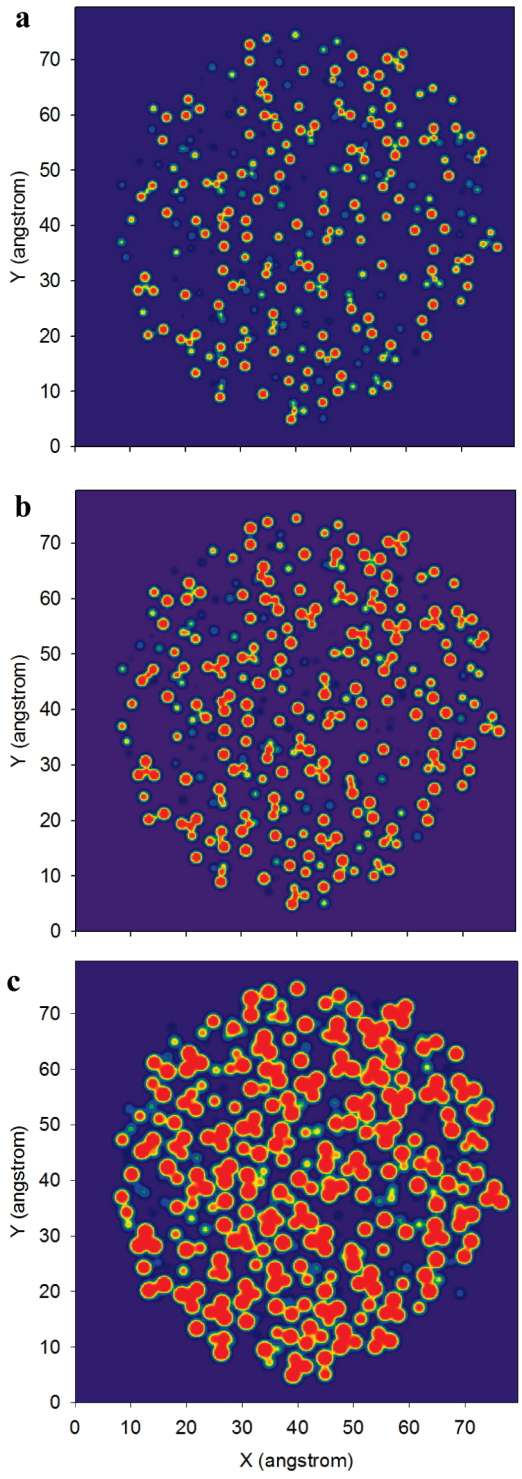


Figure 6. One of the 50 CCl_4 droplets is cut in its center, and three dielectric functions (eq 1) are plotted: a) G1−24 b) G1−12, and c) G1−4. The red color is associated with $\varepsilon(r) = \varepsilon_{in}$ and blue with $\varepsilon(r) = 1$.

ε_{in} is further reduced below 4, the whole droplet will become filled with a uniform dielectric (as the atomic radii increase and start to overlap), and the simultaneous prediction of the molecular polarizability and the refractive index will become compromised.

Also noticeable in Figure 5 is that the correlation with experiment follows the previous assessment of the models based on molecular polarizabilities: the G1−24 parametrization (Figure 5a) has a $R = 0.95$, slightly better than the G1−12 (Figure 5b) with $R = 0.94$, which is in turn significantly better than the G1−4 correlation with $R = 0.86$ (Figure 5c). However, Figure 5 shows a 0.05 systematic overestimation of the refractive indices in all cases which could correspond to a small overpolarization, a result not reflected in the gas-phase polarization tensors. The source for this deviation is not clear, but we have several hypotheses. First, the Clausius-Mossoti equation is valid for a perfect sphere, whereas we are dealing with an imperfect surface created by nanoscopic droplets. Second, we have verified that an underestimation of the droplet radius by only 3−4% (1 Å in a range of 25−35 Å) could systematically shift the calculated refractive indices by 0.05. Third, it is also possible that the liquid phase polarizability may be truly smaller than the predicted gas phase polarizability since a drop of 11% of the polarizability could explain the 0.05 shift. This would be in agreement with other studies that found similar phenomena[8,11,69] and based their reasoning on the increased Pauli exchange repulsion from the closer contact of the molecules in condensed phase. However, the magnitude of this effect differs considerably from study to study.

4.3. Hydration Free Energies. The calculation of hydration free energies is aimed at assessing whether the dielectric polarization model capable of accurately reproducing gas-phase polarizability tensors can be used "as is" in implicit solvent calculations. Because of the difference in nature and behavior between the atomic polarization radii and the atomic cavity radii used for the solute−solvent boundary, the 3-zone dielectric model is required. The hydration free energies were calculated with the G1−24, G1−12, and G1−4 polarizability models found in Table 3. For each polarizability model, charges were fitted to the vacuum-phase QM ESP as described in section 2.4.2; thus neither the various G1 molecular polarizability parameters nor the atomic charges used for the hydration calculations have been influenced by any effects of solvent. As discussed in ref 2, the high internal dielectric screening of the G1 polarizability causes the DRESP fitted charges to be of markedly higher magnitudes than if they were fitted using an internal dielectric of 1. By the same token, the differences between the various G1 models also result in different charges for each model. A comparison of the resulting G1−24, G1−12, and G1−4 charges for several example molecules and the G1−12 charges for the entire hydration data set are given in the Supporting Information. With the atomic charges in hand, each of the solute models was then used to optimize the solvent cavity atomic radii (referred to as cavity radii and noted G2 in what follows) referred to in the second Gaussian summation in eq 3. Each set of cavity radii is thus associated with a given charge set and molecular polarizability set, for example G2−12 is the set of cavity radii associated with the G1−12 molecular polarizability parameters and the charge set fitted using the G1−12 parameters. We decided to set $B = 11.8$ in all calculations, following the Grant et al.[18] suggestion as it was found to make the Bondi radii[16] optimally reproduce the hard dielectric boundary results with the same smooth boundary as used in this work. The results reported in Table 3 are split into two main categories based

Table 3. Solvent Cavity Atomic Radii (σ_{cavity}) and γ for the 3-Zone Dielectric Model Optimized on 485 Experimental Free Energy of Hydration with Different G1-n Solute Models and ΔG_{np} Sources

	model[a]					
	G2-HF	G2−4	G2−12	G2−24	G2−12SA	
	solute					
charges[b]	HF[c]	B3LYP[d]	B3LYP[d]	B3LYP[d]	B3LYP[d]	
ε_{in}[e]	1	4	12	24	12	
A[e]		10	10	4.19	10	
ref. Table 1		G1−4	G1−12	G1−24	G1−12	
ΔG_{np}	FEP[f]	FEP	FEP	FEP	SA[g]	
B	11.8	11.8	11.8	11.8	11.8	
	optimized implicit solvent parameters					Bondi
H[h]	0.98	0.95	0.97	1.02	0.98	1.20
C	1.95	2.03	2.02	1.95	2.01	1.70
N	1.74	1.74	1.74	1.68	1.69	1.55
O	1.81	1.79	1.78	1.75	1.75	1.52
S	2.60	2.27	2.29	2.33	2.41	1.80
F	2.09	2.09	2.08	2.05	2.49	1.47
Cl	2.38	2.36	2.47	2.41	2.46	1.75
Br	2.18	2.23	2.46	2.45	2.63	1.85
γ[i]					6.8	
AUE[j]	1.06	0.99	1.04	1.08	1.13	
Stdev[k]	1.00	0.96	0.99	1.00	0.90	
rms[l]	1.45	1.38	1.44	1.47	1.45	
R[m]	0.89	0.90	0.90	0.89	0.88	
RRMS[n]	0.34	0.33	0.34	0.35	0.34	
AE[o]	−0.18	−0.17	−0.17	−0.17	0.02	

[a] Tag names for each of the optimized solvent cavity radii. [b] Atomic partial charges from an ESP-fit or a DRESP fit on the given quantum method. [c] Prepolarized charges from HF/6-31G(d,p). [d] Vacuum charges from B3LYP/6-311++G-(3df,3pd). [e] A and ε_{in} of eq 3 for the solute internal dielectric. The atomic radii used in the internal dielectric are given in Table 1. [f] ΔG_{np} from free energy perturbation.[45] [g] ΔG_{np} calculated using the surface area (eq 20) with the γ term optimized. [h] Cavity atomic radii are given in angstrom. [i] Nonpolar surface tension from eq 20 in cal/Å^2. [j] Average unsigned error in kcal/mol. [k] Standard deviation of the unsigned error. [l] Root-mean-square deviation in kcal/mol. [m] Pearson correlation coefficient. [n] Relative root-mean-square deviation. [o] Average signed error in kcal/mol: experiment - calculated.

on the method used to approximate ΔG_{np}. The surface area (SA) based method follows eq 20 and required the optimization of the surface tension parameter (γ). The main effort here is however concentrated with the use of ΔG_{np} by converged free energy perturbation (FEP) calculations.[45,46] This is the first category of results that we examine below.

4.3.1. Results with FEP-Based Nonpolar Term. Two main classes of solute are studied as reported in Table 3. First, we set $\varepsilon_{in} = \varepsilon_{trans} = 1$ in eq 3, effectively turning eq 3 into a conventional 2-zone dielectric function with a nonpolarizable solute as defined previously by Grant et al.[18] For the nonpolarizable solute model, we used static atomic charges as given by ESP-fitting to the conventional overpolarized HF-6-31G(d,p) wave function (G2-HF). These charge sets are positive controls, following the traditional approaches for nonpolarizable force fields and which have been shown to produce the right degree of static polarization of the solute in water.[70] The G2-ε_{in}, with $\varepsilon_{in} = 4$, 12, and 24 (cf. Table 1), has polarizable solutes assigned charges fitted to the B3LYP/6-311++G(3df,3pd) ESP known to reproduce the gas-phase dipole moment of the molecules, being usually between 10% and 20% smaller than what is normally expected in water.

Table 4. Effects of Using a Different Solvent Cavity Radii Set (Table *3*) with the G1-12 Solute Model (Table *1*) on ΔG_{hyd}

	G2−12	G2−24	G2−4	G2-HF
AUE[a]	1.04	1.08	1.17	1.10
Stdev[b]	0.99	1.03	1.19	1.01
R[c]	0.90	0.90	0.86	0.89
RRMS[d]	0.34	0.35	0.39	0.35
AE[e]	−0.17	0.16	−0.23	0.15

[a] Average unsigned error in kcal/mol. [b] Standard deviation on the AUE in kcal/mol. [c] Pearson correlation coefficient. [d] Relative root-mean-square deviation. [e] Average signed error in kcal/mol: experiment - calculated.

We look for the polarizability to compensate for the use of gas-phase charges.

It is quite interesting to observe in Table 3 that by allowing the cavity radii to optimize in each model, the same level of error over the 485 experimental free energies of hydration is obtained for the G2-HF, G2−4, G2−12, and G2−24 solute models. The average unsigned error (AUE) compared to experiment is 1 kcal/mol with a standard deviation of 1 kcal/mol. The Pearson correlation coefficient (R) is around 0.89 in all these G2 models. The relative root-mean-square deviation (RRMS) obtained is 0.35, and the average signed error (AE) is found to be between −0.15 kcal/mol and −0.18 kcal/mol. These errors can be compared to the Rizzo et al.[47] results, on almost the same data set (460 neutral molecules included in the 485 that we use), that produce an AUE of 1.47 kcal/mol with RESP charges and R = 0.88. The reported numbers of Rizzo et al. were obtained with a SA evaluation of ΔG_{np} that allow them to subsequently optimize 14 atom typed surface tensions (γ), which improved the AUE to 1 kcal/mol while R = 0.89. For comparison, in the current study, we fit 8 atomic radii. In addition, the recent work of Mobley et al.[46] using Bondi radii and the single γ fitted by Rizzo et al. on an almost identical data set to ours obtained a root-mean-square deviation of 2.05 kcal/mol. Finally, in a different article, Mobley et al.[45] obtained a rms of 1.26 kcal/mol and R = 0.89 with explicit-solvent converged FEP calculations. The FEP based ΔG_{np} used in this work comes from this latter study. Our results are comparable or better to most other studies. We attribute the small errors to the optimization of the radii, not necessarily to the quality of the solute model. We can however examine the fitted cavity radii with the different solute models to understand the effects of the electrostatic model on the solute cavity size.

The level of solute polarization brought by the polarizable solute models (G2−4 to G2−24) seems similar to what is obtained with the G2-HF solute model. This can be assessed by comparing the atomic radii and the cross-validation error showed in Table 4 where the G1−12 solute model is used with the different G2 radii sets. The level of error produced when $\varepsilon_{in} = 4$, 12, and 24 or with the G2-HF cavity radii is similar, the G2−4 being the worst. The cross-validation results of Table 4 also show the transferability of the third zone dielectric parameters given that the solute has the physically appropriate electrostatic behavior. A possible advantage of the polarizable solute model is when the solvation free energies are computed relative to a solvent much less polar than water (e.g., a nonpolar solvent or a

nonpolar binding site in a protein). In this case, the HF based charges may not be appropriate.[71]

The fitted radii of Table 3 are significantly different from the contact Bondi radii reported in the last column. First, the H radius is a little smaller than the usual 1.1 Å contact radius in all cases (the Bondi radius of 1.2 Å for H was recognized to be a little too large and was revised to be 1.1 Å[72]). The carbon radius obtained here is much larger than the Bondi radius and makes the C−H bonds behave like a united atom model. In this perspective the carbon radius size obtained here is similar to the Nina et al.[73] carbon radius they calculated by looking at MD water charge density in explicit solvent simulations. For the other elements, we also find larger radii than Bondi, in agreement with a recent study by Nicholls et al.[31]

The larger cavity radii can be rationalized by considering the difference between contact radii (Bondi) and the cavity radii needed in implicit solvent calculations. The former is defined by crystal contacts between neighboring molecules, and the latter is defined by where the mean solvent charge density begins. In terms of the 3-zone dielectric model, the contact radii would be located in the middle of the second zone where the electronic density should be minimal given that the dielectric goes to one (no electrons to be polarized on the solute side). This is supported by the fact that Fermi repulsion between the solvent molecule electrons and the solute electrons reduces the total electronic density to its minimum exactly in the contact zone. In Figure 2a, the contact radius of an aromatic carbon atom would become 1.7 Å, exactly the Bondi radius value. Similarly the middle of the blue area in Figure 2b defines the contact line between solvent molecules and the solute.

Although we claim here that having cavity radii larger than Bondi radii may be physically motivated, it is not possible at this stage to know if this effect should be as large as we find. In particular, the fluorine radii in Table 3 are surprisingly large. This was also found by Nicholls et al.[31] where their optimal fluorine radius was 2.4 Å. Knowing that fluorine is particularly hydrophobic, this may be just another peculiar behavior of this atom. The Cl and Br radii difference in the G2−4, G2−12, and G2−24 sets uncover a drawback of using a small ε_{in}. Because the polarization radius of Cl and Br are larger in the G1−4 than in the other EPIC parametrizations, the transition zone shown in Figure 2 cannot reach $\varepsilon(r) = 1$ (in the case of Br, it only decreases to $\varepsilon(r) = 3$), and as a result the full polarizability coming from the halogen atom is not reached as the solvent cuts into the first zone dielectric function. This prevents enough solute bound charge density from building up.

A last point to mention in regard to the 3-zone dielectric function is its potential advantage in reducing the occurrence of the reentrant surface problem that often brings a lot of fluctuation in energy or force computations in proteins, for instance. The problem is the artificial formation of a cavity with high dielectric inside a protein due to the irregularity of molecular surfaces. The large size of the atomic cavity radii in the G2 sets and the use of a smooth dielectric function should contribute to create a sufficiently deep buffer of low dielectric and make implicit solvent models more stable. Indeed, the smoothness of the surface around 4-pyridone observed in Figure 2b looks like a solvent accessible surface.[17] This entire question is however left for future research.

4.3.2. Results with the Surface Area-Based Nonpolar Term. Although the use of the very computationally intense FEP-based ΔG_{np} may be better physically grounded, the obtained models cannot practically be used in a prospective manner due to the heavy computational demands for such FEP calculations. For this reason, we also optimized the cavity radii and the surface tension with the G1−12 solute models. In these calculations the solvent accessible surface area was calculated with the Bondi radii and kept constant. The results are reported in Table 3. The error levels reported are comparable to those obtained with ΔG_{np} from FEP calculation. The G2−12/SA model gives error levels a little larger than the G2−12: AUE = 1.13 kcal/mol with a standard deviation of 0.90 kcal/mol, R = 0.88, RRMS = 0.34, rms = 1.45 kcal/mol, and AE = 0.02 kcal/mol. The radii obtained for the G2−12/SA fit are similar to the G2−12 fit except for S, F, Cl, and Br. It is possible that the hydrophobicity of these atoms is overestimated by the single surface tension term used with the Bondi radii to determined ΔG_{np}.

5. Conclusions

The EPIC approach to molecular polarizability has been parametrized to include many more chemical functional groups than the previous effort.[1] This required generating a data set of 707 B3LYP/aug-cc-pVTZ molecular polarizability tensors. The ability of EPIC to account for both the average polarizability and the anisotropy of the tensor was remarkable given that the optimization of only 14 parameters (excluding water and charged species) led to a relative unsigned error in the average polarizability and anistropy of 2.6% and 5.2%, respectively (G1−12). An example of the parsimony of the atom typing is that a single radius parameter was sufficient for aromatic, nitrile, amine, aniline, or hydrazine types of nitrogen. Obtaining the same level of error with both the validation and training data sets suggests that overfitting is not an issue. With previous polarizable models, such as point-inducible dipoles, this level of accuracy could only be attained with added complexity such as anisotropic polarizable centers or molecule-specific Thole screening parameters.[5,15,28,65−67]

We found that the anisotropy could only be reproduced accurately if the interior dielectric constant was higher than 9.0. Above this value, almost any interior dielectric can also work well as long as the atomic polarization radii are appropriately adjusted. The need for a high interior dielectric raised the question of the physical soundness of the model as $\varepsilon_{in} = \varepsilon_{\infty} = n^2$ has become a dogma in the implicit solvent literature[13,30,31,39,40,74−76] whenever the dielectric constant is used to replace electronic response. The conceptual flaw of this equality comes from the fact that the interior dielectric (ε_{in}) is not uniformly distributed, whereas the refractive index (n) comes from a macroscopic measurement assuming a uniform ε_{∞} in space. To verify that the optimized models agree with experimental refractive indices, we devised a new protocol to calculate a liquid refractive index from micro-

scopic simulations. To this end, 23 organic molecules spanning the entire range of bioorganic-molecule-like refractive indices state were simulated by molecular dynamics in the liquid state and ε_∞ was calculated with the Clausius-Mossoti relation. The obtained refractive indices now come from the effective polarization of liquid configurations having intramolecular high dielectric with low-dielectric interstices. The results show a good correlation for all three G1-ε_{in} parameter sets. The highest interior dielectric ($\varepsilon_{in} = 24$) gave the best correlation with a slope of 1.00, an intercept of 0.05, and a correlation coefficient of 0.95. It is interesting to note that the polarizability anisotropy may play a role since the G1−4 parameters ($\varepsilon_{in} = 4$) gave the poorest correlation and was also the worst model for polarizability anisotropy. These results indicate that, when coupled with the appropriate radii, many choices of high ε_{in} can give results in good agreement with the experimental refractive indices.

To use the EPIC polarizable electrostatic model with implicit solvent, we have developed a smoothed-boundary 3-zone dielectric function that works with the internal dielectric continuum model. The three zones are the internal dielectric constant ε_{in}, a transition zone that tends to the vacuum dielectric ($\varepsilon_{trans} = 1$), and a third zone defined by the molecular cavity boundary, where the dielectric function reaches the bulk liquid dielectric. With this function, keeping the first zone fixed at the atomic polarization radii and ε_{in} determined for the gas phase polarizabilities, only the molecular cavity boundary needs to be parametrized. A data set of 485 experimental free energy of hydration was used to optimize the solvent cavity radii, one per element, with different charge models. The resulting level of error was smaller than found in previous implicit solvent studies with a typical average unsigned error of 1 kcal/mol, a standard deviation of about 1 kcal/mol, and a Pearson correlation coefficient of 0.9. Atomic charge sets fitted from the unpolarized gas phase B3LYP QM ESP coupled with EPIC polarization led to cavity radii comparable to those obtained with the polar-condensed-phase-like HF/6-31G(d,p) charges. The low sensitivity of the optimal cavity radii resulting from the fit with different polarizable solutes (the different G1-ε_{in}) further supports the generality of the approach. These results clearly show that EPIC can lead to accurate description of solute polarization in implicit solvent. The anisotropy of the molecular polarizability does not seem to play an important role in fitting experimental hydration free energies. However, when considering intermolecular interactions, such as in an enzyme active site, the heterogeneity of the environment and of the interactions may require an accurate directional polarizability. An important example of this is in cation-π interactions[2,77]

The proposed global optimization scheme involves several independent layers. The polarizability part is fitted on uncharged QM molecular polarizability tensors. The charges are added with the DRESP fit on *ab initio* electrostatic potentials calculated on a grid, as usual, except that here the QM method can be systematically improved since gas phase properties are needed. For implicit solvation, solvent-related radii are obtained from a fit to experimental hydration free energies. Flexibility and transferability have been demonstrated for each stage. The ease with which we could fit the polarizability of so many functional groups leads us to believe that the further extension of the parametrization and atom typing should be straightforward given more data. Moreover, the decoupling of the fitted polarization from the fitted charges as well as the physical soundness of each step makes the above parametrization scheme even more robust and general than is possible for two-body additive force fields.

This work partly addresses the question of applicability of EPIC in polar condensed phase. The calculations of the refractive indices were well behaved and show the high values for ε_{in} are not unphysical. It is not clear if the slightly larger calculated polarizabilities of the droplets were due to a change in polarizability when going from gas phase to condensed phase. Also, the level of electronic induction seen in the 3-zone implicit solvent calculations suggests that both solute and solvent polarization in polar media is well modeled. To confirm those findings, it would be interesting to perform explicit atoms simulations with the EPIC model and Poisson's equation.

The EPIC approach to polarizability has shown unprecedented accuracy and flexibility on many accounts for such a simple model. Although the optimized parameters are unconventional compared to traditional Poisson−Boltzmann applications, it is for sound physical reasons that even clarify aspects of the implicit solvent approaches. In this paper and the two previous ones,[1,2] EPIC was shown to be a powerful tool to include the effects of electronic polarization in molecular mechanics type calculations, especially appropriate to biomolecular force fields.

Acknowledgment. This work was made possible by the computational resources of the réseau québécois de calcul haute performance (RQCHP). The authors are grateful to OpenEye Inc. for free academic licenses. R.I.I. acknowledges financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC). J.-F.T. is supported by NSERC through a Canada graduate scholarship (CGS D) and by Merck & Co. through the MRL Doctoral Program I. B.R. is supported by NIH grant GM072558.

Supporting Information Available: DFT and EPIC polarizabilities for all 707 molecules examined together with the optimized coordinates of the molecules used to calculate the polarizability tensor, table of the G1−24, G1−12, and G1−4 fitted charges for several example molecules, G1−12 charges together with B3LYP/6-31++G(d,p)-optimized geometries for the 485 molecules examined for hydration free energies, calculated and experimental refractive indices, and detailed free energies of hydration. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Truchon, J.-F.; Nicholls, A.; Iftimie, R. I.; Roux, B.; Bayly, C. I. *J. Chem. Theory Comput.* **2008**, *4*, 1480–1493.

(2) Truchon, J.-F.; Nicholls, A.; Grant, J. A.; Iftimie, R. I.; Roux, B.; Bayly, C. I. *J. Comput. Chem.* in press.

(3) Applequist, J.; Carl, J. R.; Fung, K. K. *J. Am. Chem. Soc.* **1972**, *94*, 2952–2960.

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

(5) Miller, K. J. *J. Am. Chem. Soc.* **1990**, *112*, 8543–8551.

(6) Lamoureux, G.; Allouche, D.; Souaille, M.; Roux, B. *Biophys. J.* **2000**, *78*, 330A–330A.

(7) Lamoureux, G.; Roux, B. *Biophys. J.* **2001**, *80*, 328A–328A.

(8) Schropp, B.; Tavan, P. *J. Phys. Chem. B* **2008**, *112*, 6233–6240.

(9) Lamoureux, G.; MacKerell, A. D.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185–5197.

(10) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D. *J. Chem. Theory Comput.* **2005**, *1*, 153–168.

(11) Morita, A. *J. Comput. Chem.* **2002**, *23*, 1466–1471.

(12) Sharp, K.; Jean-Charles, A.; Honig, B. *J. Phys. Chem.* **1992**, *96*, 3822–3828.

(13) Tan, Y. H.; Luo, R. *J. Chem. Phys.* **2007**, *126*, 094103.

(14) Tan, Y. H.; Tan, C. H.; Wang, J.; Luo, R. *J. Phys. Chem. B* **2008**, *112*, 7675–7688.

(15) Birge, R. R. *J. Chem. Phys.* **1980**, *72*, 5312–5319.

(16) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441–451.

(17) Richards, F. M. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151–176.

(18) Grant, J. A.; Pickup, B. T.; Nicholls, A. *J. Comput. Chem.* **2001**, *22*, 608–640.

(19) Prabhu, N. V.; Zhu, P. J.; Sharp, K. A. *J. Comput. Chem.* **2004**, *25*, 2049–2064.

(20) Im, W.; Beglov, D.; Roux, B. *Comput. Phys. Commun.* **1998**, *111*, 59–75.

(21) Griffiths, D. J. *Introduction to Electrodynamics*; 3rd ed.; Prentice-Hall Inc.: Upper Saddle River, NJ, 1999.

(22) *Zap Toolkit, pre-release version*; Santa Fe, NM, USA, 2007.

(23) Weininger, D. *J. Chem. Inf. Model.* **1990**, *30*, 237–243.

(24) Weininger, D.; Weininger, A.; Weininger, J. L. *J. Chem. Inf. Model.* **1989**, *29*, 97–101.

(25) Weininger, D. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

(26) *OEChem Toolkit, version 2.2.1*; Santa Fe, NM, USA, 2007.

(27) SciPy: Open Source Scientific Tools for Python, 2001.

(28) Elking, D.; Darden, T.; Woods, R. J. *J. Comput. Chem.* **2007**, *28*, 1261–1274.

(29) Bottcher, C. J. F. *Theory of Electric Polarization*; 2nd ed.; Elsevier Scientific Publishing Company: New York, 1973; Vol. 1.

(30) Onsager, L. *J. Am. Chem. Soc.* **1936**, *58*, 1486–1493.

(31) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. *J. Med. Chem.* **2008**, *51*, 769–779.

(32) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(33) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(34) Jakalian, A.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

(35) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132–146.

(36) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(37) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.

(38) Marsaglia, G. *Annu. Math. Stat.* **1972**, *43*, 645–646.

(39) Warwicker, J.; Watson, H. C. *J. Mol. Biol.* **1982**, *157*, 671–679.

(40) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978–1988.

(41) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(42) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.

(43) Mobley, D. L.; Dill, K. A.; Chodera, J. D. *J. Phys. Chem. B* **2008**, *112*, 938–946.

(44) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(45) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. *J. Chem. Theory Comput.* **2009**, *5*.

(46) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2007**, *111*, 2242–2254.

(47) Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D. *J. Chem. Theory Comput.* **2006**, *2*, 128–139.

(48) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479–499.

(49) Pitera, J. W.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **2001**, *123*, 3163–3164.

(50) Choudhury, N.; Pettitt, B. M. *Mol. Simul.* **2005**, *31*, 457–463.

(51) Wagoner, J. A.; Baker, N. A. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 8331–8336.

(52) Nicholls, A.; Honig, B. *J. Comput. Chem.* **1991**, *12*, 435–445.

(53) Varga, R. S. *Matrix Iterative Analysis*; 2nd ed.; Springer: New York, 2000.

(54) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(55) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

(56) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Menucci, 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, USA, 2004.

(57) Kendall, R. A.; Dunning, J.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

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

(59) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. *J. Comput. Chem.* **1983**, *4*, 294–301.

(60) Woon, D. E.; Dunning, J. *J. Chem. Phys.* **1993**, *98*, 1358–1371.

(61) Hammond, J. R.; Kowalski, K.; deJong, W. A. *J. Chem. Phys.* **2007**, *127*, 144105.

(62) Dykstra, C. E.; Jasien, P. G. *Chem. Phys. Lett.* **1984**, *109*, 388–393.

(63) *Handbook of Chemistry and Physics*, 83 ed.; CRC Press: New York, 2002.

(64) Pauling, L. *The Nature of the Chemical Bond*; Cornell University Press: Ithaca: New York, 1960.

(65) Shanker, B.; Applequist, J. *J. Phys. Chem.* **1996**, *100*, 3879–3881.

(66) Bode, K. A.; Applequist, J. *J. Phys. Chem.* **1996**, *100*, 17820–17824.

(67) Harder, E.; Anisimov, V. M.; Whitrield, T.; MacKerell, A. D.; Roux, B. *J. Phys. Chem. B* **2008**, *112*, 3509–3521.

(68) Sekino, H.; Maeda, Y.; Kamiya, M.; Hirao, K. *J. Chem. Phys.* **2007**, 126, 014107.

(69) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621–627.

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

(71) Eksterowicz, J. E.; Miller, J. L.; Kollman, P. A. *J. Phys. Chem. B* **1997**, *101*, 10971–10975.

(72) Rowland, R. S.; Taylor, R. *J. Phys. Chem.* **1996**, *100*, 7384–7391.

(73) Nina, M.; Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 5239–5248.

(74) Jayaram, B.; Liu, Y.; Beveridge, D. L. *J. Chem. Phys.* **1998**, *109*, 1465–1471.

(75) Warshel, A.; Sharma, P. K.; Kato, M.; Parson, W. W. *BBA -Proteins Proteom.* **2006**, *1764*, 1647–1676.

(76) Green, D. F.; Tidor, B. *J. Phys. Chem. B* **2003**, *107*, 10261–10273.

(77) Jiao, D.; Golubkov, P. A.; Darden, T. A.; Ren, P. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 6290–6295.

CT900029D

JCTC Journal of Chemical Theory and Computation

United Atom Lipid Parameters for Combination with the Optimized Potentials for Liquid Simulations All-Atom Force Field

Jakob P. Ulmschneider*,† and Martin B. Ulmschneider‡

IWR, University of Heidelberg, Heidelberg, Germany, and Department of Chemistry, University of Utrecht, Utrecht, The Netherlands

Received February 19, 2009

Abstract: We have developed a new united-atom set of lipid force field parameters for dipalmitoylphosphatidylcholine (DPPC) lipid bilayers that can be combined with the all-atom optimized potentials for liquid simulations (OPLS-AA) protein force field. For this, all torsions have been refitted for a nonbonded 1−4 scale factor of 0.5, which is the standard in OPLS-AA. Improved van der Waals parameters have been obtained for the acyl lipid tails by matching simulation results of bulk pentadecane against recently improved experimental measurements. The charge set has been adjusted from previous lipid force fields to allow for an identical treatment of the alkoxy ester groups. This reduces the amount of parameters required for the model. Simulation of DPPC bilayers in the tension-free NPT ensemble at 50 °C gives the correct area per lipid of 62.9 ± 0.1 Å^2, which compares well with the recently refined experimental value of 63.0 Å^2. Electron density profiles and deuterium order parameters are similarly well reproduced. The new parameters will allow for improved simulation results in microsecond scale peptide partitioning simulations, which have proved problematic with prior parametrizations.

I. Introduction

In recent years computer simulations of proteins embedded in lipid bilayers have become a powerful tool to investigate this important class of proteins in its native environment. Advances in computer architecture now in principle allow classical all-atom molecular dynamics (MD) simulations of membrane proteins on the microsecond time scale, and much longer with coarse-grain models (see the recent review by Lindahl and Sansom).[1] This opens the possibility to directly observe the conformational transitions involved in protein function, such as gating and signaling. In addition, simulation time scales are now sufficient to study at atomic resolution the adsorption, folding, insertion, and self-assembly of many membrane-bound peptides, such as antimicrobials, viral channel formers, and synthetic peptides.[2−8] However, with the increased time scale of the simulations, there is a growing need to address deficiencies in the underlying models. At the time of their design most protein, water, and lipid force fields could not be tested beyond the picosecond to nanosecond range.[9] Especially tricky is the description of the membrane itself (i.e., the lipid force field) and the delicate balance of protein and lipid parameters. If such interactions are not well tuned, simulations can lead to results that are irreconcilable with experimental evidence and theoretical estimates: For example, in partitioning simulations of small hydrophobic peptides into lipid bilayers, unfolded conformations were buried in the hydrophobic core,[3] or were found to be favored over the expected transmembrane helix.[2,10] Such problems are most visible at slightly elevated temperatures where sampling is increased and for small peptides, where lipid−protein interactions dominate. For larger multispan membrane proteins, the protein−protein interactions are much stronger, making these problems difficult to detect at the time scales currently accessible. Indeed, because multiple helices are usually tightly locked through strong contacts of interdigitating hydrophobic side chains, it is in practice quite difficult to unfold membrane proteins in computer simulations. However, while the transmembrane

* Corresponding author e-mail: jakob@ulmschneider.com.
† University of Heidelberg.
‡ University of Utrecht.

10.1021/ct900086b CCC: $40.75
Published on Web 06/01/2009

part of the protein usually remains rigid, the subtle force field imbalances can manifest themselves in perturbations of the flexible parts such as extramembranous loop regions, leading to incorrect conclusions about the flexibility or function from simulation data.

Accurate and reliable lipid parameters are therefore of the utmost importance to derive meaningful data from computer simulations of membrane proteins. There are several lipid parameter sets in regular use for bilayer simulations, such as the all-atom (AA) CHARMM lipids,[11,12] including united-atom (UA) adaptations[13] and modifications.[14,15] Common UA models are the Berger[16] and GROMOS[17] lipid models. Other recent efforts have been based on the generalized AMBER force field[18,19] or polarizable models.[20] An excellent summary on the topic of lipid force field development has also been published recently.[21]

A troubling problem of several of these lipid force fields is the inability of simulated bilayers to sustain the fluid (L_α) phase area per lipid in the NPT ensemble. Instead, a dramatic lateral contraction is observed, either resulting in a bilayer that is too densely packed or even causing a transition into the ordered gel phase.[22] To compensate, a positive surface tension can be applied.[23] Finite size effects have been put forward as an explanation, yet other studies find only a small dependence of the area per lipid on the system size, especially if long-range electrostatic corrections are taken into account.[24] It seems the only significant finite size effect is on the lateral diffusion of the lipids, but not on structural properties.[25] Instead of a surface tension, another way to obtain the correct area per lipid is to simply fix the total box area to be constant (NPAT ensemble). Both the applied surface tension and NPAT simulations are problematic, since they require additional parameters. For each lipid type, lipid mixture, temperature, and embedded protein, a different value or the area or surface tension needs to be supplied, which is often not available. These parameters have also been shown to vary greatly with lipid type and hydration level.[26,27] In addition, simulations in the NPAT ensemble do not allow the membrane to stretch and breathe laterally, hindering important conformational transitions in membrane proteins or the insertion of peptides into membranes. Thus, such an approach is inappropriate for partitioning simulations. The delicate balance of large and opposing forces that goes into the lateral pressure profile indicates that a fragment-based parametrization strategy is possibly limited, and fitting to bulk lipid properties such as the area per lipid is required. Recently, Sonne et al. have described a reparametrization of the CHARMM force field to obtain greatly increased areas per lipid, though still 6% below experimental values for dipalmitoylphosphatidylcholine (DPPC).[14] Similarly, Högberg et al. have reported improved CHARMM parameters for tension-free simulations of dimyristoylphosphatidylcholine (DMPC).[15] Improved GROMOS parameters are also reported by Kukol.[28] Interestingly, even very recent new lipid parameter sets, such as the GAFF set,[19,29] or polarizable models,[20] do not provide the correct area per lipid in the tension-free NPT ensemble.

A lipid parameter set that yields reasonable areas per lipid in the NPT ensemble is the widely used UA model by Berger et al.[16] A UA description is beneficial for performance reasons: For example, an all-atom model of a DPPC molecule requires 130 atoms, while in UA models this number is reduced to 50. Thus, the calculation of lipid−lipid interactions is 6.7 times faster in the UA model. The Berger lipid paramters were essentially derived from previous united-atom optimized potentials for liquid simulations (OPLS-UA) studies by Essex et al.,[30] who described the first OPLS-UA lipid parameters by assembling parameters from studies of Jorgensen et al. on hydrocarbons,[31] ammonium ions,[32] and esters.[33] The missing types for connecting atoms between the various functional groups (choline, phosphate, glycerol) were obtained in an ad hoc fashion. For example, for C6 and C12 in DPPC, the methyls in dimethyl phosphate ($q = 0.2$, $\sigma = 3.8$ Å, $\varepsilon = 0.170$ kcal/mol) were adjusted to CH_2 united atoms by lowering the Lennard-Jones (LJ) well depth to $\varepsilon = 0.118$ kcal/mol, which is the value for a hydrocarbon CH_2 in OPLS-UA. Subsequently, Berger et al.[16] described a parameter set where the charges were replaced with entirely new values based on quantum mechanical calculations by Chiu et al.[23] The Berger set retained the LJ parameters of the Essex lipids, with the exception of the lipid tail hydrocarbon united atoms, which were refitted to reproduce thermodynamic data for liquid pentadecane. Torsions were assigned from GROMOS.[34,35]

Most commonly used simulation software cannot handle the simultaneous use of force fields that differ in their basic structure. Typically, LJ combining rules as well as 1−4 scaling factors for torsions are usually hard-coded and have to be unique throughout. This has made it difficult to perform simulations where the protein is modeled by OPLS-AA,[36] with a 1−4 scaling factor of 0.5, and the lipid parameters are modeled with the Berger set, which uses different torsions and scale factors. Some tricks have been proposed in the literature to overcome these problems, such as removing 1−4 interactions completely and refitting new torsions.[37] However, scale factors are critically important to describe both intra- and intermolecular interactions. Small or zero scale factors—where the complete 1−4 interaction is handled by the torsion potential—were found to be problematic.[36] Thus, it would be beneficial to have a set of lipid parameters with torsions refitted for a scale factor of 0.5 for both the LJ and Coulombic 1−4 interactions.

The present work has arisen from a desire for a simple UA lipid set that can be combined with the description of the protein using OPLS-AA. Starting from the Berger parameters, we report such a set. The major changes include a refitting of all torsions using quantum-mechanical profiles as the starting point, followed by further refinement in long time scale lipid bilayer simulations. In addition, the hydrocarbon lipid tail LJ parameters were reworked by long simulations of bulk liquid pentadecane. In the original study by Berger et al., a too low value of the heat of vaporization was used in the parametrization, resulting in lipid tail LJ interactions that are significantly too weak. Third, the charges on the major groups were adjusted to allow for a simpler parametrization in which the ester groups are treated equally. Finally, a key requirement of the new model was the

Table 1. Summary of Simulations of Pentadecane Performed in This Study[a]

		parameters							
		C2		C3					
	force field	σ (Å)	ε (kcal/mol)	σ (Å)	ε (kcal/mol)	time (ns)	thermostat	barostat	T (°C)
Gromacs	G96	3.96	0.091	3.96	0.136	20	Berendsen	Berendsen	25−50
Gromacs	OPLS	3.955	0.103	3.955	0.154	20	Berendsen	Berendsen	25
Hippo	OPLS	3.955	0.103	3.955	0.154	400	Andersen	MC	−20 to + 100

[a] All simulations were performed for 267 pentadecane molecules in a cubic box, with a nonbonded cutoff of 20 Å, long-range LJ cutoff corrections, and a time step of 1 fs, in the NPT ensemble at a pressure of 1 bar and the indicated temperature.

reproduction of the correct area per lipid in the NPT ensemble, without applying a surface tension.

II. Methods

All simulations were performed with the Hippo MD/Monte Carlo program (www.biowerkzeug.com) and Gromacs (version 4.0.2, www.gromacs.org).[38] The pentadecane simulations were run with 267 molecules in a cubic box in the NPT ensemble. In most of the MD simulations, the Andersen thermostat was used to maintain the temperature.[39] The pressure was kept constant at 1 bar by performing isotropic Monte Carlo pressure moves at intervals of 1000 time steps. Such moves involve the scaling of the box by a small random amount and a rejection/acceptance of the move using the Metropolis criterion.[40] Both the Andersen thermostat and Monte Carlo pressure moves have the advantage to suffer from no known artifacts, with the resulting statistics exactly representing the NPT ensemble.[40] Thus, we used these methods to obtain accurate numbers for the simulation parameters. In practice, the results are almost identical when other commonly known coupling methods are used, such as the Berendsen thermostat.[41]

The pentadecane simulations were run with Hippo using a simple nonbonded cutoff of 20 Å, with a smooth feathering to zero over the last 0.5 Å. Since there are no net charges on the aliphatic hydrocarbon united atoms, no Coulombic interactions need to be calculated, and no long-range electrostatic methods (e.g., particle-mesh Ewald (PME)) are necessary. For united-atom hydrocarbons, the electrostatic effects originating from the small polarizability need to be implicitly covered by the LJ terms. Long-range LJ interactions beyond the cutoff were included by using the usual correction terms.[40] This affects both the potential energy and—through the barostat—the system volume. Exact definitions of the cutoff correction are only available for homogeneous LJ fluids. If the system contains atom types with diverse LJ parameters, no exact description is possible, and the correction equation assumes a mix of the atomic parameters. The contribution of the cutoff correction to the energy scales with $\sim r_{\mathrm{cut}}^{-3}$. To avoid an unnecessary influence of the parametrization results on the details of the correction terms, a very large cutoff of 20 Å was chosen for the LJ interactions. This also minimizes potential differences due to the exact nature of the cutoff treatment, which is based on the distance of the center of charge groups. The time step was set to 1 fs to avoid the results being influenced by integrator errors. Simulations were 200 ns long, with the runs at 25 and 50 °C extended to 400 ns.

Control Monte Carlo simulations were also performed with Hippo using a setup identical to that of the MD runs. The results were similar to those of MD. Additional simulations of pentadecane were also performed with Gromacs using the Berendsen temperature and pressure coupling[41] and a cutoff of 20 Å. Finally, several simulations were performed with the original Berger hydrocarbon parameters.

For both pentadecane and DPPC, all bond and angle parameters were taken from previous studies.[16,30] Torsions were refitted for a 1−4 scale factor of 0.5 by matching against HF/6-31G* profiles obtained by dihedral scans on the usual lipid fragments, such as methyl acetate, glycerol, dimethyl phosphate, and choline, as described previously.[21] For the all-C2 torsion in the lipid tail, the more accurate MP2:CC data reported by Klauda et al. were used in the fitting, which results in a significant lowering of the *gauche* wells as compared to HF/6-31G*.[11] For example, in butane $\Delta E_{trans \rightarrow gauche} = 1.01$ kcal/mol for HF/6-31G*, but decreases to 0.63 kcal/mol for MP2:CC. Future improvement is possible by applying the higher theory also to the other dihedrals, but this is beyond the scope of the present study. Interestingly, dihedrals were not fitted previously for the Berger force field. In the orginal study, mainly standard torsions from the GROMOS force field were applied,[16] and a similar approach was used in a recent update of the Berger parameters by Kukol.[28] The lipid bilayer simulations were run in the NPT ensemble at 323 K, using weak temperature coupling with a coupling constant of 0.1 ps,[41] semi-isotropic weak pressure coupling, and no applied surface tension. Both systems with 50 and 128 DPPC molecules were studied. As the differences between the thermodynamic results were very small (see the Introduction), the results reported here are for the smaller system. Simulations were 100 ns long, with a time step of 2 fs. Bonds involving hydrogen atoms were constrained.[42] Water was represented using the TIP3P water model,[43] electrostatic interactions were treated by the PME method,[44] and LJ interactions used simple cutoffs ranging from 10 to 14 Å. The total sampling time reported in this work is ~3 μs, while the accumulated simulation time during the complete parametrization work was about ~15 μs.

III. Pentadecane Results

A. Pentadecane. The saturated lipid tails of DPPC are parametrized using pentadecane as the molecular fragment. Table 1 gives an overview of the simulation performed on pentadecane. A challenge in united-atom parametrization for liquid hydrocarbons is that all electrostatic effects (although minor, given their small polarizability) must be implicitly

Table 2. Density and Heat of Vaporization of Pentadecane for the Original Lipid Parameter Set

		ρ (g/cm^3)		ΔH_{vap} (kcal/mol)	
	r_{cut}[a] (Å)	25 °C	50 °C	25 °C	50 °C
Berger[b]	10		0.7457		14.77
Berger[c]	20	0.7584	0.7433	15.54	15.15
Chiu[d]	20	0.7788		18.40	
exptl[e]		0.7650	0.7478	17.41	16.96
exptl[f]				18.35	

[a] Cutoff corrections for long-range LJ interactions are included. [b] From ref 16 based on 10 ps simulations of pentadecane. [c] Based on 20 ns. [d] Reference 47. [e] Reference 45. [f] Reference 46.

incorporated through the LJ potential, whereas explicit charge dipoles can be used in all-atom descriptions. As the original saturated linear hydrocarbon OPLS-UA parameters were designed for smaller alkanes up to hexane,[31] Berger et al. obtained improved LJ parameters for pentadecane by fitting to the available liquid-phase experimental densities and heats of vaporization.[16] This was achieved by increasing the σ of both the C2 and C3 united atoms by 1.4% from 3.905 to 3.96 Å and greatly reducing ε by 23% from 0.118 to 0.091 kcal/mol for C2 and from 0.175 to 0.136 kcal/mol for C3. Combined with the proper torsion potential, these values result in a reasonable description of liquid pentadecane (see Table 2).

Since the original simulations at 50 °C were only 10 ps long, and used a rather small cutoff of 10 Å, but with LJ cutoff corrections, we performed several new pentadecane simulations with the same parameters, for 20 ns, and with a cutoff of 20 Å, including dispersion cutoff corrections. Despite the more thorough setup, we obtained almost similar results, with the density agreeing perfectly and ΔH_{vap} only

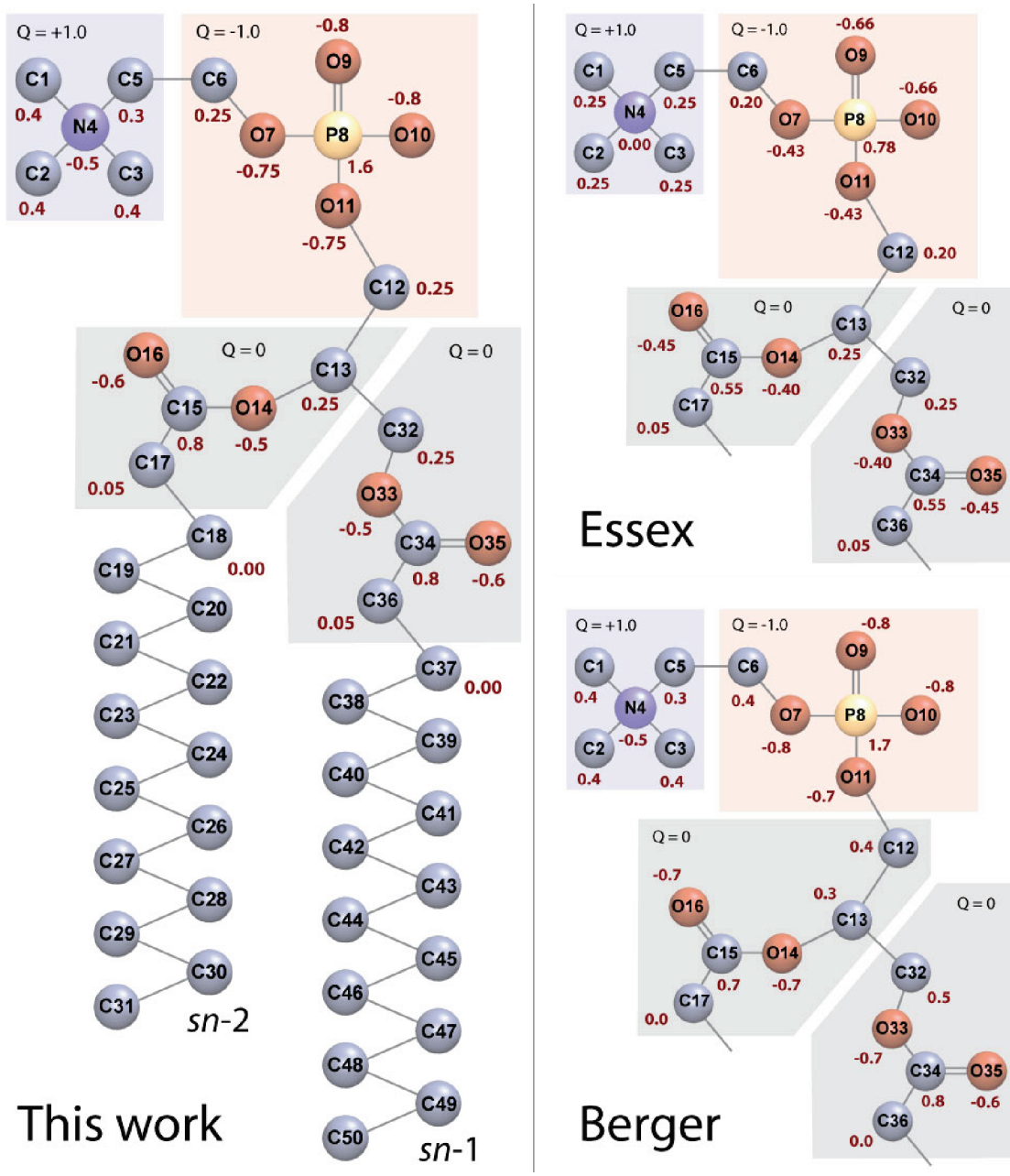


Figure 1. Illustration of a DPPC lipid molecule, with the charge groups and atomic charges indicated. Also shown are the values by Essex et al.[30] and Berger et al.[16]

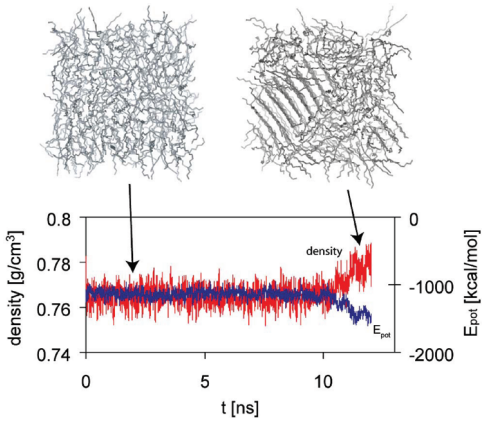


Figure 2. Example of a simulation of pentadecane that results in a freezing transition at ~11 ns.

slightly (2%) higher. This indicates that, for this almost homogeneous LJ fluid, the cutoff correction is very accurate, and it is not really necessary to use very high cutoffs of 20 Å. The density is also well matched against experimental data at both 25 and 50 °C. However, ΔH_{vap} seems to be underestimated by ~12% at both temperatures. The value of ΔH_{vap} is mainly dependent on ε, the strength of the LJ interaction. The low value of 15.5 kcal/mol reported by Berger et al. is due to the older experimental data used in that study.[16] More recent measurements indicate this value is substantially larger, ranging from 17.41 kcal/mol[45] to as high as 18.35 kcal/mol[46] at 25 °C (see Table 2). Subsequently, improved parameters for pentadecane with larger ε were reported by Chiu et al., resulting in $\Delta H_{vap} = 18.4$ kcal/mol.[47] The stronger LJ interactions will increase the attraction in the lipid tails and thus significantly affect the bilayer properties, such as lowering the area per lipid.[24] Thus, we chose to refit our new parameters to the higher values of ΔH_{vap} and over a larger temperature range.[45]

B. Melting Point Behavior. All equilibrium properties were obtained from a series of NPT simulations at temperatures ranging from 253 to 373 K. Below its melting temperature of 283.1 K, pentadecane freezes into stacked lamellae with regular hexagonal packing. Since the reproduction of liquid-state properties was the main goal here, all of the experimental values used in the fitting were from the liquid phase.

A poor parametrization of the hydrocarbon parameters can lead to freezing of the pentadecane box at temperatures above its melting point.[11] Figure 2 shows a simulation that spontaneously freezes due to incorrect parametrization, indicating that simulations should be in the multinanosecond range to detect such instabilities. Since the parameters are developed for the lipid tails, any incorrect parametrization with respect to the melting behavior will lead to bilayer simulations that show artificially high ordering, or even a gel phase. To determine the melting behavior of the present parameters, a series of simulations were performed at different temperatures to obtain T_{melt} for each parameter set. T_{melt} can only be calculated approximately due to the severe slowing of sampling near the phase transition, so there is an uncertainty of ±3 K, as illustrated in Figure 3. The melting point seems to be mainly affected by the energy difference

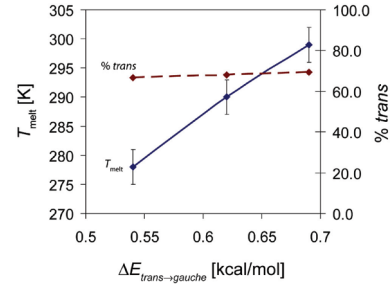


Figure 3. Dependence of the melting point T_{melt} of pentadecane on the conformational energy difference ΔE of the *trans* and *gauche* wells. An increase in ΔE of ~0.15 kcal/mol results in a positive melting point shift of ~21 K (solid line, left axis). The average population of *trans* states increases only slightly (dashed line, right axis).

Table 3. Results for the Simulations of Pentadecane[a]

	25 °C		50 °C	
	sim	exptl[b]	sim	exptl[b]
ρ (g/cm³)	0.7664	0.7650	0.7495	0.7478
ΔH_{vap} (kcal/mol)	17.46	17.41	16.88	16.96
κ (10^{-6} bar^{-1})	93.7	88.2[c]	111.8	104.1[d]
α (10^{-5} K^{-1})	88.4	89.1	92.3	93.1
C_p^{liq} (cal/(mol K))	107.5	112.3	112.2	115.3

[a] Density ρ, heat of vaporization ΔH_{vap}, isothermal compressibility κ, thermal expansion coefficient α, and liquid heat capacity C_p^{liq}. [b] Experimental data (except κ) from ref 45. [c] Reference 46. [d] Reference 72.

of the *gauche* and *trans* conformers $\Delta E_{trans \rightarrow gauche}$. By slight adjustments of the all-CH_2 torsion, a different melting point can be obtained without altering the thermodynamic properties too much. Figure 3 shows the dependence of the melting point on $\Delta E_{trans \rightarrow gauche}$. For a small increase in 0.15 kcal/mol, the melting point shifts upward by 21 K, while the average ratio of *trans* conformers of all rotatable bonds increases only minimally. These results indicate that torsional parameters must be carefully checked to accurately describe the liquid phase. Many previous lipid parametrization studies report sampling times too short to have detected such instabilities. In this study, we have verified that the melting point of both pentadecane and the DPPC bilayer is within 4 K of the experimental values.

C. Thermodynamic Properties of Pentadecane. Very good fits were obtained for the density, heat of vaporization, isothermal compressibility, thermal expansion coefficient, and constant-pressure heat capacity (Table 3). As a general trend, the density is mostly dependent on σ and the heat of vaporization on ε. We obtained the best fit for $\sigma = 3.955$ Å (C2 and C3) and $\varepsilon = 0.103$ kcal/mol (C2) and 0.154 kcal/mol (C3). This represents a lowering of the original OPLS values of ε by 12%, whereas in the Berger parametrization the reduction was 23%.[16] As a result, van der Waals interactions in the lipid tails are significantly stronger.

The density as a function of temperature is shown in Figure 4. The experimental data are represented as a solid line.[45] The standard deviations are very small due to the small fluctuation of the system box, and errors with respect to experiment are tiny, typically <0.5%. The heat of vaporization is obtained from $\Delta H_{vap} = \langle E_{gas} \rangle - \langle E_{liq} \rangle / N + RT$, where $\langle E_{gas} \rangle$ is the average potential energy of a single pentadecane

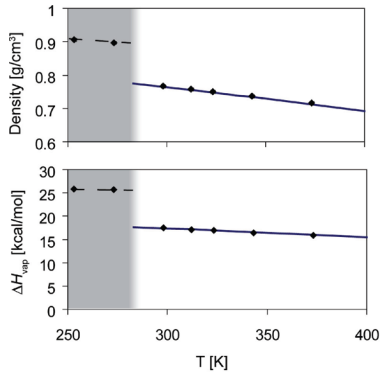


Figure 4. Density and heat of vaporization of pentadecane as a function of temperature. The results of the simulations are shown as dots. Experimental values for $T > T_{melt}$ are shown as smooth solid lines.[45] The shaded area represents the solid lamellar phase.

molecule in the gas phase (obtained from a Monte Carlo simulation) and $\langle E_{liq}\rangle/N$ is the average potential energy per molecule in the liquid phase. The agreement here is also almost perfect, and error estimates are very low since the potential energy of the system converges rapidly. Our value of $\Delta H_{vap} = 17.46$ kcal/mol at 25 °C matches the experimental one of 17.41 kcal/mol (c.f. Table 3). However, some experimental sources indicate this to be as high as 18.35 kcal/mol.[46]

While the density and ΔH_{vap} converge quite quickly, the remaining thermodynamic quantities are obtained from fluctuation formulas and require much longer time scales. To obtain fully converged results, the pentadecane simulations were extended to 400 ns each. The isothermal compressibility, κ, is computed from the volume fluctuations:

$$\kappa = \frac{\sigma_V^2}{\langle V\rangle kT}$$

where $\sigma_V = (\langle V^2\rangle - \langle V\rangle^2)^{1/2}$ is the standard deviation of the fluctuating system volume. A comparison of the calculated values versus experiment reveals that κ tends to be slightly overestimated, but the agreement nevertheless is very good (Figure 5). The thermal expansion coefficient α is given by

$$\alpha = \frac{\langle VH\rangle - \langle V\rangle\langle H\rangle}{\langle V\rangle kT^2}$$

where $H = E + PV$. Figure 5 shows the excellent correlation to the available experimental curve. Estimation of the liquid heat capacity C_p^{liq} is complicated by the improper classical description of vibrations in the fully flexible molecule. The standard procedure is to calculate only the fluctuations of the intermolecular enthalpy:

$$C_p^{inter} = \frac{\sigma_{H,inter}^2}{kT^2}$$

The full heat capacity for the liquid is then determined from $C_p^{liq} = C_p^{gas} + C_p^{inter}/N - R$, where C_p^{gas} is the experimental heat capacity for the ideal gas.[36] C_p seems to be slightly underestimated, although the agreement with experiment is still very good (Figure 5). The gas-phase contribution is quite large, typically 80%.

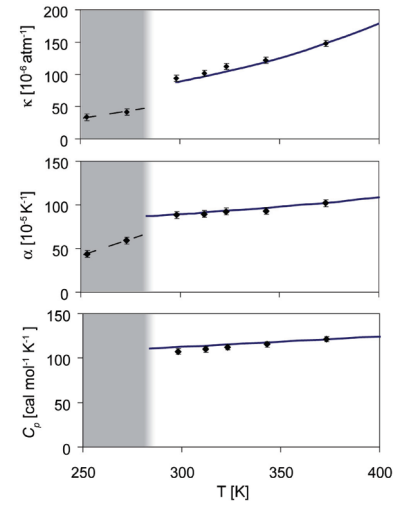


Figure 5. Isothermal compressibility κ, thermal expansion coefficient α, and liquid heat capacity C_p of pentadecane as a function of temperature. Experimental values for $T > T_{melt}$ are shown as smooth solid lines.[45,72] The shaded area represents the solid lamellar phase.

Despite the united-atom description, both the density and heat of vaporization can be reproduced with errors of <0.5% from experiment, while the other thermodynamic quantities are within ~5% of the experiments. Similar observations have been made with all-atom alkane force fields.[11,48] This also compares favorably to polarizable all-atom models for pentadecane, where the deviations from experiments were in fact slightly higher.[49] However, these parameters were designed to cover a wider range of hydrocarbons.

D. Role of the Thermostats and Barostats. The MD simulations were run with the Andersen thermostat and Monte Carlo volume moves to model the exact NPT ensemble (see the Methods). To compare to the more commonly used weak temperature and pressure coupling schemes, we repeated the simulations with these methods using Gromacs. The resulting density at 25 °C is $\rho = 7.670$ g/cm^3 (<0.07% difference), and $\Delta H_{vap} = 17.54$ kcal/mol (<0.4% difference). It seems that, for this system, the choice of the temperature and pressure coupling scheme has little to no influence on the equilibrium properties and that the values are reliable.

IV. Lipid Parameters

A. DPPC Parametrization. The next stage was the assembly of the new lipid parameter set. An illustration of a DPPC molecule is shown in Figure 1, together with the charges by Essex et al.[30] and Berger et al.[16] LJ parameters are the same on all atoms between these models, except the CH_2 and CH_3 groups in the lipid tails, which are based on the pentadecane simulations. One of the goals of the new parametrization was to have identical charges on both ester groups. This is the case in the Essex parametrization and in most other lipid force fields,[11,12,17] but not in the Berger set. Treating the ester groups similarly reduces the amount of fitting required substantially, leading to more robust and

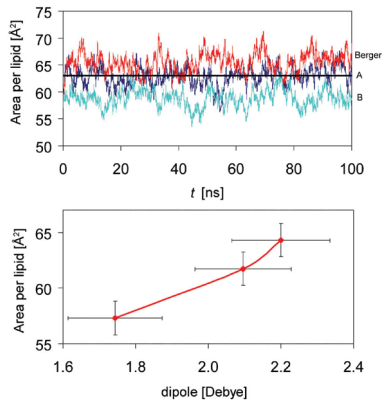


Figure 6. Upper panel: Area per lipid during the 100 ns simulations for the Berger parameters, this work (A), and a parameter set where only the charges in the headgroup were modified (B). The solid horizontal line represents the most recent experimental value of 63.0 Å^2.[51] Lower panel: Area per lipid dependence on the charge distribution of the ester groups by calculating the average ester dipole moment over all DPPC molecules over the complete simulation.

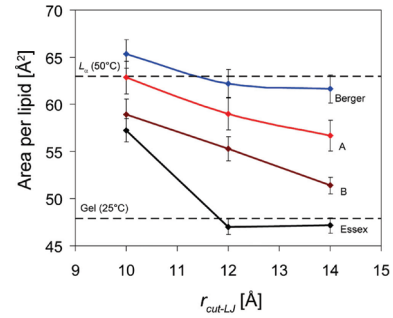


Figure 7. Dependence of the area per lipid on the van der Waals cutoff used in the simulations. Electrostatic interactions are evaluated using the PME method. The experimental values for the fluid (L_α) phase and gel phase are indicated as dashed lines.[26,51] There is a systematic decline of the area for all models as $r_{cut\text{-}LJ}$ is increased. The new model with more polar ester charges (A) results in higher areas than a model where only the headgroup is taken from the Berger set (B). The smallest decline is observed for the Berger set, due to the weak lipid tail LJ interactions. The Essex parameters show a transition into the gel phase.

transferable parameters. Although separate charge groups are no longer necessary when PME is used, it is beneficial to retain such groups to allow simple cutoff simulations. We thus started by slightly rearranging the charges of the Berger lipids to obtain the charge distribution illustrated in Figure 1. Further refinement was then based on reproducing experimental data in fully hydrated bilayer simulations.

The first test of the new parameters is whether simulated lipid bilayers retain the experimentally observed area per lipid A in the tension-free NPT ensemble. While experimental estimates of 57−71.2 Å^2 have been reported over the years (see the reviews by Nagle et al.),[26,50] we choose the more recent value of 63.0 ± 1.0 Å^2 obtained by simultaneously analyzing X-ray and neutron scattering data as the target of our parametrization.[51]

There is now a consensus that use of long-range electrostatic treatment (PME) is essential to obtain correct bilayer properties, while the use of a simple cutoff leads to bilayers that are too stretched and laterally shrunk.[52−54] This is easily understandable, given the strong charges in the lipid headgroups. Despite this, there remains a significant dependence of the area per lipid on the value of the LJ cutoff.[52] This is usually set to a low ∼10 Å for performance reasons. Given the nonisotropy of the bilayer environment, the use of simple analytic cutoff correction schemes that assume homogeneous isotropic liquids is problematic, although some advancements have been made in this direction.[55−57] The problem is acute for united-atom lipid molecules since the charges on the lipid tails are zero; i.e., there are only LJ interactions in the hydrophobic phase. With a small LJ cutoff, lipid tails interact at most with their first-shell neighbors. If the LJ cutoff is increased, the attractive nature of the LJ potential leads to more laterally compact bilayers.

The area strongly fluctuates (standard deviation of ±1.5 Å^2; see Figure 6) on the multinanosecond time scale, indicating that reliable numbers require simulations of at least 100 ns length. Figure 7 shows the obtained areas per lipid for various simulations as a function of the LJ cutoff. There is a systematic decline of ∼7−10% for all simulations as the LJ cutoff is increased from 10 to 14 Å. The lowest area is obtained for the Essex parameters, which results in an area of ∼57 Å^2, and a notably increased ordering of the lipid tails into parallel, mostly *trans* packing. For a larger cutoff, these parameters lead to a transition of the bilayer into the gel phase with the area shrinking to ∼47 Å^2. Bilayers simulated with the Berger set behave very differently. As previously observed the area is significantly larger at 65.3 Å^2,[54] and drops to only 60.9 Å^2 for the 14 Å LJ cutoff. This is a smaller decline than for the other models, and is caused by the weaker LJ interactions in the lipid tails of the Berger set. However, the much larger area is not due to the hydrocarbon parameters, but rather due to the differing charges on the polar part of the lipid molecule. The charge distribution on the headgroup developed by Chiu et al.[23] is much more polar than the values by Essex. Thus, we performed a series of bilayer simulations with the headgroup charges replaced by the Berger set, with some minor adjustments to achieve the charge group partitioning illustrated in Figure 1. Interestingly, the greatly modified phosphate and choline charges only lead to a marginal increase in the area to 58.9 Å^2 (Figures 6B and 7B). A similar insensitivity of the area per lipid on the headgroup charge distribution has been reported by Sonne et al.[14] This leaves only the ester groups (and glycerol) as the source for the higher area. As shown in Table 4, the original OPLS-UA parameters for esters are very similar to the more modern OPLS-AA ester charges obtained by Price et al.[58] In the Berger set, both the carbonyl and alkoxy O atoms are much more negative, and the carbonyl C is more positive. We therefore tried two new charge sets, one where the carbonyl C is made more positive but the oxygens are only slightly more negative (I), and one which is close to the Berger charges (II). The effect of these changes on bilayer properties is dramatic, with a strongly increased area per lipid of 62.3 Å^2 for set I and 62.9 Å^2 for set II (Figure 7A). This is very

Table 4. Electronic Net Charges on the Ester Group in Various Force Fields[a]

	O=	C	OS
OPLS-UA[b]	−0.45	0.55	−0.4
OPLS-AA[c]	−0.43	0.51	−0.33
Berger[d]	−0.7(−0.6)	0.7 (0.8)	−0.7
CHARMM for NPT[e]	−0.6	0.83	−0.47 (−0.54)
CHARMM for NPT (DMPC)[f]	−0.61	0.82	−0.54
this work (set I)	−0.55	0.7	−0.45
this work (set II)	−0.6	0.8	−0.5

[a] Charges in parentheses denote the second ester group. [b] Reference 31. [c] Reference 58. [d] Reference 16. [e] Reference 14. [f] Reference 15.

close to the recent estimate of 63.0 Å^2 by Kučerka et al.[51] Thus, we used these charges in our final parametrization.

The effect of the ester charges is also illustrated in Figure 6 (lower panel). There is a correlation between the dipole strength and the area per lipid, with the more polar ester group resulting in higher areas. We did not attempt a more thorough reparametrization of the ester charges beyond the goal of finding a charge set that is identical on both esters, but the results indicate that further investigation might bring additional improvements. Interestingly, the final charges on the esters are close to the values reported by Sonne et al. in their reparametrization of the CHARMM lipid force field (see Table 4), although the area per lipid reported in this study was significantly smaller (60.4 Å^2).[14] Almost identical charges are reported by Högberg et al., who obtained improved CHARMM parameters for DMPC.[15] Thus, it seems a general pattern that the ester charges should be more polar in lipid molecules to obtain correct bilayer properties, and are a major reason why the Berger parameters perform well. This is probably related to an increased hydration of the ester groups, but it could also be that the higher ester dipole simply covers for other imbalances in the lipid parameters. Further investigations will be performed in the future to clarify these observations.

B. Role of Headgroup Parameters. The phosphocholine headgroup of the lipid molecules exhibits a characteristic tilt toward the bilayer normal, which can be analyzed by calculating the distribution of the vector connecting the phosphorus and the nitrogen atoms, averaged over the course of the simulation and all lipid molecules. The resulting distributions for all parameter sets are shown in Figure 8. Despite the large differences in the partial charges of the phosphate and choline groups in the three models, the average orientation is perpendicular to the bilayer normal in all cases, with a tilt of ∼90°. The distribution is broad with $2\sigma = 60°$ in all cases. Similar values of 78−86° were found in previous studies with the Berger set,[19,59] and have been reported as low as ∼60° in other lipid force fields.[19] Overall, the parametrization changes in the headgroup seem to have little effect on the average headgroup orientation, which is interesting given the major differences between the Essex and Berger charges.

C. Order Parameter. The fine structure of the lipid bilayer and its ordering in the fluid phase can be measured experimentally using ^{2}H NMR spectroscopy. This involves deuterating hydrogens at selected carbon atoms and measuring the residual quadrupolar couplings of the CD bond. The

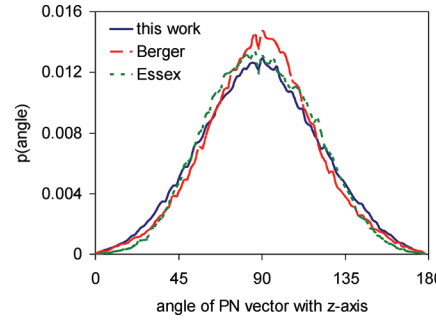


Figure 8. Distribution of the headgroup orientation vector (P−N) with respect to the bilayer normal for the individual models. The distributions are very similar, with the maximum at ∼90°.

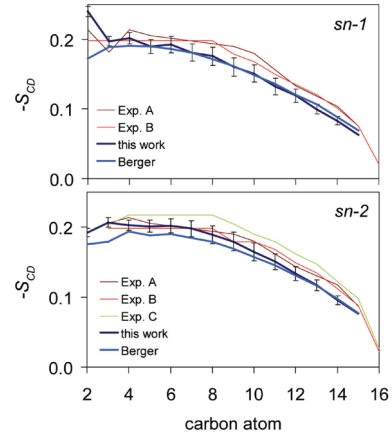


Figure 9. Deuterium order parameter for the two acyl chains of DPPC, for the simulations with the new parameter set (thick curve with error bars, estimated from block averaging), and the Berger lipids. Also shown are the experimental values of Seelig et al.[61] (A), Petrache et al.[62] (B), and Douliez et al.[63] (C).

resulting deuterium order parameter S_{CD} is a measure of the disorder and the relative orientation of the CD bond. S_{CD} is obtained from $S_{CD} = {}^2/_3 S_{xx} + {}^1/_3 S_{yy}$,[60] where the order parameter tensor is defined as

$$S_{\alpha\beta} = \frac{1}{2}\langle 3 \cos\theta_\alpha \cos\theta_\beta - \delta_{\alpha\beta}\rangle$$

with θ_α being the angle of the axis α (= x, y, z) to the bilayer normal (z axis) and the averaging taking place over molecules and time. In oriented samples, the axis of motional averaging is identical to the bilayer normal. Due to axial averaging, $S_{xx} = S_{yy}$ and $S_{xx} + S_{yy} + S_{zz} = 0$, resulting in[61]

$$S_{CD} = -S_{zz}/2 = \langle (3 \cos^2\theta_{CD} - 1)/2\rangle$$

Since no CD bonds are available on the united-atom carbons, either the deuteriums are constructed using ideal bond geometry or the vector from carbon C_{n-1} to C_{n+1} is used as the molecular axis for the nth CH_2 unit.[60]

Figure 9 shows the calculated S_{CD} parameters for both the *sn-1* and *sn-2* chains. The overall agreement with the experimental data is good.[61−63] S_{CD} values obtained for the Berger set are identical to those reported in a previous study by Patra et al.[52] The modified LJ parameters on the alkyl

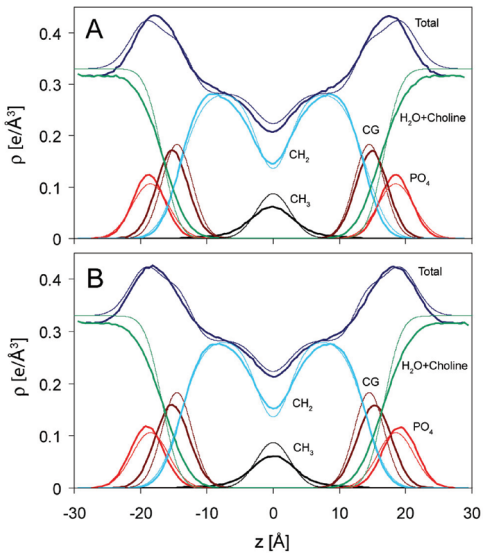


Figure 10. Electron density profiles from the simulations: (A) Berger lipids, (B) this work. The thick lines are the results of the simulations, averaged over the final 80 ns of the trajectory. The thin lines represent the experimental density determined by Kučerka and Nagle,[68] including the individual components (CG = carbonyl−glycerol).

chains will affect the torsional energetics and lipid ordering. Nevertheless, the changes are only small, with results for the *sn-1* chain similar to those of the Berger set and a slightly increased ordering in the *sn-2* chain. S_{CD} is strongly correlated to the area per lipid, with increased ordering for bilayers that are laterally too small. Values are also very sensitive to the *gauche* to *trans* energy difference of the alkyl torsions. Our results are in line with what has been reported previously on DPPC lipids.[11,14,52] A special case is the S_{CD} on C2. On the *sn-1* chain, this is in the high plateau region with $-S_{CD} = 0.215$, while it is split and much lower in the *sn-2* chain ($-S_{CD} = 0.15$ and 0.09).[61] With the new parameters, the S_{CD} on C2 is slightly higher than with the Berger parameters, which underestimates the S_{CD} on *sn-1*. On *sn-2*, both sets give a similar value that is too large compared to that from the experiments. Reproducing the splitting on *sn-2* would require abandoning the united-atom model and explicitly adding the CD bonds. Despite the absence of the deuteriums, the results show that S_{CD} can be accurately obtained even in the united-atom approximation.

D. Electron Density. The density distribution of the lipid bilayer can be measured by X-ray and neutron diffraction studies.[64] The total density is usually obtained from structure factors,[26,65,66] and the individual density contributions of the various lipid components can be extracted from structural models.[67] Computationally, the electron density is obtained by binning atoms along the direction of the membrane normal, weighted with the correct number of electrons, although it can also be calculated from the structure factors.[67] Figure 10 shows the electron density distribution of both the new model and the Berger set compared to the experimental results obtained by Kučerka and Nagle (H2 model).[68] There is a close agreement of the overall shape with the experimental data, with the characteristic headgroup peaks and the methyl troughs. As the water model underestimates the bulk

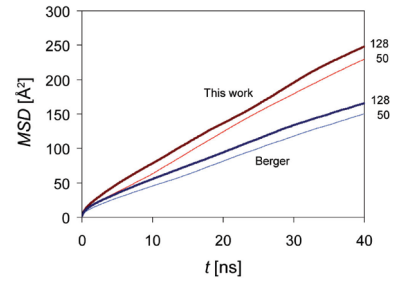


Figure 11. Average mean-square displacement of the lipid molecules as a function of time. Results for boxes containing 50 and 128 lipids are shown. The new parameters result in slightly faster lateral diffusion of the lipids, independent of the size of the simulated bilayer patch.

water density by 4% at 50 °C, the electron density differs by the same amount in the water phase. The experimental bilayer width (defined as the head-to-head distance) is 37.8 Å.[68] Our model gives a slightly smaller width of 36.6 Å, while the Berger set gives 35.6 Å. Individual components also compare well against the H2 hybrid model. Similar to what has been observed for the all-atom lipid CHARMM force field, there is an underestimation of the methyl peak in the bilayer center, which is compensated by an overestimation of the CH_2 density.[67]

E. Lateral Lipid Diffusion. Lipid bilayers are two-dimensional liquids characterized by the lateral diffusion of the individual lipid molecules. The lateral diffusion constant can be obtained from the Einstein relation

$$D_{lat} = \lim_{t\to\infty} \frac{1}{4t} \mathrm{MSD}(t)$$

where the mean-squared displacement is calculated as

$$\mathrm{MSD}(t) = \frac{1}{N}\sum_{i=1}^{N} \langle (\mathbf{r}_i(t + t_0) - \mathbf{r}_i(t_0))^2 \rangle$$

with $\mathbf{r}_i(t)$ is the position of the center of mass (CM) of lipid molecule i at time t, with averaging taking place over all time origins t_0. The results are shown in Figure 11. There is both a fast short-time diffusion caused by the fluctuation of lipid molecules in their trapped space and a slower long-time diffusion characterized by jumps of the lipid molecules from one basin to the next. Only the long-time diffusion is calculated here, by fitting the MSD over the linear phase from 20 to 30 ns. The resulting diffusion constants have to be corrected for the systematic CM motion of the two monolayers.[69] Klauda et al. have reported significant finite size effects in the lateral diffusion, with a ~3 times larger D_{lat} as the size of the lipid patch is decreased from 288 to 72 lipids.[25] To investigate a similar system size dependence, the simulations were performed with both 128 and 50 lipids. As can be seen from Figure 11, the lateral diffusion of both the new parameters and the Berger lipids depends only little on the system size, with a ~10% smaller D_{lat} for the 50-lipid systems in both cases. A slightly larger diffusion constant is found for the new parameters, with $D_{lat} = (16.5 \pm 0.9) \times 10^{-8}$ cm^2/s ($(15.3 \pm 0.8) \times 10^{-8}$ cm^2/s for 50 lipids), than for the Berger lipids, which have $D_{lat} = (11.4$

$\pm$ 0.8) $\times$ 10^{-8} cm^2/s ((10.0 $\pm$ 0.7) $\times$ 10^{-8} cm^2/s for 50 lipids). The new values closely match recent experimental estimates of $D_{lat} = 15.2 \times 10^{-8}$ cm^2/s (at 324 K) obtained from ^{1}H pulsed field gradient magic angle spinning NMR spectroscopy.[70] Earlier photobleaching experiments have reported $D_{lat} = 12.5 \times 10^{-8}$ cm^2/s at 323 K.[71] Thus, lateral diffusion is excellently reproduced by the new lipid parameters.

V. Conclusion

The new UA lipid parameters presented here have been designed to be used in combination with OPLS-AA for proteins. This avoids some of the tricks that have been necessary to use OPLS-AA in membrane protein simulations. A scale factor of 0.5 is now applied for all nonbonded 1−4 interactions throughout. In addition to refitting the torsion angles for the new scale factor, we have used the opportunity to improve on some of the other parameters of the Berger lipid force field: The acyl chains now have stronger LJ interactions, as determined from fitting to newer values of ΔH_{vap} of pentadecane. The stronger interactions do not lead to an incorrect freezing behavior, as special care was taken to ensure the melting points of both pentadecane and DPPC are not overestimated. In addition, the charge set of the Berger lipids is adjusted to allow symmetric carboxylate ester groups, simplifying the model and reducing the amount of torsions that have to be individually fitted. The area per lipid of 62.9 Å^2 is 3.5% lower than in the Berger force field, but exactly matches the more accurate recent experimental value of 63.0 Å^2 by Kučerka et al.[51] Electron density profiles and deuterium order parameters and lateral diffusion constants are also well reproduced. Most importantly, lipid bilayers can be simulated in the NPT ensemble without applying additional surface tension terms or having to fix the membrane surface area to be constant. Several recent lipid reparametrizatoin efforts have also focused on addressing this common lipid force field deficiency.[14,15] Tension-free NPT simulations are critical for studying peptide partitioning, where the membrane must be able to stretch laterally. The present work offers an accurate lipid parametrization for straightforward use with the widely employed OPLS-AA force field. The real test of the combined lipid and protein force field will be with multimicrosecond simulations of complex processes such as peptide adsorption, partitioning, and folding. To reach such time scales, the use of united-atom lipid molecules will remain crucial for the foreseeable future, given the vast computational advantage over all-atom lipids. This situation is similar to that of water molecules, where three-site models (SPC, TIP3P) are still used in the vast majority of simulations although more sophisticated, but slower alternatives exist. Additional work is currently under way to extend the present work to other classes of lipids and lipid mixtures.

Acknowledgment. This work was supported by BIOMS (J.P.U.) and the Human Frontier Science Program (M.B.U.). We thank Oliver Beckstein for helpful discussions and John F. Nagle for providing us with the experimental electron density curves.

References

(1) Lindahl, E.; Sansom, M. S. *Curr. Opin. Struct. Biol.* **2008**, *18* (4), 425.

(2) Ulmschneider, M. B.; Ulmschneider, J. P. *J. Chem. Theory Comput.* **2008**, *4* (11), 1807.

(3) Nymeyer, H.; Woolf, T. B.; Garcia, A. E. *Proteins* **2005**, *59* (4), 783.

(4) Ulmschneider, J. P.; Ulmschneider, M. B. *J. Chem. Theory Comput.* **2007**, *3*, 2335.

(5) Ulmschneider, M. B.; Ulmschneider, J. P. *Mol. Membr. Biol.* **2008**, *25* (3), 245.

(6) Ulmschneider, J. P.; Ulmschneider, M. B.; Di Nola, A. *Proteins* **2007**, *69*, 297.

(7) Im, W.; Brooks, C. L. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (19), 6771.

(8) Im, W.; Brooks, C. L., III. *J. Mol. Biol.* **2004**, *337*, 513.

(9) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. *Biophys. J.* **2008**, *94* (10), L75.

(10) Ulmschneider, J. P.; Ulmschneider, M. B. *Proteins* **2009**, *75* (3), 586–597.

(11) Klauda, J. B.; Brooks, B. R.; MacKerell, A. D.; Venable, R. M.; Pastor, R. W. *J. Phys. Chem. B* **2005**, *109* (11), 5300.

(12) Feller, S. E.; MacKerell, A. D. *J. Phys. Chem. B* **2000**, *104* (31), 7510.

(13) Henin, J.; Shinoda, W.; Klein, M. L. *J. Phys. Chem. B* **2008**, *112* (23), 7008.

(14) Sonne, J.; Jensen, M. Ø.; Hansen, F. Y.; Hemmingsen, L.; Peters, G. H. *Biophys. J.* **2007**, *92* (12), 4157.

(15) Högberg, C.-J.; Alexei, M.; Nikitin, A. M.; Lyubartsev, A. P. *J. Comput. Chem.* **2008**, *29* (14), 2359.

(16) Berger, O.; Edholm, O.; Jahnig, F. *Biophys. J.* **1997**, *72*, 2002.

(17) Chandrasekhar, I.; Kastenholz, M.; Lins, R. D.; Oostenbrink, C.; Schuler, L. D.; Tieleman, D. P.; van Gunsteren, W. F. *Eur. Biophys. J.* **2003**, *32* (1), 67.

(18) Lula Rosso, I. R. G. *J. Comput. Chem.* **2008**, *29* (1), 24.

(19) Siu, S. W. I.; Vacha, R.; Jungwirth, P.; Bockmann, R. A. *J. Chem. Phys.* **2008**, *128* (12), 125103.

(20) Davis, J. E.; Rahaman, O.; Patel, S. *Biophys. J.* **2009**, *96* (2), 385.

(21) Feller, S. *Computational Modeling of Membrane Bilayers*; Academic Press: New York, 2008; Vol. 60.

(22) Feller, S. E.; Pastor, R. W. *J. Chem. Phys.* **1999**, *111* (3), 1281.

(23) Chiu, S. W.; Clark, M.; Balaji, V.; Subramaniam, S.; Scott, H. L.; Jakobsson, E. *Biophys. J.* **1995**, *69* (4), 1230.

(24) Wohlert, J.; Edholm, O. *Biophys. J.* **2004**, *87* (4), 2433.

(25) Klauda, J. B.; Brooks, B. R.; Pastor, R. W. *J. Chem. Phys.* **2006**, *125* (14), 144710.

(26) Nagle, J. F.; Tristram-Nagle, S. *Biochim. Biophys. Acta—Rev. Biomembr.* **2000**, *1469* (3), 159.

(27) Hristova, K.; White, S. H. *Biophys. J.* **1998**, *74* (5), 2419.

(28) Kukol, A. *J. Chem. Theory Comput.* **2009**, *5* (3), 615.

(29) Rosso, L.; Gould, I. R. *J. Comput. Chem.* **2008**, *29* (1), 24.

(30) Essex, J. W.; Hann, M. M.; Richards, W. G. *Philos. Trans. R. Soc. London, Ser. B* **1994**, *344* (1309), 239.

(31) Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. *J. Am. Chem. Soc.* **1984**, *106* (22), 6638.

(32) Jorgensen, W. L.; Gao, J. *J. Phys. Chem.* **1986**, *90* (10), 2174.

(33) Briggs, J. M.; Nguyen, T. B.; Jorgensen, W. L. *J. Phys. Chem.* **1991**, *95* (8), 3315.

(34) van Gunsteren, W. F.; Krüger, P.; Billeter, S. R.; Mark, A. E.; Eising, A. A.; Scott, W. R. P.; Hüneberger, P. H.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; Biomos/Hochschulverlag AG an der ETH Zürich: Groningen, The Netherlands/Zürich, Switzerland, 1996.

(35) Egberts, E.; Marrink, S.-J.; Berendsen, H. J. C. *Eur. Biophys. J.* **1994**, *22* (6), 423.

(36) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225.

(37) Tieleman, P.; Maccallum, J.; Ash, W.; Kandt, C.; Xu, Z.; Monticelli, L. *J. Phys.: Condens. Matter* **2006**, *18* (28), S1221.

(38) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4* (3), 435.

(39) Andersen, H. C. *J. Chem. Phys.* **1980**, *72* (4), 2384.

(40) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, U.K., 1989.

(41) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81* (8), 3684.

(42) Hess, B.; Bekker, J.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463.

(43) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79* (2), 926.

(44) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98* (12), 10089.

(45) Yaws, C. L. *Chemical Properties Handbook*; McGraw-Hill: New York, 1999.

(46) Lide, D. R. *CRC Handbook of Chemistry and Physics*, 88th ed.; CRC Press: Boca Raton, FL, 2007.

(47) Chiu, S. W.; Vasudevan, S.; Jakobsson, E.; Mashl, R. J.; Scott, H. L. *Biophys. J.* **2003**, *85* (6), 3624.

(48) Thomas, L. L.; Christakis, T. J.; Jorgensen, W. L. *J. Phys. Chem. B* **2006**, *110* (42), 21198.

(49) Davis, J. E.; Warren, G. L.; Patel, S. *J. Phys. Chem. B* **2008**, *112* (28), 8298.

(50) Nagle, J. F.; Tristram-Nagle, S. *Curr. Opin. Struct. Biol.* **2000**, *10* (4), 474.

(51) Kučerka, N.; Nagle, J. F.; Sachs, J. N.; Feller, S. E.; Pencer, J.; Jackson, A.; Katsaras, J. *Biophys. J.* **2008**, *95* (5), 2356.

(52) Patra, M.; Karttunen, M.; Hyvonen, M. T.; Falck, E.; Lindqvist, P.; Vattulainen, I. *Biophys. J.* **2003**, *84* (6), 3636.

(53) Anezo, C.; de Vries, A. H.; Holtje, H.-D.; Tieleman, D. P.; Marrink, S.-J. *J. Phys. Chem. B* **2003**, *107* (35), 9424.

(54) Patra, M.; Karttunen, M.; Hyvonen, M. T.; Falck, E.; Vattulainen, I. *J. Phys. Chem. B* **2004**, *108* (14), 4485.

(55) Lagüe, P.; Pastor, R. W.; Brooks, B. R. *J. Phys. Chem. B* **2004**, *108* (1), 363.

(56) Shirts, M. R.; Mobley, D. L.; Chodera, J. D.; Pande, V. S. *J. Phys. Chem. B* **2007**, *111* (45), 13052.

(57) Klauda, J. B.; Wu, X.; Pastor, R. W.; Brooks, B. R. *J. Phys. Chem. B* **2007**, *111* (17), 4393.

(58) Price, M. L. P.; Ostrovsky, D.; Jorgensen, W. L. *J. Comput. Chem.* **2001**, *22* (13), 1340.

(59) Pandit, S. A.; Bostick, D.; Berkowitz, M. L. *Biophys. J.* **2003**, *84* (6), 3743.

(60) Egberts, E.; Berendsen, H. J. C. *J. Chem. Phys.* **1988**, *89* (6), 3718.

(61) Seelig, A.; Seelig, J. *Biochemistry* **1974**, *13* (23), 4839.

(62) Petrache, H. I.; Dodd, S. W.; Brown, M. F. *Biophys. J.* **2000**, *79* (6), 3172.

(63) Douliez, J. P.; Léonard, A.; Dufourc, E. J. *Biophys. J.* **1995**, *68* (5), 1727.

(64) Wiener, M. C.; White, S. H. *Biophys. J.* **1992**, *61* (2), 437.

(65) Benz, R. W.; Castro-Roman, F.; Tobias, D. J.; White, S. H. *Biophys. J.* **2005**, *88* (2), 805.

(66) Hub, J. S.; Salditt, T.; Rheinstädter, M. C.; de Groot, B. L. *Biophys. J.* **2007**, *93* (9), 3156.

(67) Klauda, J. B.; Kučerka, N.; Brooks, B. R.; Pastor, R. W.; Nagle, J. F. *Biophys. J.* **2006**, *90* (8), 2796.

(68) Kučerka, N.; Tristram-Nagle, S.; Nagle, J. F. *Biophys. J.* **2006**, *90* (11), L83.

(69) Lindahl, E.; Edholm, O. *J. Chem. Phys.* **2001**, *115* (10), 4938.

(70) Scheidt, H. A.; Huster, D.; Gawrisch, K. *Biophys. J.* **2005**, *89* (4), 2504.

(71) Vaz, W. L. C.; Clegg, R. M.; Hallmann, D. *Biochemistry* **1985**, *24* (3), 781.

(72) Dovnar, D. V.; Lebedinskii, Y. A.; Khasanshin, T. S.; Shchemelev, A. P. *High Temp.* **2001**, *39* (6), 835.

CT900086B

JCTC Journal of Chemical Theory and Computation

Coarse Graining of Intermolecular Vibrations by a Karhunen-Loève Transformation of Atomic Displacement Vectors

Hirohiko Houjou*

Institute of Industrial Science, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan

Received April 9, 2009

Abstract: We have formulated a procedure for evaluating the anisotropic stiffness of a molecular assembly. First, we show how to reduce the dimensions of the matrices that appear in a conventional Hessian analysis of mass-weighted coordination by using a 12-dimensional transverse-rotational basis set for expansion. This treatment yields matrix representations of the intermolecular force and inertial load of the constituent molecules. Next we applied this procedure to 2-aminopyridine dimers and numerically analyzed the low-frequency (∼THz region) normal-mode vibrations. By validating the elements of stiffness matrix, this study exemplifies a derivation of the parameters necessary for the normal-mode analysis of a large system like a crystal, without any explicit representation of the potential functions.

1. Introduction

The increasing interest in supramolecular materials and biomolecular systems demands better fundamental understanding of intermolecular interactions. Recently, terahertz (THz) spectroscopy has been used for direct observation of intermolecular interactions in such contexts as the analysis of the hydration of sugars, the arrangement of nucleobases in DNA, and the polymorphism of medicinal drugs.[1−5] The terahertz region (∼300 GHz to 3 THz) covers hydrogen bond vibrations, van der Waals interactions, overall molecular distortion, and molecular libration; hence, the comparison of theoretical and experimental vibrational spectra can provide a molecular-level picture of the macroscopic phenomena of materials. Accordingly, several quantum chemical and molecular mechanics-based approaches have been reported.[6−12] In addition, anharmonic effects should also be included to quantitatively explain experimental results.[13,14]

Because we have reached the primary stage of interpreting the contents of THz spectra, it seems important to clarify the harmonic behavior of a molecular system in which intermolecular forces dominate the optimum arrangement of the molecules in a material. Various phenomena related to intermolecular vibrations and librations have been studied by Hessian-based methods such as normal-mode analyses and lattice dynamics (phonon band calculations).[15−18] In applications of these methods to macromolecular systems, coarse graining of molecular representation has frequently been employed as an approach toward reducing computational consumption.[15c,d,19] Combined with empirical force field parameters, the coarse graining approach has been successful to some extent in reproducing the collective motions of proteins, which resonate at GHz frequencies. In these studies, the elements of the dynamical matrix were evaluated as second derivatives of a potential function whose composition included Lennard-Jones-type repulsion-dispersion and electrostatic interactions; hence, the results depend on the force field parameters used. The efforts in developing the force field seem to be devoted to constructing a set of universal parameters by decomposing the molecular interactions into components for respective atom−atom or multipole−multipole pairs.[20−23] For these efforts to succeed, some intrinsic difficulties need to be dealt with, for example, the balancing of several types of potential depth and the incorporation of secondary (cooperative) effects such as instantaneous induced polarization.

On the one hand, on the basis of the above points, it may be insufficient to apply an empirical force field to a Hessian-

* Corresponding author phone: +81(3)5452-6367; fax: +81(3)5452-6366; e-mail: houjou@iis.u-tokyo.ac.jp.

10.1021/ct900169f CCC: $40.75
Published on Web 05/29/2009

based analysis of THz spectra, in which specific intermolecular interactions are sensitively reflected. On the other hand, some groups satisfactorily reproduced THz spectra by normal-mode calculation based on ab initio molecular orbital or density functional theory methods.[7−12] However, applying these quantum chemical methods to such large systems as proteins, nucleotides, and molecular crystals is obviously difficult in view of the consumption of computational time and resources. In this study based on ab initio quantum chemical calculations, we attempt to derive the elements of a dynamical matrix for a coarse-grained representation of a molecular assembly. The normal-mode frequency calculated by the quantum chemical method reflects the curvature of the potential surface, into which the electronic effects are adequately incorporated. Because the intermolecular forces are responsible for the low-frequency vibrations, it seems quite reasonable to extract meaningful parameters from the huge amount of numerical atomic displacement data. The thus-obtained "tailor-made" parameters will lead to reliable predictions of THz spectra.

2. Theoretical Basis

In the framework of Hessian analysis of molecular vibration, the displacement vectors $\mathbf{X}$ are obtained by diagonalizing the mass-weighted Hessian matrix ($\mathbf{M}^{-1/2}\mathbf{K}\mathbf{M}^{-1/2}$) with the eigenvalue matrix $\mathbf{\Omega}$ that contains the corresponding frequency ω as follows[24]

$$(\mathbf{M}^{-1/2}\mathbf{K}\mathbf{M}^{-1/2})(\mathbf{M}^{1/2}\mathbf{X}) = (\mathbf{M}^{1/2}\mathbf{X})\mathbf{\Omega}^2 \tag{1}$$

For an isolated single molecule, the dimension of $\mathbf{M}$ and $\mathbf{K}$ is $3N$, where N is the number of constituent atoms, and the diagonalization gives six zero-eigenvalues for the molecular motions of pure translation and pure rotation. For a system composed of two molecules that have N_{I} and N_{II} atoms, respectively, the number of degrees of internal freedom is $3(N_{\mathrm{I}} + N_{\mathrm{II}}) - 6$, in which $3N_{\mathrm{I}} - 6$ and $3N_{\mathrm{II}} - 6$ modes of motion originate from the internal freedom of the respective molecule. Consequently, six residual modes are attributed to the degrees of freedom of the intermolecular vibration.

In practical normal-mode analysis for dimeric systems, the vibrations of intermolecular motions appear within a wavenumber range of 10−200 cm^{-1}, clearly separate from the wavenumber ranges of intramolecular motions that appear in the region of 400−4000 cm^{-1}. This observation suggests that there are only small coupling effects among the intermolecular and intramolecular vibrations. Therefore, the atomic displacement of an intermolecular vibration is approximately represented by a combination of several basic motions of a "frozen" molecule. The intermolecular vibrational motions are approximately represented by a linear combination of the translational and rotational (hereafter denoted as T/R) motions of the constituent molecules as a basis set. For example, the basis for symmetric transverse motion along the x-axis is hereafter denoted as $\mathrm{T}_x + \mathrm{T}_x$ (Figure 1). This representation is a kind of Karhunen-Loève (KL) expansion, a method for extracting the principal components of poly dimensional vectors.[25] In the space

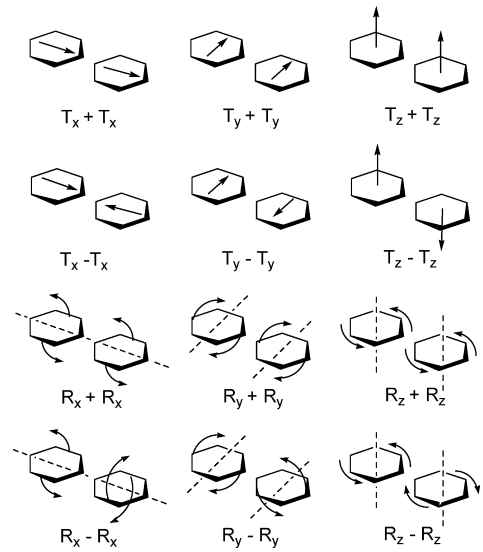


Figure 1. Schematic representation of the twelve basic motions of a molecular dimer.

spanned by the six basic motions each of molecules I and II, Hessian analysis results in six nonzero eigenvalues for intermolecular vibration modes and six zero-eigenvalues for pure translation and pure rotation of a given dimeric system.

The vectors that represent the basic motions of the monomer are normalized as follows

$$(\mathbf{T}_x\,\mathbf{T}_y\,\mathbf{T}_z) = \frac{1}{\sqrt{N}}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2}$$

$$(\mathbf{R}_x\,\mathbf{R}_y\,\mathbf{R}_z) = \begin{pmatrix} 0 & z_1\sin\theta_y & -y_1\sin\theta_z \\ -z_1\sin\theta_x & 0 & x_1\sin\theta_z \\ y_1\sin\theta_x & -x_1\sin\theta_y & 0 \\ \vdots & \vdots & \vdots \\ 0 & z_N\sin\theta_y & -y_N\sin\theta_z \\ -z_N\sin\theta_x & 0 & x_N\sin\theta_z \\ y_N\sin\theta_x & -x_N\sin\theta_y & 0 \end{pmatrix} \tag{3a}$$

$$\theta_q = \sin^{-1}\left(\sqrt{\sum_{i=1}^{N}(r_i^2 - q_i^2)}\right)^{-1} \quad q = \{x, y, z\} \tag{3b}$$

$$r_i^2 = x_i^2 + y_i^2 + z_i^2 \tag{3c}$$

where $\mathbf{T}_q$ and $\mathbf{R}_q$ ($q = x, y, z$) are column vectors of dimension $3N$, and (x_n, y_n, z_n) is the position of the n-th atom with respect to the center of mass. Thus, the $3(N_{\mathrm{I}} + N_{\mathrm{II}}) \times 12$ matrix $\mathbf{B}$ for the KL transformation of a dimeric system is written as follows, in which, e.g., $\bar{\mathbf{T}}_x$ represents $-\mathbf{T}_x$, and subscripts I and II denote the monomers I and II

$$\mathbf{B} = \frac{1}{\sqrt{2}}\begin{pmatrix} \mathbf{T}_{\mathrm{I}x}\mathbf{T}_{\mathrm{I}y}\mathbf{T}_{\mathrm{I}z}\mathbf{R}_{\mathrm{I}x}\mathbf{R}_{\mathrm{I}y}\mathbf{R}_{\mathrm{I}z} & \mathbf{T}_{\mathrm{I}x}\mathbf{T}_{\mathrm{I}y}\mathbf{T}_{\mathrm{I}z}\mathbf{R}_{\mathrm{I}x}\mathbf{R}_{\mathrm{I}y}\bar{\mathbf{R}}_{\mathrm{I}z} \\ \mathbf{T}_{\mathrm{II}x}\mathbf{T}_{\mathrm{II}y}\mathbf{T}_{\mathrm{II}z}\mathbf{R}_{\mathrm{II}x}\mathbf{R}_{\mathrm{II}y}\mathbf{R}_{\mathrm{II}z} & \bar{\mathbf{T}}_{\mathrm{II}x}\bar{\mathbf{T}}_{\mathrm{II}y}\bar{\mathbf{T}}_{\mathrm{II}z}\bar{\mathbf{R}}_{\mathrm{II}x}\bar{\mathbf{R}}_{\mathrm{II}y}\bar{\mathbf{R}}_{\mathrm{II}z} \end{pmatrix} \tag{4}$$

The orthogonality of the column vectors in the mass-weighted T/R basis $\mathbf{M}^{1/2}\mathbf{B}$ depends on the selection of the

coordinate system. These vectors can be modified to $\mathbf{M}^{1/2}\mathbf{C}$ so as to be orthonormal by using $\mathbf{S}$, the overlap matrix $\mathbf{B}^t\mathbf{M}\mathbf{B}$

$$\mathbf{M}^{1/2}\mathbf{C} = \mathbf{M}^{1/2}\mathbf{B}\mathbf{S}^{-1/2} \tag{5a}$$

$$\mathbf{S} = \mathbf{B}^t\mathbf{M}\mathbf{B} \tag{5b}$$

Here we redefine $\mathbf{X}$ as a $3(N_\mathrm{I} + N_\mathrm{II}) \times 12$ matrix that contains the atomic displacement vectors for six transverse (columns 1−3) and rotational motions (columns 4−6) and six intermolecular vibration modes (columns 7−12) of the dimer. Then, the coefficients of the KL expansion are collected in $\mathbf{\Xi}$, a 12×12 matrix. In other words, the displacement vectors of dimension $3N$ are reduced to 12 dimensions

$$\mathbf{\Xi} = \mathbf{C}^t\mathbf{M}\mathbf{X} \tag{6}$$

According to the definition of $\mathbf{S}$, we can construct a matrix $\mathbf{\Gamma}^{-1}$ that represents the inertial load of the molecules

$$\begin{aligned}\mathbf{\Gamma}^{-1} &\equiv \mathbf{N}^{1/2}\mathbf{S}\mathbf{N}^{1/2} \\ &= \mathbf{N}^{1/2}\mathbf{S}^{1/2}\mathbf{C}^t\mathbf{M}\mathbf{C}\mathbf{S}^{1/2}\mathbf{N}^{1/2} \\ &= \frac{1}{2}\begin{pmatrix} (M_\mathrm{I}+M_\mathrm{II})\mathbf{E} & 0 & (M_\mathrm{I}-M_\mathrm{II})\mathbf{E} & 0 \\ 0 & \mathbf{I}_\mathrm{I}+\mathbf{I}_\mathrm{II} & 0 & \mathbf{I}_\mathrm{I}-\mathbf{I}_\mathrm{II} \\ (M_\mathrm{I}-M_\mathrm{II})\mathbf{E} & 0 & (M_\mathrm{I}+M_\mathrm{II})\mathbf{E} & 0 \\ 0 & \mathbf{I}_\mathrm{I}-\mathbf{I}_\mathrm{II} & 0 & \mathbf{I}_\mathrm{I}+\mathbf{I}_\mathrm{II} \end{pmatrix}\end{aligned} \tag{7}$$

where M and $\mathbf{I}$ are the molecular weight and the tensor of inertia, respectively, of a monomer. In eq 7, we define the matrix $\mathbf{N}^{-1}$ as consisting of the molecular weights and the spatial extent of atoms of given monomers (θ_q^{-2}/N nearly equals the variance of the atomic location around the q-axis)

$$\mathbf{N}^{-1} = \frac{1}{2}\begin{pmatrix} (N_\mathrm{I}^1+N_\mathrm{II}^1)\mathbf{E} & 0 & (N_\mathrm{I}^1-N_\mathrm{II}^1)\mathbf{E} & 0 \\ 0 & \mathbf{\Theta}_\mathrm{I}^2+\mathbf{\Theta}_\mathrm{II}^2 & 0 & \mathbf{\Theta}_\mathrm{I}^2-\mathbf{\Theta}_\mathrm{II}^2 \\ (N_\mathrm{I}^1-N_\mathrm{II}^1)\mathbf{E} & 0 & (N_\mathrm{I}^1+N_\mathrm{II}^1)\mathbf{E} & 0 \\ 0 & \mathbf{\Theta}_\mathrm{I}^2-\mathbf{\Theta}_\mathrm{II}^2 & 0 & \mathbf{\Theta}_\mathrm{I}^2+\mathbf{\Theta}_\mathrm{II}^2 \end{pmatrix} \tag{8a}$$

$$\mathbf{E} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \mathbf{\Theta} = \begin{pmatrix} \theta_x & 0 & 0 \\ 0 & \theta_y & 0 \\ 0 & 0 & \theta_z \end{pmatrix} \tag{8b}$$

The matrix $\mathbf{\Gamma}$, named by analogy with the G-matrix in the GF method,[24] represents the measure of inertia in a given internal coordinate system. Similarly a matrix $\mathbf{\Phi}$ can be defined to represent a measure of stiffness

$$\mathbf{\Phi} = \mathbf{N}^{1/2}\mathbf{S}^{1/2}\mathbf{C}^t\mathbf{K}\mathbf{C}\mathbf{S}^{1/2}\mathbf{N}^{1/2} \tag{9}$$

Using eqs 6, 7, and 9, we can rewrite eq 1 into the following equations

$$(\mathbf{\Gamma}^{1/2}\mathbf{\Phi}\mathbf{\Gamma}^{1/2})\mathbf{L} = \mathbf{L}\mathbf{\Omega}^2 \tag{10a}$$

$$\mathbf{L} = \mathbf{\Gamma}^{-1/2}\mathbf{N}^{-1/2}\mathbf{S}^{-1/2}\mathbf{\Xi} \tag{10b}$$

Because the matrix $\mathbf{\Gamma}^{-1/2}\mathbf{\Phi}\mathbf{\Gamma}^{-1/2}$ is Hermitian, its eigenvectors $\boldsymbol{L}_i$ (i-th column vectors of $\mathbf{L}$) have to be orthogonal but not necessarily normalized. The product $\mathbf{L}^t\mathbf{L}$ gives a diagonal matrix $\mathbf{\Lambda}^2$, the elements of which represent the modal masses in the reduced T/R coordinate system

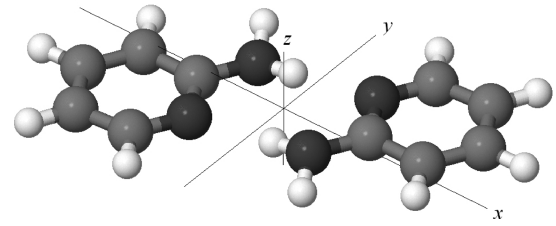


Figure 2. Coordinate system used for analysis of the intermolecular vibrations of 2-aminopyridine dimer.

$$\begin{aligned}\mathbf{\Lambda}^2 &= \mathbf{L}^t\mathbf{L} \\ &= \mathbf{\Xi}^t\mathbf{C}^t\mathbf{M}\mathbf{C}\mathbf{\Xi} \\ &= \mathbf{X}^t\mathbf{M}\mathbf{X}.\end{aligned} \tag{11}$$

Equation 11 indicates that the modal masses in the T/R coordinate system would ideally be identical to those in the full-atom coordinate system, if $\mathbf{M}^{-1/2}\mathbf{C}$ provides a sufficiently good basis of the expansion (*vide infra*). In the present study, however, we redefine $\mathbf{\Lambda}^2$ by collecting the squared norms (λ_i^2) of $\boldsymbol{L}_i$ as diagonal elements

$$\mathbf{\Lambda}^2 \equiv \begin{pmatrix} \lambda_1^2 & & & 0 \\ & \lambda_2^2 & & \\ & & \ddots & \\ 0 & & & \lambda_{12}^2 \end{pmatrix} . \tag{12}$$

Using $\mathbf{\Lambda}^{-1}$ as the normalization factor, we can obtain $\mathbf{U}$, which is approximately unitary (and could be unitary if the basis set of KL expansion is complete). The squared components of $\mathbf{U}$ represent the mixing ratio of the basic motions like $\mathrm{T}_x + \mathrm{T}_x$ in a given atomic displacement for intermolecular vibration

$$\mathbf{U} = \mathbf{\Gamma}^{-1/2}\mathbf{N}^{-1/2}\mathbf{S}^{-1/2}\mathbf{\Xi}\mathbf{\Lambda}^{-1} \tag{13}$$

Then, we can obtain the force constant matrix $\mathbf{\Phi}$ for a given internal T/R coordinate system as follows

$$\mathbf{\Phi} = \mathbf{\Gamma}^{-1/2}\mathbf{U}\mathbf{\Omega}^2\mathbf{U}^t\mathbf{\Gamma}^{-1/2} \tag{14}$$

3. Computational Details

Hydrogen-bonded dimers of 2-aminopyridine and its 5-halogenated derivatives were selected as examples for the analysis described above. The geometry of the dimers was optimized by means of the Hartree−Fock method using the 6-311G** basis set, and the structure thus obtained was used for the normal-mode vibration analysis at the same level of calculation. These molecular orbital calculations were performed with the Gaussian03W program.[26] The values of the geometry and eigenvectors were picked up from the output file and then processed with versatile spreadsheet software.

4. Results and Discussion

4.1. Derivation of Force Constants. For the optimized structure of a dimer of 2-aminopyridine, we set the coordinate system as shown in Figure 2. The values of the Cartesian coordinates were used for preparing the $\mathbf{B}$ and $\mathbf{C}$ matrices.

Table 1. Elements (in g mol^{-1} unit) of $\mathbf{\Gamma}^{-1}$

	T_x	T_y	T_z	R_x	R_y	R_z
T_x	94.1	0.00	0.00	0.00	−0.00	−0.00
T_y	0.00	94.1	0.00	−0.00	−0.00	0.00
T_z	0.00	0.00	94.1	0.00	0.00	0.00
R_x	0.00	−0.00	0.00	156.6	42.9	0.69
R_y	−0.00	−0.00	0.00	42.9	111.3	−1.29
R_z	−0.00	0.00	0.00	0.69	−1.29	267.7

Table 2. Comparison of the Modal Mass and Modal Stiffness Calculated for Internal Coordinate Systems

	ν (cm^{-1})	ω^2/N_A (s^{-2} mol)	modal mass (g mol^{-1})		modal stiffness (N m^{-1})	
			T/R	full-atom	T/R	full-atom
Twist	16.5	16.0	4.03	4.11	0.06	0.07
Buckle	28.4	47.7	4.73	4.71	0.23	0.22
Opening	63.6	238.9	4.29	4.36	1.02	1.04
Staggered	67.7	270.7	6.20	6.30	1.68	1.70
Shear	77.4	353.7	6.21	6.27	2.20	2.21
Stretch	109.4	706.0	5.68	5.70	4.01	4.02

The first six columns (for transverse and rotational motions) of **X** were made by a procedure similar to that used to make the **B** matrix, and the subsequent six columns were picked up from the output of the calculation of normal-mode vibration. Then we obtained **Ξ** according to eq 6. The 12 × 12 diagonal matrix **Ω** was made from the six eigenvalues (converted to ω^2/N_A in s^{-2} mol unit) of the normal-mode analysis. The first six elements (for transverse and rotational motions) were forced to be zero, and the latter six elements were allocated for the intermolecular vibrations.

Next, following eq 7, we constructed the $\mathbf{\Gamma}^{-1}$ matrix, which contains information on the inertial load (Table 1). Because we are studying a homodimer, only the top-left quarter of the matrix is fundamental (i.e., the full matrix is a direct sum of this table). The block related to transverse motions, i.e., T_x, T_y, and T_z, is diagonal, and the elements Γ_{ii}^{-1} substantially coincide with the molecular weight (in g mol^{-1}) of the monomer. The block related to rotational motions, i.e., R_x, R_y, and R_z, represents the tensor of inertia given in units of g Å^2 mol^{-1} (= 10^{-23} kg m^2 mol^{-1}). Then we derived $\mathbf{\Gamma}^{-1/2}$, which was used to make **L**.

By visualizing the atomic displacement, we labeled each vibration mode by one of six nicknames (Twist, Buckle, Staggered, Opening, Shear, and Stretch) that represent the character of the motion.[27] The modal mass (λ^2/g mol^{-1}) of each vibration mode for the internal T/R basis is given in Table 2. These values are in fairly good agreement within a factor of 0.98−1.00 with the modal mass calculated for the full-atom coordinate. This agreement was as expected from eq 12, which also validates the completeness of the T/R basis set. Accordingly, we can also see good agreement between the values of the modal stiffness that were calculated for respective coordinate systems.

By following eq 13, we obtained **U**. Columns 7−12 of **U** are shown in Table 3. We confirmed that the product of $\mathbf{U}^t$ and **U** approximates a unit matrix, again showing that the T/R motions that were used serve as a good basis for expansion of the intermolecular vibration.

Table 3. Elements of the Block of Intermolecular Vibrations in **U** Matrix

	Twist	Buckle	Opening	Staggered	Shear	Stretch
$T_x + T_x$	−0.005	0.005	−0.009	0.000	0.000	0.000
$T_y + T_y$	0.001	−0.001	0.008	0.000	0.000	0.000
$T_z + T_z$	−0.003	−0.009	−0.006	0.000	0.000	0.000
$R_x + R_x$	0.000	0.000	0.000	−0.095	−0.010	0.117
$R_y + R_y$	0.000	0.000	0.000	−0.680	−0.620	0.040
$R_z + R_z$	0.000	0.000	0.000	−0.201	0.312	0.772
$T_x - T_x$	0.000	0.000	0.000	0.594	−0.556	0.529
$T_y - T_y$	0.000	0.000	0.000	−0.299	0.344	0.321
$T_z - T_z$	0.000	0.000	0.000	0.215	0.301	−0.080
$R_x - R_x$	−0.959	0.192	−0.180	0.000	0.000	0.000
$R_y - R_y$	0.201	0.981	0.056	0.000	0.000	0.000
$R_z - R_z$	0.200	0.012	−0.982	0.000	0.000	0.000

Table 4. Elements (in N m^{-1}) of **Φ** in the *Gerade* Block

	A_g-like			B_g-like		
	$T_x - T_x$	$T_y - T_y$	$R_z + R_z$	$R_x + R_x$	$R_y + R_y$	$T_z - T_z$
$T_x - T_x$	37.9	0.37	30.9	4.20	3.24	−5.11
$T_y - T_y$	0.37	13.1	36.3	3.83	−0.64	0.11
$R_z + R_z$	30.9	36.3	124.8	13.7	−0.06	−3.42
$R_x + R_x$	4.20	3.83	13.7	3.92	8.40	−3.55
$R_y + R_y$	3.24	−0.64	−0.06	8.40	29.2	−11.1
$T_z - T_z$	−5.11	0.11	−3.42	−3.55	−11.1	4.60

Table 5. Elements (in N m^{-1}) of **Φ** in the *Ungerade* Block

	B_u-like			A_u-like		
	$T_x + T_x$	$T_y + T_y$	$R_z - R_z$	$R_x - R_x$	$R_y - R_y$	$T_z + T_z$
$T_x + T_x$	0.00	0.00	0.34	0.06	0.02	0.00
$T_y + T_y$	0.00	0.00	−0.29	−0.04	0.00	0.00
$R_z - R_z$	0.34	−0.29	61.89	7.67	−1.05	0.21
$R_x - R_x$	0.06	−0.04	7.67	4.07	1.93	0.01
$R_y - R_y$	0.02	0.00	−1.05	1.93	5.33	−0.05
$T_z + T_z$	0.00	0.00	0.21	0.01	−0.05	0.00

It can be seen from Table 2 that for the vibration modes of Twist, Buckle, and Opening, the atomic displacements are predominantly represented by $R_x - R_x$ (92%), $R_y - R_y$ (96%), and $R_z - R_z$ (96%) motions, respectively. Although we assigned the vibration modes at 67.7 and 77.4 cm^{-1} as Staggered and Shear motions, it is hard to distinguish one from the other by observing the visualized motion. These two modes are represented as combinations of $R_y + R_y$ and $T_x - T_x$ but with a difference in their mixing phase. The Stretch motion, which is often treated by a simple pseudodiatomic approximation, is actually dominated by $R_z + R_z$ (60%) motion for this case.

Subsequently, we obtained the **Φ** matrix according to eq 14. Tables 4 and 5 summarize the elements of **Φ** in the *gerade* block and the *ungerade* block, respectively, where the elements (Φ_{ij}) are given in units of N m^{-1}. For the terms related to rotational motions, the values correspond to a torque measured in N m^{-1} Å^2 (= 10^{-20} N m). Because the molecular system belongs to the C_i point group, the off-diagonal elements of **Φ** between the motions of the *gerade* representation ($T_x - T_x$, $T_y - T_y$, $T_z - T_z$, $R_x + R_x$, $R_y + R_y$, and $R_z + R_z$) and the motions of the *ungerade* representation ($T_x + T_x$, $T_y + T_y$, $T_z + T_z$, $R_x - R_x$, $R_y - R_y$, and $R_z - R_z$) are definitely zero. Furthermore, if we neglect the pyramidization of the amino groups, the molec-

Table 6. Comparison of Force Constants and Wavenumbers of Basic Motions

	wavenumber (cm^{-1})		force constant (N m^{-1})	
	T/R basis	rigid body	T/R basis	rigid body
$T_x - T_x$	82.6	90.9	18.9	22.9
$T_y - T_y$	48.5	53.0	6.5	9.6
$T_z - T_z$	28.8	29.4	2.3	2.4
$R_x + R_x$	20.6	18.5	2.0	1.6
$R_y + R_y$	66.7	30.9	14.6	3.1
$R_z + R_z$	88.9	18.4	62.4	2.7
$R_x - R_x$	21.0	26.1	2.0	3.1
$R_y - R_y$	28.5	0.0	2.7	0.0
$R_z - R_z$	62.6	61.9	30.9	30.2

ular dimer has nearly C_{2h} symmetry. Thus, dividing the table into blocks for the A_g-like motions ($T_x - T_x$, $T_y - T_y$, and $R_z + R_z$) and the B_g-like motions ($R_x + R_x$, $R_y + R_y$, and $T_z - T_z$) is informative. For the A_g vs A_g and B_g vs B_g blocks, the diagonal elements show relatively large positive values, but some off-diagonal elements are as large as the diagonal ones. The off-diagonal elements in the A_g vs B_g block are relatively small but are not quite zero, suggesting a mixing of the A_g and B_g representations due to the deformation from a strict C_{2h} symmetry. In particular, the cross term (13.7 N m^{-1}) between $R_x + R_x$ and $R_z + R_z$ is exceptionally large. For the B_u vs B_u and A_u vs A_u blocks, the diagonal terms of $T_x + T_x$, $T_y + T_y$, and $T_z + T_z$ are naturally close to zero, because these bases represent the transverse motion of the molecular dimer. The off-diagonal terms in the A_u vs B_u block are relatively small, but again we observe an exceptionally large coupling constant (7.67 N m^{-1}) between $R_x - R_x$ and $R_z - R_z$.

4.2. Comparison with the Rigid Body Approximation. In a previous study, we investigated the mechanical nature of multiply hydrogen-bonded systems by means of ab initio quantum chemical calculations.[28] By a multivariate analysis of the curvature of the potential surface for 20 dimeric molecular systems, we derived a set of force constants for translational shear in the x, y, and z directions. The force constants only roughly reproduced the frequencies of the intermolecular vibration modes of a molecular dimer, which in turn indicated the significance of the off-diagonal terms of the dynamical matrix when evaluating the frequency of intermolecular vibration. In the present study, we succeeded in explicitly obtaining the full elements of the stiffness matrix for an internal coordinate system reduced by KL transformation. Here we attempt to compare the present results with those of our previous calculations. If we neglect the off-diagonal terms, we can calculate the wavenumber for each mode of the basic motions as follows

$$\tilde{\nu}_i = \frac{1}{2\pi c}\sqrt{\frac{\Phi_{ii}}{\Gamma_{ii}}} \qquad (15)$$

Table 6 lists the wavenumbers calculated with eq 15. The values (in cm^{-1}) for $R_x - R_x$ (21.0), $R_y - R_y$ (28.5), and $R_z - R_z$ (62.6) motions are in good agreement with those for Twist (16.5), Buckle (28.4), and Opening (63.6) motions, respectively, as expected from the predominant component of these normal modes (Table 3). The wavenumbers for $T_x - T_x$ (82.6), $T_y - T_y$ (48.5), $R_y + R_y$ (66.7), and $R_z + R_z$ (88.9) are within a range similar to those of Shear (67.7), Stagger (77.4), and Stretch (109.4) modes but are not exactly identical. We interpret this mismatch in frequencies as a result of appreciable coupling terms, namely, $T_x - T_x$ vs $R_z + R_z$ (30.9 N m^{-1}) and $T_y - T_y$ vs $R_z + R_z$ (36.3 N m^{-1}), as given in Table 4.

On the other hand, based on simple rigid body mechanics, the wavenumber of homodimeric molecule is given as follows

$$\tilde{\nu}_i = \frac{1}{2\pi c}\sqrt{\frac{(M_{\mathrm{I}} + M_{\mathrm{II}})K_i}{M_{\mathrm{I}}M_{\mathrm{II}}}} = \frac{1}{2\pi c}\sqrt{\frac{2K_i}{M}} \qquad (16)$$

where K_i (i denotes the type of motion) is the force constant of the intermolecular interaction, and M ($= M_{\mathrm{I}} = M_{\mathrm{II}}$) is either the molecular weight (for transverse motion) or moment of inertia divided by 1 $Å^2$ (for rotational motion). Therefore, we can compare a series of force constants for 12 basic motions by simply calculating $\Phi_{ii}/2$. Table 6 compares the force constants and wavenumbers calculated by means of the method described in this paper (eq 15) and those calculated by rigid body mechanics (eq 16). For the rigid body mechanical calculation, we used a set of force constants that we developed for NH···N hydrogen bonds.[28] Note that for transverse motion (e.g., $T_x - T_x$), the force constants (18.9, 6.5, 2.3 N m^{-1}) derived from the T/R based expansion are in good agreement with those (22.9, 9.6, 2.4 N m^{-1}) from a rigid body approximation. As for rotational motions, the force constants for antisymmetric combinations (e.g., $R_x - R_x$) show good agreement between the two methods. A deviation found in the $R_y - R_y$ mode may be attributed to the intentional neglect of the pyramidization of the amino group in our former work. On the contrary, a similar comparison for symmetric combinations (e.g., $R_x + R_x$) shows a considerable disagreement between the results obtained by means of the two methods. This disagreement demonstrates the difficulty in deriving force constants analytically for molecules with complicated shapes.

4.3. Effects of Molecular Inertial Load on Frequency. Among the various intermolecular vibration modes, Stretch motion is of special interest because of its spectroscopic accessibility as well as its relevance to the strongest force of intermolecular interaction.[29−31] Conventionally, the relation between the frequency and force constants is analyzed based on the pseudodiatomic approximation or on its modified formulation. Although the latter method is applicable to asymmetric top dimers, it can handle only the force constant of stretching motion.[29] As demonstrated above, however, the frequency of the Stretch mode is determined as a result of coupling among $T_x - T_x$, $T_y - T_y$, and $R_z + R_z$ motions, all of which belong to the A_g representation on the assumption that the dimer has C_{2h} symmetry. These three basic motions give rise to two other motions with A_g representation, that is, Shear mode vibrations and rotation of the dimer around the z-axis (R_z). Our present study provides a comprehensive interpretation of the relation between the frequency and force constants of intermolecular vibration.

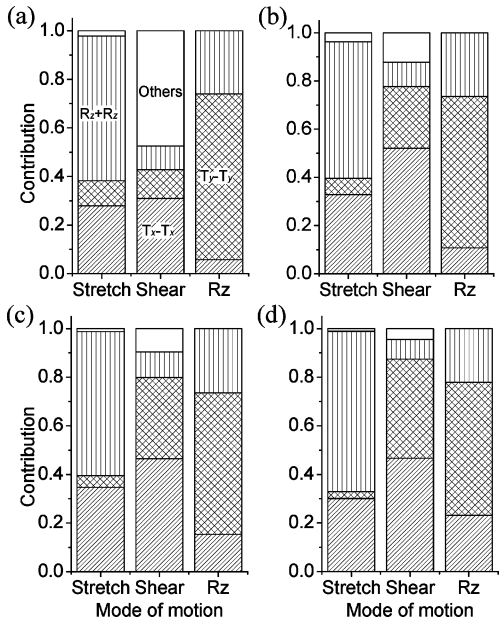


Figure 3. Contributions of $T_x - T_x$, $T_y - T_y$, and $R_z + R_z$ motions to the vibration modes of (a) 2-aminopyridine and (b) 5-fluoro-, (c) 5-chloro-, and (d) 5-bromo-2-aminopyridines.

Previously, we pointed out that the frequencies of the intermolecular vibration modes of 5-halogenated-2-aminopyridine do not follow the pseudodiatomic model.[28] For example, the frequency of the Stretch mode was nearly insensitive to changes in the molecular weight. Here we attempt to clarify the relation between the frequencies and molecular weight of 2-aminopyridines on the basis of our present formulation. Figure 3(a)-(d) shows the contribution (U_{ij}^2) of the $T_x - T_x$, $T_y - T_y$, and $R_z + R_z$ motions to Stretch and Shear mode vibrations and R_z motion, for 2-aminopyridine and its 5-halogenated derivatives. As can be seen from these figures, each of these motions is sufficiently described as a combination of three basic motions with A_g representation, although their contributions differ slightly depending on the molecule. This result suggests that the frequency of the Stretch and Shear modes can be approximately evaluated by diagonalizing the 3×3 partial matrix of $\mathbf{\Gamma}^{1/2}\mathbf{\Phi}\mathbf{\Gamma}^{1/2}$ that corresponds to the cross terms among the $T_x - T_x$, $T_y - T_y$, and $R_z + R_z$ bases.

Figure 4 shows the values of selected elements in the stiffness matrix $\mathbf{\Phi}$ as a function of $(M_0/M)^{1/2}$, where M_0 and M are the molecular weights of 2-aminopyridine and 5-halogenated-2-aminopyridine, respectively. The diagonal elements related to transverse displacement, namely, $\Phi_{Tx\text{-}Tx,Tx\text{-}Tx}$ and $\Phi_{Ty\text{-}Ty,Ty\text{-}Ty}$ (Figure 3(a)), are scarcely influenced by substitution at the 5-position, and their cross term $\Phi_{Tx\text{-}Tx,Ty\text{-}Ty}$ (Figure 3(b)) is nearly equal to zero. These results indicate that the restoring force along the *x*- and *y*-axes mainly originates in double NH···N hydrogen bonds, and any effects from halogen substitution are negligibly small. As for $\Phi_{Rz+Rz,Rz+Rz}$, the force constant for the twisting distortion, the value steeply increases with increasing molecular weight. Because $\Phi_{Rz+Rz,Rz+Rz}$ is proportional to the torque around the *z*-axis, this constant should increase with the square of distance (R_{HB}) between the center of mass and the centers of hydrogen bonding sites.[32] Figure 4(a) shows the

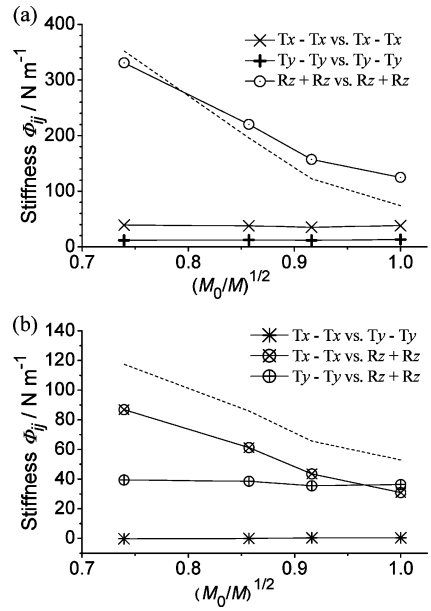


Figure 4. (a) Diagonal and (b) off-diagonal elements of the stiffness matrix ($\mathbf{\Phi}$) as functions of $(M_0/M)^{-1/2}$. Dashed lines indicate the values of $\Phi_{Tx\text{-}Tx,Tx\text{-}Tx}R_{HB}^2$ (in (a)) and $\Phi_{Tx\text{-}Tx,Tx\text{-}Tx}R_{HB}$ (in (b)) for comparison with a model based on rigid mechanics.

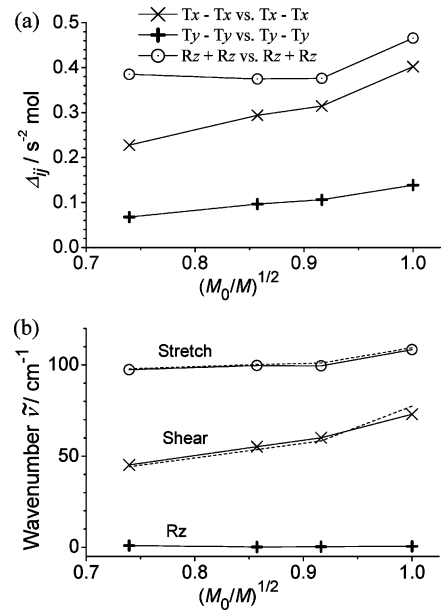


Figure 5. (a) The diagonal elements and (b) the eigenvalues of a 3×3 partial matrix of $\mathbf{\Delta}$. The frequencies of the Stretch and Shear modes calculated for the full-atomic representation are also imposed on (b).

plot of $\Phi_{Tx\text{-}Tx,Tx\text{-}Tx}R_{HB}^2$, which is in fairly good agreement with $\Phi_{Rz+Rz,Rz+Rz}$, again suggesting that hydrogen bonding is a predominant interaction for intermolecular forces. Similarly, the plot of $\Phi_{Tx\text{-}Tx,Tx\text{-}Tx}R_{HB}$ shows behavior parallel to that of $\Phi_{Tx\text{-}Tx,Rz+Rz}$.

Figure 5(a) shows some elements of $\mathbf{\Gamma}^{1/2}\mathbf{\Phi}\mathbf{\Gamma}^{1/2}$ ($\equiv \mathbf{\Delta}$), a dynamical matrix for a given T/R coordinate. The elements related to transverse motion, namely, $\Delta_{Tx\text{-}Tx,Tx\text{-}Tx}$ and $\Delta_{Ty\text{-}Ty,Ty\text{-}Ty}$, are nearly proportional to $(M_0/M)^{1/2}$, which is a result expected from the pseudodiatomic model. The behavior

of the value of $\Delta_{Rz+Rz,Rz+Rz}$ is not simple, but it is rather insensitive to the change of molecular weight. This constancy is interpretable from the compensation of the changes in $\Phi_{Rz+Rz,Rz+Rz}$ and $\Gamma^{-1}{}_{Rz+Rz,Rz+Rz}$, both of which are roughly proportional to $R_{HB}{}^2$. Figure 5(b) shows the eigenvalues of a 3×3 matrix that contains the elements related to $T_x - T_x$, $T_y - T_y$, and $R_z + R_z$ in $\mathbf{\Delta}$. One of the eigenvalues nearly equals zero, suggesting its correspondence to R_z motion. The other two eigenvalues are in excellent agreement with those of the Stretch and Shear modes that were calculated by using full-atom coordination. Consequently, the apparent complicated behavior of the Stretch mode frequency observed for the 5-halogenated-2-aminopyridine dimer is a result of the constant stiffness ($\sim$20 N m^{-1}) of double NH$\cdots$N hydrogen bonds and the sequential transfer of the center of mass. This result clearly shows that our approach is useful for understanding the fundamentals of intermolecular vibrations even for asymmetric top dimers with various moments of inertia.

5. Conclusions

This report presents a procedure to evaluate the elements of the stiffness matrix relevant to intermolecular force based on normal-mode calculations for full-atomic representation. We utilized a variation of Karhunen-Loève transformation that is quite useful for extracting the characteristics of serial data such as digitalized information. We developed a quantitative representation for characterizing atomic displacement during intermolecular vibration, which had conventionally been classified according to visualization of molecular motion. By using a compressed 12-dimensional space of molecular motion, we obtained the elements of the stiffness matrix, which yields a dynamical matrix when combined with the inertial load of the molecules in question. Note also that the treatment based on KL transformation is not confined to a rigid molecule approximation. It would be possible to take into account the mixing of intra- and intermolecular vibrations in the THz region by adding some eigenvectors for the single molecular normal mode to the basis set of the KL expansion. Accordingly, the method presented in this study is quite versatile for evaluating the parameters necessary for the coarse-grained normal-mode calculation of a large molecular assembly including a periodic system as well as for obtaining deep insight into the nature of intermolecular vibration.

References

(1) For recent review: Plusquellic, D. F.; Siegrist, K.; Heilweil, E. J.; Esenturk, O. *Chem. Phys. Chem.* **2007**, *8*, 2412–2431.

(2) Heyden, M.; Bründermann, E.; Heugen, U.; Niehues, G.; Leitner, D. M.; Havenith, M. *J. Am. Chem. Soc.* **2008**, *130*, 5773–5779.

(3) Fischer, B. M.; Walther, M.; Jepsen, P. U. *Phys. Med. Biol.* **2002**, *47*, 3807–3814.

(4) Taday, P. F.; Bradley, I. V.; Arnone, D. D.; Pepper, M. *J. Pharm. Sci.* **2003**, *92*, 831–838.

(5) Aaltonen, J.; Allesø, M.; Mirza, S.; Koradia, V.; Gordon, K. C.; Rantanen, J. *Eur. J. Pharm. Biopharm.* **2009**, *71*, 23–37.

(6) Day, G. M.; Zeitler, J. A.; Jones, W.; Rades, T.; Taday, P. F. *J. Phys. Chem. B* **2006**, *110*, 447–456.

(7) Takahashi, M.; Ishikawa, Y.; Nishizawa, J.; Ito, H. *Chem. Phys. Lett.* **2005**, *401*, 475–482.

(8) Rungsawang, R.; Ueno, Y.; Tomita, I.; Ajito, K. *J. Phys. Chem. B* **2006**, *110*, 21259–21263.

(9) Jepsen, P. U.; Clark, S. J. *Chem. Phys. Lett.* **2007**, *442*, 275–280.

(10) (a) Siegrist, K.; Bucher, R.; Mandelbaum, I.; Walker, A. R. H.; Balu, R.; Gregurick, S. K.; Plusquellic, D. F. *J. Am. Chem. Soc.* **2006**, *128*, 5764–5775. (b) Zhang, H.; Siegrist, K.; Plusquelic, D. F.; Gregurick, S. K. *J. Am. Chem. Soc.* **2008**, *130*, 17846–17857. (c) Zhang, H.; Zukowski, E.; Balu, R.; Gregurick, S. K. *J. Mol. Graphics Modell.* **2009**, *27*, 655–663.

(11) Saito, S.; Inerbaev, T. M.; Mizuseki, H.; Igarashi, N.; Note, R.; Kawazoe, Y. *Chem. Phys. Lett.* **2006**, *423*, 439–444.

(12) (a) Fedor, A. M.; Korter, T. M. *Chem. Phys. Lett.* **2006**, *429*, 405–409. (b) Korter, T. M.; Balu, R.; Campbell, M. B.; Beard, M. C.; Gregurick, S. K.; Heilweil, E. J. *Chem. Phys. Lett.* **2006**, *418*, 65–70. (c) Allis, D. G.; Fedor, A. M.; Korter, T. M.; Bjarnason, J. E.; Brown, E. R. *Chem. Phys. Lett.* **2007**, *440*, 203–209. (ca) Allis, D. G.; Korter, T. M. *Chem. Phys. Chem.* **2006**, *7*, 2398–2408. (d) Allis, D. G.; Prokhorova, D. A.; Korter, T. M. *J. Phys. Chem. A* **2006**, *110*, 1951–1959.

(13) Brauer, B.; Gerber, R. B.; Kabelác, M.; Hobza, P.; Bakker, J. M.; Riziq, A. G. A.; de Vries, M. S. *J. Phys. Chem. A* **2005**, *109*, 6974–6984.

(14) Špirko, V.; Šponer, J.; Hobza, P. *J. Chem. Phys.* **1997**, *106*, 1472–1479.

(15) For examples: (a) Ma, J. *Structure* **2005**, *13*, 373–380. (b) Bahar, I.; Rader, A. J. *Curr. Opin. Struct. Biol.* **2005**, *15*, 586–592. (c) Li, G.; Cui, Q. *Biophys. J.* **2004**, *86*, 743–763. (d) Li, G.; Cui, Q. *Biophys. J.* **2002**, *83*, 2457–2474. (e) Kindt, J. T.; Schmuttenmaer, C. A. *J. Chem. Phys.* **1997**, *106*, 4389–4400.

(16) Dove, M. T. In *Introduction to Lattice Dynamics*; Cambridge University Press: New York, 1993.

(17) (a) Neto, N.; Righini, R.; Califano, S.; Walmsley, S. H. *Chem. Phys.* **1978**, *29*, 167–179. (b) Schettino, V.; Califano, S. *J. Mol. Struct.* **1983**, *100*, 459–483.

(18) Kearley, G. J.; Johnson, M. R.; Tomkinson, J. *J. Chem. Phys.* **2006**, *124*, 044514.

(19) (a) Durand, D.; Trinquier, G.; Sanejourand, Y.-H. *Biophys. J.* **1994**, *34*, 759–771. (b) Tama, F.; Gadea, F. X.; Marques, O.; Sanejourand, Y.-H. *Prot. Struct. Funct. Gen.* **2000**, *41*, 1–7.

(20) Willock, D. J.; Price, S. L.; Leslie, M.; Catlow, C. R. A. *J. Comput. Chem.* **1995**, *16*, 628–647.

(21) (a) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765. (b) 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.

(22) MacKerell, A. D.; Wiórkiewicz-Kuczera, J.; Karplus, M. *J. Am. Chem. Soc.* **1995**, *117*, 11946–11975.

(23) Sun, H. *J. Phys. Chem. B* **1998**, *102*, 7338–7364.

(24) Wilson, E. B.,Jr.; Decius, J. C.; Cross, P. C. In Molecular Vibrations, Dover Publications, Inc.: New York, 1980.

(25) Ryckelynck, D.; Chinesta, F.; Cueto, E.; Ammar, A. *Arch. Comput. Meth. Engng.* **2006**, *13*, 91–128.

(26) 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, 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.

(27) Diekmann, S. *EMBO J.* **1989**, *8*, 1–4.

(28) Houjou, H.; Koga, R. *J. Phys. Chem. A* **2008**, *112*, 11256–11262.

(29) (a) Legon, A. C.; Millen, D. J. *Chem. Rev.* **1986**, *86*, 635–657. (b) Legon, A. C.; Millen, D. J. *Acc. Chem. Res.* **1987**, *20*, 39–46. (c) Millen, D. J. *Can. J. Chem.* **1985**, *63*, 1477–1479.

(30) (a) Müller, A.; Talbot, F.; Leutwyler, S. *J. Chem. Phys.* **2000**, *112*, 3717–3725. (b) Müller, A.; Talbot, F.; Leutwyler, S. *J. Chem. Phys.* **2002**, *116*, 2836–2847. (c) Müller, A.; Talbot, F.; Leutwyler, S. *J. Am. Chem. Soc.* **2002**, *124*, 14486–14494. (d) Müller, A.; Talbot, F.; Leutwyler, S. *J. Chem. Phys.* **2001**, *115*, 5192–5202.

(31) Melandri, S.; Sanz, M. E.; Caminati, W.; Favero, P. G.; Kisiel, Z. *J. Am. Chem. Soc.* **1998**, *120*, 11504–11509.

(32) This simple model assumes the potential function of $V(x, \theta) = (1/2)\Phi_{Tx\text{-}Tx}(x + R_{HB}\theta)^2$. The second derivatives are given as $\partial^2V/\partial\theta^2 = \Phi_{Tx\text{-}Tx}R_{HB}^2$ and $\partial^2V/\partial x\partial\theta = \Phi_{Tx\text{-}Tx}R_{HB}$.

CT900169F

JCTC Journal of Chemical Theory and Computation

Many-Body Perturbation Theory Extended to the Quantum Mechanics/Molecular Mechanics Approach: Application to Indole in Water Solution

Adriano Mosca Conte,[†] Emiliano Ippoliti,[‡] Rodolfo Del Sole,[†] Paolo Carloni,*[,‡,§] and Olivia Pulci[†]

NAST, ETSF, CNR INFM-SMC, Department of Physics, Universita' di Roma Tor Vergata, Via della Ricerca Scientifica 1, Roma, Italy, Democritos, SISSA—Scuola Internazionale Superiore di Studi Avanzati, via Beirut 2-4, I-34014 Trieste, Italy, and Italian Institute of Technology, SISSA Unit, Via Beirut 2—4, Trieste, Italy

Received December 3, 2008

Abstract: Optical properties of aromatic chromophores are used to probe complex biological processes, yet how the environment tunes their optical properties is far from being fully understood. Here we present a method to calculate such properties on large-scale systems, like biologically relevant molecules in aqueous solution. Our approach is based on many-body perturbation theory combined with a quantum mechanics/molecular mechanics (QM/MM) approach. We include quasiparticle and excitonic effects for the calculation of optical absorption spectra in a QM/MM scheme. We apply this scheme, together with the well-established TDDFT approach, to indole in water solution. Our calculations show that the solvent induces a red shift in the main spectral peak of indole, in quantitative agreement with the experiments, and they point to the relevance of both the electrostatic and geometrical origin of the shift.

1. Introduction

Optical properties of aromatic chromophores embody a key facet of cell biology, allowing for a precise interrogation of a variety of biochemical events, including signaling, metabolism, and aberrant processes. These range from probing transient interactions between biomolecules (proteins and nucleic acids), to protein dynamics and fibrillation and plaque formation in neurodegenerative diseases. Understanding how the environment tunes such optical properties is therefore crucial, yet this information is so far mostly lacking. A powerful tool to address this issue is given by the so-called quantum mechanics/molecular mechanics (QM/MM) methods.[1,2] In this approach, the aromatic moiety may be treated at the quantum mechanical level, while the environment is described with an effective potential: the influence of the MM (presumably very complex and very large) environment is basically included as an external potential and, in case the chromophore is covalently bound to MM region, by a mechanical coupling with the environment.

Most often the QM approach is solved within density functional theory (DFT)[3,4] to study ground state properties, and time-dependent DFT (TDDFT)[5,6] when excited states are involved, as in the case of the optical properties.[7–9] TDDFT is computationally very efficient, yet its predictive power depends dramatically on the system and on the functional used to reproduce the exchange and correlation interactions.

Several approaches, including post-Hartree—Fock ones[10] (configuration interaction and similar methods), have been already used to predict optical properties of biomolecules. Many-body perturbation theory (MBPT),[11,12] is an attractive alternative, although of course it comes with higher computational cost than TDDFT. In fact, TDDFT[13] scales with N^3 ("N" is the number of atoms), while MBPT scales with N^6 (or N^4 in the case where the Haydock algorithm[14] is used). However, biophysical applications of one of the most widely used schemes of MBPT, the combination of the GW

* To whom correspondence should be addressed. E-mail: carloni@sissa.it.

[†] Universita' di Roma Tor Vergata.

[‡] Democritos.

[§] Italian Institute of Technology.

10.1021/ct800528e CCC: $40.75
Published on Web 05/29/2009

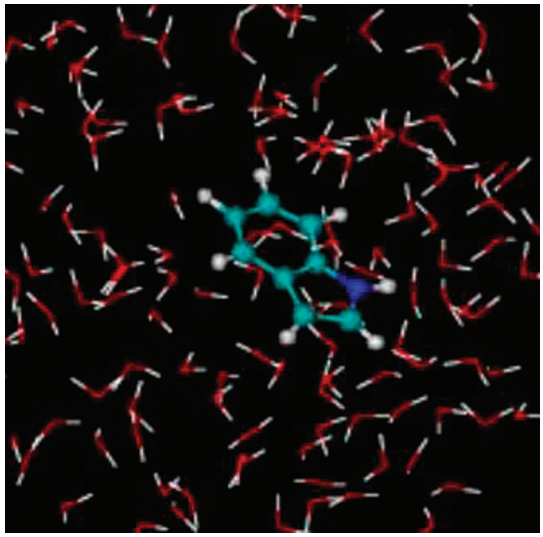


Figure 1. Indole in water solution. Colors correspond to the following atomic species: blue, N; cyan, C; white, H; red, O.

method[11] with the Bethe−Salpeter equation (BSE),[12] are, so far, lacking. The GW method is used for the evaluation of the single quasiparticle energies, and the BSE to introduce excitonic effects. The determination of a more accurate long-range exchange-correlation kernel of TDDFT is also based on MBPT.[15] Keeping in mind this for future biological applications, it is imperative to assess the accuracy of the MBPT/MM approach versus the more conventional TDDFT/MM one.[7,8]

The main assumption in interfacing a QM/MM method with TDDFT or MBPT approaches is that the optical properties of the chromophore do not involve the electronic structure of the MM part. Hence, special care has to be devoted to the choice of the two regions.

Here we present MBPT/MM calculations on the indole ring of the tryptophan residue (Figure 1). This system appears ideal for such an approach in several respects. First, it is very relevant biologically, as the indole ring has been exploited as a spectroscopic tool to monitor changes in proteins[16] and to yield information about local structure and dynamics. In fact, its spectral signatures allow it to be used as a structural probe in proteins. Second, it contains a relatively small number of atoms ($N = 16$), which can be treated quite easily at the GW-BSE level. Next, the optical gap of liquid water (7 eV)[17] is larger than the gap of the indole molecule (4.3 eV).[18] Under 7 eV, the spectra of indole and water do not overlap, and it is justified to treat the solvent in a classical scheme. Finally, CASPT2 calculations[18] and experimental data[19] are available and allow us to compare the changes of the optical properties upon passing from the gas phase to aqueous solution.

2. Methods and Computational Details

We performed QM/MM Car−Parrinello[20] simulations of indole in water by the fully Hamiltonian QM/MM scheme.[2] Such a scheme has been applied to a variety of biological systems.[21] The biomolecule was treated, at this step, at the DFT level, while the solvent was described by the TIP3P water model[22] and the van der Waals parameters on each indole atomic site in the interaction potential were employed by using the Amber force field.[23] This approach allows for an explicit treatment of solvation, in contrast to previous studies.[18,19,24]

Indole single quasiparticle energies have been then evaluated at the GW level for several snapshots. Finally, we solved the BSE to calculate the average absorption spectrum and compared the results with the ones obtained within TDDFT. We calculated the indole absorbance in water as well as in gas phase. The shift in the spectra gives the solvatochromism.

In this section we review the main aspects of the GW methods and of the Bethe−Salpeter equation and explain how it is possible to include these methods in a QM/MM scheme. In 1965, Hedin[25] formulated a set of five equations that link together five important functions: the Green's function, G; the self-energy, Σ; the vertex function, Γ; the polarizability, P; and the screened potential, W. The poles of the Green's function in frequency space are the energies of the states of an ionized system referred to the ground-state energy. Here the ionized system is a system to which an electron has been subtracted or added. These excitation energies are called quasiparticle energies ε_j^{QP} and can be derived by solving the following eigen-problem:

$$\left[-\frac{\hbar^2\nabla^2}{2m} + U^{QM}(\mathbf{r}) + U^{QM/MM}(\mathbf{r}) + V_H(\mathbf{r})\right]\phi_j^{QP}(\mathbf{r}) + \int d^3\mathbf{r}'\,\Sigma(\mathbf{r},\mathbf{r}',\varepsilon_j^{QP})\,\phi_j^{QP}(\mathbf{r}') = \varepsilon_j^{QP}\phi_j^{QP}(\mathbf{r}) \quad (1)$$

This expression is derived from the Dyson equation in the Lehmann representation.[11] V_H is the Hartree potential and U^{QM} is the Coulomb potential generated by the QM ions, while $U^{QM/MM}$ is the potential felt by the electrons due to the point charges of the MM part.

According to Hedin equations, the self-energy is a functional of G, W, and Γ. Several calculations have demonstrated that Γ can be approximated to a delta function for most of the systems.[26] Under this condition, the time-Fourier transform of the proper exchange-correlation self-energy, $\Sigma(\mathbf{r},\mathbf{r}',\omega)$, is a convolution of the Green's function, $G(\mathbf{r},\mathbf{r}',\omega)$, with the screened Coulomb potential, $W(\mathbf{r},\mathbf{r}',\omega)$: $\Sigma = iGW$. This is the reason for the name "GW approximation". In this work, Σ is calculated as the convolution of the noninteracting electron system's Green function (G_0) and the screened Coulomb potential (W_0) built from the Kohn−Sham (KS) eigenvalues and eigenvectors of the QM/MM system, $\varepsilon_j^{QM/MM}$ and $\phi_j^{QM/MM}$, respectively. More explicitly, G_0 in frequency space is

$$G_0(\mathbf{r},\mathbf{r}',\omega) = \sum_j \phi_j^{QM/MM}(\mathbf{r})\,\phi_j^{QM/MM*}(\mathbf{r}') \left[\frac{\theta(\varepsilon_j^{QM/MM} - \varepsilon_F)}{\omega - \varepsilon_j^{QM/MM} + i\eta} + \frac{\theta(\varepsilon_F - \varepsilon_j^{QM/MM})}{\omega - \varepsilon_j^{QM/MM} - i\eta}\right] \quad (2)$$

where ε_F is the Fermi energy. The screened potential is evaluated as $W = \varepsilon^{-1}v$, where v is the bare Coulomb potential and ε is the microscopic dielectric function at the independent particle level.

Equation 1 has the same form as the KS equation[4] in the presence of an external field, where the exchange-correlation potential $V_{xc}^{DFT}(\mathbf{r})$ is replaced by the self-energy $\Sigma(\mathbf{r},\mathbf{r}',\varepsilon_j^{QP})$, which acts as a nonlocal, energy-dependent potential. Therefore, the eigenvalue problem described above can be solved perturbatively considering the KS Hamiltonian as an unperturbed Hamiltonian and $\Sigma - V_{xc}$ as a perturbative term. The quasiparticle eigenvalues are obtained in first-order perturbation theory:

$$\varepsilon_j^{QP} = \varepsilon_j^{QM/MM} + \frac{\langle \phi_j^{QM/MM} | \Sigma(\varepsilon_j^{QM/MM}) - V_{xc} | \phi_j^{QM/MM} \rangle}{1 - \langle \phi_j^{QM/MM} | \frac{d\Sigma(\omega)}{d\omega}\Big|_{\omega=\varepsilon_j^{QM/MM}} | \phi_j^{QM/MM} \rangle} \quad (3)$$

All the Coulomb interactions, and hence also the one induced by the classical MM region, are included in the KS eigenvalues $\varepsilon_j^{QM/MM}$ and eigenvectors $|\phi_j^{QM/MM}\rangle$.

The quasiparticle energies obtained by the GW method and the QM/MM eigenfunctions $\phi_j^{QM/MM}(\mathbf{r})$ are then used to calculate the optical absorption spectrum through the solution of the BSE equation:

$$P(1,1',2,2') = P_{IQP}(1,1',2,2') + \int P_{IQP}(1,1',3,3')\,\Xi(3,3',4,4')\,P(4,4',2,2')\,d(3,3',4,4') \quad (4)$$

where $1 \equiv \mathbf{r}_1, t_1$ (and similarly for 1′, 2, 2′, etc.) and P_{IQP} is a generalized four-point irreducible polarizability and describes the propagation of independent quasielectron and quasihole couples (for a review, see ref 12). All the interactions are contained in the kernel Ξ, defined as

$$\Xi(1,1',2,2') = \delta(1,1')\,\delta(2,2')\,v(1,1') - \delta(1,1')\,\delta(1',2')\,W(1,1') \quad (5)$$

The kernel Ξ is made of two parts, v and W, resulting from the functional derivative of the Hartree potential and of the self-energy with respect to the single-particle Green's function, respectively. v is the electron−hole exchange, and W is the term responsible for bound excitons.

In practice, to solve the BSE, the problem is recast into an effective two-body Hamiltonian form:

$$H_{exc}^{(n_1,n_2),(n_3,n_4)} = (\varepsilon_{n_2}^{QP} - \varepsilon_{n_1}^{QP})\delta_{n_1,n_3}\delta_{n_2,n_4} + (f_{n_1} - f_{n_2}) \int d\mathbf{r}_1 \int d\mathbf{r}'_1 \int d\mathbf{r}_2 \int d\mathbf{r}'_2\, \phi_{n_1}^{QM/MM*}(\mathbf{r}_1)\,\phi_{n_2}^{QM/MM}(\mathbf{r}'_1) \times \Xi(\mathbf{r}_1,\mathbf{r}'_1,\mathbf{r}_2,\mathbf{r}'_2)\,\phi_{n_3}^{QM/MM}(\mathbf{r}_2)\,\phi_{n_4}^{QM/MM*}(\mathbf{r}'_2) \quad (6)$$

where f_n is the occupation number of level n. Hence, using the spectral representation[27] for the inverse of a matrix, the interacting polarization can be obtained by solving an effective eigenvalue problem:

$$\sum_{n_3,n_4} H_{exc}^{(n_1,n_2),(n_3,n_4)} A_\lambda^{(n_3,n_4)} = E_\lambda A_\lambda^{(n_1,n_2)} \quad (7)$$

The optical spectrum assumes then the following form:

$$\varepsilon_M(\omega) = 1 - \lim_{q\to 0} v(\mathbf{q}) \sum_{\lambda,\lambda'} \left[\sum_{n_1,n_2} \langle n_1 | e^{-i\mathbf{qr}} | n_2 \rangle \times \frac{A_\lambda^{(n_1,n_2)}}{\omega - E_\lambda + i\eta} N_{\lambda,\lambda'}^{-1} \times \sum_{n_3,n_4} \langle n_4 | e^{-i\mathbf{qr}} | n_3 \rangle A_{\lambda'}^{*(n_1,n_2)} (f_{n_3} - f_{n_4}) \right] \quad (8)$$

where

$$N_{\lambda,\lambda'}^{-1} = \sum_{n_1,n_2} A_\lambda^{*(n_1,n_2)} A_{\lambda'}^{(n_1,n_2)} \quad (9)$$

The BSE can be physically interpreted as adding the electron−hole interaction to the energy of the excited state of the system. An electron−hole pair is called an exciton and its contribution to the optical spectrum (obtained by solving the BSE) gives the so-called excitonic effects. Since the electron−hole interaction is attractive, it affects the DFT+GW optical spectrum mainly with a red shift of the peaks. Moreover, electron−hole bound states (the excitons) occur below the single-particle gap. The difference between the DFT+GW electronic gap and the excitonic optical gap measures the exciton binding energy. The quantum plus classical external potential is not explicitly present in the BSE but indirectly determines all its ingredients via the quasiparticle energies and wave functions.

To take into account solvation and temperature (300 K) effects, we performed a 20 ps hybrid QM/MM Car−Parrinello simulation of a system where the QM part was the indole molecule and the MM part was 1376 water molecules.[28]

For 10 snapshots of the QM/MM dynamics (one every 2 ps) we computed the optical spectra at the independent particle level (DFT-IPA) and within TDDFT.[29] The running average of the spectra indicates that, at the DFT level and for a dynamics of 20 ps, the convergency of the spectrum with the number of snapshots is achieved after six snapshots. We confirmed this statement by a comparison with a spectrum averaged over 120 snapshots. A better sampling could be obtained by performing several dynamics and repeating the test by using TDDFT and/or GW-BSE, but this goes beyond our computational possibilities. Hence, the subsequent GW and BSE calculations have been performed on only six snapshots (atomic coordinates are in the Supporting Information).

3. Results and Discussion

HOMO−LUMO Gaps. Our calculated DFT and GW HOMO−LUMO gaps[30] averaged over the QM/MM configurations, resulted to be 3.8 eV (with standard deviation $\sigma = \pm 0.1$ eV) and 7.2 eV ($\sigma = \pm 0.2$), respectively. The HOMO−LUMO gap calculated by the GW method corresponds to the electronic gap and not to the optical gap. The GW correction to the electronic gap, during the dynamics, ranges from 3.3 to 3.5 eV. Therefore, its value can in principle be calculated for just one snapshot, and used, with an error of 0.2 eV, for all the other frames. This fact, already found for liquid water,[17] confirms that one can strongly reduce the computational effort, by performing a GW calculation for just one snapshot. However, this error would be too large for our purpose of evaluating the solvent shift.

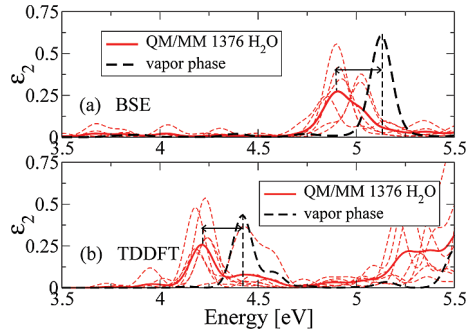


Figure 2. BSE (a) and TDDFT (b) spectra of indole in water. The tiny red dashed lines are the spectrum of each snapshot. The red solid line is obtained by an average over these spectra. The black dashed line is for indole in vapor phase.

As a consequence, in this work the GW corrections of the electronic gap have been calculated for each single snapshot.

Optical Spectrum. We next calculated the low-energy range of the optical spectrum of indole, by GW-BSE[31] and TDDFT, always as a result of an average over the QM/MM snapshots. In Figure 2 we report our results together with the calculated absorption spectrum in the gas phase. In our TDDFT calculations, the most intense peak (which corresponds to the ^1L_a transition located at 4.77 eV in the experimental spectrum)[19] is mainly due to an HOMO−LUMO transition, in agreement with previous calculations performed by CASSCF-CASPT2 methods.[18,32] The analysis of the charge distribution of the electronic levels reveals the $\pi-\pi^*$ character of this transition and an electronic charge depletion around the nitrogen atom on passing from the HOMO to the LUMO.

We notice that, in both TDDFT and GW-BSE approaches, this peak is red-shifted on passing from gas phase to the water solution. The value we calculate for such a red shift is ∼0.2 eV in both cases, in good agreement with the experiment (0.18 eV).[19] A similar trend was obtained by previous theoretical calculations of indole in water based on CASSCF-CASPT2 methods.[18] In these approaches, the indole configuration was the one obtained after a geometry optimization, with the solvent simulated by a continuum model, with a cavity containing the molecule. The CASSCF-CASPT2 prediction for the solvent shift is about 0.06 eV only. Such an underestimation may depend on the different geometrical conformation of indole molecule in the gas phase and in water, caused by the interaction with the solvent (which was not considered explicitly in).[18] In fact, in our calculations, comparison of the indole geometries in all the snapshots with that of the stable conformation in the gas phase (defined as the equilibrium geometry in vacuum) shows a slight loss of planarity (the dihedral angle between the planes of the two rings of indole is below 6°). Moreover, compared to the gas-phase configuration, the C−C bond lengths of indole in water can differ up to 4% and their average value is 1%−2% larger than the one calculated for the relaxed configuration. The angles differ by 1°−4° with respect to the gas-phase configuration. These geometrical changes are small but can significantly affect the optical properties, since they involve the breaking of some symmetries.

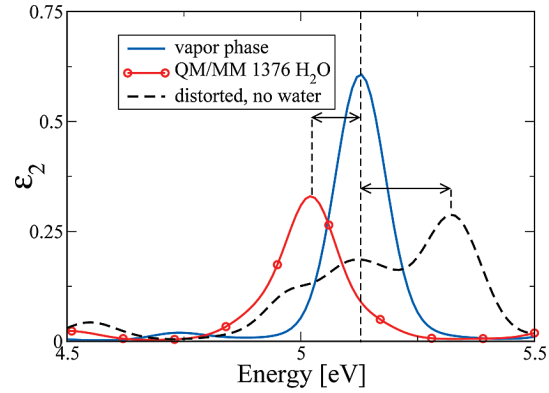


Figure 3. BSE optical spectra: solid blue line, indole in vapor phase; black dashed line, indole without water molecules, with atomic coordinates taken from a snapshot corresponding to 13.08 ps of the dynamics; circles, indole in water, with the spectrum calculated for the same snapshot.

To quantify the contribution of these effects on the solvent shift, we performed GW-BSE calculations of the optical absorption spectrum of indole switching on and off the water field, in order to separate the geometry effect from the electrostatic one. The results are presented for a single snapshot in Figure 3. The corresponding solvent shift goes from −0.1 eV with water field to +0.2 eV (hence, a blue shift) without water field. This emphasizes the importance of taking into account explicitly the electrostatic interaction with the solvent, since the geometry distortion alone would give, at least for this snapshot, a wrong sign.

In our calculations, TDDFT underestimates the energy of the ^1L_a peak[19] with respect to experiments, both in gas phase and in solution by ∼0.4 eV, and GW-BSE overestimates them by ∼0.3 eV. Concerning the gas phase, a result closer to the experiment is obtained by using a B3LYP functional[33] with a localized basis set; the underestimation in this case drops to 0.05 eV. We note that for this system the predictive power of TDDFT is comparable to MBPT. This was expected, since TDDFT is usually very efficient for small molecules such as indole (except in some particular cases).[34] Therefore, beside being a relevant biological molecule, indole is also a good system for the main purpose of this work, that is, the introduction and validation of a new scheme (GW-BSE/MM) by comparing the results with the experiments and with well-working methods on a relevant biological system.

As expected,[24] CASSCF is worse in predicting the energy position of the first peak of the absorption spectrum. It overestimates it by almost 1 eV or more, while CASPT2 is more accurate (∼0.13 eV for the gas phase). All the experimental and theoretical values obtained by different methods concerning the indole ^1L_a transition are summarized and compared in Table 1.

Role of the Solute−Solvent H-Bond. The indole N−H moiety is the only group that can form a hydrogen bond with the solvent molecules. In this section, we investigate qualitatively the role of the H-bond for the solvatochromism by performing TDFFT calculations of the optical absorption spectrum. These calculations are of course much cheaper than the GW calculations.

Table 1. Theoretical and Experimental Transition Energy (in eV) of the ^1L_a Peak of Indole Spectrum in Gas Phase and in Water and Their Difference

	MBPT/MM[a]	TDDFT/MM[a]	CASPT2[b]	EXP[c]
gas phase	5.1	4.4	4.73	4.77
in water	4.9	4.2	4.67	4.59
solvent shift	0.2	0.2	0.06	0.18

[a] This work. [b] Reference 19. [c] Reference 18.

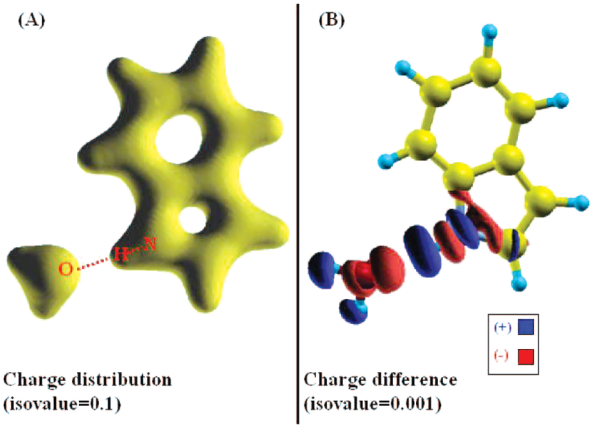


Figure 4. Indole and the water molecule involved in the H-bond. (A) The electronic charge distribution. (B) The difference in the charge distribution between the entire system and the two separated subsystems: isolated indole and isolated water molecule. This enables us to visualize the increasing of the polarization induced by the H-bond.

The geometry of the H-bond can be evinced by the calculation of distribution functions calculated from the QM/MM trajectory. The first maximum and minimum of the radial distribution function (rdf) between N and water oxygens O_w are 2.96 and 3.20 Å, respectively. The maximum of N−H···O_w angle, as obtained from the correspondent angular distribution function, is 155° (radial and angular distribution functions are available in the Supporting Information). This result agrees with force-field-based molecular dynamics simulation of indole in a large water box (∼10 000 solvent molecules) performed here[35] as well as with previous results reported in the literature.[36]

We then selected a representative QM/MM snapshot (Figure 4), in which the N−H···O_w distance is 2.91 Å, very close to that of the first rdf peak. In addition, in this snapshot the N−H···O_w angle is 155°; this value is similar to the maximum of the correspondent angular distribution function.

We performed on this snapshot TDDFT calculations of the optical spectrum. The calculations differed for the MM external potential $U^{QM/MM}$: $U^{QM/MM}$ is either generated by a water molecule (WAT hereafter) involved in the H-bond (i), or by all water molecules (ii), or it is zero (iii). By comparing the spectrum of i with that of iii, we conclude that the H-bond induces a small red shift (<0.1 eV) on the ^1L_a peak in this representative snapshot (Figure 5). By comparing the spectrum of ii with that of iii we observe a shift larger than 0.3 eV. On the basis of these calculations, we suggest therefore that the H-bond plays a role in the solvatochromism and that a considerable effect is given by the bulk solvent. The latter has to be included to obtain a quantitative evaluation of the solvent shift.

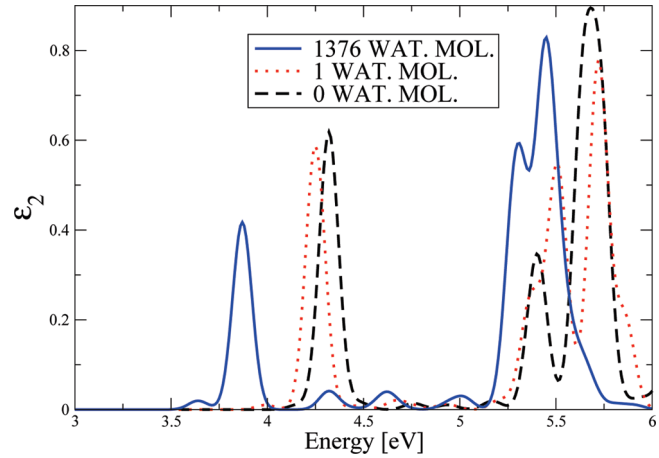


Figure 5. TDDFT absorption spectrum of indole in water. Spectra are calculated for a selected snapshot (see the text). Water molecules are treated classically.

Of course, solvent polarization and charge transfer effects associated with the H-bond might affect the solvatochromism. We addressed this issue by switching on the QM character of WAT in calculation ii. In other words, we performed a calculation in which WAT molecule was treated at the TDDFT level as indole, and all the other were treated classically. A Löwdin charge analysis[37] points to the absence of charge transfer effects between indole and WAT, while polarization effects are present (see Figure 4). The change in absorption peak ^1L_a relative to calculation ii is very small (less than 0.02 eV). The same results were obtained by including also a second water molecule. Thus, polarization effects, although sizable, do not affect largely the absorption spectrum. Our conclusions agree with TDDFT/MM calculations performed for another organic molecule in water solution: acetone.[38] Like indole, this molecule forms H-bonds with the solvent. The calculated absorption spectrum is almost the same when the solvation sphere of 12 water molecules surrounding acetone are treated by a QM or MM approach. The approximation of treating the solvent at a classical level allows us to avoid small box size effects and sample much longer time scales than using a fully ab initio scheme. This is absolutely crucial already for this small system. It is expected to be even more relevant for large biological systems.

4. Summary

In this paper, we have included many-body perturbation theory in a QM/MM scheme. We have applied it, together with a TDDFT/MM approach, to study the optical properties of indole in water solution. Both methods reproduce quantitatively the red shift induced by the solvent. The GW-BSE/MM method can be applied to biomolecules in aqueous solution (i.e., in laboratory-realizable conditions), although with a larger computational cost (for this particular case, MBPT requires 8 times more CPU time than TDDFT to calculate the indole optical spectrum).

Many works put in evidence the GW-BSE method to be a possible alternative to TDDFT to treat large size materials

or charge transfer systems.[39] We expect the GW-BSE, whose range of applications has been here extended to molecules in solution and in different chemical environments, to be able to study long-range charge-transfer molecules in their biochemical environment.[40,41] Moreover, a better exchange and correlation kernel can be derived from MBPT to improve TDLDA/GGA.[15]

Our calculations show that the solvent shift is a consequence of the combination of two effects: the geometrical distortion of indole molecule in the solvent and the electrostatic interaction with the water molecules' electric dipoles. Both effects, and their sum, depend on the particular configuration of the system; this emphasizes the need of including *both* altogether and of averaging over several snapshots.

This work opens the way to further applications of MBPT/MM to other biorelevant molecules, such as fluorescent probes in their target proteins, for which the evaluation of the optical shift enables the understanding of the nature of their environment.

Acknowledgment. We acknowledge support from EU e-13 ETSF Project 211956. Computer resources from INFM "Progetto Calcolo Parallelo" at CINECA are gratefully acknowledged. We also thank L. Guidoni for interesting discussions.

Supporting Information Available: PDB file including most of the snapshots of the QM/MM trajectory used for the calculations; radial distribution function calculated from the QM/MM trajectory between the nitrogen of the indole and water oxygens; angular distribution function for the angle formed by the nitrogen of the indole, its hydrogen, and the oxygens in the corresponding first shell of water. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Carter, E. A.; Hynes, J. T. *J. Chem. Phys.* **1991**, *94*, 5961–1129.

(2) Laio, A.; VandeVondele, J.; Röthlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941.

(3) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.

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

(5) Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997.

(6) Marques, M. A. L.; Gross, E. K. U. Time-Dependent Density Functional Theory. In *A Primer in Density Functional Theory*; Fiolhais, C., Nogueira, F., Marques, M. A. L., Eds.; Springer-Verlag: Berlin, 2003; Vol. 620, pp 144−184.

(7) Sulpizi, M.; Carloni, P.; Hutter, J.; Röthlisberger, U. *Phys. Chem. Chem. Phys.* **2003**, *5*, 4798.

(8) Sulpizi, M.; Röhrig, U. F.; Hutter, J.; Röthlisberger, U. *Int. J. Quantum Chem.* **2004**, *101*, 671.

(9) Moret, M.-E.; Tapavicza, E.; Guidoni, L.; Röhrig, U. F.; Sulpizi, M.; Tavernelli, I.; Röthlisberger, U. *CHIMIA Int. J. Chem.* **2005**, *59*, 493–498.

(10) Frutos, L. M.; Andruniów, T.; Santoro, F.; Ferré, N.; Olivucci, M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 7764.

(11) Fetter, A.; Walecka, J. *Quantum Theory of Many-Particle Systems*; McGraw-Hill: San Francisco, 1971; pp 1−197.

(12) Onida, G.; Reining, L.; Rubio, A. *Rev. Mod. Phys.* **2002**, *74*, 601.

(13) TDDFT, as implemented in the CPMD code[42] used in this work, uses the Tamm−Dancoff approximation.

(14) Haydock, R. *Comput. Phys. Commun.* **1980**, *20*, 11.

(15) Botti, S.; Shindlmayr, A.; Del Sole, R.; Reining, L. *Rep. Prog. Phys.* **2007**, *70*, 357, and references therein.

(16) Creed, D. *Photochem. Photobiol.* **1984**, *39*, 537.

(17) Garbuio, V.; Cascella, M.; Reining, L.; Del Sole, R.; Pulci, O. *Phys. Rev. Lett.* **2006**, *97*, 137402, and references therein.

(18) Serrano-Andres, L.; Roos, B. O. *J. Am. Chem. Soc.* **1996**, *118*, 185, and references therein.

(19) Lami, H. *J. Chem. Phys.* **1977**, *67*, 3274.

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

(21) Peraro, M. D.; Ruggerone, P.; Raugei, S.; Gervasio, F. L.; Carloni, P. *Curr. Opin. Struct. Biol.* **2007**, *17*, 149.

(22) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

(23) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. *Comput. Phys. Commun.* **1995**, *91*, 1.

(24) Rogers, D. M.; Hirst, J. D. *J. Phys. Chem.* **2003**, *107*, 11191.

(25) Hedin, L. *Phys. Rev.* **1965**, *139*, A796.

(26) Aryasetiawan, F.; Gunnarsson, O. *Rep. Prog. Phys.* **1998**, *61*, 237–312.

(27) Albrecht, S.; Reining, L.; Del Sole, R.; Onida, G. *Phys. Rev. Lett.* **1998**, *80*, 4510.

(28) The QM/MM code[2] combines the CPMD3.11.1 and the P3M version of the GROMOS code.[43] This is a dual simulation box QM/MM method: QM solute and MM solvent are defined on periodically replicated simulation cells of different sizes. The QM cell size is $11.8 \times 10.8 \times 12.5$ Å^3 and it is immersed in a box of $33.7 \times 35.1 \times 35.6$ Å^3 containing 1376 TIP3P water molecules. Electrostatic interaction between periodic images of the QM part is decoupled by the scheme of Martyna−Tuckerman.[44] The electrostatic interaction of the classical part with periodic boundary conditions is treated by the particle−particle, particle−mesh method: we checked that the size of the box was large enough to prevent the solute from interacting with its images, estimating such interaction energy in the presence of the water and comparing it with the thermal energy fluctuation. We used BLYP pseudopotentials corrected for a better description of van der Waals interactions,[45] 70 Ry energy cutoff (indole's N−H group forms H-bonds with water), and 0.1 fs time step for the dynamics. A Nose−Hoover thermostat is applied throughout all simulations to keep the temperature constant.

(29) All TDDFT calculations are obtained within the Tamm−Dancoff approximation as implemented in the CPMD 3.11.1 package. For the exchange-correlation kernel, we tested different functionals (LDA, BLYP, BP, PBE), the choice being limited by the use of a plane-wave basis set. On the other hand, plane-wave-based calculations have some advantages when performing molecular dynamics (summarized in ref 46), particularly the ability to efficiently calculate the forces on atoms that are most important in our QM/MM scheme. The same corrected Troullier−Martins pseudopotentials and elec-

trostatic treatment of the classical part as for the dynamics have been used.

(30) GW calculations have been done by using 12 077 plane-waves and 500 electronic bands. Following ref 47, we used periodic boundary conditions and a cutoff in real space for the Coulomb potential to prevent that periodic images interact with each other. The screening function is calculated within a plasmon pole approximation. For MBPT calculations, we used codes developed within the ETSF (www.etsf.eu/resources/software/codes).

(31) The BSE was solved by diagonalizing the excitonic Hamiltonian with the inclusion of the coupling part. We used 12 077 plane-waves for the wave functions, 257 for the local fields, and 5100 quasiparticle transitions. We use periodic boundary conditions with a cutoff in real space for the Coulomb potential.[47]

(32) Borin, A. C.; Serrano-Andres, L. *Chem. Phys.* **2000**, *262*, 253.

(33) Rogers, D. M.; Besley, N. A.; O'Shea, P.; Hirst, J. D. *J. Phys. Chem. B* **2005**, *109* (48), 23061.

(34) Sun, M.; Ding, Y.; Cui, G.; Liu, Y. *J. Phys. Chem. A* **2007**, *111*, 2946.

(35) The indole molecule with a geometry taken after a geometry optimization by using CPMD with 70 Ryd energy cutoff was inserted in a box of 9575 water molecules of around 70 Å edges. The AMBER force field and TIP3P force fields were used for indole and water, respectively. A time step of 1.5 fs was used. The SHAKE algorithm was used to fix the lengths of covalent bonds that involve hydrogens. We used a Langevin dynamics with a collision frequency of 5 ps^{-1} and the isotropic position scaling algorithm to regulate pressure as implemented in the Amber package to sample in a NPT ensemble. Electrostatic interactions were evaluated using the particle mesh Ewald method, with a cutoff for the real part of 12 Å. The same value was used for the cutoff of the van der Waals interactions.

(36) Simonson, T.; Wong, C. F.; Brünger, A. T. *J. Phys. Chem. A* **1997**, *101* (10), 1935.

(37) Löwdin, P.-O. *J. Chem. Phys.* **1950**, *18*, 365.

(38) Röhrig, U. F.; Frank, I.; Hutter, J.; Laio, A.; VandeVondele, J.; Röthlisberger, U. *ChemPhysChem* **2003**, *4*, 1177.

(39) Reining, L.; Olevano, V.; Rubio, A.; Onida, G. *Phys. Rev. Lett.* **2002**, *88*, 066404.

(40) Cai, Z.-L.; Sendt, K.; Reimers, J. R. *J. Chem. Phys.* **2002**, *117*, 5543.

(41) Dreuw, A.; Head-Gordon, M. *J. Am. Chem. Soc.* **2004**, *126*, 4007.

(42) Hutter, J.; Alavi, A.; Deutsch, T.; Ballone, P.; Bernasconi, M.; Focher, P.; Goedecker, S.; Tuckerman, M.; Parrinello, M. *CPMD 3.11.1*; Copyright IBM Corp, 1990−2006; Copyright MPI für Festkörperforschung, Stuttgart, 1997−2001, 2006.

(43) Luty, B. A.; Davis, M. E.; Tironi, I. G.; van Gunsteren, W. F. *Mol. Simul.* **1994**, *14*, 11.

(44) Martyna, G. J.; Tuckerman, M. E. *J. Chem. Phys.* **1999**, *110*, 2810.

(45) von Lilienfeld, O. A.; Tavernelli, I.; Röthlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 153004.

(46) Marx, D.; Hutter, J. Ab initio molecular dynamics: Theory and Implementation. In *Modern Methods and Algorithms of Quantum Chemistry*; Grotendorst, J., Ed.; John von Neumann Institute for Computing: Juelich, 2000; Vol. 1, pp 301−449.

(47) Onida, G.; Reining, L.; Godby, R. W.; Del Sole, R.; Andreoni, W. *Phys. Rev. Lett.* **1995**, *75*, 818–821.

CT800528E

JCTC Journal of Chemical Theory and Computation

A Simple One-Body Approach to the Calculation of the First Electronic Absorption Band of Water

Ricardo A. Mata,*,† Hermann Stoll,‡ and B. J. Costa Cabral†,‡

Grupo de Física-Matemática da Universidade de Lisboa, Avenue Professor Gama Pinto 2, 1649-003 Lisboa, Portugal, Institut für Theoretische Chemie, Universität Stuttgart, Pfaffenwaldring 55, D-70569 Stuttgart, Germany, and Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade de Lisboa, 1749-016, Lisboa, Portugal

Received January 27, 2009

Abstract: A one-body decomposition approach for investigating the electronic absorption spectra of molecular systems was proposed and applied to water clusters $(H_2O)_N$ including up to $N =$ 80 water molecules. Two specific aspects of the present implementation are the inclusion of the coupling between excited states and a simplified representation for the *N*-body Coulombic effects. For smaller clusters, the results based on the one-body decomposition scheme are in good agreement with full EOM-CCSD calculations. Two different regimes can be identified in the electronic absorption profile of larger water clusters. The first low-energy regime is dominated by local excitonic states on the cluster surface, whereas the higher-energy excitations associated with the second one are of delocalized nature.

1. Introduction

Electronic properties of water are of fundamental importance for understanding chemical reaction mechanisms and kinetics in solution. However, they are not very well understood.[1,2] Specifically, the relationship between the electronic and topological properties of the hydrogen-bond network[3,4] and the influence of ionic solvation on the electronic properties of bulk or interfacial water[2] deserve further investigation. Therefore, several recent works on the electronic properties of water were reported.[3,5−8] Emphasis was placed on polarization effects in liquid phase,[5] electron binding energies,[5] dynamic polarizability,[8] and electronic absorption spectrum of water.[3,6−8] On the other hand, some relevant studies were also dedicated to water clusters with emphasis on cooperativity effects[9−11] and electronic properties.[12,13] The interest in clusters is motivated by the well-established fact that many of the cooperative effects characterizing the complex hydrogen-bond network of liquid water can also play a major role in determining the structure and electronic properties of clusters. Moreover, there is also theoretical evidence that even for relatively small clusters, the number of hydrogen bonds for the water molecules in the interior of the cluster is quite similar to what is found in condensed phases of water, whereas a more labile network can be observed for water molecules closer to the surface.[14] Consequently, it should be expected that the investigation of the electronic properties of water clusters can also provide relevant information on the relationship between the changing hydrogen-bond network associated with different molecular environments and the electronic properties of water.

In this work we will focus on the prediction of the first electronic absorption band of water clusters. Our main purpose is to adopt a simple, general, and accurate theoretical procedure for calculating the electronic absorption spectra of water clusters. In addition, we have investigated the dependence of the calculated spectra on the cluster size as well as the role played by water molecules at different regions of the aggregates in the absorption process. The present theoretical procedure is based on the many-body decomposition for the energy of a molecular aggregate.[15−29] In agreement with previous many-body decomposition schemes relying on the multilayer fragment molecular orbital

* Corresponding author e-mail: rmata@cii.fc.ul.pt.
† Grupo de Fisica-Matematica da Universidade de Lisboa.
‡ Universität Stuttgart.
‡ Departamento de Química e Bioquímica, Faculdade de Ciencias, Universidade de Lisboa.

10.1021/ct9001653 CCC: $40.75
Published on Web 05/29/2009

method[18,19,27] we are providing further evidence that reliable estimates of the excitation energies of water clusters can be carried out by using a one-body expansion of the cluster energy. Some specific aspects of the present implementation for calculating the electronic spectra of water clusters are the explicit consideration of the coupling between excited states and inclusion of different simplified representations for the *N*-body Coulombic effects. This approach is applied for the calculation of the electronic spectra for optimized structures of water clusters including up to 80 water molecules.

2. Theoretical Methods and Computational Details

2.1. One-Body Treatment of Excitations. Due to the high-order scaling of computational costs in ab initio methods, their use is prohibitive for more than a few water molecules. An alternative approach for the study of larger water clusters is to use a many-body expansion, which potentially reduces the scaling of the method to $\mathcal{O}(N^X)$, where X is the order to which the expansion is truncated. Recently, Chiba et al.[19] have proposed an extension of their FMO method for use in TD-DFT calculations. Here we present a many-body approach for investigating the electronic excitations of large water clusters with an explicit treatment of the coupling between excited states.

We follow the model of Harvey et al.[12] to treat the excitonic states. A matrix Hamiltonian is constructed in the representation of the basis of the uncoupled water molecules. Each of the basis functions represents the system where a single water molecule is excited

$$\Psi_i = \Phi_i^*(\mathbf{r}_i, \mathbf{R}_i) \prod_{j \neq i} \Phi_j(\mathbf{r}_j, \mathbf{R}_j) \tag{1}$$

where i stands for the index of the excited water molecule, Φ_i^* is the excited state of that molecule, Φ_j is the ground state of a neighboring water, $\mathbf{r}_i$ are the electronic coordinates, and $\mathbf{R}_i$ are the nuclear coordinates. Molecules in the excited state will be marked with an asterisk.

In contrast with the model of Harvey et al.,[12] which was based on a semiempirical model for both ground and excited electronic states, we will apply a one-body approximation to the calculation of the ab initio energy of each state. The diagonal elements of the Hamiltonian matrix H are given by

$$H_{ii} = E^{1B}(i^*) = E(i^*) + \sum_{j \neq i}^{N-1} E(j) \tag{2}$$

where $E(...)$ stands for the energy of the excited (or ground state) molecule at a given level of theory. In this work we will use the EOM-CCSD method. The terms $E(i^*)$ are taken from a EOM-CCSD calculation on the excited state of monomer i and the CCSD energies for ground-state energies $E(j)$.

A common approach to increase the convergence of the many-body expansion is to perform each monomer calculation in a point charge field representing the environment molecules. This leads to the approximate inclusion of *N*-body Coulombic effects. The diagonal terms in this case are given by

$$H_{ii} = \bar{E}^{1B}(i^*) = \bar{E}(i^*) + \sum_{j \neq i}^{N-1} \bar{E}(j) - C(i^*) \tag{3}$$

The $\bar{E}(i)$ terms are defined just as before, except that the calculations are performed with an operator added to the one-electron Hamiltonian

$$h_i^{PC} = \sum_{j \neq i} \sum_{\alpha \in j} \frac{q_\alpha}{r_{1\alpha}} \tag{4}$$

where j runs over surrounding monomers, α runs over all atoms in the monomer unit j (in case atom-centered point charges are used), q_α are the charges, and $r_{1\alpha}$ is the distance between an electron and a point charge. The $C(i^*)$ term in eq 3 is a correction energy added to avoid double counting. Since the energy sum in eq 3 runs over all monomers, and each single term already contains the interaction between the monomer and the other units, the interactions are double counted. This correction term will be later discussed.

The off-diagonal elements H_{ij} of the Hamiltonian, which give the coupling between two excited states, are approximated by the interaction of the transition dipoles $\mathbf{d}_{01}$ of the two excitations

$$H_{ij} = \frac{1}{R_{ij}^3}[\mathbf{d}_{01}^i \cdot \mathbf{d}_{01}^j - 3(\mathbf{d}_{01}^i \cdot \mathbf{R}_{ij})(\mathbf{d}_{01}^j \cdot \mathbf{R}_{ij})] \tag{5}$$

where $\mathbf{R}_{ij}$ is the distance vector between the two molecules i and j. The same approach was used by Harvey et al.,[12] where the amplitude of the transition dipole vectors was given by an analytical expression. In our case, the dipole moments are taken from the excited-state calculations performed on each monomer in eq 3 and placed at the center of mass. In the EOM-CCSD case, we use the geometric average of the right and left transition dipole moments

$$\mathbf{d}_{01}^i = \frac{\langle \Phi_i | \hat{\mathbf{d}} | \Phi_i^* \rangle + \langle \Phi_i^* | \hat{\mathbf{d}} | \Phi_i \rangle}{2} \tag{6}$$

where $\hat{\mathbf{d}}$ is the local dipole operator at center i.

Diagonalization of the Hamiltonian matrix gives N energies, one for each state. In order to obtain the excitation energies, one subtracts the ground-state result, which is also obtained under a one-body treatment

$$\bar{E}^{1B}(0) = \sum_i^N \bar{E}(i) - C(0) \tag{7}$$

where $C(0)$ is again a term to avoid double counting of particle interactions due to the point charge field.

Up to this point we have not discussed the point charge field to be used or the form of the correction terms, which are thereof dependent. In our calculations, the charges q_α in eq 4 are taken from the TIP3P model. This may seem at first a crude approximation, since one neglects polarization effects (the point charge field is fixed) and the fact that the density distribution of an excited water molecule is different

from its ground state. The reasons behind this choice are manifold. First, in the TIP3P model, which is an effective pair-potential for liquid water, the charges already include, at least partially, polarization effects in average way. Previous studies have also shown that these charges give reasonable results in expanding the total energy under a many-body approximation.[15,20] Including the full Hartree-Fock potential of neighboring molecules or simply replacing them by these atom-centered point charges has been shown to provide similar accuracy. We expect that this effect is also true when applied in excited-state calculations. Second, by using the same point charge field in all calculations, the correction term for all excited states is the same and equal to the one in the ground state. Under this approximation, including the correction in eq 3 amounts to adding a constant value to the diagonal terms and, therefore, has no effect on the eigenvectors of the Hamiltonian. It also cancels out when computing the difference between the ground and the excited state. In short, if one is only interested in excitation energies, this term can be neglected. It is unclear which approach leads to a higher accuracy, whether neglecting the correction term or including the excitation effect on the charge distribution. The first approximation, however, has the advantage of a nearly linear scaling computational cost. We consider it almost linear, since for larger structures the addition of point charges into the Hamiltonian will become a bottleneck (but only for extremely large structures). It also has a very small prefactor. If the changes in the charge distribution are considered, an extra calculation has to be performed for each monomer, which significantly affects the performance. However, as strongly indicated by some test calculations (see section 3.1), the first approximation already gives highly accurate results, with errors below the ones estimated for the level of theory used as a reference (EOM-CCSD/aug-cc-pVTZ). Also, as previous works on the absorption spectra of water show, the first absorption maximum of liquid water can be well reproduced by treating a single water quantum mechanically.[8] This is due to the localized character of the first excitation.

At this stage, we would like to point out the similarities and differences between our method and other related work. The diagonal terms H_{ii} are calculated as in the FMO method,[18,19,27,28] with the exception that the embedding is simplified, including only fixed point charges to mimic electrostatic effects. In the original method, Coulomb operators taken from converged monomer densities are used. In the work of Hirata et al.,[26] the embedding was based on self-consistent dipole moments. We find that the former approach significantly increases the computational effort. In the case of water, it has been seen to give only marginal improvements for ground-state energies.[15,20] The use of point dipoles per molecule, on the other hand, may be unsuitable to describe the electrostatics of hydrogen-bonded systems. Our main goal, however, is not to review the way the electrostatic embedding is performed but, instead, to expand the applicability of FMO-based formulations to cases where identical chromophores are present. A many-body expansion by itself cannot describe excitonic coupling effects.

The elongation method of Aoki and co-workers[23−25] does account for excitonic coupling. However, their formulation does not include the electrostatic environment in the intrafragment excitation and is also not easily extendable to an arbitrary correlation method. It is based on a CI-type formalism. Also, the coupling elements have to be limited to neighboring fragments due to the computational costs involved in their calculation. However, in a system with identical chromophores, excitonic coupling can lead to mixing of states throughout the system, regardless of its size. The symmetry-adapted-cluster/symmetry-adapted-cluster configuration interaction theory[29] is the most complete formulation, including long-range electrostatic embedding as well as state coupling. However, it is questionable whether it can be computationally efficient in the case of irregular clusters.

All excited-state calculations were performed with the equation-of-motion coupled cluster singles and doubles (EOM-CCSD) method.[30] The basis set chosen was the augmented correlation consistent triple-ζ valence basis set (aug-cc-pVTZ) of Kendall et al.[31] The one-body calculations were performed with in-house Python programs interfaced to the Molpro program package.[32]

2.2. Cluster Structures. One of the main purposes of the present work is to analyze the accuracy of the one-body energy decomposition scheme for predicting the absorption spectra of large water clusters. Since the optimization of very large clusters with ab initio methods is not affordable, we have performed calculations with the polarizable molecular mechanics potential AMOEBA.[33] The choice of this potential was driven by previous works on the properties of water clusters and liquid water that supports the reliability of the model. It should be observed that the force field has been mainly parametrized for reproducing the structure, energetics, and/or vibrational properties of the water monomer, dimer, and small water clusters. The successful applications of the AMOEBA potential to describe both the gas and condensed phases are basically related to the explicit inclusion of polarization effects, which favors transferability to different environments and thermodynamic conditions.

In order to discuss the accuracy of the AMOEBA potential for generating the structure of water clusters, comparison was made with ab initio structures. The use of this force field for the optimization of water clusters has been already tested for a few conformations of clusters ranging up to the hexamer. Further information can be found in ref 33. In this section we reevaluate the potential, taking special care to the evolution of its performance with increasing cluster size.

For $(H_2O)_N$ (with $N = 4-10$), three structures were chosen and optimized at the local second-order Møller−Plesset perturbation (LMP2) level of theory.[34] The basis set used was the Dunning cc-pVTZ basis set.[35] Previous calculations showed that the use of diffuse functions has little impact on the optimized structures. Up to the octamer, the starting cluster geometries were taken from ref 20. The clusters with 9 and 10 water molecules were optimized from starting structures taken from Monte Carlo simulations.[36] The LMP2 structures were then reoptimized with the AMOEBA potential, without any constraints, to an rms gradient per atom of

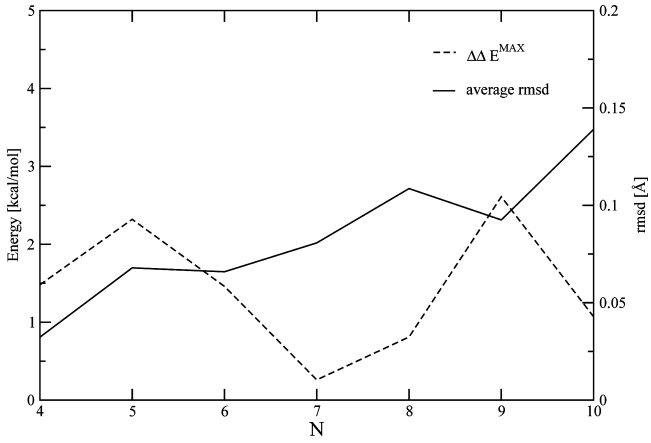


Figure 1. Largest deviations in energy differences of optimized cluster geometries with LMP2/cc-pVTZ and the AMOEBA potential ($\Delta\Delta E^{MAX}$ in kcal/mol) and the average root-mean-square distance (average rmsd in angstroms). Three conformers have been used for each point.

0.0001 kcal/mol/Å. All molecular mechanics calculations were carried out with the TINKER program.[37]

When the AMOEBA and LMP2/cc-pVTZ results are compared, the energetic ordering between different conformers is kept in all but two cases. The maximum deviation in the relative energies as a function of the cluster size is plotted in Figure 1. The results were also compared by superposing the optimized structures, based on mass-weighted coordinates, and computing the rms distance between all atoms. The average values are plotted in the same figure. Both analyses confirm the good performance of the potential in the optimization of small-sized water clusters. The rms distance obtained from superposing the two sets of structures is particularly small. The energy difference is somewhat harder to analyze since there are large fluctuations, but seem to point to a maximum error around 2−3 kcal/mol. All of these results seem to validate the use of the force field in sampling and optimization of large water cluster structures. In all of the following sections, and for consistency, only AMOEBA-optimized structures will be used.

3. Results and Discussion

3.1. Water Dimer. As a first test, calculations were performed on the water dimer, treating each monomer sequentially while representing the other molecule with a different embedding scheme. We neglect the coupling between the excited states since the transition dipole moments are close to perpendicular in the dimer orientation, and it would therefore be close to zero. The embedding schemes used were the same as in a previous work.[20] We compare results for calculations with no embedding potential ("no embedding"), with symmetrical orthogonalization of the monomer orbitals (with level shifting as in eq 7 in ref 20) and a Coulomb potential representing the other monomer ("J(0)"), with the full Hartree−Fock potential ("HF(0)") or the same potential iterated ("HF(1)"), and the calculation with TIP3P charges. The results are shown in Table 1.

The set of results which are in closest agreement to the full calculation are obtained with a simple TIP3P point charge

Table 1. Excitation Energies (in eV) for the AMOEBA-Optimized Water Dimer, Using Different Embedding Schemes

	monomer 1	monomer 2
no embedding	7.587	7.588
J(0)	7.817	8.079
HF(0)	7.734	8.100
HF(1)	7.669	8.152
TIP3P	7.613	8.015
full	7.571	8.018

embedding. This might seem at first surprising, but it should be noted that introduction of the accurate Coulomb or full Hartree−Fock potentials necessitates the use of level-shift operators which essentially freeze the surroundings of a given monomer.[20] The frozen surroundings, in turn, will lead to an upward shift of both ground and excited states, but more so for the excited state since it is less localized (cf. below). The TIP3P results seem to agree well with the full calculation by neglecting the Pauli repulsion of the surroundings and a rather fortunate error cancellation between the lack of exchange contributions and an approximate description of Coulomb effects. The only way to improve on the results with a Fock potential embedding would be to use different frozen surroundings for the ground and the excited state.

3.2. Water Tetramers and Pentamers. In order to validate our approach to the computation of the first absorption band of water, we tested our method in some smaller clusters, where it is still possible to apply the full quantum treatment. We first look at water tetramers, using AMOEBA structures taken from section 2.2.

In Table 2 we report the first four excitation energies, computed in a full EOM-CCSD calculation or with the use of a one-body approximation. In the latter, we considered two cases. In the first set we considered the off-diagonal elements of the excitonic states Hamiltonian to be zero. This amounts to neglecting the coupling between the excited states and simply taking the excitation energies from the one-body calculation. In the second set, we introduced the off-diagonal elements and obtained the energies by diagonalization of the matrix.

Both sets of results from the one-body approach replicate well the EOM-CCSD estimates. However, there is a relevant difference between the two. Neglecting the excited-states coupling, many are degenerate due to the structural symmetry of the cluster. We would like to point out that this error would not be corrected by including higher-order body terms in estimating H_{ii} (this is further discussed in the Conclusions section). It is an effect due to the decoupling of excitations. By the use of the nondiagonal matrix elements, we are able to correctly reproduce the mixing of states which leads to a lift of the degeneracy. The results with diagonalization of the Hamiltonian are on the average closer to the full EOM-CCSD values (the error is halved in cases where the simple one-body approach has a significant deviation). The largest deviation is around 0.07 eV, well below the error estimate of the full method.

In order to confirm that the diagonalization procedure does correctly mix the one-body excitonic states (as defined in eq 1), we have examined the eigenvectors of the solutions

Table 2. First Excitation Energies (in eV) for Three Selected Water Tetramer Clusters

ring 1			ring 2			cage		
full	1B ($H_{ij} = 0$)	1B ($H_{ij} \neq 0$)	full	1B ($H_{ij} = 0$)	1B ($H_{ij} \neq 0$)	full	1B ($H_{ij} = 0$)	1B ($H_{ij} \neq 0$)
8.094	8.103	8.081	7.937	8.040	7.990	7.908	7.933	7.908
8.094	8.103	8.081	7.942	8.040	8.019	7.939	7.933	7.947
8.102	8.103	8.114	8.108	8.057	8.083	8.203	8.264	8.270
8.117	8.103	8.137	8.130	8.057	8.103	8.381	8.353	8.358

Table 3. Relative Weights of Each Monomer for a Given Excitation (Given in Percent)[a]

		relative weight	
exc. energy	molecule	EOM-CCSD	1B ($H_{ij} \neq 0$)
7.908	1	0.0	0.0
	2	9.8	1.6
	3	45.1	49.2
	4	45.1	49.2
7.939	1	6.8	1.4
	2	0.0	0.0
	3	46.6	49.3
	4	46.6	49.3
8.203	1	0.0	0.0
	2	93.0	98.4
	3	3.5	0.8
	4	3.5	0.8
8.381	1	92.8	98.6
	2	2.5	0.0
	3	2.3	0.7
	4	2.3	0.7

[a] The excitation energy (in eV) is taken from the full EOM-CCSD calculation.

for the "cage" tetramer. These allow us to identify the molecules which mostly contribute to a given excited state. We compare these findings to a EOM-CCSD calculation using Pipek−Mezey localized orbitals[38] in the occupied space. By considering the largest coefficients (above 0.1), we calculated their weight in the excitation, labeling each coefficient according to the occupied orbital from which the excitation is made. Results are shown in Table 3.

Table 3 shows a strong agreement between the weights computed with localized orbitals and the ones derived from the transition dipole moments. The first two states are a combination of the individual excitations of molecules 3 and 4 (which in Table 2 are shown to give degenerate one-body excited states 7.933 eV above the ground state). The neighbors only contribute marginally to these states. The other higher-energy states are almost pure one-body states. The differences in the relative weights are relatively small; the largest difference is found in the two first excitations, where our analysis somewhat overestimates the contribution of molecules 3 and 4. Also, although the transition dipole interaction gives no contribution from molecule 2 to the highest-energy excitation, the EOM-CCSD coefficients give a balanced weight for the "spectrator" molecules. But even with these small discrepancies, there is a general agreement. The results for the other tetramers were similar and encourage us to make further use of this analysis to study the local (or delocalized) character of the excitations in larger clusters. Also, if the states are local (with a sparse eigenvector), we may be able to identify the region where the excitation takes place.

As a further test to our approach, we also computed the spectra for the water pentamers. The structures, optimized with the AMOEBA force field as detailed in section 2.2, are shown in Figure 2. One structure of particular interest is the one labeled "4coord", where a central water molecule is hydrogen-bonded to four neighbors (two times as acceptor and as a donor). This is the conformer which most closely resembles the liquid structure of water and therefore will become of greater relevance as one increases the cluster size.

The results for the three conformers of the water pentamer are shown in Table 4. Overall, the results are quite similar to the ones obtained for the tetramers. Except for the highest excitation energy of the "tetra + 1" conformer, which shows a 0.14 eV deviation from the full result, all other values have errors well below 0.10 eV (the average absolute deviation is 0.03 eV). Another fault should, however, be noted. Contrary to the tetramers, the energetic ordering of the excitations (comparing full results to the one-body approximation) is not preserved. In Table 4, we chose to order the excitations according to the weight analysis, and not by the energies. In some cases, where the excitations are near-lying, the ordering had to be inverted. This happens for the third and fourth excitations of the "ring" conformer and for the second and third excitations of "tetra + 1" and "4coord" conformers. Although undesirable, such a problem should be expected due to the small differences between the energies of these conformers and also by the simplified treatment we opted to use, by limiting the results to one-body terms. Inspecting the relative weights of each molecule in the excitations, one concludes that the estimates agree better in the case of localized excitations. In the delocalized cases, our results tend to underestimate delocalization effects. This is particularly so when three molecules have a significant weight on the state. The one-body approximation, although it may identify the most significant contributions, deviates

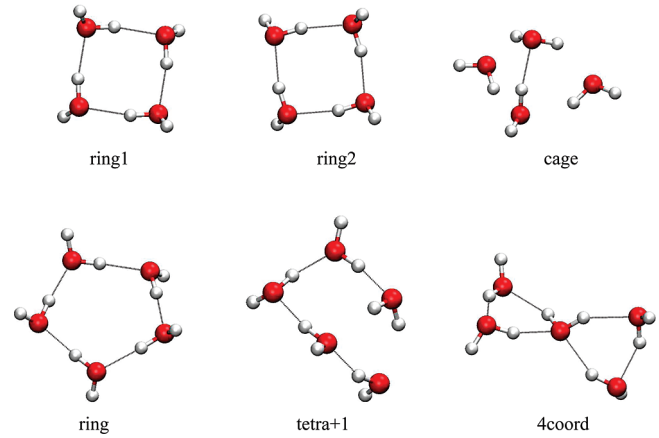


Figure 2. Water tetramer and pentamer structures.

Table 4. Excitation Energies (in eV) and Relative Weights of Each Monomer (Given in Percent), Calculated at the EOM-CCSD/aug-cc-pVTZ Level and Using a 1B Approximation with Coupling Terms

		ring		tetra + 1		4coord	
	molecule	EOM-CCSD	1B ($H_{ij} \neq 0$)	EOM-CCSD	1B ($H_{ij} \neq 0$)	EOM-CCSD	1B ($H_{ij} \neq 0$)
exc. energy(1)		7.915	7.950	7.870	7.925	7.990	8.044
	1	60.5	40.7	0.0	0.4	78.6	48.4
	2	0.0	0.7	0.0	0.4	0.0	1.0
relative weight	3	0.0	0.0	0.0	0.1	3.8	1.1
	4	0.0	0.3	1.5	1.8	0.0	13.8
	5	39.5	58.3	98.5	97.3	12.8	35.7
exc. energy(2)		8.078	8.030	8.089	8.130	8.032	8.065
	1	10.8	40.1	0.0	3.2	2.9	9.7
	2	18.0	1.2	22.7	18.6	1.5	7.4
relative weight	3	51.5	30.4	77.3	77.7	0.0	0.2
	4	0.0	0.1	0.0	0.1	27.7	28.8
	5	19.7	28.2	0.0	0.4	67.9	53.9
exc. energy(3)		8.090	8.078	8.115	8.098	8.054	8.061
	1	4.3	16.1	1.8	0.5	0.0	0.7
	2	19.4	1.0	74.1	76.6	66.1	79.0
relative weight	3	63.1	64.6	24.1	20.4	0.0	0.0
	4	0.0	6.2	0.0	2.4	33.9	20.2
	5	13.2	12.1	0.0	0.1	0.0	0.1
exc. energy(4)		8.096	8.063	8.281	8.288	8.081	8.078
	1	2.6	0.7	96.1	95.7	11.8	40.5
	2	18.1	52.8	2.0	2.4	34.8	12.6
relative weight	3	0.0	1.8	0.0	1.6	0.0	0.1
	4	75.9	44.4	1.9	0.1	40.1	37.2
	5	3.4	0.3	0.0	0.2	13.4	9.6
exc. energy(5)		8.110	8.097	8.437	8.294	8.630	8.591
	1	10.3	2.4	2.4	0.1	0.0	0.7
	2	38.2	44.3	0.0	2.0	0.0	0.0
relative weight	3	0.0	3.1	0.0	0.2	97.0	98.6
	4	37.0	49.1	97.6	95.7	0.0	0.0
	5	14.5	1.1	0.0	2.0	3.0	0.7

somewhat from the reference distribution. This problem could be solved by including higher-order terms in the coupling. For the case in study, and taking into account that the comparison is at best qualitative, the accuracy seems to be adequate, and suffices for an analysis of the excitonic states character.

We now look at the error in the highest-lying excitation of the "tetra + 1" structure. The state is localized, both in the full and one-body results. This indicates that the problem is not due to the coupling between states (the use of transition dipole moments interaction) but, instead, to the description of the one-body state. The excitation takes place at the monomer located in the tetramer ring acting as a donor to the external water molecule. The latter is also oriented toward another molecule in the ring, forming a trimer ring. However, comparison to a regular trimer ring shows that the extra molecule is oriented unfavorably relative to the first monomer. In this particular situation, the use of atom-centered point charges could be insufficient to describe the effect of this neighboring molecule.

3.3. Large Water Clusters ($N \geq 10$). In this section, we apply our approach to the spectra of larger clusters, looking into optimized structures of various sizes. We included water clusters $(H_2O)_N$ with sizes $N = 10$, 20, 30, 40, 60, and 80. In order to sample a significant conformational space, NVT molecular dynamics simulations were carried out for each cluster size at 250 K. The temperature was chosen so that no significant vaporization would occur, while giving enough energy for the system to explore different conformations. The time step used was 1 fs, using a modified Beeman algorithm for integration. Each system was thermalized for 50 ps, and afterward structures were sampled in 10 ps intervals during a 1 ns production run. This sums up to a total of 100 structures for each given size. Some of these simulations had to be repeated (with a different starting structure) since a water molecule would disattach from the cluster. The selected structures were then optimized, at the same level of theory as in the simulation, with use of the AMOEBA potential.[33] The optimization convergence criteria was set to an rms gradient per atom of 0.0001 kcal/mol/Å. We then performed a one-body EOM-CCSD calculation (with excitonic coupling) on each of the selected structures, taking the first N absorption values. In order to represent these in a suitable graphical form, we replaced each value (or peak) by a normalized Gaussian with a variance of 0.0025 eV^{-2}. The results are shown in Figure 3.

Before we discuss Figure 3 in more detail, there is some additional information that should be taken into account. The predicted excitation energy of the AMOEBA-optimized water monomer with EOM-CCSD/aug-cc-pVTZ is 7.6 eV. The absorption peak of gas-phase water is around 7.4 eV.[39,40] For liquid water, the first maximum of the one-photon absorption spectrum is around 8.2 eV,[41] which corresponds to a shift of 0.8 eV. In ice, this value is about 8.7 eV, a shift of 1.3 eV relative to the gas phase. If we consider that our combination of methods would lead, in the limit, to the same

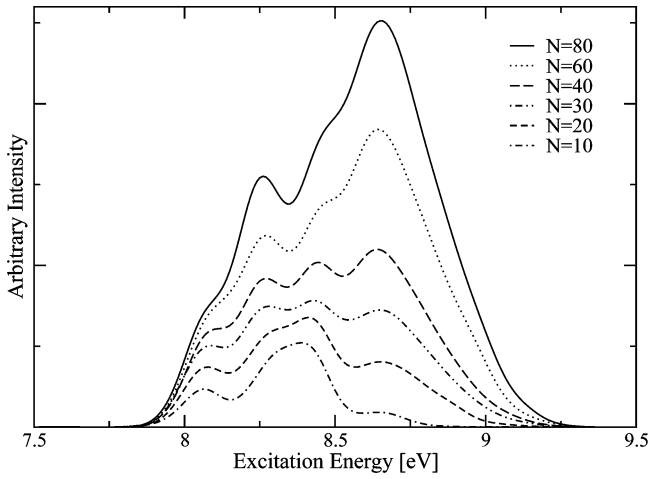


Figure 3. Absorption peaks for water clusters of varying sizes. Each peak has been replaced by a normalized Gaussian ($\sigma^2 = 0.0025$ eV^{-2}).

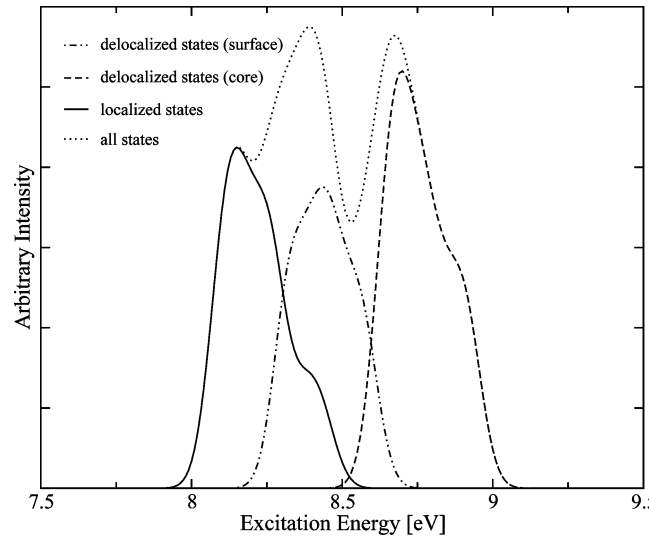


Figure 4. Absorption profile for a $(H_2O)_{40}$ cluster. The spectra is decomposed into localized states (found in the surface) and delocalized states, with or without significant elements from core water molecules. Each peak has been replaced by a normalized Gaussian ($\sigma^2 = 0.0025$ eV^{-2}).

shifts when describing the liquid and the solid water phase, our estimates would be 8.4 and 8.9 eV for the absorption maxima, respectively.

There are some identifiable trends as one increases the size of the cluster. With $N = 10$, one observes a broad range of excitations between 8.0 and 8.5 eV. There is also a small shoulder around 8.7 eV. All of the excitations are above 7.9 eV, which is expected since all of the waters will have at least one hydrogen-bond to a neighboring molecule. As the cluster size increases, the shoulder starts increasing, until it becomes the dominant band with $N = 80$. The value of 8.7 eV is between our estimates for the absorption maximum (in this level of theory) of liquid water and ice. This is also expected since the structures have been optimized, which leads to a rigid hydrogen-bond network (just as in the case of ice). The reason why the two values do not coincide is, however, uncertain. Such a small difference (0.2 eV) could be either due to the fact that EOM-CCSD predicts a different shift than experiment, to our model, to the sampling, or even the molecular mechanics potential used. In any case, the agreement seems reasonable, taking into account that we are studying clusters and the various approximations used.

The second maximum (to the red) seems to converge with increasing cluster size at a value around 8.2−8.3 eV. The first absorption band of water, both in the liquid and in the ice, is known to have no structure, so this splitting must be connected to surface effects. In order to better understand the structure of the band, we have analyzed the eigenvectors of the states computed for a structure of $(H_2O)_{40}$. Directly inspecting the vectors, we have found that the majority of states is significantly delocalized, with more than five significant contributions. We will consider significant eigenvector elements the ones above a 0.1 threshold (thus contributing more than 1%). The only localized states are found on the surface of the cluster. We confirmed this behavior in several other structures. In order to separate the contribution of surface molecules to the total spectra, we divided the excitations into classes. In the first group we include excitations localized exclusively on surface waters (looking only into the significant vector elements as defined above) and with few significant elements (≤ 5). The second group includes excitations which are delocalized, but still exclusively with surface water contributions. The third group is built from the remaining excitations (which are all delocalized and with significant core contributions). The decomposition of the spectra is shown in Figure 4.

The total spectra is divided into two distinguishable bands, one showing two maxima, the other with some structure on the higher excitation energies end. The band of lower excitation energies is the one extracted from the total graph when taking into account only local excitations found on the surface. These are, as expected, somewhat in between the value for the monomer and the one found in bulk water. The second maximum is mainly due to delocalized excitations over the surface. The maxima to the blue side of the spectra (at 8.7 eV) can be attributed to delocalized core excitons. The shoulder, around 8.9 eV, is hard to characterize but is also visible in the individual spectra of larger clusters. It is mainly dominated by the contribution of the water molecules closest to the center of mass, although it does delocalize significantly. The largest maximum represents the dominant character of the excitations present in the cluster. The classification is rather crude but does give some explanation for the clusters' elaborate UV spectra. Perhaps the most reliable (or even physical) classifications would be the localized surface states and delocalized core states. To classify the transitions in between, as done in Figure 4, is somewhat arbitrary but is an interesting way to look at the transition between the two regimes.

4. Conclusions

We have presented a new simple approach to the treatment of the first excitation band of water, which can be generally applied to very large systems. The method, by making use of only one-body embedded excited-state calculations, requires very little computational effort. For a system of N

water molecules, one only needs to perform N monomer calculations and diagonalize an $N \times N$ matrix. The errors are estimated to be small, even below the ones inherent to the high-end method we used as a benchmark in this work (EOM-CCSD/aug-cc-pVTZ). The method was used to compute and analyze the spectra of medium- to large-sized water clusters (up to 80 water molecules). The results show that for sizable structures, two dominant bands will be observed. The lower energy regime will be dominated by excitonic states of local character on the surface of the cluster. The remaining excitations show significant delocalization, with contributions from several different one-body states. However, the delocalization effect on the excitation energies, as estimated by our method, is rather small.

The approach, as formulated in this work, can and should be further improved before application to other systems. Here we give a short account of some of the developments we are currently pursuing. The diagonal elements of the Hamiltonian as well as the ground-state energy should be computed with at least two-body contributions. As previously remarked, the FMO method has been extended to the computation of excitation spectra. However, it is only applicable to excitations in a single molecule which is energetically well separated from the remaining possible excitations in the system. This is mainly due to difficulties in defining the excitations as belonging to a given molecule when computing two-body terms. At the one-body level, this problem is only partly avoided. Although all excitations are strictly localized, close energy-lying states may still swap, raising serious problems in the use of the many-body expansion. This was not a serious issue in this study since the second excitation is about 2 eV higher in energy.

The use of local occupied spaces can alleviate these problems. One possibility is to restrict the coefficients of a given excited state by grouping molecular orbitals according to the monomers where their main centers of charge are found. This would be equivalent to the use of local correlation domains.[30,42] The other possibility is to analyze the excited-state coefficients and derive monomer weights. Both alternatives could allow the inclusion of two-body terms. However, other problems remain to be solved at this level. A balanced virtual space has to be used in order to avoid inconsistencies in the computation of the one- and two-body terms. Preliminary calculations show that this is the main obstacle to a well-defined and convergent expansion. On the other hand, if the two-body terms are added, the correction terms introduced in eqs 3 and 7 are no longer needed. Another advantage lies in the fact that two-body contributions will account for polarization effects on neutral waters due to the excitation of a neighboring molecule. As previously discussed, this is absent in the one-body model.

Other improvements to be made are on the coupling between excited states. The use of the transition dipole seems to be adequate in the case studied, but the implementation of a more general scheme should be considered in future applications.

Acknowledgment. The authors thank Dr. Tatiana Korona for assistance in the use of the EOM-CCSD code, as well as Dr. Jeremy Harvey for some helpful comments on his work. R.A.M. gratefully acknowledges a Research Grant from Fundação para a Ciência e Tecnologia (reference SFRH/BPD/38447/2007).

References

(1) Winter, B.; Weber, R.; Widdra, W.; Dittmar, M.; Faubel, M.; Hertel, I. *J. Phys. Chem. A* **2004**, *108*, 2625.

(2) Winter, B.; Faubel, M. *Chem. Rev.* **2006**, *106*, 1176.

(3) Hermann, A.; Schmidt, W. G.; Schwerdtfeger, P. *Phys. Rev. Lett.* **2008**, *100*, 207403.

(4) Estácio, S. G.; Martiniano, H. F. M. C.; do Couto, P. C.; Cabral, B. J. C. In *Solvation Effects in Molecules and Biomolecules*; Canuto, S., Ed.; Elsevier: Heidelberg, 2008; Chapter 5, p 115.

(5) Millot, C.; Cabral, B. J. C. *Chem. Phys. Lett.* **2008**, *460*, 466.

(6) Brancato, G.; Rega, N.; Barone, V. *Phys. Rev. Lett.* **2008**, *100*, 107401.

(7) Lu, D.; Gygi, F.; Galli, G. *Phys. Rev. Lett.* **2008**, *100*, 147601.

(8) Mata, R. A.; Cabral, B.; Millot, C.; Coutinho, K.; Canuto, S. *J. Chem. Phys.* **2009**, *130*, 014505.

(9) Xantheas, S. S. *J. Chem. Phys.* **1994**, *100*, 7523.

(10) Liu, K.; Cruzan, J. D.; Saykally, R. J. *Science* **1996**, *271*, 929.

(11) Cruzan, J. D.; Braly, L. B.; Liu, K.; Brown, M. G.; Loeser, J. G.; Saykally, R. J. *Science* **1996**, *271*, 59.

(12) Harvey, J. N.; Jung, J. O.; Gerber, R. B. *J. Chem. Phys.* **1998**, *109*, 8747.

(13) Fredj, Y. M. E.; Harvey, J. N.; Gerber, R. B. *J. Phys. Chem. A* **2004**, *108*, 4405.

(14) Galamba, N.; Cabral, B. J. C. *J. Am. Chem. Soc.* **2008**, *130*, 17955.

(15) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 46.

(16) Sorkin, A.; Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 683.

(17) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *3*, 46.

(18) Chiba, M.; Fedorov, D. G.; Kitaura, K. *Chem. Phys. Lett.* **2007**, *444*, 346.

(19) Chiba, M.; Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2007**, *127*, 104108.

(20) Mata, R. A.; Stoll, H. *Chem. Phys. Lett.* **2008**, *465*, 136.

(21) Fedorov, D. G.; Kitaura, K. *J. Phys. Chem. A* **2007**, *111*, 6904.

(22) Cui, J.; Liu, H.; Jordan, K. D. *J. Phys. Chem. B* **2006**, *110*, 18872.

(23) Kurihara, Y.; Aoki, Y.; Imamura, A. *J. Chem. Phys.* **1997**, *107*, 3569.

(24) Pomogaev, V.; Gu, F. L.; Pomogaeva, A.; Aoki, Y. *Int. J. Quantum Chem.* **2009**, *109*, 1328.

(25) Pomogaev, V.; Pomogaeva, A.; Aoki, Y. *J. Phys. Chem. A* **2009**, *113*, 1429.

(26) Hirata, S.; Valiev, M.; Dupuis, M.; Xantheas, S. S.; Sugiki, S.; Sekino, H. *Mol. Phys.* **2005**, *103*, 2255.

(27) Chiba, M.; Fedorov, D. G.; Kitaura, K. *J. Comput. Chem.* 200829, 2667.

(28) Fukunaga, H.; Fedorov, D. G.; Chiba, M.; Nii, K.; Kitaura, K. *J. Phys. Chem. A* **2008**, *112*, 10887.

(29) Nakatsuji, H.; Miyahara, T.; Fukuda, R. *J. Chem. Phys.* **2007**, *126*, 084104.

(30) Korona, T.; Werner, H.-J. *J. Chem. Phys.* **2003**, *118*, 3006–3019.

(31) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.

(32) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M. et al. MOLPRO, version 2008. 1, a package of ab initio programs, 2008. http://www.molpro.net.

(33) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933.

(34) Schütz, M.; Hetzer, G.; Werner, H.-J. *J. Chem. Phys.* **1999**, *111*, 5691–5705.

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

(36) do Couto, P. C.; Cabral, B. J. C.; Canuto, S. *Chem. Phys. Lett.* **2006**, *429*, 129.

(37) Ponder, J. W. In TINKER: Software Tools for Molecular Design, 4.2 ed.; Washington University School of Medicine: Saint Louis, MO, 2003.

(38) Pipek, J.; Mezey, P. G. *J. Chem. Phys.* **1989**, *90*, 4916–4926.

(39) Cheng, B.-M.; Chew, E. P.; Liu, C.-P.; Bahou, M.; Lee, Y.-P.; Yung, Y.; Grestell, M. F. *Geophys. Res. Lett.* **1999**, *26*, 3657.

(40) Cai, Z.-L.; Tozer, D. J.; Reimers, J. R. *J. Chem. Phys.* **2000**, *113*, 7084.

(41) Heller, J. M., Jr.; Hamm, R. N.; Birkhoff, R. D.; Painter, L. R. *J. Chem. Phys.* **1974**, *60*, 3483.

(42) Boughton, J. W.; Pulay, P. *J. Comput. Chem.* **1993**, *14*, 736–740.

CT9001653

JCTC Journal of Chemical Theory and Computation

Electronic Energy Transfer in Condensed Phase Studied by a Polarizable QM/MM Model

Carles Curutchet,*,† Aurora Muñoz-Losa,‡ Susanna Monti,§ Jacob Kongsted,|| Gregory D. Scholes,† and Benedetta Mennucci‡

Department of Chemistry, 80 St. George Street, Institute for Optical Sciences, and Centre for Quantum Information and Quantum Control, University of Toronto, Toronto, Ontario, M5S 3H6 Canada, Dipartimento di Chimica e Chimica Industriale, Università di Pisa, via Risorgimento 35, 56126 Pisa, Italy, Istituto per i Processi Chimico-Fisici (IPCF-CNR), Area della Ricerca, via G. Moruzzi 1, I-56124 Pisa, Italy, and Department of Physics and Chemistry, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

Received March 23, 2009

Abstract: We present a combined quantum mechanics and molecular mechanics (QM/MM) method to study electronic energy transfer (EET) in condensed phases. The method introduces a quantum mechanically based linear response (LR) scheme to describe both chromophore electronic excitations and electronic couplings, while the environment is described through a classical polarizable force field. Explicit treatment of the solvent electronic polarization is a key aspect of the model, as this allows account of solvent screening effects in the coupling. The method is tested on a model perylene diimide (PDI) dimer in water solution. We find an excellent agreement between the QM/MM method and "exact" supermolecule calculations in which the complete solute−solvent system is described at the QM level. In addition, the estimation of the electronic coupling is shown to be very sensitive to the quality of the parameters used to describe solvent polarization. Finally, we compare ensemble-averaged QM/MM results to the predictions of the PCM-LR method, which is based on a continuum dielectric description of the solvent. We find that both continuum and atomistic solvent models behave similarly in homogeneous media such as water. Our findings demonstrate the potential of the method to investigate the role of complex heterogeneous environments, e.g. proteins or nanostructured host materials, on EET.

1. Introduction

Electronic energy transfer (EET) is a fundamental nonradiative process involving de-excitation of a donor molecule and concomitant electronic excitation of a nearby acceptor.[1] EET is used to harvest light in photosynthesis with near-perfect efficiency[2−4] and is intrinsic to many applications in materials and life sciences. Some examples include the design of artificial light-harvesting antennae,[5−8] the optimization of organic light-emitting diodes,[9−12] and the measurement of distances in biological systems.[13−15]

Much progress related to EET was incubated more than 50 years ago, when Förster proposed an elegant theory relating experimental observables to the mechanisms of EET.[16] Despite the success of the Förster theory in explaining EET dynamics in a wide variety of systems, researchers have identified a number of situations in which such a theory can fail.[17] Non-Förster effects include the breakdown of the point dipole approximation used to describe the electronic coupling,[18−21] distance-dependent dielectric screening effects,[22,23] and the need to evaluate the spectral overlap factor from homogeneously broadened absorption and emission spectra and subsequently averaging over inhomogeneous, i.e.

* Corresponding author e-mail: ccurutch@chem.utoronto.ca.
† University of Toronto.
‡ Università di Pisa.
§ Istituto per i Processi Chimico-Fisici (IPCF-CNR).
|| University of Southern Denmark.

10.1021/ct9001366 CCC: $40.75
Published on Web 06/04/2009

static, disorder.[24−26] In addition, in multichromophoric aggregates it is necessary to account for effective donor/acceptor states shared over multiple strongly coupled chromophores.[26−28] Photosynthetic proteins represent a paradigmatic case in which such effects are significant. Moreover, recent experimental evidence points to the importance of wavelike coherent energy transfer in such systems, in contrast to the traditionally assumed Förster incoherent hopping mechanism, suggesting that the protein plays an active role in protecting electronic coherences between the electronic states of the pigments.[29,30] In this context, a combination of both elaborate experimental and theoretical approaches are needed in order to assess the significance of such non-Förster effects and, in particular, to gain fundamental insights into the role of the protein structure and dynamics on the overall process.

In the past decade, much theoretical effort has been directed toward the accurate prediction of electronic couplings from quantum mechanical (QM) methods, thus overcoming the point dipole approximation.[18−21,31,32] The accuracy of these approaches depends mainly on the quality of the QM level of theory and basis set adopted.[33] However, detailed theoretical insights on the role of solvation on EET have eluded researchers for a long time. Typically, the simple screening factor introduced by Förster, $1/n^2$, where n is the refractive index of the medium, is used. The significance of this factor is evident, as it can reduce the EET rate by a factor of $\sim$4 in typical environments.

A significant advance in the field has been the development of a QM method to study EET between molecules in condensed phase.[31,34] This method is based on a linear response (LR) approach (either within Hartree−Fock and density functional theory or semiempirical approaches) and introduces the effect of the environment in terms of the polarizable continuum model (PCM). In the PCM model,[35] the solvent is represented as a polarizable continuum medium characterized by its macroscopic dielectric properties, whereas the solute, located in a molecular-shaped cavity inside the dielectric, is described at a full QM level. Such a methodology allows for a consistent treatment of solvent effects on both the evaluation of the excited states and the electronic coupling and in addition properly accounts for molecular shape, thus overcoming a fundamental limitation of Förster's screening factor. By applying this method, we recently discovered that the molecular shape indeed has a strong influence on the screening of the electronic couplings between photosynthetic pigments, leading to an exponential attenuation of the screening at separations less than about 20 Å.[22,23]

While the PCM-LR methodology is well-suited to study EET in homogeneous solvents, the treatment of solvation in light-harvesting proteins is more challenging. Measurements of time-dependent fluorescence Stokes shifts[36] and molecular dynamics (MD) simulations[37] have shown that polar solvation dynamics in such systems are position-dependent and highly heterogeneous. Depending on the particular protein or protein site, for example, static dielectric permittivities ranging from 4 to 40 have been estimated from MD simulations.[38−41] In other words, the fine-tuning of the transition energies of the pigments that modulates the EET pathways as well as the screening effect on the couplings arise from different local pigment−protein interactions that cannot be captured by a continuum model. In addition, a continuum model directly provides ensemble-averaged quantities meaning that it is unable to describe the disorder in the transition energies and the electronic couplings due to the fluctuating environment, which are particularly important in the description of EET dynamics in multichromophoric arrays.[42,43]

In this paper we present a new combined quantum mechanical and molecular mechanical (QM/MM) method to study EET that overcomes the limitations of continuum models. In a QM/MM scheme, one part of the system (in this case the chromophores) is fully described at the QM level, whereas the solvent molecules or surrounding environment are described by a classical force field. QM/MM methods have successfully been applied to describe solvent effects on a variety of molecular properties including solvatochromic shifts in optical spectra.[44−49] The method we present here follows the same strategy used with PCM and it introduces a LR scheme to describe both chromophore excitations and EET couplings, while the environment is described through a MM force field. It is important to stress that an explicit treatment of electronic polarization in the environment is essential in order to account for screening effects on EET couplings: this effect is here recovered by using a polarizable MM force field. Below we will refer to this method as QM/MMpol.

We validate the method by comparing the results to "exact" calculations of the excitonic splitting in a model perylene diimide (PDI) dimer in water solution, in which the complete solute−solvent system is described at the QM level. In addition, QM/MMpol results averaged over solvent configurations sampled from MD simulations are compared to the predictions of the PCM-LR method, showing that both approaches are consistent and describe a similar distance-dependent decay of the solvent screening factor. Our results thus demonstrate the potential of the method to investigate the role of complex heterogeneous environments on EET.

The paper is organized as follows. In Section 2 we describe the theory underlying the method. In Section 3, we describe the computational details of the MD simulations and the QM/MMpol and PCM calculations. In addition, we report the derivation of the polarizable force field used to describe water in the QM/MMpol calculations. In Section 4, we first discuss the validity of the QM/MMpol model by comparing the results with exact QM supermolecule calculations of the complete solute−solvent system. Then, we compare solvent effects induced on transition properties as well as on the electronic couplings as described by the QM/MMpol and PCM-LR methods. Finally, in Section 5 we report our conclusions and give some perspectives on the potential of the method.

2. Methodology

2.1. Effective Hamiltonian for QM/MMpol. QM/MMpol and PCM belong to the same family of the so-called

"focussed models". The most important characteristic of both of them is in fact that the system is divided in two parts (or layers) which are described at different level of accuracy. The target layer (the solute plus eventually some solvent molecules) is described at the QM level (either ab initio or semiempirical), while the rest (the solvent) is approximated using a MM or a continuum description. In all cases, the formalism of *in vacuo* QM molecular calculations can be maintained if we introduce an effective Hamiltonian, H_{eff}, which includes an explicit term representing the solute−solvent interaction (and the energy of the MM region in the case of QM/MMpol). Introducing the standard Born−Oppenheimer approximation, the solute electronic wave function will satisfy the following equation

$$\hat{H}_{eff}|\Psi\rangle = (\hat{H}_0 + \hat{H}_{env})|\Psi\rangle = E|\Psi\rangle \quad (1)$$

where H_0 is the gas phase Hamiltonian of the QM solute system, and the operator H_{env} introduces the coupling between the solute and the solvent. What distinguishes MMpol from PCM is exactly the form of the operator H_{env}.

In the QM/MMpol approach we adopt here, the MM system is described by a classical polarizable force field based on the induced dipole model. In particular, electrostatic forces are described by atomic partial charges, whereas polarization is explicitly treated by adding isotropic polarizabilities at selected points in the solvent molecules. We thus have

$$\hat{H}_{env} = \hat{H}_{QM/MM} + \hat{H}_{MM} \quad (2)$$

and the solute−solvent interaction $\hat{H}_{QM/MM}$ and MM energy $\hat{H}_{MM}$ terms are given by

$$\hat{H}_{QM/MM} = \hat{H}^{el}_{QM/MM} + \hat{H}^{pol}_{QM/MM} = \sum_m q_m \hat{V}(r_m) - \frac{1}{2}\sum_a \mu_a^{ind} \hat{\mathbf{E}}_a^{solute}(r_a) \quad (3)$$

$$\hat{H}_{MM} = \hat{H}^{el}_{MM} + \hat{H}^{pol}_{MM} = \sum_m \sum_{n>m} \frac{q_m q_n}{r_{mn}} - \frac{1}{2}\sum_a \mu_a^{ind} \sum_m \frac{q_m(\mathbf{r}_a - \mathbf{r}_m)}{|\mathbf{r}_a - \mathbf{r}_m|^3} \quad (4)$$

where $\hat{V}(r_m)$ and $\hat{\mathbf{E}}_a^{solute}(r_a)$ are the electrostatic potential and electric field operators due to electrons and nuclei of the QM (solute) system at the MM sites, and the indexes *m (n)* and *a* run over the MM charges q_m and induced dipoles μ_a^{ind} located at r_m and r_a, respectively.

In eq 3, $H^{el}_{QM/MM}$ and $H^{pol}_{QM/MM}$ describe the interaction between the QM system and the MM charges and induced dipoles, respectively. On the other hand, in eq 4 H^{el}_{MM} describes the electrostatic self-energy of the MM charges, while H^{pol}_{MM} represents the polarization interaction between such charges and the induced dipoles. We recall that the H^{el}_{MM} term enters in the effective Hamiltonian only as a constant energetic quantity, while the H^{pol}_{MM} contribution is explicitly considered in the corresponding Fock operator because of the explicit dependence of the induced dipoles on the QM wave function. In addition, here we do not consider short-range dispersion and repulsion contributions in $H_{QM/MM}$ and H_{MM}, as in most combined QM/MM methods these are described by empirical potentials independent of the QM electronic degrees of freedom, thus not affecting our results.

The dipoles induced on each MM polarizable site are given by

$$\mu_a^{ind} = \alpha_a(\mathbf{E}_a^{solute} + \mathbf{E}_a^{solvent}\{\mathbf{q};\mu^{ind}\}) \quad (5)$$

where we have assumed a linear approximation, neglected any contribution of magnetic character related to the total electric field, and used an isotropic polarizability (α_a) for each selected point in the MM part of the system. In eq 5, $\mathbf{E}_a^{solvent}$ refers to the total solvent electric field calculated at site *a* and contains a sum of contributions from the point charges and the induced dipole moments in the MM part of the system. Such a field (and hence the induced dipole) depends on all other induced dipole moments in the solvent. This means that eq 5 must be solved iteratively within each SCF iteration. As an alternative, mutual polarization between the dipoles can be solved through a matrix inversion approach, where eq 5 is reformulated into a matrix equation

$$\boldsymbol{\mu}^{ind} = \mathbf{BE} \quad (6)$$

where the matrix **B** is of dimension 3N⊗3N, with N being the number of polarizable sites, and the vector **E** collects the electric field from the solute and the solvent permanent charge distribution. The form of matrix **B** will be determined uniquely by the position of the polarizable sites and the polarizability values.

Equation 6 is a further direct link between QM/MMpol and PCM; in PCM in fact the polarization of the solvent is expressed in terms of a set of apparent point charges placed on the surface of the molecular cavity embedding the QM system. These apparent charges are, exactly as the MMpol induced dipoles, determined by the electric field from the solute (or the electrostatic potential in more recent formulations of the model) calculated at the positions of the charges, namely

$$\mathbf{q}^{PCM} = -\mathbf{K}\mathbf{f}^{solute} \quad (7)$$

As the MMpol matrix **B**, also **K** is a square matrix (the dimension being now equal to Nts⊗Nts, where Nts is the number of apparent charges): it only depends on the geometrical cavity parameters and the dielectric constant of the solvent.

In both MMpol and PCM the addition of $\hat{H}_{env}$ to the solute Hamiltonian automatically leads to a modification of the solute wave function which has now to be determined by solving the effective eq 1. This can be done using exactly the same methods used for isolated molecules; here in particular we shall mainly focus on the standard Self Consistent Field (SCF) approach (either in its Hartree−Fock or DFT formulation). Due to the presence of $\hat{H}_{env}$ the modified SCF scheme is generally known as Self Consistent Reaction Field (SCRF), which emphasizes the mutually polarized solute−solvent system obtained at the end of the SCF. Historically the term SCRF has been coined for the

QM/Continuum approach, but here, due the parallelism between the two schemes, it will be used indistinctly for both.

2.2. QM/MMpol Linear Response. In the following we develop the working expressions to include the effects of the polarizable MM environment in a TD-DFT linear response scheme. Its extension to the Hartree−Fock or semiempirical level (TD-HF, CIS, or ZINDO) is straightforward. In this framework, the excitation energies of a molecular system can be determined by solving[50]

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{A}^* \end{pmatrix}\begin{pmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{pmatrix} = \omega_n \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{pmatrix} \tag{8}$$

where the matrices **A** and **B** form the Hessian of the electronic energy, and the transition vectors ($\mathbf{X}_n$ $\mathbf{Y}_n$) correspond to collective eigenmodes of the density matrix with eigenfrequencies ω_n. The Coulombic and exchange-correlation (XC) kernels produce both diagonal and off-diagonal contributions to **A** and **B**, correcting the transitions between occupied and unoccupied levels of the ground-state potential into the true transitions of the system. The effect of the polarizable environment on the **A** and **B** matrices can be included by considering the MM dipoles induced by the density matrix associated with the transition vectors ($\mathbf{X}_n$ $\mathbf{Y}_n$). The extension of this linear response scheme to include the effect of the MMpol environment is analogous to the inclusion of PCM solvent effects, which have been described in detail in ref 51. In the present case, the electronic part of the polarization response of the environment is represented by a set of point dipoles induced by the appropriate density matrix instead of a set of apparent surface charges displaced on the cavity surface as described in the PCM model.

By using the usual convention with respect to labeling the molecular orbitals (i.e., (i, j,...) for occupied; (a, b,...) for virtual), the matrices **A** and **B** thus become

$$A_{ai,bj} = \delta_{ab}\delta_{ij}(\varepsilon_a - \varepsilon_i) + K_{ai,bj} + C^{pol}_{ai,bj}$$

$$B_{ai,bj} = K_{ai,jb} + C^{pol}_{ai,bj} \tag{9}$$

where ε_r are the orbital energies, and $K_{ai,bj}$ and $C^{pol}_{ai,bj}$ are the coupling matrix and the polarizable MM matrix, respectively

$$K_{ai,bj} = \int d\mathbf{r}\int d\mathbf{r}'\phi_i(\mathbf{r}')\phi_a^*(\mathbf{r}')\left(\frac{1}{|\mathbf{r}'-\mathbf{r}|} + g_{xc}(\mathbf{r}',\mathbf{r})\right)\phi_j(\mathbf{r})\phi_b^*(\mathbf{r}) \tag{10}$$

$$C^{pol}_{ai,bj} = -\sum_k\left(\int d\mathbf{r}\phi_i(\mathbf{r})\phi_a^*(\mathbf{r})\frac{(\mathbf{r}_k - \mathbf{r})}{|\mathbf{r}_k - \mathbf{r}|^3}\right)\mu_k^{ind}(\phi_j\phi_b^*) \tag{11}$$

and g_{xc} is the exchange eventually plus correlation (if a DFT description is used) kernel. In eq 11 the k index runs on the total number of polarizable MM sites. We note that for CIS and ZINDO the whole matrix **B** is neglected, and for ZINDO the xc terms in eq 10 become zero.

2.3. Electronic Energy Transfer Coupling. We begin by considering two solvated chromophores, A and D, with a common resonance frequency, ω_0, when not interacting. When the interaction is turned on, their respective transitions are no longer degenerate. Instead, two distinct transition frequencies ω_+ and ω_- appear, and the excited states become delocalized over the two monomers. The splitting between these defines the energy transfer coupling, V

$$V = \frac{(\omega_+ - \omega_-)}{2} \tag{12}$$

Such a splitting can be evaluated by computing the excitation energies of the $D\oplus A$ system through a TD-DFT scheme, as shown in the previous section. This procedure to estimate electronic couplings is known as the "supermolecule" approach.

An approximate solution to eq 12, however, can be obtained by introducing a perturbative approach which considers the D/A interaction as a perturbation and defines the zero-order resulting eigenvectors, ($\mathbf{X}_+$ $\mathbf{Y}_+$) and ($\mathbf{X}_-$ $\mathbf{Y}_-$), as linear combinations of the unperturbed Kohn−Sham orbitals of the isolated D and A systems.[52] In analogy to the PCM-LR method for EET,[31] this approximation allows the estimation of the splitting and the corresponding coupling from the transition densities calculated for the noninteracting D and A. To first order, the electronic coupling, V, is obtained as a sum of two terms

$$V = V_s + V_{\text{explicit}} \tag{13}$$

$$V_s = \int d\mathbf{r}\int d\mathbf{r}'\rho_A^{T^*}(\mathbf{r}')\left(\frac{1}{|\mathbf{r}'-\mathbf{r}|} + g_{xc}(\mathbf{r}',\mathbf{r})\right)\rho_D^T(\mathbf{r}) - \omega_0\int d\mathbf{r}\rho_A^{T^*}(\mathbf{r})\rho_D^T(\mathbf{r}) \tag{14}$$

$$V_{\text{explicit}} = -\sum_k\left(\int d\mathbf{r}\rho_A^{T^*}(\mathbf{r})\frac{(\mathbf{r}_k - \mathbf{r})}{|\mathbf{r}_k - \mathbf{r}|^3}\right)\mu_k^{ind}(\rho_D^T) \tag{15}$$

where ρ_D^T and ρ_A^T indicate transition densities of the solvated D and A, respectively, in the absence of their interaction.

In eq 13, V_s describes a chromophore−chromophore Coulomb and exchange-correlation interaction corrected by an overlap contribution. This term is the only one present in vacuum but can be significantly modified upon solvation due to changes induced by the environment in the transition densities, such effect in general leading to an enhancement of the coupling.[23] On the other hand, $V_{explicit}$ describes the interaction between D and A mediated by the polarizable environment. This contribution typically reduces, i.e. *screens*, the overall interaction and is given by the interaction between ρ_A^T and the MM dipoles induced by ρ_D^T (the same result is obtained by exchanging A with D as the term is symmetric with respect to the transition densities). As introduced in our previous studies,[22,23] we can now define the solvent screening factor as

$$s = \frac{V}{V_s} = \frac{V_s + V_{\text{explicit}}}{V_s} \tag{16}$$

which can be directly compared to the $s = 1/n^2$ factor used in Förster's model.

Finally, we note that in the case of an asymmetric system, where the transition energies ω_D and ω_A of the noninteracting monomers are not equal, such asymmetry has to be taken

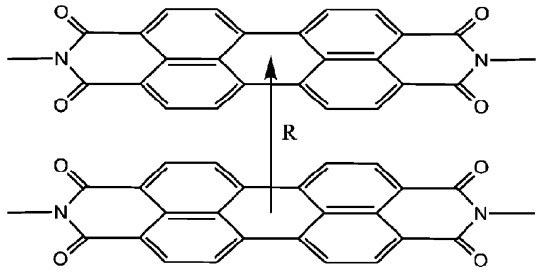


Figure 1. Structure of the perylene diimide (PDI) dimer considered in this work. Two interchromophoric distances (R=3.5 and R=7.0 Å) corresponding to a face-to-face orientation were considered.

into account when deriving the coupling from the energy splitting. This case is relevant for the "supermolecule" calculations we present in Section 4.1. The corresponding expression can be derived from the secular determinant describing Frenkel excitons by considering two-level chromophores[53]

$$V = \frac{1}{2}\sqrt{(\omega_+ - \omega_-)^2 - (\omega_D - \omega_A)^2} \qquad (17)$$

It is obvious that for identical ω_D=ω_A we recover the initial expression, eq 12. The contribution of other electronic states to the splitting may not be neglected, but this problem is minimized in the case of nearly identical PDI molecules as considered in this work, where the asymmetry arises only from slightly different local interactions with the solvent.

3. Computational Details

3.1. Molecular Dynamics Simulations. The systems made of two PDI stacked molecules placed in parallel orientation (see Figure 1) at a distance of 3.5 and 7.0 Å (hereafter named mod35 and mod70, respectively) were inserted in rectangular parallelepiped boxes and solvated with polarizable POL3 water molecules[54] removing those waters falling within 1.5 Å from the adducts. All simulations were performed using the AMBER9 package with the general amber force field (GAFF)[55] to describe the solute. Before production simulations were started, the cell size for both adducts was adjusted in a series of minimizations and short NVT molecular dynamics simulation runs in order to achieve the correct density of the water molecules filling the simulation box. The final box dimensions and the corresponding water molecules were $37.4 \times 28.4 \times 27.4$ Å^3 (937 waters) and $37.0 \times 28.4 \times 30.2$ Å^3 (1055 waters) for mod35 and mod70, respectively.

Then pre-equilibration in the NVT ensemble was performed first at high temperature ($T = 600$ K), in order to randomize water positions, and then at lower temperature ($T = 298$ K). The Andersen temperature coupling scheme[56] with a relaxation time of 0.4 ps was employed. The solute was kept fixed at the initial geometry during all the simulations. The time step was set to 1 fs. Periodic boundary conditions were applied, and the particle mesh Ewald method[57] was used to deal with electrostatic forces. Starting from the last obtained equilibrium configuration, production runs were performed in the NVT ensemble for a total simulation time of 2 ns. Configurations were saved every picosecond for subsequent QM and QM/MMpol calculations. In particular, QM/MM results presented on Section 4.2 correspond to 100 structures saved every 20 ps, which we found to be enough to obtain converged results.

3.2. Definition and Determination of Partial Atomic Charges and Distributed Polarizabilities Used in QM/MMpol Calculations. As described in the previous section, MD simulations were performed adopting the polarizable POL3 water model.[54] While this model has been optimized to reproduce bulk structural properties of water, it substantially underestimates the water polarizability (see ref 58 for a detailed discussion). It is therefore not adequate to study electronic properties, so we derived new sets of partial atomic charges and distributed polarizabilities to be used in the QM/MMpol calculations. In particular, the distributed atomic dipole−dipole polarizabilities were calculated using the LoProp[59] approach as implemented in the Molcas[60] program, whereas the atomic charges were fitted to the electrostatic potential following the ESP method implemented in the GAUSSIAN package.[61] The calculations were performed at either the level of Hartree−Fock or DFT employing the B3LYP density functional. The basis set used were either the 6-31G(d) or the aug-cc-pVTZ. However, for the LoProp to be properly defined, these basis sets were first transformed into the atomic natural orbital form by a linear transformation which does not affect the orbital optimization. The expansion points used for water were either (i) at the atomic nuclei or (ii) at the atomic nuclei and the bond midpoints. However, test calculations indicated that the effect of the local properties was equally described using either approaches, and since the number of polarizable sites is reduced by expanding only at the atomic nuclei, this approach was taken by us in the final and reported calculations. In all property calculations a single water monomer at the POL3 geometry was considered. This means that intramolecular polarization is included automatically in the derived force field parameters and the multipole-induced dipole, and, in the QM/MMpol calculations, induced dipole−induced dipole interactions are thereby only explicitly considered between different water molecules. In this way only intermolecular distances are relevant, and damping functions related to the electric field were therefore not considered. In addition, the induced dipoles were in the QM/MMpol calculations solved by a matrix inversion procedure, as indicated by eq 6, and no artificially large water induced dipoles were observed. The sets of atomic charges and isotropic polarizabilities (in atomic units) obtained were as follows: i) HF/6-31G(d): $\alpha_O = 2.94$, $\alpha_H = 1.18$, $q_O = -0.78$, $q_H = 0.39$; ii) B3LYP/aug-cc-pVTZ: $\alpha_O = 5.74$, $\alpha_H = 2.31$, $q_O = -0.64$, $q_H = 0.32$. We note that the HF/6-31G(d) set of parameters is very similar to the POL3 water model, and test calculations showed very similar results using one or the other. However, we preferred to adopt the HF set in order to perform a fair comparison to full QM CIS/6-31G(d) calculations, as described below. All QM/MMpol calculations were performed in a locally modified version of Gaussian.

3.3. PCM Calculations. The only parameters needed in the PCM model are the positions and radii of the spheres

determining the cavity embedding the molecule and the environment optical (ε_{opt}) and static (ε_0) permittivities. PCM cavities have been constructed by applying the united atom topological model and the atomic radii of the UFF[62] force field as implemented in the GAUSSIAN 03 code.[61] Transition energies, transition densities, and electronic couplings have been obtained by considering a single cavity enclosing the *D*/*A* pair for the 3.5 Å interchromophoric distance and two distinct cavities (one for each chromophore) for the 7.0 Å distance. As here the PCM has to mimic water, we have used the permittivities experimentally known for water, namely $\varepsilon_{opt} = 1.776$ and $\varepsilon_0 = 78.39$.

4. Results

4.1. Comparison between QM/MMpol and Full QM Supermolecule Calculations. In this section, we validate the QM/MMpol model by comparison with exact supermolecule calculations of the complete solute−solvent system described at a full quantum-mechanical level. Such full QM calculations are computationally very expensive, so we limit the analysis to a single solute−solvent configuration extracted from the MD simulations. This is in contrast to the ensemble averaged quantitities that we will present in the next section. Moreover, even for a single snapshot one must include a very large number of solvent molecules in order to get converged results due to the long-range nature of electrostatic interactions. For the PDI dimer in water, test calculations indicate that convergence below 1% error on electronic couplings and solvent screening factors is achieved by including the water molecules located inside a hypothetical sphere of cutoff radius of ∼19 Å from the center of the dimer; this corresponds to ∼800 water molecules. We also note that the convergence is faster for transition energies and dipoles. Due to the high computational cost of the full QM calculations, for this first part of the analysis we have to use smaller cutoff values (in the range 7−11 Å).

For the same reason, MMpol and full QM calculations are performed with a single level of theory (CIS/6-31G(d)), whereas in the next section ZINDO and time-dependent density functional theory (TD-DFT) results will also be presented. The main reason for choosing CIS instead of TDDFT is related to the problem of the large number of artificial low-lying charge-transfer states that the surrounding waters introduced in full QM TD-DFT calculations. Such a problem arises from the well-known deficiencies of most of the present exchange-correlation functionals[63] and strongly complicates the determination of the dimer states of interest. In contrast, at the CIS level, the states of interest are the first two excited states.

As already noted in the Introduction, a proper description of electronic polarization is crucial in order to describe the effect of the environment on the electronic coupling. It is well-known that extended basis sets are needed in QM methods in order to properly describe polarization. Again, the cost of the full QM calculations limits us to use a split-valence 6-31G(d) basis set, which is expected to recover only part of the polarization effect on the coupling. In order to be consistent in the treatment of polarization in the full QM and the QM/MMpol calculations, we have thus developed a MM polarizable force-field of water derived from HF/6-31G(d) calculations, as described in Section 3.2. This allows us to perform a fair and consistent comparison between the MM and QM descriptions of the environment. We are however aware of the deficient treatment of polarization effects at this level of theory; this is clearly illustrated by the fact that the experimental polarizability of water is approximately two times the value predicted at the HF/6-31G(d) level. Thus, we have also developed an accurate force field of water from B3LYP/aug-cc-pVTZ calculations, which in turn provides a set of atomic polarizabilities that accurately describe the experimental polarizability of water.

Table 1. CIS/6-31G(d) Electronic Coupling and Solvent Screening Factors Calculated with the Perturbative QM/MMpol Model and from Exact Supermolecule Calculations of the Full QM Solute-Solvent System[b]

	MMpol (HF/6-31G(d) derived force field)				
cutoff (Å)[a]	V_s	$V_{explicit}$	V	s	full QM V
7	585	−5	580	0.99	569
8	590	−25	565	0.96	550
9	604	−45	560	0.93	547
10	605	−56	549	0.91	540
11	610	−70	541	0.89	534

	MMpol (B3LYP/aug-cc-pVTZ derived force field)				
cutoff (Å)[a]	V_s	$V_{explicit}$	V	s	full QM V
7	599	−27	572	0.95	569
8	610	−68	542	0.89	550
9	629	−105	524	0.83	547
10	633	−126	507	0.80	540
11	639	−148	492	0.77	534

[a] Increasing cutoff distances correspond to the consideration of 21, 46, 80, 108, and 159 number of water molecules in the system. [b] Results corresponding to an interchromophoric distance of 7.0 Å are given as a function of the cutoff distance determining the number of waters included in the calculation. Two different sets of MMpol parameters (charges and atomic polarizabilities), derived at the HF/6-31G(d) and B3LYP/aug-cc-pVTZ level, are used to describe the MM environment. All couplings are in cm^{-1}.

In Table 1 we show the results obtained from the full QM CIS/6-31G(d) calculations as well as the perturbative couplings obtained from CIS/6-31G(d)/MMpol calculations by adopting the two different polarizable force fields. The calculations have been performed for a dimer at an interchromophoric distance of 7.0 Å, because, at this distance, solvent screening effects are expected to be more significant than at smaller separations. Full QM couplings are derived from the splitting of the dimer excited states corrected by the mismatch in the donor−acceptor energies, as indicated by eq 17. This implies a QM calculation on the complete dimer-solvent system to obtain ω_+ and ω_- as well as two calculations on the donor-solvent (acceptor-solvent) system, where the other chromophore has been removed, in order to estimate ω_D (ω_A).

The computed couplings in Table 1 clearly show than when a consistent treatment of polarization is used in the QM and MM descriptions through the HF-derived force field, the MMpol model accurately reproduces solvent effects on the coupling, with error in the electronic coupling being always <3%, which corresponds to <15 cm^{-1}. Such behavior

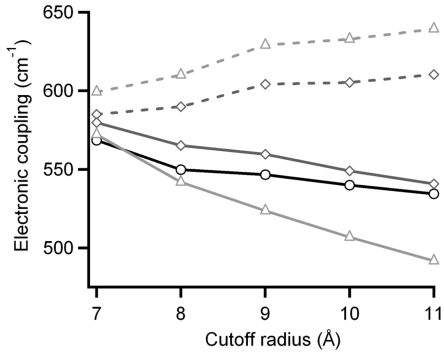


Figure 2. CIS/6-31G(d) electronic couplings obtained from full QM calculations (circles) and from QM/MMpol calculations using the HF/6-31G(d) (diamonds) and the B3LYP/aug-cc-pVTZ (triangles) force fields, represented as a function of the cutoff radius determining the number of water molecules in the system. Solid lines correspond to the total electronic coupling, V, and dashed lines correspond to the unscreened Coulombic contribution, V_s.

is achieved because the QM/MMpol model seems to capture both implicit and explicit screening effects on the coupling in a balanced way, as illustrated in Figure 2. That is, when the number of solvent molecules in the system is increased, the change in the transition densities upon solvation is reflected in a significant enhancement of the direct Coulombic interaction between D/A, V_s, which passes from 585 to 610 cm^{-1}. At the same time, however, the solvent-mediated term goes from -5 to -70 cm^{-1}, thus resulting in an overall reduction or screening of the total interaction. Thus, the MMpol model also provides physical insights on the solvent-induced changes on the coupling, which are not possible from the supermolecule calculations.

It is also worth mentioning that throughout the cutoff range considered, where the number of solvent molecules is increased from 21 to 159, the error in the estimated coupling stays relatively constant, close to $\sim$10 cm^{-1}, suggesting that the small differences observed between the two approaches could arise from short-range dispersion/repulsion effects neglected in the MMpol model. In addition, we want to note that at the distance considered here (R = 7 Å), the differences observed between MMpol and full QM results arise because of the different treatment of the environment and not because of the perturbative approach used to estimate the coupling. This has been checked by computing the exact supermolecule coupling also from QM/MMpol calculations, and the differences between the exact (eq 17) and the approximate values (eq 13) were always less than 1%.

On the other hand, the results obtained with the DFT-derived force field illustrate the importance of accurately describing solvent polarization in the estimation of the coupling. As noted above, this set of atomic polarizabilities describe a water molecular polarizability ($\alpha = 1.53$ Å^3) close to the experimental value ($\alpha_{exp} = 1.44$ Å^3), whereas the HF-derived set accounts for only one-half of it ($\alpha = 0.78$ Å^3). As a result, both the enhancement of the Coulombic coupling and the screening contribution induced by the solvent are strongly enlarged with these new polarizabilities, as reflected in Table 1. For the largest cutoff considered, the change involves 49 cm^{-1}, corresponding to a 10% change in the total coupling. Moreover, this change represents a 111% enlargement of the solvent-mediated contribution, which significantly modifies the solvent screening factor from $s = 0.89$ to 0.77. We note that the effect of the charges used to describe water is far less significant than the set of atomic polarizabilities. For instance, test calculations adopting the DFT set of polarizabilities but the charges of the POL3 force field induced only minor $\sim$1% changes in the total coupling.

It is also interesting to investigate the differential solvatochromic shifts on the donor and acceptor transition energies obtained from a QM and MMpol descriptions of the environment, as this has important consequences on the localization/delocalization character of the dimer excited states. As we consider a single solvent configuration in this section, the slightly different solvent structure surrounding D and A induces a mismatch in their transition energies. In Table 2 we show the transition energies obtained from the QM/MMpol and full QM calculations for the donor (ω_D) and acceptor (ω_A) as well as the corresponding energy difference between them (ω_D - ω_A). The results indicate that the MMpol model correctly describes the lowest-energy chromophore in all cases. In addition, the difference ω_D -ω_A is reproduced reasonably well, in particular when the HF-derived parameters consistent with the QM level of theory are used. This is illustrated by the fact that the observed errors in this latter case, about $\sim$0.01 eV, are reasonably small compared to the absolute magnitude of the solvatochromic shifts, which are 0.05 (0.08) eV and 0.07 (0.10) eV for the donor (acceptor) as given by the QM/MMpol and full QM calculations, respectively. These shifts are estimated from the reference vacuum transition energy, equal to 3.31 eV. The ability of the MMpol model to describe the relative fine-tuning of the transition energies induced by different local environments is in fact an important feature of the model, because in multichromophoric protein systems such local interactions can strongly modulate the EET pathways.[64]

4.2. Comparison between QM/MMpol and PCM. Once the QM/MMpol approach is validated with exact QM supermolecule results, we proceed to compare solvent effects on both transition properties and electronic couplings as described by the MMpol and PCM methods. In order to explore the distance-dependence behavior of solvent screening effects, two interchromophore distances, 3.5 Å and 7 Å, are considered. For this analysis we have used the B3LYP/aug-cc-pVTZ set of charges and polarizabilities, which accurately describe the aqueous polarizable environment as discussed in the previous section. In addition, results are averaged over 100 different solvent configurations extracted from the MD trajectories, which we found to be enough to obtain converged results. The standard deviation in transition energies, dipoles, and total electronic couplings was less than 2% in all cases.

Transition energies and transition dipole moments calculated in vacuo and in aqueous solution are reported in Table 3 for both QM/MMpol and PCM models at ZINDO, CIS, and TD-B3LYP levels. We note that due to the ensemble-averaging, equivalent properties are obtained for D and A, so a single set of values is reported.

Table 2. CIS/6-31G(d) Donor and Acceptor Transition Energies Calculated from the QM/MMpol Model and from Full QM Calculations of the Chromophore-Solvent System[b]

	MMpol (HF force field)			MMpol (B3LYP force field)			full QM		
cutoff (Å)[a]	ω_D	ω_A	ω_D - ω_A	ω_D	ω_A	ω_D - ω_A	ω_D	ω_A	ω_D -ω_A
7	3.28	3.27	0.01	3.26	3.26	0.01	3.27	3.25	0.01
8	3.28	3.26	0.02	3.24	3.24	0.01	3.26	3.23	0.03
9	3.27	3.24	0.04	3.23	3.20	0.03	3.26	3.21	0.05
10	3.27	3.23	0.03	3.22	3.19	0.03	3.25	3.21	0.05
11	3.26	3.23	0.03	3.21	3.19	0.02	3.24	3.21	0.04

[a] Increasing cutoff distances correspond to the consideration of 21, 46, 80, 108, and 159 number of water molecules in the system. [b] Results given as a function of the cutoff distance determining the number of waters included in the calculation. Two different sets of MMpol parameters (charges and atomic polarizabilities), derived at the HF/6-31G(d) and B3LYP/aug-cc-pVTZ level, are used to describe the MM environment. All energies are in eV.

Table 3. Transition Energies (in eV) and Transition Dipole Moments (in Debye) for PDI in Vacuum and in Aqueous Solution Calculated at the ZINDO, CIS/6-31G(d), and TD-B3LYP/6-31G(d) Levels from the QM/MMpol and PCM Methods

	ZINDO		CIS		TD-B3LYP	
	ΔE	μ^T	ΔE	μ^T	ΔE	μ^T
vacuum	2.60	10.9	3.31	9.9	2.43	8.5
MMpol	2.39	12.0	3.18	10.7	2.31	9.7
PCM	2.37	12.1	3.19	10.6	2.32	9.7

As expected from the $\pi-\pi^*$ nature of the excitation, a red shift is found passing from vacuum to aqueous solution (0.21, 0.13, and 0.12 eV for ZINDO, CIS and TD-B3LYP, respectively). Similar red shifts were obtained in a previous study of PDI in toluene.[33] The $\pi\rightarrow\pi^*$ character of the transition also explains the transition dipole moment orientation along the main longitudinal axis of the planar structure of the PDI, shown in Figure 1. The solvent does not affect the orientation of the transition dipole moment but increases its magnitude by about 10, 8, and 14% for ZINDO, CIS, and TD-B3LYP, respectively.

In Table 4 we show the ZINDO, CIS, and TD-B3LYP electronic couplings obtained in vacuo and in water with either the MMpol or the PCM models. The tables report the total value of the coupling and its two contributions, V_s and $V_{explicit}$ (see eqs 14 and 15). We recall that V_s contains different terms, Coulombic, exchange (eventually plus correlation in the case of TD-DFT), and overlap. In all cases overlap is negligible, whereas exchange(-correlation) terms represent ~5% of the total coupling in the CIS results at R = 3.5 Å, and ~1% in all other cases.

As we have shown in a previous study,[33] TD-B3LYP tends to underestimate the electronic coupling by about 20−25% when compared to more accurate SAC-CI results, while ZINDO presents a more unpredictable behavior depending on the particular system under study. We found that ZINDO coupling values for PDI are in good agreement with SAC-CI values. In the same study we have also found that the coupling values do not significantly depend on the chosen basis set. On the basis of such findings, we have limited here our analysis to the 6-31G(d) basis set; this in fact represents a good compromise between accuracy and computationally efficiency (we recall that each MMpol result implies an average on 100 calculations).

As expected, when the solvent is introduced we observe a net decrease of the coupling (about 25, 20, and 15% for ZINDO, CIS, and TD-B3LYP, respectively); this is due to the screening effect here quantified in terms of the factor s that relates the total coupling with the unscreened Coulombic contribution (eq 16). In addition, the solvent screening factor s decreases in all cases when the *D*/*A* separation is increased. At R = 3.5 Å, in fact, the solvent cannot penetrate between the two chromophores, and as a result the screening is reduced. Thus, the factor s is lower for the larger distance (0.72 and 0.67 for 3.5 and 7.0 Å, respectively for TD-B3LYP results).

Let us now compare MMpol with PCM. Concerning transition energies and dipole moments reported in Table 3, MMpol and PCM show very similar behaviors. This seems to indicate that in the present system, possible effects due to specific short-range solute−solvent interactions (such as hydrogen-bonding) do not significantly affect the transition properties and that an averaged picture as the one represented by the PCM is not only accurate enough but also realistic in terms of the description of the main solvent effects.

Moving to the coupling values, once again MMpol compare well with PCM values at all QM levels, with PCM results being always slightly smaller than MMpol. This behavior is due to larger screening contributions obtained by PCM with respect to MMpol (the Coulombic terms are in fact very similar in the two models). This is reflected in the factor s which is always smaller in the PCM description, that is, the screening effect of the solvent in PCM is always larger than in MMpol.

In order to better appreciate these similarities between MMpol and PCM, we have, however, to analyze some numerical aspects. In the PCM model we have to define the cavity embedding the molecular system; this is generally done by defining the cavity as an envelope of spheres centered on selected atoms. As in all continuum models, PCM results will depend on the parameters used to define such a cavity (in particular, the radii used for the spheres). To check how this dependency can affect the results presented in Table 4 we have repeated the PCM calculations using different cavities obtained by scaling the default radii by different factors, namely $f = 1.1$, 1.2, and 1.3. The results obtained for the coupling are reported in Table 5 (for this analysis only ZINDO and TDDFT are presented)

By comparing the results of Tables 4 and 5, two different effects of such a tuning of the cavity are observed at the

Table 4. ZINDO, CIS/6-31G(d), and TD-B3LYP/6-31G(d) Electronic Couplings and the Corresponding Screening Factors *s* Calculated for the PDI Dimer in Aqueous Solution from the QM/MMpol and PCM Methods[a]

	ZINDO				CIS/6-31G(d)				TD-B3LYP/6-31G(d)			
	V_s	$V_{explicit}$	V	s	V_s	$V_{explicit}$	V	s	V_s	$V_{explicit}$	V	s
						Distance 3.5 Å						
vacuum	1359		1359		1642		1642		1086		1086	
MMpol	1579	−543	1036 (−24)	0.65	1805	−457	1348 (−18)	0.75	1282	−356	926 (−15)	0.72
PCM	1603	−608	994 (−27)	0.62	1813	−504	1309 (−20)	0.72	1294	−390	904 (−17)	0.69
						Distance 7.0 Å						
vacuum	567		567		571		571		399		399	
MMpol	693	−257	436 (−23)	0.63	661	−215	446 (−22)	0.68	508	−168	340 (−15)	0.67
PCM	711	−280	431 (−24)	0.61	660	−228	432 (−24)	0.65	514	−182	332 (−17)	0.65

[a] The values in parentheses refer to percent variations with respect to vacuum. All couplings are in cm^{-1}.

Table 5. PCM EET Couplings Calculated for the PDI Dimer in Aqueous Solution at the ZINDO and TD-B3LYP/6-31G(d) Levels[a]

	ZINDO				TD-B3LYP/6-31G(d)			
f	V_s	$V_{explicit}$	V	s	V_s	$V_{explicit}$	V	s
			Distance 3.5 Å					
1.1	1563	−544	1020 (−25)	0.65	1264	−351	913 (−16)	0.72
1.2	1534	−489	1045 (−23)	0.68	1241	−317	924 (−15)	0.74
1.3	1510	−442	1068 (−21)	0.71	1223	−288	935 (−14)	0.76
			Distance 7.0 Å					
1.1	688	−265	424 (−25)	0.62	499	−172	327 (−18)	0.65
1.2	670	−250	420 (−26)	0.63	487	−158	324 (−19)	0.67
1.3	656	−243	413 (−27)	0.63	476	−150	319 (−20)	0.67

[a] The three sets of data correspond to three different cavities (see text for details). All couplings are in cm^{-1}.

two interchromophore distances. For the distance of 3.5 Å, the total coupling increases with the factor *f*, while for the distance of 7.0 Å, the total coupling decreases when the factor *f* is larger. To understand this difference we have to recall that, for the 3.5 distance, the PCM calculations are done with a single cavity that embeds the two monomers; there is no solvent between the molecules. So, when the factor *f* increases, both coupling terms decrease but the screening one to a greater extent. In other words, by enlarging the cavity there is a smaller enhancement of the Coulombic interaction upon solvation, but the decrease of screening effects is even greater so that the balance between both contributions leads to a small overall increase in the coupling. It seems that the best agreement in the total coupling between PCM and MMpol is obtained for *f* equal to 1.2; however, if we analyze the single components and the screening factor *s*, the best agreement is found for $f = 1.1$. On the other hand, for the distance of 7.0 Å, we have two separated cavities for the two monomers so we can expect that the dominant changes at large *f* are in the Coulombic contribution and not in the screening one as found for the short distance. In addition, at this distance the solvent screening factor is significantly less sensitive to the particular definition of the PCM cavity through the factor *f*.

Finally, we discuss in some detail the computational cost associated with the introduction of solvent effects through the PCM and MMpol methods. At the CIS/6-31G(d) level, the total CPU time associated with the calculation of the coupling relative to the vacuum calculation was 1.4/3.2 (R = 3.5 Å) and 1.5/8.6 (R = 7.0 Å) for MMpol/PCM, respectively. This illustrates the fact that the cost associated with the QM/MMpol method is quite insensitive to the D−A separation, as the number of induced dipoles considered is kept relatively constant. In contrast, the cost associated with the PCM calculation increases substantially when passing to R = 7.0 Å, because in this case two different cavities host the chromophores, thus increasing the number of apparent surface charges displaced on the cavity surface. This is because in both methods, the added computational cost is mainly originated by the matrix inversion step needed to obtain the **B** and **K** matrices in eqs 6 and 7, respectively. Note also that if higher QM levels of theory are used, the relative cost added to the vacuum calculation is expected to be smaller. We also remark that the above timings refer to a single QM/MMpol calculation, whereas in general one has to perform a proper ensemble-average over several solute−solvent configurations. In this work, we considered 100 solute−solvent structures, and coupling values (transition energies) fluctuated over a ∼40 cm^{-1} (∼50 meV) range in the above-mentioned CIS calculations. Nevertheless, convergence in this system was very fast, and averaging over 25 structures already gave results converged below 1 cm^{-1} (2 meV).

To summarize, we find that the QM/MMpol method describes solvent effects both on transition properties and on the coupling in a very similar way to PCM despite the completely different description of the solvent characterizing the two models. Moreover, it is very remarkable that the *change* in the solvent screening factor when the interchromophoric distance is enlarged from 3.5 Å to 7.0 Å is almost the same as predicted by a continuum dielectric and an explicit discrete representation of the environment, a finding which strongly supports the distance-dependent screening function we have recently proposed based on PCM calculations.[22,23]

5. Conclusions and Perspectives

We have presented a novel polarizable QM/MM method to study EET in condensed phase. The method is based on a linear response approach, and it has been implemented at the semiempirical (ZINDO), Hartree−Fock (CIS), and density functional theory (TD-DFT) levels. This approach allows an atomistic description of environment effects on all quantities determining EET, i.e. chromophores' transition energies and dipoles, and on the electronic couplings. The method has been tested on a model PDI dimer in water

solution; we have found an excellent agreement between the QM/MMpol method and "exact" supermolecule calculations in which the complete solute−solvent system is described at the QM level. Our results also indicate that the estimation of the electronic coupling is extremely sensitive to the treatment of the solvent polarization and that an accurate set of parameters for the polarizable force field is necessary. On the other hand, the accuracy of the results predicted using the PCM model are found to be very sensitive to the exact shape and size of the molecular cavity imposing in this respect the problem of defining the most physically correct cavity. Finally, we have compared QM/MMpol results averaged over solvent configurations sampled from MD simulations to the corresponding ones obtained using the PCM-LR method, which is based on a continuum dielectric description of the solvent. We have shown that both continuum and atomistic solvent models describe similar solvent effects on EET in homogeneous media such as water. Most notably, both approaches describe a consistent decay of solvent screening as a function of donor−acceptor separation. This latter finding strongly supports the empirical distance-dependent screening function we recently derived from PCM-LR calculations.[22,23]

All these results demonstrate the reliability and robustness of the polarizable QM/MM method, and they make us confident in its potential to investigate the role of complex heterogeneous environments on EET, e.g. proteins or nanostructured host materials, where a continuum description of the environment represents an important limitation.

Acknowledgment. The work in Toronto was supported by the Natural Sciences and Engineering Research Council of Canada. G.D.S. acknowledges the support of an E. W. R. Steacie Memorial Fellowship. A.M.L. thanks support from the Spanish Ministerio de Ciencia e Innovación (Programa Nacional de Recursos Humanos del Plan Nacional I-D+I 2008-2011). J.K. thanks the Danish Natural Science Research Council/The Danish Councils for Independent Research and the Villum Kann Rasmussen foundation for financial support. B.M. wishes to thank Gaussian Inc. for financial support.

References

(1) Scholes, G. D. *Annu. Rev. Phys. Chem.* **2003**, *54*, 57.

(2) Fleming, G. R.; Scholes, G. D. *Nature* **2004**, *431*, 256.

(3) Sundstrom, V.; Pullerits, T.; van Grondelle, R. *J. Phys. Chem. B* **1999**, *103*, 2327.

(4) van Amerongen, H.; Valkunas, L.; van Grondelle, R. *Photosynthetic Excitons*; World Scientific Publishers: Singapore, 2000.

(5) Gust, D.; Moore, T. A.; Moore, A. L. *Acc. Chem. Res.* **2001**, *34*, 40.

(6) Jolliffe, K. A.; Bell, T. D. M.; Ghiggino, K. P.; Langford, S. J.; Paddon-Row, M. N. *Angew. Chem., Int. Ed.* **1998**, *37*, 915.

(7) Balzani, V.; Campagna, S.; Denti, G.; Juris, A.; Serroni, S.; Venturi, M. *Acc. Chem. Res.* **1998**, *31*, 26.

(8) Holten, D.; Bocian, D. F.; Lindsey, J. S. *Acc. Chem. Res.* **2002**, *35*, 57.

(9) Lee, J.-I.; Kang, I.-N.; Hwang, D.-H.; Shim, H.-K.; Jeoung, S. C.; Kim, D. *Chem. Mater.* **1996**, *8*, 1925.

(10) List, E. J. W.; Holzer, L.; Tasch, S.; Leising, G.; Scherf, U.; Müllen, K.; Catellani, M.; Luzzati, S. *Solid State Commun.* **1999**, *109*, 455.

(11) Wang, H.-L.; McBranch, D. W.; Klimov, V. I.; Helgeson, R.; Wudl, F. *Chem. Phys. Lett.* **1999**, *315*, 173.

(12) Brédas, J. L.; Beljonne, D.; Coropceanu, V.; Cornil, J. *Chem. Rev.* **2004**, *104*, 4971.

(13) Jares-Erijman, E.; Jovin, T. M. *Nat. Biotechnol.* **2003**, *21*, 1387.

(14) Lippincott-Schwartz, J.; Snapp, E.; Kenworthy, A. *Nat. Rev. Mol. Cell. Biol.* **2001**, *2*, 444.

(15) Weiss, S. *Nat. Struct. Biol.* **2000**, *7*, 724.

(16) Förster, T. *Ann. Phys.* **1948**, *2*, 55.

(17) Beljonne, D.; Curutchet, C.; Scholes, G. D.; Silbey, R. J. *J. Phys. Chem. B* **2009**, *113*, 6583.

(18) Krueger, B. P.; Scholes, G. D.; Fleming, G. R. *J. Phys. Chem. B* **1998**, *102*, 5378.

(19) Beljonne, D.; Cornil, J.; Silbey, R.; Millie, P.; Bredas, J. L. *J. Chem. Phys.* **2000**, *112*, 4749.

(20) Wong, K. F.; Bagchi, B.; Rossky, P. J. *J. Phys. Chem. A* **2004**, *108*, 5752.

(21) Beenken, W. J. D.; Pullerits, T. *J. Chem. Phys.* **2004**, *120*, 2490.

(22) Scholes, G. D.; Curutchet, C.; Mennucci, B.; Cammi, R.; Tomasi, J. *J. Phys. Chem. B* **2007**, *111*, 6978.

(23) Curutchet, C.; Scholes, G. D.; Mennucci, B.; Cammi, R. *J. Phys. Chem. B* **2007**, *111*, 13253.

(24) Pullerits, T.; Freiberg, A. *Chem. Phys.* **1991**, *149*, 409.

(25) Pullerits, T.; Hess, S.; Herek, J. L.; Sundstrom, V. *J. Phys. Chem. B* **1997**, *101*, 10560.

(26) Scholes, G. D.; Jordanides, X. J.; Fleming, G. R. *J. Phys. Chem. B* **2001**, *105*, 1640.

(27) Sumi, H. *J. Phys. Chem. B* **1999**, *103*, 252.

(28) Jang, S.; Newton, M. D.; Silbey, R. J. *Phys. Rev. Lett.* **2004**, *92*, 218301.

(29) Engel, G. S.; Calhoun, T. R.; Read, E. L.; Ahn, T.-K.; Mancal, T.; Cheng, Y.-C.; Blankenship, R. E.; Fleming, G. R. *Nature* **2007**, *446*, 782.

(30) Lee, H.; Cheng, Y.-C.; Fleming, G. R. *Science* **2007**, *316*, 1462.

(31) Iozzi, M. F.; Mennucci, B.; Tomasi, J.; Cammi, R. *J. Chem. Phys.* **2004**, *120*, 7029.

(32) Madjet, M. E.; Abdurahman, A.; Renger, T. *J. Phys. Chem. B* **2006**, *110*, 17268.

(33) Muñoz-Losa, A.; Curutchet, C.; Fdez. Galván, I.; Mennucci, B. *J. Chem. Phys.* **2008**, *129*, 034104.

(34) Curutchet, C.; Mennucci, B. *J. Am. Chem. Soc.* **2005**, *127*, 16733.

(35) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.

(36) Cohen, B. E.; McAnaney, T. B.; Park, E. S.; Jan, Y. N.; Boxer, S. G.; Jan, L. Y. *Science* **2002**, *296*, 1700.

(37) Golosov, A. A.; Karplus, M. *J. Phys. Chem. B* **2007**, *111*, 1482.

(38) King, G.; Lee, F. S.; Warshel, A. *J. Chem. Phys.* **1991**, *95*, 4366.

(39) Smith, P. E.; Brunne, R. M.; Mark, A. E.; Van Gunsteren, W. F. *J. Phys. Chem.* **1993**, *97*, 2009.

(40) Pitera, J. W.; Falta, M.; van Gunsteren, W. F. *Biophys. J.* **2001**, *80*, 2546.

(41) Simonson, T.; Brooks, C. L. *J. Am. Chem. Soc.* **1996**, *118*, 8452.

(42) Fidder, H.; Knoester, J.; Wiersma, D. A. *J. Chem. Phys.* **1991**, *95*, 7880.

(43) Scholes, G. D.; Fleming, G. R. *J. Phys. Chem. B* **2000**, *104*, 1854.

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

(45) Osted, A.; Kongsted, J.; Mikkelsen, K. V.; Astrand, P.-O.; Christiansen, O. *J. Chem. Phys.* **2006**, *124*, 124503.

(46) Nielsen, C. B.; Christiansen, O.; Mikkelsen, K. V.; Kongsted, J. *J. Chem. Phys.* **2007**, *126*, 154112.

(47) Öhrn, A.; Karlström, G. *Mol. Phys.* **2006**, *104*, 3087.

(48) Lin, Y.-l.; Gao, J. *J. Chem. Theory Comput.* **2007**, *3*, 1484.

(49) Muñoz-Losa, A.; Fdez. Galván, I.; Aguilar, M. A.; Martín, M. E. *J. Phys. Chem. B* **2007**, *111*, 9864.

(50) Casida, M. E. *In Recent Advances in Density Functional Methods*; Chong, D. P., Ed.; World Scientific: Singapore, 1995; Part I.

(51) Cammi, R.; Mennucci, B. *J. Chem. Phys.* **1999**, *110*, 9877.

(52) Hsu, C.-P.; Fleming, G. R.; Head-Gordon, M.; Head-Gordon, T. *J. Chem. Phys.* **2001**, *114*, 3065.

(53) Tretiak, S.; Middleton, C.; Chernyak, V.; Mukamel, S. *J. Phys. Chem. B* **2000**, *104*, 4519.

(54) Caldwell, J. W.; Kollman, P. A. *J. Phys. Chem.* **1995**, *99*, 6208.

(55) Case, D. A.; Darden, T. A.; Cheatham, T. E., III.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*, 9th ed.; University of California: San Francisco, 2006.

(56) Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384.

(57) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.

(58) Ponder, J. W.; Case, D. A.; Valerie, D. *Adv. Protein Chem.* **2003**, *66*, 27.

(59) Gagliardi, L.; Lindh, R.; Karlström, G. *J. Chem. Phys.* **2004**, *121*, 4494.

(60) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neograd y, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222.

(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.; BurantJ. 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.

(62) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024.

(63) Magyar, R. J.; Tretiak, S. *J. Chem. Theory Comput.* **2007**, *3*, 976.

(64) Müh, F.; Madjet, M. E.-A.; Adolphs, J.; Abdurahman, A.; Rabenstein, B.; Ishikita, H.; Knapp, E.-W.; Renger, T. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 16862.

CT9001366

JCTC Journal of Chemical Theory and Computation

On the Potential Use of Squaraine Derivatives as Photosensitizers in Photodynamic Therapy: A TDDFT and RICC2 Survey

Angelo Domenico Quartarolo,[†] Emilia Sicilia,[†] and Nino Russo*,[†]

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 April 23, 2009

Abstract: A time-dependent density functional theory (TDDFT) and the second-order approximated coupled-cluster model with the resolution of identity approximation (RICC2) studies are reported here for some classes of squaraine derivatives. These compounds have a sharp electronic band, ranging from the visible to near-red part of the spectrum, with an high molar absorption coefficient. These features make them potential photosensitizers in the photodynamic therapy of cancer (PDT), in which a light source, a photosensitizer, and molecular oxygen ($^3O_2$) are combined to give cytotoxic singlet oxygen ($^1O_2$) as a final result in a photochemical process. For the examined structures, the introduction of different substituents (electron donating, electron withdrawing, or fused rings) in the parent molecule, in order to give different squaraine derivatives, changes the maximum absorption wavelength (λ_{max}) from 620 to 730 nm, giving a near-red absorbing photosensitizer that can better penetrate human tissue to damage tumor cells. Theoretical results, obtained from both TDDFT/PBE0 and RICC2, are able to reproduce qualitatively the substitution effect on λ_{max}, resulting in a useful tool for testing different structure modifications and, in general, for the molecular design of PDT photosensitizers. Calculated vertical excitation energies (singlet−singlet transitions) generally agree with experimental data within 0.3 eV. The singlet oxygen generation ability of these compounds requires that their triplet energy, for a type II reaction mechanism, should be greater than 0.98 eV. Theoretical triplet energies from the RICC2 method suggests that this requisite is fulfilled for all compounds, though the results are generally overestimated with respect to experiment by 0.7 eV, whereas TDDFT/PBE0 triplet energies, which are underestimated within 0.2 eV in few cases, lie close to the above-mentioned limit and can be considered suitable for PDT applications.

1. Introduction

Squaraine dyes are a class of organic compounds derived from the 1,3-condensation reaction between squaric acid and electron-rich compounds and are characterized by a sharp and intense electronic absorption band in the near-red part of the visible region (600−700 nm).[1,2] Recently, these compounds have been investigated for use in many research areas, for example, technological applications, such as dye-sensitized solar cells,[3] optical storage devices,[4] and fluorescent probes for detecting metal ions[5,6] and in photosensitizer drugs in photodynamic therapy of cancer (PDT).[7−9] This latter application is a noninvasive medicine treatment for different tumoral diseases, like skin or bladder cancer, or for psoriasis and age-related macular degeneration.[10−14] The basic principle of PDT is given by an appropriate combination of a light source, a chemical dose containing photosensitizer molecules and dioxygen, that is largely present in the human cell environment and can promote selective cellular

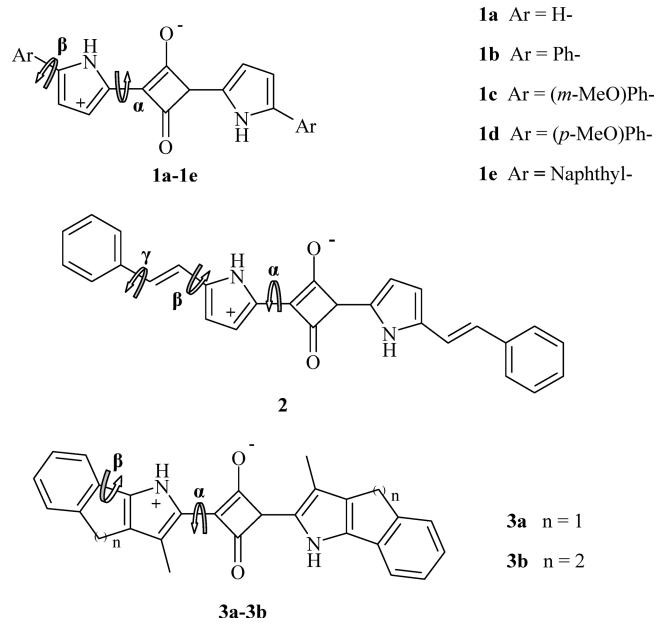
* Corresponding author: Telephone: +39-0984-492048. Fax: +39-0984-492044. E-mail: nrusso@unical.it.

[†] Università della Calabria.

10.1021/ct900199j CCC: $40.75
Published on Web 06/25/2009

damage through irradiation.[15] There are two reaction mechanisms through which the photosensitizer drug can produce a cytotoxic effect against cancer cells.[16−19] The first pathway, the so-called type I mechanism, involves radical oxygen species generated, for example, from an electron transfer between a photosensitizer in an excited state and an organic substrate followed by the interaction with dioxygen. In the second mechanism, or type II mechanism, the photosensitizer is promoted by irradiation to its first singlet excited state (S_1) and then a fast depletion from this state, via an intersystem spin crossing decay, to the first excited triplet state (T_1) occurs. An energy transfer process can occur between the photosensitizer T_1 state and ground-state molecular oxygen ($^3O_2$) leading to the formation of singlet molecular oxygen ($^1O_2$), which represents the final cytotoxic agent.[20] One of the main features for an optimal molecule to act as a type II PDT drug is the presence of an intense electronic absorption band in its spectrum, falling inside the therapeutic window (600−900 nm), where tissue penetration by light is greater. This goal is usually achieved by further extending the electronic delocalization of π-molecular systems and/or by introducing suitable functional groups in order to reduce the HOMO−LUMO energy gap and shifting the electronic absorption band in the near-red part of the visible spectrum.[21] On the other hand, the efficiency of the intersystem spin crossing mechanism is enhanced by the presence of heavy atoms, for example, bromine or transition metals, that as a result of spin−orbit effects increase the triplet quantum yield. Moreover, other important experimental factors affect the efficiency of a PDT photosensitizer such as high singlet oxygen quantum yield, long triplet state lifetime, and water solubility, as they have to work in biological systems, low dark toxicity, and preferential localization in the tumoral tissue.[15] Photosensitizers, currently investigated for PDT applications, belong mainly to the class of porphyrin-like systems (e.g., expanded porphyrins, phthalocyanine, and porphycene derivatives)[22] or, in part, to nonporphyrin systems (e.g., psoralens, and phenothiazines dyes).[23] Photofrin, a porphyrin derivative, has been approved in many countries for the treatment of early stage lung cancer.[24,25] Currently, other polypyrrolic macrocycles as lutetium texaphyrin (Lutex),[26] an expanded metal porphyrin-like molecule, or a benzoporphyrin derivative (Verteporphyrin)[27] are in different stages of clinical trials. Recently, many studies have regarded the synthesis and photochemical characterization of new nonporphyrin systems for PDT application. Some examples are given by difluoro-boron(III) dipyrromethenes,[28,29] green perylenediimides,[30] and squaraine dyes.[31] Previous works regarding the synthesis, photophysical properties, and in vitro biological studies of halogenated (brominated and iodinated) squaraine dyes by Ramaiah et al. proved the DNA damage through singlet oxygen generation and that it takes place, for the nonhalogenated form, through the type I mechanism.[32] For these compounds, which exist in solution as either neutral or protonated forms depending on the pH, the maximum absorbance wavelength falls in the range of 500−600 nm. Moreover, the influence of heavy atom on singlet oxygen generation has been studied also, for example, in works concerning squarilium cyanine dyes containing sulfur and selenium[33] or in iodinated squaraine-rotaxanes derivatives.[34] In this theoretical work we will focus our attention on two series of symmetrical squaraine derivatives (Scheme 1 and 2), synthesized in the past few years by Bonnett et al.[35] and Beverina et al.[36] These compounds have been specifically designed to further red-shift the maximum absorption wavelength by the addition of different functional groups and to increase the solubility in water.[36] Structures and electronic spectra of these two series of compounds have been calculated by means of density functional theory (DFT) and its time-dependent formulation (TDDFT)[37] as well as by the second-order approximated coupled-cluster model with the resolution of identity approximation (RICC2).[38] The main aim of this work was to reproduce the electronic band changes as a function of the substituent groups and to evaluate the triplet energy of each molecule, which is a basic requirement in order to consider it as a type-II PDT photosensitizer.

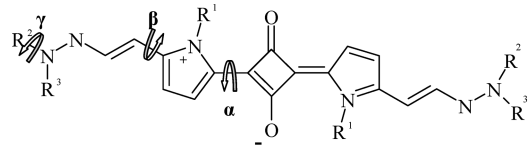
Scheme 1. Two-Dimensional Plot for Structures 1a−e, 2, and 3a−b



2. Computational Details

The TURBOMOLE V5.10 software package has been used for the TDDFT and RICC2 calculations.[39] Geometry optimizations, without imposing symmetry constraints, as well as vibrational frequency analysis were carried out at the density functional level of theory in conjunction with the nonempirical PBE0 hybrid functional that adds up a fixed amount of Hartree−Fock exchange energy (25%) to the gradient corrected PBE exchange correlation functional.[40,41] The split valence basis set plus polarization functions (SVP) of Ahlrichs et al.[42] was used for all atoms for the structure optimizations and vibrational frequency analysis. Vertical excitation energies were calculated by means of two different methodologies: time-dependent density functional linear response theory (TD-DFRT)[38,43] and RI-CC2.[38,44,45] In both cases, singlet and triplet vertical transitions for each molecule were obtained starting from the PBE0/SVP optimized

Scheme 2. Two-Dimensional Plot of Structures 4a−c, 5a−b, and 6

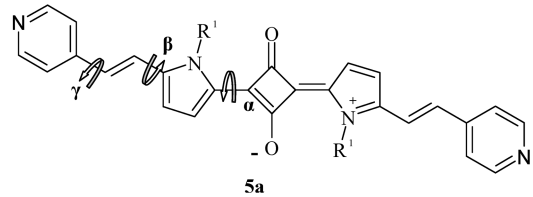


4a $R^1 = CH_3$ $R^2 = R^3 = Ph$

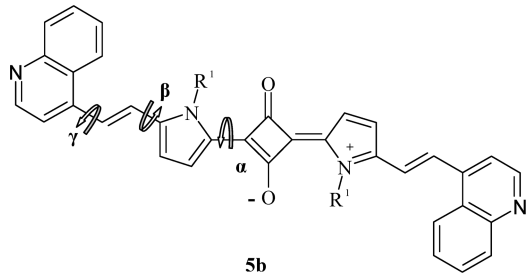
4b $R^1 = CH_3$, $R^2 = R^3 = CH_3$

4c $R^1 = CH_3$, $R^2 = H$, $R^3 = p$-Br-Ph

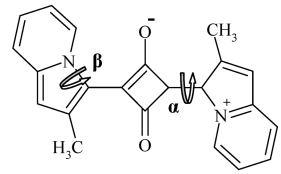
4a-4c



5a



5b



6

structures. In order to assess the influence of the chosen basis set upon excitation energies for TD-DFT calculations, different quality basis sets, with an increasing number of basis functions, were initially tested on molecule 6. For this aim the following basis sets were employed: split valence basis sets without and with polarization functions on hydrogen atoms (SV(P) and SVP),[42] double- and triple-ζ valence basis set plus one (DZP and TZVP)[46] and two polarization functions for each atom (TZVPP)[46] and the correlated consistent polarized valence double and triple-ζ basis sets of Dunning et al. (cc-pVDZ, cc-pVTZ).[47] The latter have been employed also with the addition of s and p diffuse functions (aug-cc-pVDZ and aug-cc-pVTZ). The reliability of excitation energies of organic and inorganic dyes, obtained from the adopted split valence basis set (SVP) and hybrid functional (PBE0), has been also proved in recent works and yields a mean absolute error (MAE) within 0.3−0.4 eV.[48−51] For RICC2 singlet and triplet excitation energy calculations, only SVP and TZVP basis sets have been tested, since the basis sets increasing size becomes more computationally demanding. In this case, single point calculations were made on PBE0/SVP optimized geometries. The influence of solvent effects on geometries and excitation energies has been estimated with the COSMO (conductor-like screening model) approach,[52,53] where the solute molecule is embedded within a dielectric of permittivity ε, that represents the solvent. The inclusion of bulk solvent effects can quantitatively improve excitation energies, though in vacuo results qualitatively account for the observed experimental trends. The dielectric constant values of chloroform ($\varepsilon = 4.9$, for compounds 4a−c, 5a−b and 6) and dichloromethane ($\varepsilon = 8.93$, for compounds 1a−e, 2, 3a−b) and default parameters for the cavity generation have been used for solvent calculations.

3. Results and Discussion

3.1. Structures. Two classes of squaraine derivatives have been studied in this work. In the first series (see Scheme 1), synthesized and characterized by Bonnett et al.,[35] suitable substituent groups are introduced in the position 2 of the pyrrole moieties in order to red-shift the λ_{max} with respect to the parent molecule 1a. The reported molecules include electron-donating phenyl (1b), phenyl substituted or naphthyl groups (1c, 1d and 1e), styryl substituent containing a phenyl terminal group (2a), and saturated 5- and 6-term rings between the benzenoid and pyrrole parts (3a−b). In the second series, we have considered the squaraine derivatives (see Scheme 2), synthesized by Beverina et al.,[36] where the pyrrole parts were functionalized by arylhydrazone groups (4a−c) or, as for the case of 2a, by extending the π conjugation system with heteroaromatic rings, specifically a pyridine molecule for 5a and a quinoline ring in 5b, and last through an indolizine ring attached directly to the squaric core structure (6). All reported structures (see Figures 1 and 2 in the Supporting Information) have been considered to assume an *anti* geometry configuration on the basis of

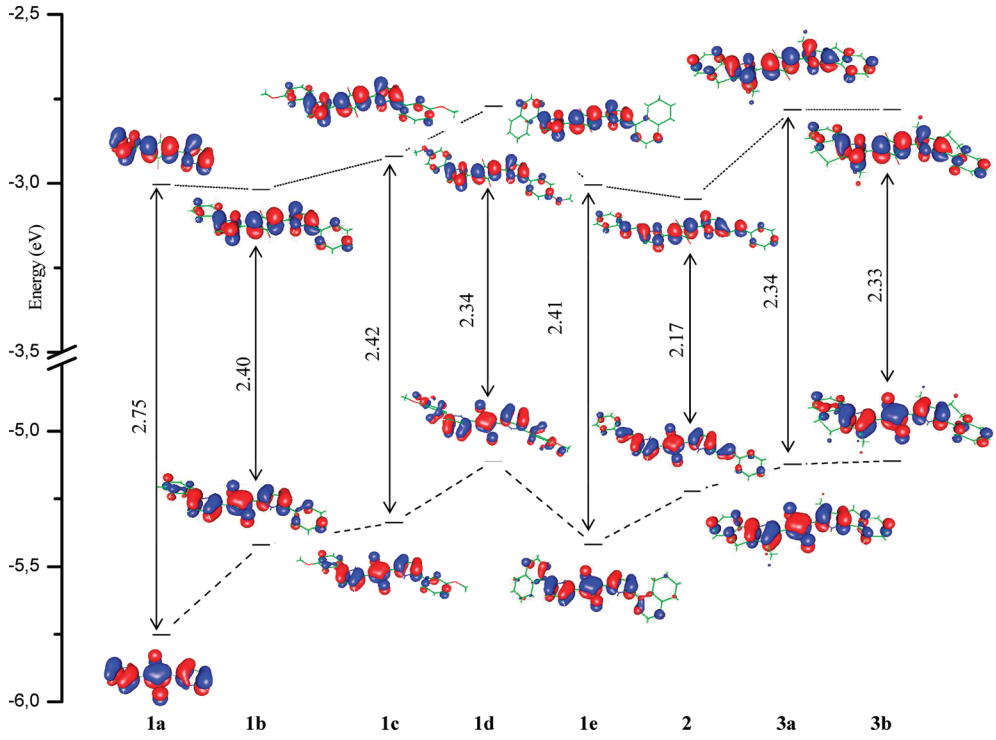


Figure 1. HOMO and LUMO isodensity molecular surfaces and energy level diagram for 1a−e, 2 and 3a−b.

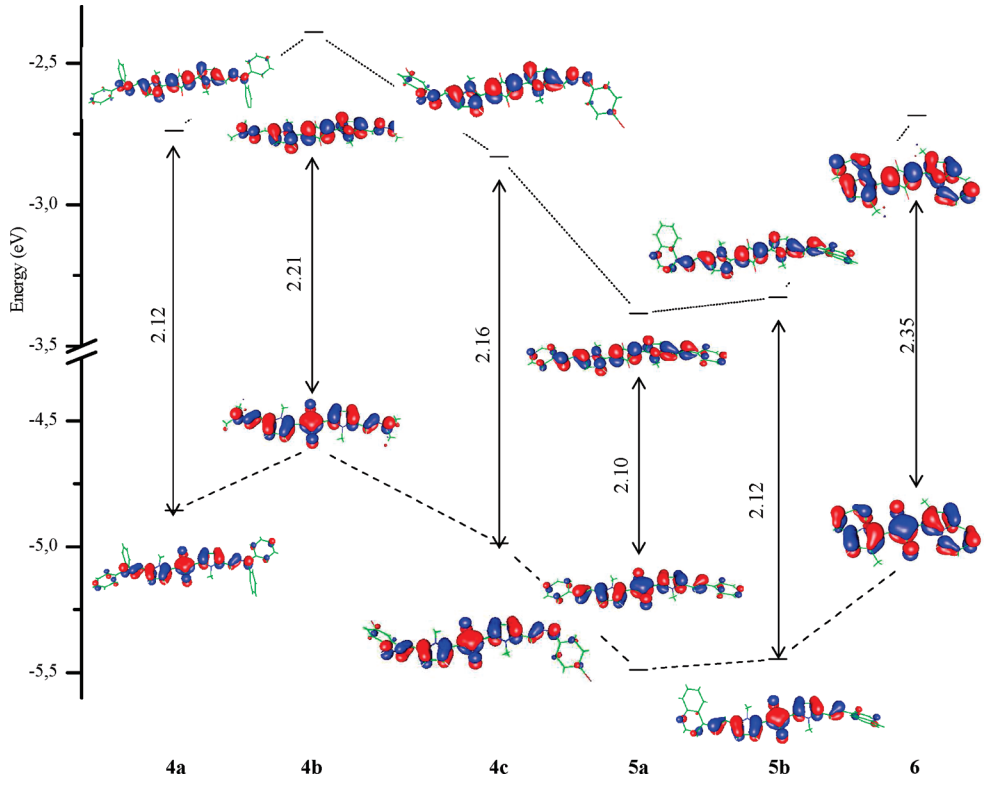


Figure 2. HOMO and LUMO isodensity molecular surfaces and energy level diagram for 4a−c, 5a−b and 6.

preliminary calculations on 1a and 6 compounds. In fact, after geometry optimizations, this conformation resulted to be energetically more stable than the corresponding *syn* geometries, even if by few kcal/mol energy units (including zero-point vibrational correction energies). The stabilization is mainly due to two reasons: (a) the formation of two intramolecular hydrogen bonds between the pyrrole hydrogen atoms and the oxygen atoms; and (b) the steric hindrance effects. The possible deviation from planarity of all compounds has been analyzed by defining three main dihedral angles (α, β, and γ) as reported in Schemes 1 and 2. The dihedral angle α connects the pyrrole ring to the squaric part through the carbon−carbon bond, which showed theoretically a partial double character (∼1.38−1.40 Å). The dihedral angle β represents the distortion between the substituent group at position 2 of the pyrrole ring and the ring itself or,

Table 1. Dihedral Angles α, *β*, *γ*, as Indicated in Schemes *1* and *2*

molecule	dihedral angles α	*β*	*γ*
1a	0.0	–	–
1b	0.2	14.4	–
1c	0.1	17.0	–
1d	0.1	13.2	–
1e	0.6	37.7	–
2	1.0	3.5	4.2
3a	0.9	0.2	–
3b	0.5	12.4	–
4a	1.5	176.1	95.2 (5.8)[a]
4b	1.5	179.3	5.0
4c	2.3	174.5	37.1
5a	1.1	176.1	4.9
5b	0.5	171.2	35.2
6	0.0	0.0	–

[a] Dihedral angles for the two phenyl groups.

in the case of the condensed structures (3a−b and 6), the degree of coplanarity of the fused rings. For structures 2 and 5a−b, the elongation of the electronic delocalization through the formation of carbon−carbon or nitrogen−carbon bonds with the presence of a terminal heteroaromatic ring gives rise to another degree of conformational freedom represented by the dihedral *γ*. The dihedral angle values (α, *β*, and *γ*) for each molecule are summarized in Table 1. In all compounds, the squaric core part is nearly coplanar with the *N*-substituted pyrrole ring (α value is lesser than 3°). The dihedral *β* angle depends on the nature of the substitution on the pyrrole rings. For compounds 1b−e, which a phenyl (substituted) or naphthyl ring is bound to the pyrrole one, the deviation ranges between 13° and about 40° (Table 1), depending on the different steric hindrance of the substituent group. For compounds reported in Scheme 2 and compound 2, which present a chain elongation starting from the position 2 of the pyrrole ring, the *β* value (−N−C−C−N− or −N−C−C−C−) denotes a planar conformation for this part of the molecules. Condensed structures are strictly coplanar except for compound 3b where the cyclohexenyl ring has a *β* value equal to 12.4°. Further substitution along the molecular chain, represented by the dihedral angle *γ*, is relevant in the case of hindered substituent groups (4c and 5b, *γ* values ca. 37° and 35°) or in situations where there are two adjacent phenyl groups, as is the case of 4a, which has two different *γ* values (95° and 6°) in order to minimize sterical repulsion between phenyl rings. The nature of the substituent groups as well as distortion from planarity can influence the electronic delocalization and, as a consequence, the λ_{max} value, as shown in the next paragraph.

3.2. Electronic Spectra. The electronic spectra of 2,4-bis-pyrrolyl squaraine derivatives are characterized, in the visible part of the electromagnetic spectrum, by a sharp and intense absorption band due to the *π* electronic system, which is highly delocalized. Formally the core part of these compounds can be described by three resonance structures with 12 *π*-electrons and a positive charge delocalized through the two symmetrical moieties. For all compounds UV−vis spectra are available;[35,36] moreover, for the compounds reported in Scheme 2, singlet oxygen quantum yield measurements and two-photon spectroscopical studies were also made.[31,36] In this section, we will outline the results obtained by TDDFT calculations of the λ_{max} in vacuo and in solvent environments and outline how structural changes are theoretically reproduced in comparison to the experimental behavior, also using more refined methodology as coupled-cluster methods. The basis set influence on λ_{max} was tested for compound 6, the results are reported in Table 2 and are relative to PBE0 calculations. It is evident that the increasing size of the basis set improves little the agreement with the experimental λ_{max} (684 nm), even by using the augmented triple-*ζ* basis set of Dunning et al.[47] In terms of eV units, the benefit is 0.04 and is well below the TDDFT method error deviation, which is typically within 0.4 eV.[48−50] So a good compromise between computational timings and numerical reproduction of λ_{max} can be obtained by just employing a SVP. Similar consideration can be made for the influence of the basis set on λ_{max} for the RICC2 methodology. In fact, from Table 2, it is evident that with this method the SVP basis sets also give results that differ by about 10 nm from that obtained by using the triple-*ζ* valence basis set. The excitation energies (in eV and nm) and oscillator strengths for the two series of squaraine derivatives (Scheme 1 and 2) obtained at both TDDFT and RICC2 level of theory in vacuo and in the presence of the solvent (TDDFT only) are reported in Table 3. All structures showed one excitation energy in the visible region composed mainly by a transition from the highest occupied molecular orbital (HOMO) to the lowest unoccupied one (LUMO). We will first analyze in vacuo TDDFT results. Considering the compound 1a as a reference structure for the evaluation of the λ_{max} shift as a function of substituents, it can be noted that the introduction of an electron-donating group (1b−e) increases the λ_{max} by 80−90 nm. Comparing the two isomeric forms 1c and 1d, the latter resulted more wavelength red-shifted by 18 nm as can be confirmed also by the experimental difference between the two isomers. This different behavior can be explained taking into account the energetic trend and the isodensity electron density surfaces of the HOMO (*π* orbital character) and LUMO (*π** orbital character) orbitals (see Figure 1). The molecular orbitals considered are those mainly involved in the electronic transition that gives rise to the maximum wavelength absorption band.

Table 2. Basis Set Influence for the First Excitation Energy Δ*E* (eV, nm) and Oscillator Strength *f* of Compound 6 from TDDFT and RICC2 Calculations

basis set	TDDFT[a] Δ*E*, eV	Δ*E*, nm	*f*	RICC2[a] Δ*E*, eV	Δ*E*, nm	exptl.[b] Δ*E*, eV	Δ*E*, nm
SV(P)	2.22 (2.08)	559	0.9762	2.08	596	1.81	684
SVP	2.22 (2.08)	559	0.9770	2.08	597		
DZ	2.20	563	0.9741				
DZP	2.21	560	0.9685				
TZVP	2.20	564	0.9736	2.04	607		
TZVPP	2.20	565	0.9579				
cc-pVDZ	2.22	559	0.9635	2.08	595		
cc-pVTZ	2.20	563	0.9585	2.05	605		
aug-cc-pVDZ	2.19	567	0.9560	2.04	609		
aug-cc-pVTZ	2.18	568	0.9548	2.02	612		

[a] Single point calculations from PBE0/SVP optimized structures. [b] In dichloromethane; from reference 36.

Table 3. Calculated Singlet Excitation Energies ΔE (eV and nm (in parentheses)) and Oscillator Strength f for All Studied Compounds[a]

	TDDFT,[b] vacuum		TDDFT,[b] c-pcm		RICC2,[b] vacuum	
molecule	ΔE	f	ΔE	f	ΔE	exptl.[c]
1a	2.73 (454)	0.832	2.78 (445)	0.909	2.61 (475)	/
1b	2.33 (533)	1.578	2.35 (528)	1.714	2.31 (537)	2.00 (621)
1c	2.33 (531)	1.661	2.35 (527)	1.792	2.31 (537)	1.98 (625)
1d	2.26 (549)	1.835	2.26 (548)	1.983	2.24 (552)	1.93 (643)
1e	2.27 (546)	1.701	2.28 (543)	1.803	2.29 (542)	2.02 (613)
2	2.12 (586)	2.368	2.14 (580)	2.526	2.16 (574)	1.84 (673)
3a	2.28 (544)	1.849	2.31 (537)	1.950	2.23 (555)	1.90 (654)
3b	2.25 (552)	1.746	2.27 (546)	1.840	2.22 (559)	1.88 (660)
4a	2.04 (609)	2.619	2.03 (611)	2.750	2.01 (616)	1.70 (728)
4b	2.24 (553)	2.029	2.23 (556)	2.228	2.18 (569)	1.80 (688)
4c	2.09 (593)	2.337	2.09 (593)	2.506	2.11 (586)	1.73 (717)
5a	2.07 (600)	2.356	2.12 (586)	2.516	2.13 (581)	1.83 (678)
5b	2.03 (612)	2.376	2.07 (600)	2.520	2.12 (585)	1.80 (688)
6	2.22 (559)	0.9770	2.26 (549)	1.033	2.08 (593)	1.81 (684)
MAE	0.31		0.32		0.30	

[a] In vacuo and solution (C-PCM) from TDDFT and in vacuo from RICC2 calculations. The mean absolute deviation (MAE) for excitation energies relative to TDDFT (in vacuo and solution) and RICC2 methods are given in eV. [b] Single point calculations from optimized structures at PBE0/SVP level of theory. [c] In chloroform for 1a−e, 2, and 3a−b; in dichloromethane for 4a−c, 5a−b, and 6.

The HOMO orbital for all compounds is characterized by an high electron density in the squaric core part, while for the LUMO orbital, electronic density is depleted from the oxygen atoms. By viewing the HOMO and LUMO orbital energies of 1d, both resulted destabilized with respect to 1c, but the LUMO is less destabilized in energy than the HOMO by ca. 0.1 eV, so λ_{max} increases. As can also be seen in Figure 1, oxygen electronic density belonging to the *m*-methoxy group in 1c is not involved in the electronic delocalization system. Moreover, the *m*-methoxy substitution in 1c leaves λ_{max} practically unchanged in comparison to the phenyl substituted case 1b (533 vs 531 nm), in qualitative agreement with the experimental values (621 vs 625 nm). Going to the next compound in Table 3, it can be seen that in 1e, the introduction of the naphthyl substituent group on the pyrrole ring, notwithstanding its strong electron-donating character, decreases slightly the value of λ_{max} (546 nm) with respect to 1d (549 nm), in qualitative agreement with the experimental λ_{max} trend. For 1e, the HOMO−LUMO energy gap is lower than that of 1d by 0.07 eV, the decreased λ_{max} could be due to the more distorted structure ($\beta = 38°$) which, therefore, reduces the overlapping of the molecular orbitals and a greater HOMO energy stabilization with respect to the LUMO orbital energy. The elongation of the molecular chain through a carbon−carbon double bond with a terminal phenyl group (compound 2) increases λ_{max} (586 nm). This fact is consistent with the experimental red-shift evidence and the decreasing of the HOMO−LUMO gap characterized by a strong energy destabilization of the HOMO orbital by ca. 0.2 eV with respect to 1e analogue orbital (see Figure 2). As for compounds of Scheme 1, the main results regarding the first lowest excitation energy (λ_{max} and oscillator strength) are summarized in Table 3. For the condensed ring structures (3a−b), theoretical λ_{max} are both red-shifted in comparison with the phenyl substituted form 1b with values, respectively, of 544 and 553 nm. In particular compound 3b, containing a six-term ring, is more red-shifted than 3a in agreement with the experiment. However, for these condensed structures, λ_{max} is lower in comparison to the more extended form 2, confirming that elongation is a better strategy for red-shifting of λ_{max}. The HOMO−LUMO energy gap trend for 3a−b also is consistent with TDDFT excitation energies and confirmed by experiment. For the compounds just examined, the oscillator strengths, corresponding to the spin- and dipole-allowed electronic transition in the visible range between 0.832 (1a) and 2.368 (2), the deviation of the calculated λ_{max} from the experimental one has a minimum value for 1e (0.25 eV) and a maximum value for 3a (0.38 eV). Proceeding to the second series of squaraine derivatives (see Scheme 2), we can note some structural differences and analogies with compounds in Scheme 1. In compounds 4a−c, the electronic delocalization was obtained by functionalizing the pyrrole ring by aryl- or alkyl-hydrazone groups. In this study, the nitrogen pyrrole atoms are always attached to a methyl group, although compound 4a was synthesized also with the group R^1 (see Scheme 1) constituted by a triethyleneglycolic chain in order to increase water solubility. Similarly for compounds 4b and 5a−b, the latter being synthesized only in this functionalized form, we adopted methyl groups as R^1 substituents. This choice was justified by the fact that a preliminary calculation for the evaluation of λ_{max} of compound 4a showed that its λ_{max} is unaffected if R^1 is present as a methyl or triethyleneglicolic group, as clearly evidenced also by experimental data. Analyzing first the squaraine derivatives containing arylhydrazone groups (4a−c), we noted that the highest calculated λ_{max} (609 nm) corresponds to compound 4a for which R^1 and R^2 terms are both electron-donating phenyl groups. The presence of weak electron-donating methyl groups in 4b gave a decreased value of λ_{max} (553 nm), whereas the situation represented by compound 4c, where R^2 is given by an hydrogen atom and R^3 by a *para*-bromine substituted phenyl group, is intermediate between that of 4a and 4b ($\lambda_{max} = 593$ nm). These results reflect the order of the HOMO−LUMO energy gaps depicted in Figure 2, the lowest energy gap (highest λ_{max}) corresponds to compound 4a (2.12 eV). In compounds 5a and 5b, the electron conjugation is extended from each pyrrole ring through a carbon−carbon double bond with, respectively, a pyridine and a quinoline terminal group (see Scheme 2). Experimentally the introduction of these weak electron-withdrawing heteroaromatic groups has the effect to reduce λ_{max} (678 nm for 5a, 688nm for 5b) with respect to the case of 4a, which contains strong electron-donating phenyl groups.

From TDDFT results, we note that the relative magnitude order and difference of λ_{max} 5a and 5b is well reproduced, and the maximum deviation of λ_{max} from the experimental data is 0.24 eV. Despite these results, by comparing compounds 4a and 4c (containing phenyl groups) with 5a−b, the experimental decrease of λ_{max} is not so clearly found from TDDFT excitation energies. Oscillator strengths of compounds 4a−c and 5a−b are in the range between 2.0 and 2.6 and generally more intense than the first series of squaraine derivatives (1a−e, 2, and 3a−b). In compound 6, the annelated structure, resulting from the further π-delocalization of the pyrrole ring, gives a calculated λ_{max} of 559 nm, which, among all examined compounds, presents the maximum energy difference from the experimental (about 0.4 eV). In this case, RICC2 calculations (see Table 3) gives an improved calculated λ_{max} with respect to the experimental value (593 vs 684 nm). The correlation of λ_{max} with the dihedral angle α (Scheme 2, compound 6), for both sides of the squaric core part, has been examined in order to verify if TDDFT deviation of λ_{max} stems from the simultaneous presence at room temperature of the different conformers. The next most stable conformer was found to be the *syn* conformer (2.7 kcal/mol above in energy) with λ_{max} identical (2.21 eV) to that of the *anti* conformer. The weighted contribution of λ_{max} to other conformers can be neglected on the basis of their higher energy (>20 kcal/mol) with respect to the *anti* energy minimum structure. For compound 6, where the π-system is more delocalized, it seems that RICC2 method better describes the electronic transition contribution of the double excitation character, which completely lacks in the TDDFT method. From our calculations, the percentage weight of these transitions is about 10%. However, in all the other cases, the difference of calculated λ_{max}, between TDDFT and RICC2 methods, is not so drastic, changing from in 0.02 to 0.08 eV (absolute values). Taking into account solvent effects through the C-PCM method, similar consideration holds for the excitation energies (no drastic changes in λ_{max}), while oscillator strengths are intensified in value (see Table 3). A quantitative assessment of the reliability of TDDFT and RICC2 results can be made by considering the mean absolute error (MAE) for the whole series of studied compounds. The value of MAE is nearly 0.3 eV (Table 3), so in this case, there is not an evident difference between the theoretical results obtained from both methodologies.

3.3. Triplet Energies. As pointed out in the Introduction, in oxygen-dependent type II reactions, the efficiency of a PDT photosensitizer drug is measured by its single oxygen quantum yield. The energy transfer from the triplet state of the photosensitizer to the ground-state molecular oxygen, in order to be an efficient process, needs an appropriate triplet energy for the photosensitizer. This value should be equal or greater to 0.98 eV, which corresponds to the experimental triplet molecular oxygen energy (or $^3\Sigma_g \rightarrow {}^1\Delta_g$ electronic transition). For compounds 4a−c, 5a−b and 6 also have been previously reported, a qualitative comparative study has been reported on their ability to generate singlet oxygen, by monitoring the time disappearance of the absorption band at 415 nm of 1,3-diphenylisobenzofuran, which can react with a photosensitizer dye to form an endoperoxide species. The experimental results proved that the above-mentioned compounds and in particular 4a and 4c (containing bromine atoms) are able to give a singlet oxygen yield. Triplet energies can be theoretically calculated from TDDFT and RICC2 methodologies as triplet excitation energies referred to the singlet ground state. As reported in the results of Table 4, the TDDFT computations indicate that all the studied compounds have triplet energies below the limit of 0.98 eV. In particular, in vacuo triplet energies are in the range between 0.61 (5b) and 0.95 eV (1a). Bulk solvation effects slightly increase the triplet excitation energies by 0.02−0.08 eV. It can be argued, from TDDFT, that only 1a−e, 3a−b, and 6 compounds lie close to the mentioned energetic gap. On the other hand, RICC2 calculations showed that for the studied compounds the triplet energies are between 1.13 (5a) and 1.36 eV (1a), fulfilling one of the requirements to act as a PDT drug. In order to better understand the origin of this great discrepancy between TDDFT and RICC2 results, we have studied a series of molecular systems where the triplet energies have been experimentally evaluated.[54−56] The results, collected in Table 5, clearly showed that the vertical triplet energies, from RICC2 calculations, are overestimated by about 0.6−0.7 eV in comparison with experimental results obtained from phosphorescence spectra, whereas the TDDFT triplet energies, from PBE0

Table 4. Triplet Energies in Vacuo and with C-PCM Solvation Model from TDDFT and in Vacuo from RICC2 Calculations

	TD-DFT[a]		RICC2[a]
molecule	ΔE, vacuum	ΔE, c-pcm[b]	ΔE, vacuum
1a	0.95	1.00	1.36
1b	0.85	0.88	1.28
1c	0.86	0.89	1.28
1d	0.85	0.87	1.27
1e	0.86	0.90	1.29
2	0.69	0.72	1.18
3a	0.87	0.90	1.28
3b	0.84	0.88	1.28
4a	0.65	0.70	1.18
4b	0.69	0.72	1.18
4c	0.65	0.71	1.17
5a	0.58	0.66	1.13
5b	0.61	0.69	1.15
6	0.91	0.96	1.27

[a] Single point calculations from optimized structures PBE0/SVP. [b] In chloroform ($\varepsilon = 4.9$) for 1a−e, 2, and 3a−b; in dichloromethane ($\varepsilon = 8.93$) for 4a−c, 5a−b, and 6.

Table 5. Triplet Energies $\Delta E_{S_0 - T_1}$ (eV) for Some Aromatic Hydrocarbons and the Free Base Porphyn (FBP) from TDDFT (PBE0) and RICC2 Calculations in Comparison with Experimental Values

	$\Delta E_{S_0 - T_1}$ (eV)		
molecule	PBE0[a]	RICC2[a]	exptl.[b]
benzene	3.68	4.42	3.69
naphthalene	2.65	3.35	2.65
anthracene	1.72	2.42	1.82
pyrene	2.05	2.71	2.08
FBP	1.38	2.24	1.58

[a] Single point calculations with TZVP basis set from PBE0/TZVP optimized geometries. [b] Experimental data taken from ref 54, except for FBP from refs 55 and 56.

calculations, are slightly underestimated, the maximum deviation being 0.2 eV for the free base porphyn.

4. Conclusions

In this paper, the theoretical electronic spectra in the visible region for some classes of squaraine derivatives have been simulated by adopting two approaches: time-dependent density functional methods (TDDFT) and coupled-cluster model with the resolution of identity approximation (RICC2). First, the geometrical structures corresponding to energy minima have been examined and, in particular, their possible conformational dihedral changes that can affect the maximum absorption wavelength, due to more or less effective molecular orbital overlapping. For example, for naphthyl substituted squaraine (1e), the presence of strong electron-donating gives λ_{max} a decreased value with respect to the phenyl substituted counterpart, as a consequence of the distortion (about 40°) from the planarity of the rest of the molecule. The maximum absorption wavelength shifts, as a function of the substituent nature within each different class of squaraine derivatives, are qualitatively well reproduced in comparison with experimental results by the two theoretical approaches. By comparing the two theoretical approaches, singlet excitation energies show, for both TDDFT and RICC2 results, an absolute mean error of roughly 0.3 eV. In this case, their performance in predicting electronic spectra is comparable. On the other hand, triplet energies calculated by RICC2 are strongly overestimated (about 0.7 eV) with respect to the experiment than those obtained by PBE0 calculations. All studied compounds show an absorption electronic band that falls in the so-called therapeutic window (550−800 nm) for PDT treatment. Notwithstanding, including bulk solvation effects, only a few compounds (1a−e, 3a−b, and 6) have a triplet energy close to that of molecular oxygen and, consequently, could be active as type II PDT photosensitizers. The limits of hybrid functionals in TDDFT are currently under investigation through the development of new DFT functionals (so-called long-range corrected hybrid functionals) in order to improve the accuracy of excitation energies, in particular, for charge transfer electronic transitions where TDDFT gives unreliable results.[57−59]

Acknowledgment. Financial support from the Università degli Studi della Calabria and Regione Calabria (POR Calabria 2000/2006, misura 3.16, progetto PROSICA) is gratefully acknowledged.

Supporting Information Available: All in vacuo optimized structures with Cartesian coordinates for all compounds reported in Scheme 1 and 2. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Schmidt, A. H. In *Oxocarbons*, West, R. Ed.; Academic Press: New York, 1980; pp 1−185.

(2) Law, K. J. *Chem. Rev.* **1993**, *93*, 449–486.

(3) Yum, Jun-Ho; Walter, P.; Huber, S.; Rentsch, D.; Geiger, T.; Nüesch, F.; De Angelis, F.; Grätzel, M.; Nazeerudin, M. K. *J. Am. Chem. Soc.* **2007**, *129*, 10320–10321.

(4) Emmelius, M.; Pawlowsky, G.; Vollmann, H. W. *Angew. Chem., Int. Ed. Engl.* **1989**, *28*, 1445–1471.

(5) Ajayaghosh, A. *Acc. Chem. Res.* **2005**, *38*, 449–459.

(6) Basheer, M. C.; Alex, S.; Thomas, K. G.; Suresh, C. H.; Dasa, S. *Tetrahedron* **2006**, *62*, 605–610.

(7) Ramaiah, D.; Joy, A.; Chandasekhar, N.; Eldho, N. V.; Das, S.; George, M. V. *Photochem. Photobiol.* **1997**, *65*, 783–790.

(8) Ramaiah, D.; Eckert, I.; Arun, K. T.; Weidenfeller, L.; Epe, B. *Photochem. Photobiol.* **2002**, *76*, 672–677.

(9) Reis, V.; Serrano, J. P. C.; Almeida, P.; Santos, P. F. *Synlett.* **2002**, 1617–1620.

(10) Juzeniene, A.; Peng, Q.; Moan, J. *Photochem. Photobiol. Sci.* **2007**, *6*, 1234–1245.

(11) MacDonald, I. J.; Dougherty, T. J. *J. Porphyrins Phthalocyanines* **2001**, *5*, 105–129.

(12) Van Tenten, Y.; Schuitmaker, H. J.; De Wolf, A.; Willekens, B.; Vrensen, G. F. J. M.; Tassignon, M. J. *Exp. Eye Res.* **2001**, *72*, 41–48.

(13) Dolmans, D. E. J. G. J.; Fukumura, D.; Jain, R. K. *Nat. Rev. Cancer* **2003**, *3*, 380–387.

(14) Dougherty, T. J.; Gomer, C. J.; Henderson, B. W.; Jori, G.; Kessel, D.; Korbelik, M.; Moan, J.; Peng, Q. *J. Natl. Cancer Inst.* **1998**, *90*, 889–905.

(15) Bonnett, R. In *Chemical Aspects of Photodynamic Therapy*, Gordon & Breach Science Publishers: Amsterdam, 2000; pp 1−289.

(16) DeRosa, M. C.; Crutchley, R. J. *Coord. Chem. Rev.* **2002**, *233−234*, 351–371.

(17) Schweitzer, C.; Schmidt, R. *Chem. Rev.* **2003**, *103*, 1685–1757.

(18) Schmidt, R. *Photochem. Photobiol. A* **2006**, *82*, 1161–1177.

(19) Woodhams, J. H.; MacRobert, A. J.; Bown, S. G. *Photochem. Photobiol. A* **2007**, *6*, 1246–1256.

(20) Oleinick, N. L.; Morris, R. L.; Belichenko, I. *Photochem. Photobiol. A* **2002**, *1*, 1–21.

(21) Jyothish, K.; Arun, K. T.; Ramaiah, D. *Org. Lett.* **2004**, *22*, 3965–3968.

(22) Sternberg, E. D.; Dolphin, D.; Brückner, C. *Tetrahedron* **1998**, *54*, 4151–4202.

(23) Wainwright, M. *Chem. Soc. Rev.* **1996**, *25*, 351–359.

(24) Wöhrle, D.; Hirth, A.; Bogdahn-Rai, T.; Schnurpfeil, G.; Shopova, M. *Chem. Bull.* **1998**, *47*, 807–816.

(25) Sharman, W. M.; Allen, G. M.; VanLier, J. E. *Drug Discovery Today* **1999**, *4*, 507–517.

(26) Young, S. W.; Woodbourn, K. W.; Wright, M. *Photochem. Photobiol. A* **1996**, *63*, 892–897.

(27) Chowdhary, R. K.; Ratkay, L. G.; Canaan, A. J.; Waterfield, J. D.; Richter, A. M.; Levy, J. G. *Biopharm. Drug Dispos.* **1998**, *19*, 395–400.

(28) Gorman, A.; Killoran, J.; O'Shea, C.; Kenna, T.; Gallagher, W. M.; O'Shea, D. F. *J. Am. Chem. Soc.* **2004**, *126*, 10619–10631.

(29) Killoran, J.; Allen, L.; Gallagher, J. F.; Gallagher, W. M.; O'Shea, D. F. *Chem. Commun.* **2002**, *17*, 1862–1863.

(30) Yukruk, F.; Dogan, A. L.; Canpinar, H.; Guc, D.; Akkaya, E. U. *Org. Lett.* **2005**, *7*, 2885–2887.

(31) Beverina, L.; Crippa, M.; Landenna, M.; Ruffo, R.; Salice, P.; Silvestri, F.; Versari, S.; Villa, A.; Ciaffoni, L.; Collini, E.; Ferrante, C.; Bradamante, S.; Mari, C. M.; Bozio, R.; Pagani, G. A. *J. Am. Chem. Soc.* **2008**, *130*, 1894–1902.

(32) Ramaiah, D.; Eckert, I.; Arun, K. T.; Weidenfeller, L.; Epe, B. *Photochem. Photobiol. A* **2004**, *79*, 99–104.

(33) Santos, P. F.; Reis, L. V.; Almeida, P.; Oliveira, A. S.; Vieira Ferriera, L. F. *J. Photochem. Photobiol. A* **2003**, *160*, 159–161.

(34) Arunkumar, E.; Sudeep, P. K.; Kamat, P. V.; Nolla, B. C.; Smith, B. D. *New J. Chem.* **2007**, *31*, 677–683.

(35) Bonnett, R.; Motevalli, M.; Siu, J. *Tetrahedron* **2004**, *60*, 8913–8918.

(36) Beverina, L.; Abbotto, A.; Landenna, M.; Cerminara, M.; Tubino, R.; Meinardi, F.; Bradamante, S.; Pagani, G. A. *Org. Lett.* **2005**, *7*, 4257–4260.

(37) Casida, M. E. In *Recent Advances in Density Functional Methods*; Chong, D. P. Ed.; World Scientific: Singapore, 1995; Part I, pp 155−192.

(38) Christiansen, O.; Koch, H.; Jørgensen, P. *Chem. Phys. Lett.* **1995**, *243*, 409–418.

(39) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

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

(41) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982–9985.

(42) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571–2577.

(43) Bauernschmitt, R.; Ahlrichs, R. *Chem. Phys. Lett.* **1996**, *256*, 454–464.

(44) Hättig, C.; Weigend, F. *J. Chem. Phys.* **2000**, *113*, 5154–5161.

(45) Hättig, C.; Hald, K. *Phys. Chem. Chem. Phys.* **2002**, *4*, 2111–2118.

(46) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *100*, 5829–5835.

(47) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023. http://bse.pnl.gov/bse/portal (accessed January, 2009).

(48) Quartarolo, A. D.; Russo, N.; Sicilia, E. *Chem.— Eur. J.* **2006**, *12*, 6797–6803.

(49) Jacquemin, D.; Perpete, E.; Scuseria, G. E.; Ciofini, I.; Adamo, C. *J. Chem. Theory Comput.* **2008**, *4*, 123–135.

(50) Lanzo, I.; Quartarolo, A. D.; Russo, N.; Sicilia, E. *Photochem. Photobiol. A* **2009**, *8*, 386–390.

(51) Jacquemin, D.; Perpete, E.; Ciofini, I.; Adamo, C. *Acc. Chem. Res.* **2009**, *42*, 326–332.

(52) Klamt, A.; Schüürmann, G. *J. Chem. Soc. Perkin Trans. 2* **1993**, *5*, 799–805.

(53) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3094.

(54) Turro, N. J. In *Modern Molecular Photochemistry*; University Science Books, 1991, pp 1−292.

(55) Gouterman, M.; Khalil, G. *J. Mol. Spectrosc.* **1974**, *53*, 88–100.

(56) Radziszewski, J. G.; Waluk, J.; Nepras, M.; Michl, J. *J. Phys. Chem.* **1991**, *95*, 1963–1969.

(57) Tawada, Y.; Tsuneda, T.; Yanagisawa, S. *J. Chem. Phys.* **2004**, *18*, 8425–8433.

(58) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.

(59) Chai, J.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 84106.

CT900199J

JCTC Journal of Chemical Theory and Computation

Order−Disorder Transition and Phase Separation in the MgB_2 Metallic Sublattice Induced by Al Doping

S. Brutti* and G. Gigli

Dipartimento di Chimica, Dipartimento di Chimica, Sapienza Università di Roma, P.le Aldo Moro 5 00185 Roma, Italy

Received January 20, 2009

Abstract: MgB_2 is a superconductor constituted by alternating Mg and B planar layers: doping of both the sublattices has been observed experimentally to destroy the outstanding superconductive properties of this simple material. In this study we present the investigation by first principles methods at atomistic scale of the phase separation induced by aluminum doping in the MgB_2 lattice. The calculations were performed by Density Functional Theory in generalized gradient approximation and pseudopotentials. Orthorhombic oP36 supercells derived by the primitive hR3 MgB_2 cell were built in order to simulate the aluminum−magnesium substitution in the 0−50% composition range. The computational results explained the occurrence of a phase separation in the $Mg_{1-x}Al_xB_2$ system. The miscibility gap is predicted to be induced by an order−disorder transition in the metallic sublattice at high Al concentration. Indeed at 1000 K aluminum substitution takes place on random Mg sites for concentration up to 17% of the total metallic sites, whereas at Al content larger than 31% the substitution is energetically more favorable on alternated metallic layers (Mg undoped planes alternate with Mg−Al layers). The formation of this Al-rich phase lead at 50% doping to the formation of the double omega $Mg_{1/2}Al_{1/2}B_2$ ordered lattice. From 17 to 31% the two phases, the disordered $Mg_{1-x}Al_xB_2$ ($x < 0.17$) and the ordered $Mg_{1/2+y}Al_{1/2-y}B_2$ ($y < 0.19$) lattices, coexist. This phase separation is driven by the balance of the enthalpy and entropy contributions to the Gibbs energy. Present DFT-GGA calculations indicate that this thermodynamically predicted suppression of the Al doping disorder in the metallic sublattice of MgB_2 occurs in parallel with the collapse of the superconductive properties of the material.

Introduction

Magnesium diboride, MgB_2, was observed to undergo a superconductive transition at $T_c = 39$ K, one of the highest known transition temperatures for a noncopper-oxide material.[1] This finding boosted the interest in this simple compound, and stimulated research efforts mainly focused on the physical and thermodynamic properties (i.e., energetic stability, thermal conductivity, elastic and electric properties, electronic and crystal structure, e.g. refs 2 and 3) as well as fabrication, in particular, film growth processing (e.g., ref 4).

The electronic structure of MgB_2 is now well understood: the Fermi surface consists of two three-dimensional sheets due to the π bonding and antibonding bands and two nearly cylindrical sheets due to the two-dimensional σ bands. Its superconductivity properties arise from a phonon mediated mechanism, with different coupling strengths, with the two mentioned σ and π electronic bands, which leads to the uncommon appearance of two distinct superconducting gaps.[5,6] In the past few years the substitution of Mg with Al, Ca, Li and Sc, and B with C has been investigated to understand the evolution of the pairing process (e.g., refs 7−10). The various attempts to obtain compounds with higher T_c derived by doping from MgB_2 have been up to the present unsuccessful, suggesting that MgB_2 may represent a unique combination of events where electronic and dynamic properties reach an extremely favorable balance.[11] Moreover

* Corresponding author e-mail: sergio.brutti@uniroma1.it.

10.1021/ct900167k CCC: $40.75
Published on Web 06/05/2009

MgB_2, due to its very simple crystal and electronic structure, is even a fortunate example of a two-gap superconductor where the effect of doping on the electronic structure can be accurately calculated by first principles methods: this makes the MgB_2 system a typical case study.[12]

There is experimental evidence[13−16] of a phase separation in the $Mg_{1-x}Al_xB_2$ system in the x=0.1−0.4 range. This separation is apparently related to the precipitation of a ternary Al-rich phase with a superstructure related to the MgB_2 lattice with doubled c-axis, similar to the crystalline double-ω phase $Mg_{0.5}Al_{0.5}B_2$.[17] The crystal chemistry and its correlation with the functional properties of the substituted MgB_2 phase by aluminum and other doping atoms has been the object of a number of computational investigations.[18−23] Nevertheless a detailed interpretation at the atomistic level of both the lattice distortion related to the phase separation and the driving forces that lead to the establishing of this multiphase equilibrium are still missing. This is the object of this paper where we present the results of the investigation by first principles methods of the lattice stability of the $Mg_{1-x}Al_xB_2$ (0<x<0.5) system. The electronic structure calculations have been carried out by means of Density Functional Theory (DFT), using supercells. The goal is the analysis of the phase separation due to the modulation of the composition of MgB_2 by Al doping. This will be characterized by considering the lattice configurations at different doping concentrations in order to give a structural interpretation at the atomistic level. The driving force that, from a thermodynamic point of view, promotes the phase separation will be analyzed in terms of free energy by distinguishing enthalpy and entropy contributions.

Computational Method

The study of the lattice stability of the $Mg_{1-x}Al_xB_2$ (0<x<0.5) system has been carried out by spin unpolarized electronic structure calculations within the density functional theory (DFT) approach, in the generalized-gradient approximation (GGA-PW91),[24] and using pseudopotentials for core electrons.

MgB_2 has an hP3 primitive cell (C32 strukturbericht designation) with space group *P6/mmm* (no. 191) and Mg and B atoms in the 1a (0 0 0) and 2d ($^1/_3,^2/_3,^1/_2,^2/_3,^1/_3,^1/_2$) sites, respectively. The corresponding standard hexagonal axes ($\hat{A}_1,\hat{A}_2,\hat{A}_3$) in terms of Cartesian versors ($\hat{x},\hat{y},\hat{z}$) are

$$\hat{A}_1 = {}^1/_2 a\hat{x} - {}^{\sqrt{3}}/_2 a\hat{y}$$

$$\hat{A}_2 = {}^1/_2 a\hat{x} - {}^{\sqrt{3}}/_2 a\hat{y}$$

$$\hat{A}_3 = c\hat{z}$$

The study of the aluminum doping has been carried out by using the supercells approach: in particular a oP36 supercell has been derived from the primitive hP3 MgB_2 lattice by the following transformation

$$\hat{A}_1^* = \hat{A}_1 - \hat{A}_2$$

$$\hat{A}_2^* = -3 \cdot (\hat{A}_1 - \hat{A}_2)$$

$$\hat{A}_2^* = -2 \cdot \hat{A}_3$$

where ($\hat{A}_1^*,\hat{A}_2^*,\hat{A}_3^*$) are the novel axes derived from the primitive vectors ($\hat{A}_1,\hat{A}_2,\hat{A}_3$). The resultant supercell axes in terms of Cartesian versors and primitive cell parameters are

$$\hat{A}_1^* = \sqrt{3} \cdot a\hat{x}$$

$$\hat{A}_2^* = -3 \cdot a\hat{y}$$

$$\hat{A}_3^* = -2 \cdot c\hat{z}$$

Both the primitive cell and the supercell are presented in Figure 1a and b.

In summary the orthorhombic supercell results from the merge of 12 primitive hexagonal cells: the new periodic unit consists of an alternated couple of metallic and boron layers. The oP36 structure has 12 metallic sites (Figure 1c). By following the designation outlined in Figure 1c for the metallic sites in the oP36 lattice, the corresponding site coordinates are as follows: (1) 0 0 0; (2) 1/2; 1/6 0; (3) 0 1/3 0; (4) 1/2; 1/2 0; (5) 0 2/3 0; (6) 1/2 5/6 0; (a) 0 0 1/2; (b)1/2 1/6 1/2; (c) 0 1/3 1/2; (d) 1/2, 1/2, 1/2; (e) 0 2/3 1/2; (f) 1/2, 5/6, 1/2. As a consequence 7 different compositions in the $Mg_{1-x}Al_xB_2$ (0<x<0.5) ternary system can be simulated (i.e., 0, 1/12, 2/12, 3/12, 4/12, 5/12, 6/12) as well as the lattice of the double omega $Mg_{0.5}Al_{0.5}B_2$ phase (Figure 1d). The seven intermediate compositions are obviously obtained by substituting the necessary number of Mg atoms by Al. It is to be noted that the number of configurations that represents the occurring substitutions, increases as the concentration of the doping Al increases. In the oP36 supercell, the p aluminum and $(n-p)$ magnesium atoms are accommodated in $n = 12$ metallic sites: the resulting number of configurations is $\zeta = (n!)/(p!(n-p)!)$. From a crystallographic point of view each configuration represents an unique chemical prototype. As a consequence at a given doping concentration

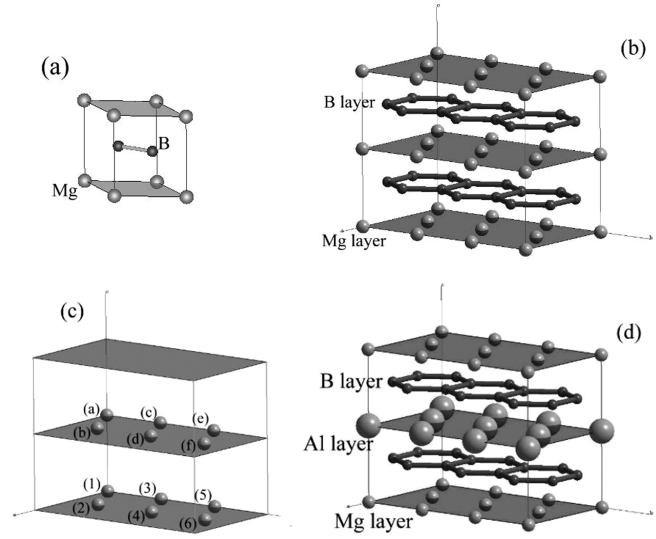


Figure 1. (a) MgB_2 primitive cell hP3; (b) $(MgB_2)_{12}$ supercell oP36; (c) 12 Mg sites in the oP36 supercell; (d) $(Mg_{0.5}Al_{0.5}B_2)_{12}$ double omega-based structure supercell oP36.

Table 1. Irreducible Supercells Simulated by DFT Calculations

x Al content in the $Mg_{1-x}Al_xB_2$ system	number of Al doping atoms	oP36 supercell designation	doping character	Al substituted metallic sites
0.0	0	oP36(0)[a]	-	-
0.083	1	oP36(1)	-	(1)
0.166	2	oP36(2V)oP36(2L)	in-volume in-layer	(1)-(d) (1)-(4)
0.25	3	oP36(3V)oP36(3L)	in-volume in-layer	(1)-(d)-(4) (1)-(4)-(2)
0.333	4	oP36(4V)oP36(4L)	in-volume in-layer	(1)-(d)-(4)-(a) (1)-(4)-(2)-(6)
0.416	5	oP36(5V)oP36(5L)	in-volume in-layer	(1)-(d)-(4)-(a)-(3) (1)-(4)-(2)-(6)-(3)
0.50	6	oP36(6V)oP36(6L)[b]	in-volume in-layer	(1)-(d)-(4)-(a)-(3)-(f)(1)-(4)-(2)-(6)-(3)-(5)
1.0	12	oP36(12)[c]	-	all

[a] This configuration corresponds to the hP3 MgB_2 phase. [b] This configuration corresponds to the double omega $Mg_{0.5}Al_{0.5}B_2$ phase. [c] This configuration corresponds to the hP3 AlB_2 phase.

the number of inequivalent supercells to be simulated becomes rather large, the configurations being 1, 12, 66, 220, 495, 792, 924, for 0, 1, 2, 3, 4, 5, 6 Al substitutions, respectively. Symmetry considerations can lead to a reduction of the supercells to be considered: as an example, in the case of our oP36 supercell, the symmetry-irreducible structural permutations, for 0, 1, 2, 3, 4 substitutions, are 1, 1, 7, 13 and 36, respectively. However, although the resulting number of inequivalent configurations is reduced, they are still too many to allow a complete and easy screening of all the configurations.

On considering these constraints, in order to draw a reliable picture at the atomistic scale of the $Mg_{1-x}Al_xB_2$ ($0<x<0.5$) ternary system, two symmetry-inequivalent configurations were considered for each composition. This choice aimed at taking into account the alternative and concurrent in-layer and in-volume doping. Indeed for a multiple aluminum substitution, the doping can occur on a single metallic layer (in-layer doping), leaving the other Mg layer in the oP36 structure unaltered, or on both metallic planes (in-volume doping). Among all the possible in-volume or in-layer substituted structures, we apply another "selection rule", by adopting the configuration in which the interatomic Al−Al distance is maximized. This rather arbitrary assumption aims to mimic a "minimum-perturbation criteria" of the hosting MgB_2 lattice, by minimizing the direct interaction between the aluminum atoms and the resulting chemical pressure. The thirteen supercells selected with these criteria for the DFT calculations are summarized in Table 1.

DFT calculations have been performed by using the PWscf code, included in the QUANTUM-ESPRESSO package:[25] post-SCF calculations were carried out by using the routines included in the same package. Vanderbilt ultrasoft pseudopotentials were adopted for Mg, Al, and B including 8, 3, and 3 electrons in the valence shells, respectively. The adopted pseudopotentials were retrieved from the database provided by the PWscf compilers.[25]

The Irreducible Brillouin Zone (IBZ) was sampled by adopting Γ-centered uniform Monkhorst-Pack (MP) k-points grids.[26] A $7 \times 7 \times 7$ grid with 64 irreducible k-points has been adopted for the oP36 supercells. The tetrahedron method[27] has been adopted for deriving the electron occupancies in the IBZ integration at the Fermi level.

The adopted pseudopotentials were checked on the boron, magnesium, and aluminum isolated atoms and elemental bulks in order to assess a satisfactory kinetic energy cutoff, E_{cut}, for the plane wave basis set calculations: a total energy convergence threshold of 0.3 mRy at^{-1} was achieved using energy cutoffs of 29 and 116 Ry for the electron wave functions and charge density, respectively. The adopted total energy convergence threshold in the self-consistent field calculations was 10^{-8} Ry. All the elemental Al(fcc), Mg(hcp),[28] and β-B[29] lattices were fully relaxed, optimizing the cell parameters and the atomic positions: the relevant convergence threshold on the pressure was set at 0.2 kbar. The MP grid meshes, and irreducible k-points adopted for elemental bulk calculations were $15 \times 15 \times 15$ (120), $17 \times 17 \times 17$ (297), and $5 \times 5 \times 5$ (63) for Al, Mg, and B, respectively. It is to be noted that for the β-B structure the triclinic lattice proposed in ref 29 derived from the classical trigonal structure was adopted.

The computed cell parameters are in agreement with the experimental literature values within 0.5% and the cohesion energies within 2.5% in the case of Al and Mg, whereas triclinic boron cohesion energy is overestimated by about 40 kJ mol at^{-1}, approximately 7% of the experimental value.[30] However it is to be noted that recently van Setten and co-workers[29] reported a detailed study about the stability of elemental boron polymorphic structures and in particular the β-B lattice. These authors observed that more than 16 boron allotropes exist and that, from a theoretical point of view, a complete picture of the lattice stabilities of this simple system is still far from being assessed due to the complex structure of the β-boron primitive lattice.

As a consequence we performed a further check of the reliability of the computational condition adopted by simulating the MgB_2 hP3 primitive lattice (MP grid $15 \times 15 \times 15$ with 216 irreducible k-points). The optimized cell parameters *a* and *c* are 3.0750 ± 0.0001 Å and 3.5360 ± 0.0002 Å to be compared to the experimental values of 3.083 ± 0.001 Å and 3.520 ± 0.001 Å:[28] the disagreement is in both cases smaller than 0.5%. For what concerns the enthalpy of formation the benchmark is the experimental value -41.5 ± 0.5 kJ mol at^{-1}.[3] It is to be noted that two different energy reference states can be adopted in order to derive the enthalpy of formation of the MgB_2 phase: (i) the sum of the isolated atoms total energies or (ii) the sum of the elemental bulk total energies. In the first case the auxiliaries thermodynamic properties (i.e., the enthalpy of formation of the isolated atoms) can be retrieved from a standard thermodynamic database.[30] The computed formation enthalpy for the hP3 MgB_2 phase are -42.5 ± 2.0 and -18.5 ± 0.1 kJ mol at^{-1},

Table 2. Computational Results: Structural Parameters and Energetics

	in-volume Al doped supercells				in-plane Al doped supercells			
x Al in the $Mg_{1-x}Al_xB_2$ system		*a*[Å]	*c*[Å]	$\Delta_fH°_{0K}$[kJ mol at^{-1}]		*a*[Å]	*c*[Å]	$\Delta_fH°_{0K}$[kJ mol at^{-1}]
0.0	oP36(0)	3.076[a]	3.537[a]	−42.3[a]				
0.083	oP36(1)	3.072	3.498	−42.6				
0.166	oP36(2 V)	3.067	3.460	−42.8	oP36(2 L)	3.067	3.462	−43.0
0.25	oP36(3 V)	3.064	3.420	−42.9	oP36(3 L)	3.062	3.432	−43.6
0.333	oP36(4 V)	3.060	3.384	−43.1	oP36(4 L)	3.054	3.409	−44.4
0.416	oP36(5 V)	3.053	3.361	−43.3	oP36(5 L)	3.046	3.388	−45.1
0.50	oP36(6 V)	3.044	3.343	−43.2	oP36(6 L)	3.037[b]	3.372[b]	−45.6[b]
1.0	oP36(12)	3.001[c]	3.282[c]	−37.1[c]				

[a] Undoped MgB_2 structure [oP36(0)]; experimental values[28] $a = 3.083$ Å; $c = 3.521$ Å. [b] Double-ω $Mg_{0.5}Al_{0.5}B_2$ structure; experimental values[28] $a = 3.044$ Å; $c = 3.356$ Å. [c] Fully doped AlB_2 structure[oP36(12)]; experimental values[28] $a = 3.005$ Å; $c = 3.257$ Å.

for the (i) and (ii) reference state, respectively. As expected the unsatisfactory prediction of the β-B cohesion energy directly propagates to the heat of formation of the hP3 phase resulting in a large underestimation of the experimental value. On the other hand by adopting the isolated atoms as a reference state the experimental formation enthalpy and the DFT data are in excellent agreement within 1.0 kJ mol at^{-1}. As a consequence for the supercell calculations we preferred to adopt as energy reference state the isolated atoms rather than the elemental bulk lattices.

Similarly to the cases of the bulk elements the supercells were fully relaxed, optimizing both the atomic positions and the cell parameters. In the optimization of the atomic positions, the total energy convergence threshold was set at 10^{-5} Ry, and the convergence threshold on forces acting on the atoms was set at 10^{-4} Ry. The cell parameters were relaxed with respect to a convergence threshold on the pressure set at 0.2 kbar, to be satisfied by each component of the 3×3 stress matrix.

Using these computational conditions the total energy for all the oP36 supercells was converged to better than 0.3 mRy at^{-1} in comparison with data obtained with a $5 \times 5 \times 5$ MP grid and $E_{cut} = 28$ Ry.

Calculations Results

The computational results are summarized in Table 2. The cell parameters are those of the corresponding reduced primitive hP3 lattice: these have been obtained by using the inverse crystal transformation discussed in the previous section. The energy stabilities are presented as heat of formation at 0 K. The comparison with previous experimental data from the literature is satisfactory both for what concerns the structural trends of the *a* and *c* cell parameters in the entire composition range (see refs 14 and 15) and for the energetic stabilities of the MgB_2, $Mg_{0.5}Al_{0.5}B_2$, and AlB_2 phases (see refs 3, 12, and 31). The lattice parameters matched in all cases the experimental values within 0.5%.

From a structural point of view the differences between in-plane and in-volume doped supercells appears minor in the case of the *a*-axis trends, whereas the *c*-axis comparison evidences a more compact structure for the in-volume doping case. These prediction are somewhat expected since the *a*-axis modifications are limited by the rigid covalent B−B honeycomb network, whereas the *c*-axis variations are related to changes in the stacking among the alternated metallic-boron planar layers. Indeed DFT calculations correctly predict that the metal−boron layer distances along the *c*-axis are expanded in the case of MgB_2 (oP36(0)) and compressed for the AlB_2 lattice (oP36(12)). In the intermediate cases the in-volume doping allows the relaxation of all the interlayer distances, whereas the in-plane substitution keeps unaltered one Mg-layer resulting in an uncompressed B−Mg−B stacking only partially compensated by the relaxation of the adjacent B-(Mg,Al)-B stacking planes.

From a thermodynamic point of view the comparison between the formation energy of the in-volume and in-plane doped supercells reported in Figure 2a makes clear that these last lattices are in all cases energetically more stable, suggesting that the Al atoms in the MgB_2 lattice cluster preferentially on single metallic layers leaving unaltered alternated Mg-planes. However in order to draw a conclusion about the thermodynamic equilibria between the in-plane/in-volume substituted phases, the entropy contribution must be taken into account, and therefore some further discussion concerning the configuration entropy is needed.

As discussed in the previous section the substitution of the magnesium atoms by aluminum in the oP6 lattice can occur in 12 metallic sites leading to a number of structural configurations increasing with the doping concentration. Similarly to the case of random or ordered alloys, the resulting entropy contribution deserves to be accounted for predicting the thermodynamic stability of the various lattices that is driven at a given temperature by the Gibbs energy. Therefore, by neglecting the vibrational contributions at finite temperature for the oP36 lattices formation enthalpy and entropy, both partially compensated by the elemental bulk atoms concurrent effects, the Gibbs energy of formation for a generic *i* supercell can be approximated by $\Delta_f G^o_{T,i} = \Delta_f H^o_{0K,i}(DFT) - T \cdot S_{conf,i}$, where $S_{conf,i}$ is the corresponding *i*-th configurational entropy.

The configurational entropy is $S_{conf} = R \cdot \ln \zeta$, where ζ are the structural permutations of the Al and Mg atoms in the N metallic sites in the oP36 lattice at a given doping concentration. Under periodic conditions at the continuum limit the configurational entropies for the in-volume and in-plane doped structures are given at any concentration by

$$S_{conf}^{involume} = -\frac{R}{3} \cdot [(1 - \chi_{Al}) \cdot \ln(1 - \chi_{Al}) + \chi_{Al} \cdot \ln \chi_{Al}]$$

$$S_{conf}^{inplane} = -\frac{R}{6} \cdot [(1 - 2 \cdot \chi_{Al}) \cdot \ln(1 - 2 \cdot \chi_{Al}) + 2 \cdot \chi_{Al} \cdot \ln(2 \cdot \chi_{Al})]$$

where R is the gas constant, and χ_{Al} is the molar fraction of aluminum in the system. The $^1/_3$ and $^1/_6$ factors normalize the configurational entropies per mole of atoms. The plots of the configurational entropies in the two cases are presented in Figure 2b as a function of the doping concentration.

The formation enthalpies at intermediate concentrations were obtained by a linear interpolation of the in-volume and in-plane DFT values reported in Table 2 for the 0.083, 0.166, 0.25, 0.333, 0.416, 0.5 Al content in the $Mg_{1-x}Al_xB_2$ system.

An example of the resulting Gibbs energy plot at 1000 K is presented in Figure 2c. The crossing between the two Gibbs energy curves at 1000 K for the in-plane and in-volume doped structures indicates the presence of two single phase stability regions. Indeed starting from the MgB_2 pure lattice for Al contents <17% the Gibbs energy of the in-volume doped structures is larger than the corresponding one for the in-plane substituted lattices. In this region the doping is predicted to occur randomly driven by the larger configuration entropy of the in-volume doping compared to the in plane substitution. A reverse picture is observed for Al concentration >31% where the Gibbs energy curve of the in-plane doped lattices is more negative than the corresponding one for the in-volume doping. In this region DFT calculations predicts the clustering of the Al atoms on single planes alternated with unaltered Mg-layers: this ordering transition is driven by the larger formation enthalpy of the in-plane doping compared to the in volume doped lattices. In the intermediate compositions 17−31 Al at.% the Gibbs energy plot predicts the occurring of a phase separation stability field. Indeed in this doping range it is possible to draw a common tangent line to the two Gibbs energy curves. This tangent line represents the total Gibbs energy of a system constituted by a mechanical mixture of the two in-volume and in-plane lattices, both at fixed Al concentration: 17 and 31 at.% Al, respectively. In the 17−31% composition range the mixture is predicted to be the thermodynamically stable system being its Gibbs energy smaller than those of the in-volume or in-plane doped lattices at the same overall Al doping concentration. A similar analysis can be repeated at any temperature: the resulting phase diagram is shown in Figure 3a.

In summary, going from pure MgB_2 to $Mg_{0.5}Al_{0.5}B_2$, our calculations predict that at small Al concentrations the C32

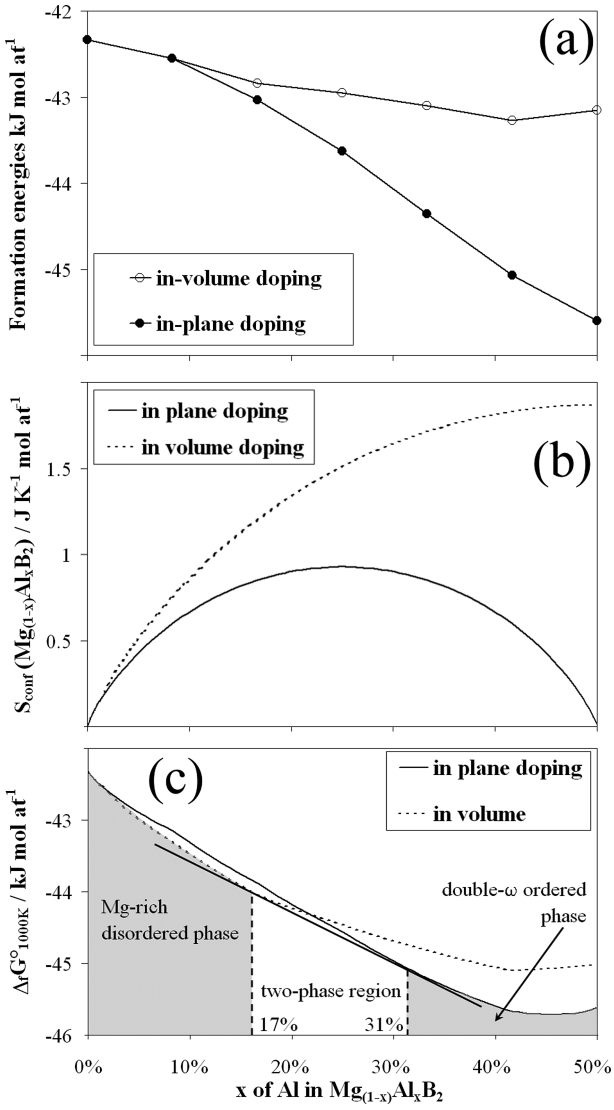


Figure 2. (a) DFT formation energies for the in-plane and in-volume aluminum doped supercells. (b) Configurational entropy for the in-plane or in-volume doping of the MgB_2 phase in the 0−50 Al at% composition range. (c) Predicted Gibbs energy plot and phase diagram at 1000 K for the $Mg_{1-x}Al_xB_2$ system.

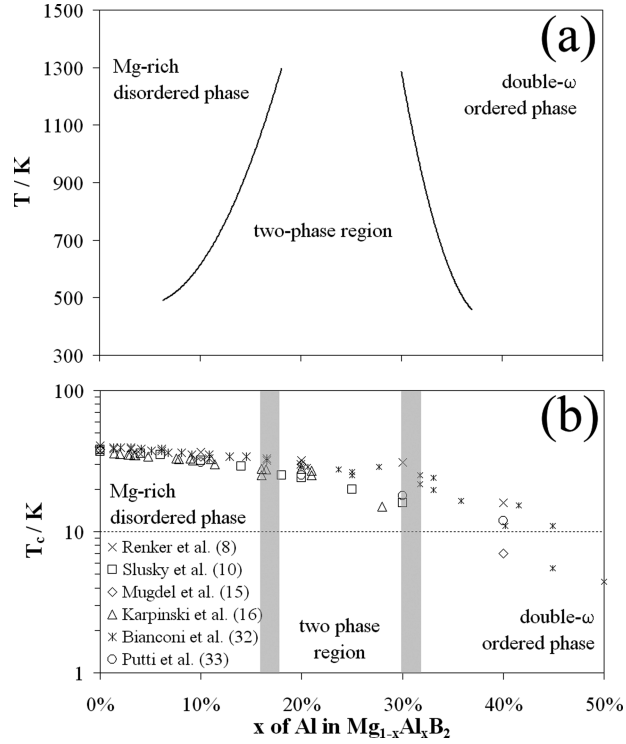


Figure 3. (a) Predicted phase diagram for the $Mg_{1-x}Al_xB_2$ system in the 0−50 at.% Al composition range. (b) Experimental literature superconductive T_c for the $Mg_{1-x}Al_xB_2$ system. The two gray areas represent the predicted intervals of the two-phase region boundaries in the high temperatures range 1000−1300 K. This temperature range is typically used in the synthesis of the $Mg_{1-x}Al_xB_2$ samples.

lattice, driven by configurational entropy, would host randomly the doping atoms resulting in a disordered metallic sublattice. At large Al content the ordering of the doping atoms in a double-ω-like structure is energetically more favored. At intermediate concentration the two disordered and ordered structures coexist: the phase borders are temperature dependent and range between 4−18% and 37−30% for the disordered/two-phase and the two-phase/ordered boundaries, respectively, in the temperature range 450−1300 K.

Discussion and Conclusions

In this paper we presented the analysis of the lattice stability of the $Mg_{1-x}Al_xB_2$ system by DFT and supercells approach. The computational results predict, in agreement with the experimental literature, the occurrence of a phase separation in the $Mg_{1-x}Al_xB_2$ system. The miscibility gap is predicted to be induced by a disorder−order transition in the metallic sublattice at high Al concentration. Indeed at 1000 K aluminum substitution takes place on random Mg sites for concentration up to 17% of the total metallic sites, whereas at Al content larger than 31% the substitution is energetically more favored on alternated metallic layers (Mg undoped planes alternate with Mg−Al layers). The formation of this Al-rich phase leads at 50% doping to the formation of the double omega $Mg_{1/2}Al_{1/2}B_2$ ordered lattice. At 1000 K from 17 to 31% the two phases, the disordered $Mg_{1-x}Al_xB_2$ (x=0.17) and the ordered $Mg_{1/2+y}Al_{1/2-y}B_2$ (y=0.19) lattices coexist.

This phase separation is energetically driven by the balance of the enthalpy and entropy contributions to the Gibbs energy and occurs in parallel with the variation of the superconductor properties of the overall $Mg_{1-x}Al_xB_2$ ($0<x<0.5$) system. In Figure 3b the experimental superconductive T_c reported in the literature for the $Mg_{1-x}Al_xB_2$ system are shown.[8,10,15,16,32,33] The two gray areas in Figure 3b represent the predicted intervals of the two-phase region boundaries in the high temperatures range 1000−1300K. This temperature range is that typically used in the synthesis of the $Mg_{1-x}Al_xB_2$ samples.

The experimental superconductive T_c decreases going from pure MgB_2 to the double-ω $Mg_{0.5}Al_{0.5}B_2$ structure. However two different decreasing trends are clearly evident for aluminum concentration <17% and >31%. The $Mg_{1-x}Al_xB_2$ system for $x > 0.31$ shows a decreasing trend of the superconductive critical temperature larger than the corresponding trend in the composition range $0<x<0.17$. It is to be noted that our computational results predict the formation at large aluminum concentrations of a double-ω-like structure with doubled c-axis compared to the starting hP3 lattice. In this structure unaltered magnesium layers are alternated with partially substituted metallic planes containing both Al and Mg atoms.

On the contrary, at small aluminum concentrations, DFT-GGA calculations predict a random substitution, even on adjacent metallic planes, of Mg by aluminum atoms: this spacial organization is favored by the configurational entropy. The stability field of this disordered phase occurs in the range 0−17 atom % of Al substitution. Apparently in the same composition range the T_c decrease is smooth and does not show any drastic suppression of the superconductivity even at large Al doping.

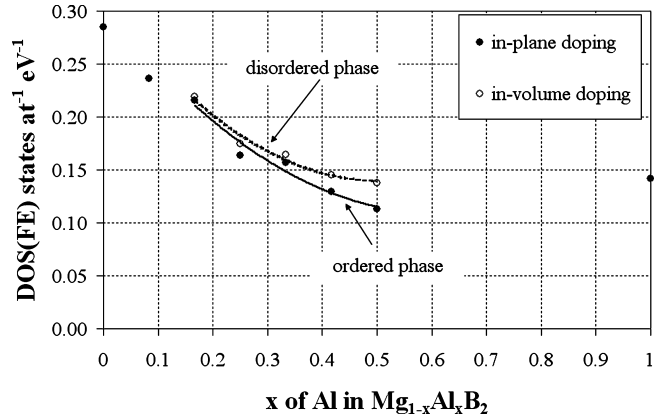


Figure 4. Variation of the density of the states at the Fermi energy with the aluminum content.

In the intermediate region 17−31 atom % of Al a large scattering of the T_c values is observed. In this composition range, our calculations predict the occurring of a phase separation: this two-phase stability field has also observed experimentally by many authors (see as e.g. refs 10, 14, and 15). On a qualitative basis we expect that this intermediate system, constituted by both the disordered and ordered phases, shows a more complicated and less easily predictable superconductive character. Indeed the specific morphology of the samples, the synthesis and annealing procedures, and the experimental methods could lead to large scattering in the measured low-temperature magnetization curves and therefore to the experimental T_c.

On the basis of all these experimental evidence we can conclude that present DFT-GGA calculations suggest that the suppression of the Al doping disorder in the metallic sublattice of MgB_2 driven by thermodynamics occurs in parallel with the collapse of the superconductive properties of the material.

Moreover it can be of interest to discuss this peculiar parallelism, i.e. the segregation of the ordered phase and the fading of the superconductor character of the $Mg_{1-x}Al_xB_2$ system, also in view of the accepted mechanism of the depression of the T_c due to the saturation of the σ bands of the MgB_2 by an increased number of valence electrons. Indeed it is now well understood that the Al doping fills the empty σ bands right above the Fermi level in the undoped MgB_2 thus leading to a decrease of the density of states (DOS) at the Fermi energy (E_F)[9,13,32] and to a topological change in the shape of the Fermi surfaces. Our calculations predict that the DOS at the Fermi energy for the in-plane and in-volume doped structures shows different decreasing trends: the pertinent variation with the aluminum content is shown in Figure 4. Indeed the density of states at E_F decreases from 0.28 states $eV^{-1}at^{-1}$ for the undoped oP36(0) lattice to 0.24 states $eV^{-1}at^{-1}$ for the oP36(1) structure. On further increasing the aluminum substitution, the DOS values at E_F progressively split for the in-volume and in-plane doped structures. Indeed in the so-called "disordered" structure, a larger DOS at the Fermi level is found, whereas in the "ordered" lattice a smaller one is calculated. This difference

increases monotonically from the 1% to the 22% for the oP36(2 L) and the oP36(6 L) lattices, respectively. Considering the different stability fields of the two disordered and ordered phases, it is likely to put into correlation the aforementioned variation of the DOS values at E_F with the decreasing trends of the T_c in the two composition ranges 0−17% and 31−50% of aluminum substitution.

Another effect that usually plays a role in the suppression of the T_c in multiband superconducting materials is the increase of the scattering due to nonmagnetic impurities. It has been reported[9] that in the case of the MgB_2 system the role of nonmagnetic impurities on the scattering is very weak,[34] particularly for Al. The only potential effect that could provide interband scattering is related to the buckling of the boron-layers and the subsequent out of plane disorder. Our calculation suggests that the buckling in both the cases of the in-plane or in-volume doping is similar and limited to a corrugation of the boron honeycomb smaller than <0.1 Å. On passing one can speculate that, by combining this result with the aforementioned trends of the DOS at E_F, the experimental spread in T_c values observed in the mixed phase region of the phase diagram (Figure 3) could be correlated to the splitting in the DOS values, rather than due to different impurity scattering.

However any further quantitative analysis or conclusion about the effect of the out of plane disorder here sketched is beyond the capabilities of the computational approach adopted and would require the calculation of the phonon spectra and of the electron−phonon coupling constants.

Acknowledgment. This research has been carried out using the CASPUR consortium computational resources under a "BANDO B 2007" grant entitled "Electronic and dynamic properties of MgB_2 and related composition modulated superstructures".

References

(1) Service, R. F. *Science* **2001**, *291*, 1476.

(2) Kortus, J. *Phys. C (Amsterdam, Neth.)* **2007**, *456*, 54.

(3) Balducci, G.; Brutti, S.; Ciccioli, A.; Gigli, G.; Manfrinetti, P.; Palenzona, A.; Butman, M.; Kudin, L. *J. Phys. Chem. Solids* **2005**, *66*, 292.

(4) Tomsic, M.; Rindfleisch, M.; Yue, J.; McFadden, K.; Phillips, J.; Sumption, M. D.; Bhatia, M.; Bohnenstiehl, S.; Collings, E. W. *Int. J. Appl. Ceram. Technol.* **2007**, *4*, 250.

(5) Canfield, P. C.; Crabtree, G. W. *Phys. Today* **2003**, *56*, 34.

(6) Ummarino, G. A.; Gonnelli, R. S.; Bianconi, A. *J. Supercond.* **2005**, *18*, 791.

(7) Slusky, J. S.; Rogado, N.; Regan, K. A.; Hayward, M. A.; Kalifah, P.; Inumaru, T. He.; Loureiro, S. M.; Haas, M. K.; Zandbergen, H. W.; Cava, R. J. *Nature* **2001**, *410*, 343.

(8) Agrestini, S.; Metallo, C.; Filippi, M.; Campi, G.; Manipoli, C.; De Negri, S.; Giovannini, M.; Saccone, A.; Latini, A.; Bianconi, A. *J. Phys. Chem. Solids* **2004**, *65*, 1479.

(9) Kortus, J.; Dolgov, O. V.; Kremer, R. K.; Golubov, A. A. *Phys. Rev. Lett.* **2005**, *94*, 027002.

(10) Kasinathan, D.; Lee, K. W.; Pickett, W. E. *Phys. C (Amsterdam, Neth.)* **2005**, *424*, 116.

(11) Profeta, G.; Continenza, A.; Bernardini, F.; Satta, G.; Massidda, S. *Int. J. Mod. Phys. B* **2002**, *16*, 1563.

(12) Bernardini, F.; Massidda, S. *Europhys. Lett.* **2006**, *76*, 491.

(13) Cava, R. J.; Zandberegen, H. W.; Inumaru, K. *Phys. C (Amsterdam, Neth.)* **2003**, *385*, 8.

(14) Palmisano, V.; Simonelli, L.; Puri, A.; Fratini, M.; Busby, Y.; Parasiades, P.; Liarokapis, E.; Brunelli, M.; Fitch, A. N.; Bianconi, A. *J. Phys.: Condens. Matter* **2008**, *20*, 434222.

(15) Mugdel, M.; Awana, V. P. S.; Kishan, H.; Bhalla, G. L. *Phys. C (Amsterdam, Neth.)* **2007**, *467*, 31.

(16) Karpinski, J.; Zhigadlo, N. D.; Schuck, G.; Kazakov, S. M.; Batlogg, B.; Rogacki, K.; Puzniak, R.; Jun, J.; Muller, E.; Wagli, P.; Gonnelli, R.; Daghero, D.; Ummarino, G. A.; Stepanov, V. A. *Phys. Rev. B* **2005**, *71*, 174506.

(17) Margadonna, S.; Prassides, K.; Arvanitidis, I.; Pissas, M.; Papavassiliou, G.; Fitch, A. N. *Phys. Rev. B* **2002**, *66*, 014518.

(18) Barabash, S. V.; Stroud, D. *Phys. Rev. B* **2002**, *66*, 012509.

(19) Profeta, G.; Continenza, A.; Bernardini, F.; Monni, M.; Massidda, S. *Supercond. Sci. Technol.* **2003**, *16*, 137.

(20) Profeta, G.; Continenza, A.; Massidda, S. *Phys. Rev. B* **2003**, *68*, 144508.

(21) Singh, P. P. *Bull. Mater. Sci.* **2003**, *26*, 131.

(22) Liu, J.; Zhao, Y.; Yi, L. *Comm. Theo. Phys. (Beijing)* **2008**, *49*, 504.

(23) de la Pena, O.; Aguayo, A.; de Coss, R. *Phys. Rev. B* **2002**, *66*, 012511.

(24) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244.

(25) Baroni, S.; Dal Corso, A.; de Gironcoli, S.; Giannozzi, P.; Cavazzoni, C.; Ballabio, G.; Scandolo, S.; Chiarotti, G.; Focher, P.; Pasquarello, A.; Laasonen, K.; Trave, A.; Car, R.; Marzari, N.; Kokalj, A. http://www.pwscf.org accessed 13/05/2009.

(26) Monkhorst, H. J.; Pack, J. D. *Phys. Rev. B* **1976**, *13*, 5188.

(27) MacDonald, A. H.; Vosko, S. H.; Coleridge, P. T. *J. Phys. C Solid State Phys.* **1979**, *12*, 2991.

(28) Villars, P. *Pearson's Handbook of Crystallographic Data*; ASM International: Materials Park, OH, 1997 (lattice structures given in alphabetical order of the corresponding chemical formula).

(29) van Setten, M. J.; Uijttewaal, M. A.; de Wijs, G. A.; de Groot, R. A. *J. Am. Chem. Soc.* **2007**, *129*, 2458.

(30) Gurvich, L. V.; Iorish, V. S. *IVTANTHERMO-Database of thermodynamic properties of individual substances*. Thermocentre of the Russian Academy of Science. Begell House: New York, 1993 (Software Vers. 1.01).

(31) Domalski, E. S.; Armstrong, G. T. *J. Res. Nat. Bureau Stand. A* **1967**, *71*, 307.

(32) Bianconi, A.; Agrestini, S.; Di Castro, D.; Campi, G.; Zangari, G.; Saini, N. L.; Saccone, A.; De Negri, S.; Giovannini, M.; Profeta, G.; Continenza, A.; Satta, G.; Massidda, S.; Cassetta, A.; Pifferi, A.; Colapietro, M. *Phys. Rev. B* **2002**, *65*, 174515.

(33) Putti, M.; Affronte, M.; Manfrinetti, P.; Palenzona, A. *Phys. Rev. B* **2003**, *68*, 094514.

(34) Mazin, I. I.; Andersen, O. K.; Jepsen, O.; Dolgov, O. V.; Kortus, J.; Golubov, A. A.; Kuzmenko, A. B.; van der Marel, D. *Phys. Rev. Lett.* **2002**, *89*, 107002.

CT900167K

JCTC Journal of Chemical Theory and Computation

Chemical Detail Force Fields for Mesogenic Molecules

Ivo Cacelli, Antonella Cimoli, Luca De Gaetani, Giacomo Prampolini, and Alessandro Tani*

Dipartimento di Chimica e Chimica Industriale, Università di Pisa, via Risorgimento 35, I-56126 Pisa, Italy

Received January 2, 2009

Abstract: Intra- and intermolecular potential energy surfaces of the 4,4′-di-*n*-heptyl azoxybenzene molecule have been sampled by *ab initio* calculations and represented through a force field suitable for classical bulk simulations. The parametrization of the molecular internal flexibility has been performed by a fitting procedure based on single molecule Hessian, gradients and torsional energies, computed using density functional theory. The intermolecular part of the force field has been derived as a pure pair potential, by fitting the dimer potential energy surface sampled by the Fragmentation Reconstruction Method. Preliminary molecular dynamics runs have been performed on systems of 210 and 600 molecules at atmospheric pressure and different temperatures, showing the presence of ordered and isotropic phases. Several properties have been computed, all resulting in a good agreement with the corresponding experimental data.

1. Introduction

Molecular dynamics (MD) simulations[1,2] have rapidly become a powerful tool in the study of soft matter.[3−6] The massive increase of computational resources has nowadays made possible performing molecular dynamics (MD) atomistic simulations on bulk phases up to thousands of medium sized molecules for several tens of nanoseconds. At the same time, there has been a growing effort to search for "realistic" force fields (FF), capable of retaining most of the detail which specifies the chemical identity of the bulk phase forming molecules. One possibility is to adopt a FF parametrization based on quantum mechanical (QM) calculations[7] on both monomer and dimer of the target molecule. This *ab initio* derived (ABD) model potential can then be employed in MD simulations for the calculation of the bulk properties. A scheme of such approach can be summarized as follows:

1) QM calculations of intermolecular and intramolecular potentials with quantum mechanical methods

2) Parameterization of the computed energies (and possibly, energy derivatives) with an analytical model potential suitable for computer simulations

3) MD simulations and comparison of the resulting macroscopic properties with the relevant experimental data

QM based parametrizations of the intramolecular FF part have been proposed by several groups and employed in widely used force fields.[8−14] In a recent work,[15] parameters for bonded interactions and partial charges of the azobenzene group (AB) have been derived from *ab initio* molecular dynamics reference calculations. A classical FF, including this description of the AB unit, was applied to the liquid crystal 8AB8, where 8-carbon chains are linked to the phenyl rings of AB via ether bonds. However, the phase transition temperatures were missed by almost 100 K.

In this paper, the equilibrium internal coordinates and the force constants of the HAB molecule are obtained by fitting DFT optimized energies, gradients and Hessian matrix using the JOYCE[16] program, recently[17] developed in our laboratory.

The QM route to intermolecular FF is much more involved because it requires the creation of a large energy vs geometry database, obtained by suitable post Hartree−Fock calculations capable of capturing a relevant fraction of the correlation energy, in order to properly account for the dispersion energy. Moreover, from early simulations on liquid argon[1] and argon−krypton mixtures,[18] it was found that the inclusion of three-body interactions in the intermolecular potential is a necessary condition to obtain accurate results of pressure and liquid−vapor equilibria.

* Corresponding author e-mail: tani@dcci.unipi.it.

10.1021/ct900002p CCC: $40.75
Published on Web 05/27/2009

Unfortunately, the inclusion of three-body interactions in the QM database as well as in the MD simulation, is computationally prohibitive in view of the dimensions of the molecules that typically can form mesophases. On the other hand, it has been demonstrated that the use of standard FFs, based on two-body effective potentials that include three-body effects in an average way, does not lead to accurate results in the case of liquid crystal forming molecules.[19] However, in a previous work on benzene,[20] we have shown that an accurate two-body potential is capable of accounting for several bulk properties, both in crystalline and liquid phases. The overall performance of that ABD FF was comparable to that obtained by employing the well-known OPLS effective potential,[21] which was tuned to reproduce experimental density and vaporization enthalpy of liquid benzene. In the case of larger molecules, e.g. polymers or liquid crystals, a corresponding tuning of the FF parameters can be very time-consuming for the long equilibration time, and this route is rather hard to be pursued.[19] On the other hand, the transferability of such parameters from smaller molecules does not allow for retaining a sufficient degree of chemical specificity. Moreover, FF parameters are tuned for a well-defined thermodynamic state (usually the isotropic liquid), so it is unlikely that their quality stays the same in different thermodynamic conditions.

For these reasons in this paper we neglect three-body effects and consider only two-body intermolecular interactions. Even with this reduction, a reliable QM sampling of a dimer potential energy surface (PES) is very computational demanding for the large number of dimer arrangements to be considered. For this step, we employ the Fragmentation Reconstruction Method (FRM),[7,22] which allows us to calculate the dimer intermolecular energies of large molecules with reduced effort, while preserving high accuracy. This method is based on fragmenting each monomer into moieties and reconstructing the intermolecular energy as a sum of the interaction energies of all pairs of resulting fragments (see Section 3.2).

In previous applications on medium size molecules, the indirect reconstruction route of FRM was adopted for *n*-pentyl-4′-cyanobiphenyl (5CB),[22,23] while the direct route was followed for *n*-pentyloxy-4′-cyanobiphenyl (5OCB)[17] (see the following section for details). The MD results obtained for 5CB and 5OCB seem to indicate that closer agreement with the experimental data is obtained if the potential is parametrized through the direct route. In fact, the overall behavior of the 5CB intermolecular model FF appeared to be slightly too attractive, resulting in an overestimation of the density by almost ∼6%,[23] which was shown to have dramatic effects on the translational dynamics.[24] On the other hand, when the direct reconstruction is employed, as for 5OCB, the density is much better reproduced (∼2%) and, consequently, a more accurate description of the translational dynamics is achieved.[17]

We believe additional tests of our approach to force fields are in order, to prove that it is worth the extra computational effort it entails, compared to the straightforward use of literature potentials.

In this paper a new application of the FRM is proposed for the mesogenic molecule 4,4′-di-*n*-heptylazoxybenzene (HAB) via the direct method for the parametrization of the intermolecular FF. HAB can be considered a good test case, as its molecule contains new chemical motifs that force us to explore diverse fragmentation schemes. In addition, its smectic phase has a reasonably large range of stability, although this is not true for the nematic one. The fairly large amount of experimental data available on HAB (see e.g. ref 25) provide the necesssary information for a thorough test of our model force field.

A successful validation, obtained from a detailed calculation of several relevant experimental observables, would support the reliability of the approach, confirming the possibility of exploiting its predictive capabilities in the calculations of properties not easily accessible to experiments and/or of not yet synthesized materials.

The paper is organized as follows: Section 2 contains the main computational details of both QM and MD calculations; the results of the FF parametrization are reported in Section 3, together with the discussion of preliminary simulation runs. Main conclusions are collected in the last section.

2. Methods and Computational Details

2.1. Intramolecular FF. QM calculations for both intra- and intermolecular PES sampling were performed with the GAUSSIAN 03 package.[26] In all single molecule calculations, the density functional B3LYP method[27] was used with a correlation consistent basis set, cc-pvDz. The absolute energy minimum was obtained by a complete geometry optimization. Vibrational frequencies, gradients and Hessian matrix were computed only for this conformation. Energy profiles for flexible dihedrals were obtained by performing geometry optimizations without any restriction but the investigated torsional angle, which is increased in a stepwise manner. The torsional energy fitting was performed within the Frozen Internal Rotation Approximation.[28]

From this database, the intramolecular FF has been parametrized with the JOYCE program,[16] through a least-squares minimization of the functional I^{intra}

$$I^{intra} = \sum_{g=0}^{N_{geom}} W_g\left[U_g - E_g^{intra}\right]^2 + \sum_{K\le L}^{3N-6} \frac{2W''_{KL}}{(3N-6)(3N-5)} \times \left[H_{KL} - \left(\frac{\partial^2 E^{intra}}{\partial Q_K \partial Q_L}\right)\right]_{g=0}^2 \quad (1)$$

where N_{geom} is the number of the sampled conformations, Q_K is the K^{th} normal coordinate, and U_g is the DFT computed energy in the g^{th} geometry. The QM Hessian matrix H_{KL} and the FF Hessian are evaluated at the absolute minimum $g = 0$. Two different weights W_g have been chosen: 0.0076 for the conformations with a low internal energy (≃5 kJ/mol) and 0.0019 for all the others, in order to obtain a more accurate description of the more lowest energy geometries. The diagonal and off diagonal elements of the weight matrix W'' were set to 0.05 Å^4 amu^2 and 0.025 Å^4 amu^2, respectively. Details of the fitting procedure can be found in ref 28.

The employed intramolecular ABD-FF E^{intra} is in diagonal form (i.e., without coupling terms between internal coordinates)

$$E^{intra} = E^{stretch} + E^{bend} + E^{Rtors} + E^{Ftors} + E^{LJintra} \quad (2)$$

The first three terms have an harmonic expression

$$E^{stretch} = \frac{1}{2}\sum_{\mu}^{N_s} k_\mu^s (r_\mu - r_\mu^0)^2; E^{bend} = \frac{1}{2}\sum_{\mu}^{N_b} k_\mu^b (\theta_\mu - \theta_\mu^0)^2;$$
$$E^{Rtors} = \frac{1}{2}\sum_{\mu}^{N_{Rt}} k_\mu^t (\phi_\mu - \phi_\mu^0)^2 \quad (3)$$

where k_μ^s, k_μ^b, k_μ^t and r_μ^0, θ_μ^0, ϕ_μ^0 are the force constants and equilibrium values for stretching, bending, and rigid torsional internal coordinates, respectively. For flexible dihedrals a sum of cosines is used, namely

$$E^{Ftors} = \sum_{\mu}^{N_{Fdihedrals}} \sum_{j=1}^{N_\mu} k_{j\mu}^d [1 + \cos(n_\mu^j \delta_\mu - \gamma_\mu^j)] \quad (4)$$

where $k_{j\mu}^d$ is the force constant, δ_μ is the flexible dihedral, and n_μ^j and γ_μ^j are the multiplicity and a phase factor for the j^{th} cosine. N_μ is the number of cosine functions employed for dihedral μ. Finally, in the last term of eq 2, $E^{LJintra}$ is the standard 12−6 Lennard-Jones potential between the interaction sites of the same molecule.

2.2. Intermolecular FF. As mentioned above, the FRM approach can be implemented in an indirect and a direct way. In both cases, the molecules of the target dimer are fragmented into the same number of moieties. In the indirect approach the PES of all resulting fragment-fragment pairs were computed at the QM level for many geometrical arrangements. All PESs were fitted with complex site−site analytical model functions[22] including several polynomials, exponentials, and Gaussian functions, with and without angular dependence. The intermolecular PES of the target dimer was eventually reconstructed by summing up all the fragment-fragment analytical model functions. It must be stressed that only during the reconstruction step the fragments are arranged in the same geometry in which they are reciprocally placed when considered as moieties of the whole molecules.

Conversely, in the direct FRM route, the fitting step of the fragment-fragment PES is avoided, and the sampling procedure is directly performed on the whole dimer. For each dimer arrangement, the interaction energy is computed at the QM level by summing the appropriate fragment-fragment contributions obtained by QM calculation, rather than summing the energy contribution obtained by previous fragment-fragment fittings, as in the indirect route.

Here, the HAB intermolecular PES was sampled through FRM,[7,22] computing the fragment-fragment interaction energies of all dimers with the direct route, in the supermolecule approach with a MP2 method, and considering the BSSE by the standard counterpoise correction.[29] A suitably 6-31G* modified basis set was used, where the exponents of the *d* polarization functions are decreased to $\alpha_d = 0.25$, following the suggestion of Hobza and co-workers.[30,31] As to the benzene dimer, which can be considered as a prototype for the aromatic interactions, the interaction energies computed at the MP2/6-31G*(0.25) level were shown[20,30,31] to well reproduce the results of high quality calculations.[32,33] In our case, the results of former FRM applications[20,22,34−36] suggest that this choice is a good compromise between the accuracy required in the PES description and the high number of energies required for an accurate sampling.

The intermolecular parameters were obtained from a least-squares fitting procedure, by minimizing the functional

$$I^{inter} = \frac{\sum_{k=1}^{N_{geom}} [(U_k^{FRM} - E_k^{inter})^2] e^{-\alpha U_k^{FRM}}}{\sum_{k=1}^{N_{geom}} e^{-\alpha U_k^{FRM}}} \quad (5)$$

where N_{geom} is the number of geometries considered for the HAB dimer, U_k^{FRM} is the energy of the k^{th} dimer arrangement computed by FRM/MP2, and E_k^{inter} is the value of the fitting model function for the geometry *k*

$$E_k^{inter} = \sum_{i=1}^{N_{sites}} \sum_{j=1}^{N_{sites}} [E_{ij}^{LJ} + E_{ij}^{Coul}]_k \quad (6)$$

where E_{ij}^{LJ} and E_{ij}^{Coul} are the standard 12−6 Lennard-Jones potential and the charge−charge interaction between a pair of sites *i*,*j* of two different molecules. As for other applications,[7,20] the minimization procedure of functional I^{inter} was performed by imposing for all geometries a Boltzmann-like weight with $\alpha = 1.6$ kJ/mol^{-1}.

2.3. Bulk Simulations. The ABD-FF, obtained as sketched above, has been employed for preliminary MD simulations carried out with a parallel version of the Moscito4.0[37] package on systems of 210 and 600 HAB molecules. In all runs bond lengths were kept fixed at their equilibrium value using the SHAKE algorithm[38] which allowed us to use a time step of 2 fs. Charge−charge long-range forces were treated with the particle mesh Ewald method,[39,40] using a convergence parameter α of $5.36/2R_c$ and a 4^{th} order spline interpolation, while the short-range interactions were truncated at $\mathbf{R}_c = 10$ Å, employing standard corrections for energy and virial.[1] In the NPT ensemble, temperature and pressure were kept constant using the weak coupling scheme of Berendsen et al.[41]

The equilibration of the resulting trajectories was assessed by monitoring the mass density (ρ), the positional and orientational order vs the simulation time. The phase director **n** was identified as the eigenvector corresponding to the largest eigenvalue of the Saupe ordering matrix $\hat{\mathbf{Q}}$, whose elements are

$$Q_{ab} = \left\langle \frac{1}{2}(3u_a u_b - \delta_{ab}) \right\rangle$$

where the mean value $\langle ... \rangle$ is obtained averaging on all molecules composing the system, and **u** (a = x,y,z) is the unit vector of HAB molecular long axis. The orientational order was measured through the second rank order parameter P_2, computed as the maximum eigenvalue of $\hat{\mathbf{Q}}$. Positional

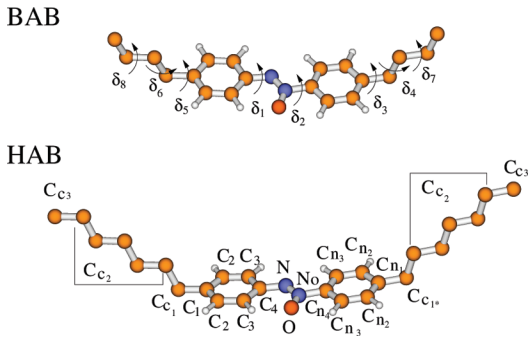


Figure 1. Model adopted for the HAB molecule (bottom panel) and flexible dihedral definition in the BAB homologue, employed for intramolecular parametrization (top panel). All aromatic hydrogens have been labeled after the carbon atom they belong to, i.e. H_{n3} is hydrogen bonded to the C_{n3} atom, etc.

order was monitored by computing the positional order parameter τ

$$\tau = \left\langle \left| \exp\left(2\pi i \frac{\mathbf{r}\cdot\mathbf{e}}{d}\right) \right| \right\rangle \tag{7}$$

where $\mathbf{r}$ is the position of the molecular center of mass, $\mathbf{e}$ is the unit vector normal to the smectic layers, and d is the layer separation. The latter is unknown and is optimized in order to maximize τ, since this ensures that the average we obtain has the same periodicity as the translational distribution function.[42]

Once equilibrated, HAB systems at several temperatures were simulated in the NVE ensemble to calculate some dynamical properties, as the isotropic translational diffusion coefficient D and the shear viscosity η_S. The former is computed as proportional to the long time limit of the center of mass mean square displacement (MSD)

$$D = \lim_{t\to\infty} D(t) = \lim_{t\to\infty} \frac{1}{6t}\langle [r(t) - r(0)]^2 \rangle \tag{8}$$

where $\langle ... \rangle$ means a double average over all configurations and molecules. The shear viscosity is computed as

$$\eta_S = \lim_{t\to\infty} \frac{V}{6k_B T} \int_0^t C_\sigma(t')dt' \tag{9}$$

where V is the volume of the simulation box, k_B is the Boltzmann constant, and $C_\sigma(t')$ is the correlation function of the off-diagonal elements of the stress tensor $\hat{\boldsymbol{\sigma}}$.[43] Due to the long time scales that characterize the collective dynamics of these systems, it has been necessary to extrapolate the value of η_S from a fitting with a double exponential function, as done previously in other applications.[24]

3. Results and Discussion

3.1. Intramolecular Parametrization. The HAB molecule is described with a full atomic model, except for the aliphatic hydrogens which are grouped with the chain carbon they are bonded to, in a united atom (UA) description, for a total of 37 interaction sites per molecule (see Figure 1).

Table 1. BAB: Optimized Stretching Parameters

stretching					
bond	r_0 (Å)	k^s (kJ/(mol Å^2))	bond	r_0 (Å)	k^s (kJ/(mol Å^2))
C_{c1}-C_1	1.511	2699	C_{c1*}-C_{n1}	1.512	2699
C_1-C_2	1.403	2766	C_{n2}-C_{n1}	1.405	2766
C_2-C_3	1.392	3377	C_{n3}-C_{n2}	1.394	3377
C_3-C_4	1.415	3083	C_{n3}-C_{n4}	1.397	3083
C_2-H_2	1.093	3368	C_{n2}-H_{n2}	1.093	3368
C_3-H_3	1.087	3368	C_{n3}-H_{n3}	1.088	3368
C_4-N	1.400	2841	C_{c3}-C_{c2}	1.530	2644
N-N_O	1.281	4124	C_{c1}-C_{c2}	1.541	2308
N_O-O	1.254	3960	C_{c1*}-C_{c2}	1.541	2308
N_O-C_{n4}	1.467	1925	C_{c2}-C_{c2}	1.532	2308

The intramolecular contribution to the FF, E^{intra}, was parametrized on an HAB smaller homologue, namely 4,4′-dibutylazoxybenzene (BAB, see Figure 1), which differs from the target molecule for six methylene units (three for each side chain). This seemed a reasonable choice since it reduces the computational time and the UA parameters describing the flexibility of the longer chains (HAB) are not expected to be much different from those obtained for the smaller homologue (BAB).

The internal coordinates can be classified into flexible and rigid ones. Rigid coordinates are bond lengths, bond angles, and those dihedral angles determining the aromatic ring planarity, and they are all described with the harmonic type potential reported in eq 3. Conversely dihedrals δ_1-δ_8 reported in Figure 1 are to be considered flexible coordinates: the harmonic approximation fails in describing their potential, since the energy profile allows them to assume several values between 0° and 360° at room temperature and they will be described through the cosine expansion (4). Among these, δ_1 is the angle between the aromatic ring bonded to the site labeled N and the azoxy bridge plane; δ_2 is the analogue of δ_1 referred to the other ring; δ_3 and δ_5 are the angles formed by the aromatic ring plane and the plane containing the first two atoms of the aliphatic chain. Finally, δ_4 and δ_6-δ_8 are the dihedral angles driving the flexibility of the aliphatic chains. The Hessian matrix for the BAB minimum energy conformation has been computed together with the torsional energy profile for dihedrals δ_1 to δ_6, which have been sampled with steps of 30° in the [0−180] range.

No sampling has been performed for δ_7 and δ_8, and their description was made through the parameters reported for *n*-butane,[9] assuming the transferability of this torsional potential. Thus, the JOYCE fitting procedure was applied according to eq 1, with $N_{geom} = 42$, yielding an overall standard deviation of 0.044 kJ/mol, with maximum error on energies of 1.7 kJ/mol (optimized parameters are reported in Tables 1−4).

The torsional profile for δ_1-δ_8 is reported in Figure 2 which shows the very good agreement between the DFT computed energies and the ABD torsional curves. The lowest energy of δ_1 and δ_2 is at 0° with a torsional barrier of ∼25 kJ/mol at 90° meaning that in the equilibrium geometry the azoxy group and the aromatic rings are expected to be coplanar. On the contrary, δ_3 and δ_5 prefer a 90° conformation, as found at the aliphatic-aromatic linkage in more simple compounds as ethylbenzene.[44] The energy profile vs both

Table 2. BAB: Optimized Bending Parameters

Bending					
Angle	θ_0 (°)	k^b (kJ/(mol·rad^2))	Angle	θ_0 (°)	k^b (kJ/(mol·rad^2))
$C_{c1}\widehat{C_1}C_2$	121.2	896.7	$C_{c1*}\widehat{C_{n1}}C_{n2}$	120.9	896.7
$C_1\widehat{C_{c1}}C_{c2}$	113.2	760.6	$C_{n1}\widehat{C_{c1*}}C_{c2}$	113.1	760.6
$C_2\widehat{C_1}C_2$	117.7	449.5	$C_{n2}\widehat{C_{n1}}C_{n2}$	118.0	434.1
$C_1\widehat{C_2}C_3$	122.3	434.1	$C_{n1}\widehat{C_{n2}}C_{n3}$	121.5	434.2
$C_2\widehat{C_3}C_4$	119.6	968.7	$C_{n2}\widehat{C_{n3}}C_{n4}$	118.9	968.8
$C_3\widehat{C_4}C_3$	118.1	245.2	$C_{n3}\widehat{C_{n4}}C_{n3}$	121.0	245.2
$C_1\widehat{C_2}H_2$ $C_3\widehat{C_2}H_2$ $C_2\widehat{C_3}H_3$ $C_4\widehat{C_3}H_3$	120.0	317.5	$C_{n1}\widehat{C_{n2}}H_{n2}$ $C_{n3}\widehat{C_{n2}}H_{n2}$ $C_{n2}\widehat{C_{n3}}H_{n3}$ $C_{n4}\widehat{C_{n3}}H_{n3}$	120.0	317.5
$C_3\widehat{C_4}N$	120.0	431.0	$N_O\widehat{C_{n4}}C_{n3}$	120.0	666.5
$C_4\widehat{N}N_O$	121.2	785.2	$N\widehat{N_O}C_{n4}$	115.1	103.6
$N\widehat{N_O}O$	128.0	705.2	$C_{n4}\widehat{N_O}O$	116.9	757.0
$C_{c3}\widehat{C_{c2}}C_{c2}$	113.2	836.9	$C_{c2}\widehat{C_{c2}}C_{c3}$	113.1	836.9
$C_{c1*}\widehat{C_{c2}}C_{c2}$	113.2	836.9	$C_{c1}\widehat{C_{c2}}C_{c2}$	113.3	836.9

Table 3. BAB: Optimized Rigid Torsion Parameters

Rigid torsions					
Dihedral Angle	ϕ_0 (°)	k^t (kJ/(mol·rad^2))	Dihedral Angle	ϕ_0 (°)	k^t (kJ/(mol·rad^2))
$C_{c1}\widehat{C_1}\widehat{C_2}C_3$	180.0	100.4	$C_{c1*}C_{n1}\widehat{C_{n2}}C_{n3}$	180.0	100.4
$C_{c1}\widehat{C_1}\widehat{C_2}H_2$	0.0	59.05	$C_{c1*}C_{n1}\widehat{C_{n2}}H_{n2}$	0.0	59.05
$C_2\widehat{C_1}\widehat{C_2}C_3$ $C_1\widehat{C_2}\widehat{C_3}C_4$ $C_2\widehat{C_3}\widehat{C_4}C_3$	0.0	87.54	$C_{n2}C_{n1}\widehat{C_{n2}}C_{n3}$ $C_{n1}C_{n2}\widehat{C_{n3}}C_{n4}$ $C_{n2}C_{n3}\widehat{C_{n4}}C_{n3}$	0.0	87.54
$C_2\widehat{C_1}\widehat{C_2}H_2$ $C_1\widehat{C_2}\widehat{C_3}H_3$ $C_4\widehat{C_3}\widehat{C_2}H_2$ $C_3\widehat{C_4}\widehat{C_3}H_3$	180	65.37	$C_{n2}C_{n1}\widehat{C_{n2}}H_{n2}$ $C_{n1}C_{n2}\widehat{C_{n3}}H_{n3}$ $C_{n4}C_{n3}\widehat{C_{n2}}H_{n2}$ $C_{n3}C_{n4}\widehat{C_{n3}}H_{n3}$	180	65.37
$H_2\widehat{C_2}\widehat{C_3}H_3$	0.0	36.16	$H_{n2}C_{n2}\widehat{C_{n3}}H_{n3}$	0.0	36.16
$C_2\widehat{C_3}\widehat{C_4}N$	180.0	30.18	$C_{n2}C_{n3}\widehat{C_{n4}}N_O$	180.0	189.1
$H_3\widehat{C_3}\widehat{C_4}N$	0.0	59.22	$H_{n3}C_{n3}\widehat{C_{n4}}N_O$	0.0	59.22
$C_4\widehat{N}\widehat{N_O}O$	0.0	222.5	$C_4N\widehat{N_O}C_{n4}$	180.0	248.7

chain dihedrals, δ_4 and δ_6, has an absolute minimum at 180° (*trans* conformation) and a relative one at ±60° *gauche* conformations) separated by rather high barriers. It is worth noticing that, as expected, the torsional potential curves of the chain dihedrals δ_4 and δ_6 are in excellent accord with those of *n*-butane (see Figure 2), confirming that this potential can be confidently transferred to all aliphatic chain dihedrals.

Extending the fitted intramolecular potential to the HAB molecule is now straightforward: by looking at the bottom panel of Figure 2 it is clear that all the inner chain dihedrals behave in similar manner, since the fitted torsional profile for δ_4 and δ_6 exactly traces out the δ_7, δ_8 shape, which is transferred from *n*-butane.

A fortiori, the force constants of more rigid internal coordinates, as methylene-methylene bond stretching and angle bending, are also expected to be transferable from those computed for the BAB smaller chains.

However, some intramolecular LJ terms were added to the FF. Previous experience on the UA modeling of 5CB molecule[23] stressed the need for such terms to both prevent an unphysical curling of the aliphatic chain on the core and on itself and to correctly take into account intramolecular interaction between the aromatic carbons and the methyl group.[45] Pair interaction parameters were added between aromatic carbons and chain UA sites at least 5 bonds apart and between methylic and methylenic chain groups 6 bonds apart. Finally pair interaction parameters have been also added to describe the interaction between the $H_{2(n2)}$ and the methylenic group bonded to $C_{c1(*)}$ (see Table 5).

3.2. The FRM Approach. The interaction potential has been calculated with the FRM, which relies on the assumption that the interaction energy of a dimer can be approximated to a good accuracy as a sum of energy contributions between the pairs of fragments which concur to form both molecules. By

Table 4. BAB: Optimized Flexible Torsion Parameters

flexible torsions					
dihedral angle	n	k^d (kJ/mol)	dihedral angle	n	k^d (kJ/mol)
$C_3\widehat{C_4NN_o}$ (δ_1)	0	0.998	$(O)\widehat{NN_oC_{n4}C_{n3}}$ (δ_2)	0	1.997
	2	−5.739		2	−3.205
	4	0.174		4	0.502
	6	−0.047		6	−0.025
$C_{n2}\widehat{C_{n1}C_{c1}\cdot C_{c2}}$ (δ_3)	0	1.997	$C_2\widehat{C_1C_{c1}C_{c2}}$ (δ_5)	0	1.997
	2	1.358		2	1.358
	4	−0.325		4	−0.325
	6	−0.099		6	−0.099
$C_1\widehat{C_{c1}C_{c2}C_{c2}}$ (δ_6)	0	0.998	$C_{n1}\widehat{C_{c1}\cdot C_{c2}C_{c2}}$ (δ_4)	0	0.998
	1	3.657		1	3.657
	2	1.995		2	1.995
	3	7.504		3	7.504
	4	−0.263		4	−0.263
$C_{c3}\widehat{C_{c2}C_{c2}C_{c1}\cdot}$ (δ_7)	0	−2.106	$C_{c3}\widehat{C_{c2}C_{c2}C_{c1}}$ (δ_8)	0	−2.106
	1	4.330		1	4.330
	2	1.738		2	1.738
	3	7.520		3	7.520
	4	0.126		4	0.126
	5	0.172		5	0.172
	6	0.241		6	0.241

way of example, let us consider a simple *AB* molecule where *A* and *B* are two moieties connected by a single bond. The whole molecule can be formally written as

$$AB \simeq AH_a + H_bB - H_aH_b$$

where the two intruder atoms H_a and H_b are first included to saturate the *A* and *B* fragments and then removed as a H_aH_b molecule. The spatial position of the fragments is the same as in the whole molecule, whereas the location of the intruder atoms H_a and H_b is determined by the equilibrium geometry of the saturated fragments BH_a and H_bB, which normally correspond to stable molecules. For the success of the method it

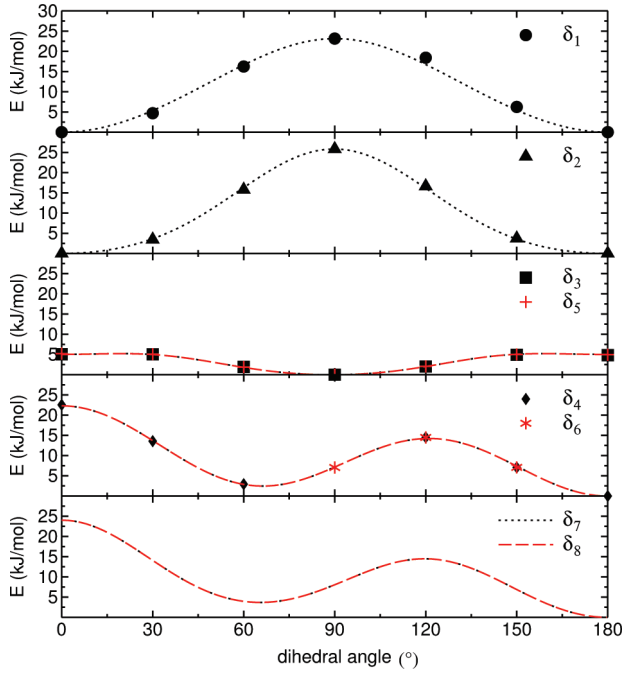


Figure 2. BAB (4,4′-dibutylazoxybenzene) - proper torsions, the symbols represent the sampled points while the dashed lines are the curves obtained with the model parametrized potential *V*. Only the region 0°−180° is shown because of symmetry reasons.

Table 5. Intramolecular LJ Parameters[a]

intramolecular LJ parameters		
site couple	σ (Å)	ε (kJ/mol)
$C_{c2}^{***} \cdots C_a$	3.30	0.69
$C_{c2}^{**} \cdots C_a$	3.56	0.69
$C_{c2}^{*} \cdots C_a$	3.56	0.69
$C_{c3} \cdots C_a$	3.56	0.69
$C_{c3} \cdots C_{c2}^{c1(*)}$	4.34	0.021
$C_{c2}^{c1(*)} \cdots H_{2(n2)}$	4.80	0.004

[a] C_a indicates an aromatic carbon site; C_{c2}^{*} is the C_{c2} site bonded to the C_{c3} one; C_{c2}^{**} and C_{c2}^{***} are the sites that follow along the aliphatic chain. Finally $C_{c2}^{c1(*)}$ represents the C_{c2} bonded to $C_{c1(*)}$.

must be verified that the electronic density on the *A* (*B*) moiety is as close as possible to that in the *AB* and AH_a molecules. This is of particular importance for the electronic density of the π orbitals in aromatic systems. Let us now consider a second molecule *CD* interacting with *AB* (of course *CD* can be the previous *AB* molecule moved to a different spatial location). A similar fragmentation scheme leads to

$$CD \simeq CH_c + H_dD - H_cH_d$$

The complete interaction energy af the $AB\cdots CD$ system can be recovered by summing up the interaction energy of all the involved pairs, including the fragments only formed by intruder atoms

$$\begin{aligned} &E(AB\cdots CD) = \\ &E(AH_a\cdots CH_c) + E(AH_a\cdots H_dD) - E(AH_a\cdots H_cH_d) + \\ &E(H_bB\cdots CH_c) + E(H_bB\cdots H_dD) - E(H_bB\cdots H_cH_d) - \\ &E(H_aH_b\cdots CH_c) - E(H_aH_b\cdots H_dD) + E(H_aH_b\cdots H_cH_d) \end{aligned} \quad (10)$$

The sign of each terms is determined by the single sign in the above fragmentation schemes for AB and CD. From the last equation, the computational advantage of the FRM approach is apparent: the interaction energy is written as a sum of interaction energies of pairs well smaller than the whole $AB\cdots CD$ system. For this calculation a suitable QM method able to recover a large part of the dispersion energy has to be employed. Since such QM methods scale at least as the fourth power of the number of electrons, the computational gain can be further appreciated. Previous applications of the FRM approach[17,35] have shown that, with a suitable choice of the method and basis set, a reliable database for many geometrical arrangements can be obtained and employed for FF parametrization.

3.3. Intermolecular Parametrization. The direct FRM route has been applied to the HAB molecule to sample its intermolecular PES, following the fragmentation scheme reported in Figure 3.

The reliability of the fragmentation using H_2 as intruder molecules was already validated.[35] Here, more attention has been paid to verify possible fragmentation schemes of the central aromatic core. With this aim preliminary calculations[46] have been performed on the HAB's smaller homologue, 4,4′-dimethylazoxybenzene. We found a cut along the N−C_4 bond (see Figure 1) preferable to that along N_O-C_{n4}, as the electronic density distribution of the resulting frag-

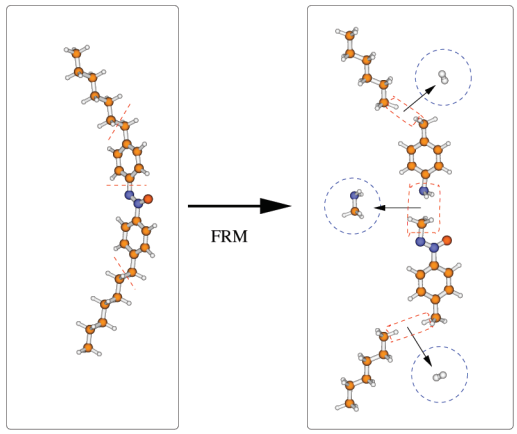


Figure 3. Fragmentation route for HAB. The fragmentation has been performed by cutting bonds as shown by the red dashed lines (left panel). The four resulting fragments are shown on the right, together with the intruder fragments, which are encircled in blue.

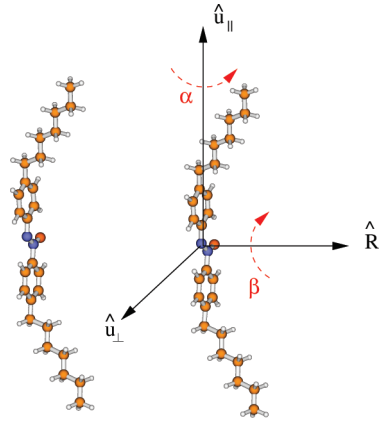


Figure 4. FRM sampling. Dimer arrangements are generated by displacing one HAB molecule with respect to the other by translation along $\hat{R}$, $\hat{u}_{\perp}$, or $\hat{u}_{\parallel}$ and/or rotation around α or β.

ments is more similar to that exhibited by the moieties in the whole molecule. Thus the final fragments are two hexanes, one 4-methylbenzenamine and one methyl,4-(methyl-NNO-azoxy)phenyl. It may be worth noting that the intruder molecule arising from the chosen scheme in the central core is a CH_3NH_2 molecule, as shown in Figure 3.

The low symmetry of the target dimer makes sampling the PES far from straightforward. However, a classification of the many dimer arrangements is still possible in terms of displacement vectors ($\hat{R}$, $\hat{u}_{\perp}$, and $\hat{u}_{\parallel}$) and Euler angles (α and β), as reported in Figure 4. Face-to-face (FF) configurations can be obtained by shifting one molecule along the vector $\hat{R}$, with $\alpha = 0°$ and $\beta = 0°$ or 180°. In the former case, the vectors N_O-O point in the same direction, and the geometries so obtained are labeled as parallel face-to-face (p-FF). In the latter, the vectors N_O-O point in opposite directions resulting in antiparallel (a-FF) arrangements. If both molecules are rotated by $\alpha = 90°$, the cores are found in a side-by-side disposition (p-SS or a-SS, whether $\beta = 0°$ or 180°). Side-to-face (p-SF and a-SF) geometries are obtained with a rotation of $\alpha = 90°$ of one molecule. If the β angle is set to 90°, the two cores draw a cross and the x-FF, x-SF, and x-SS arrangements can be created by applying a further rotation around the β-rotated long molecular axis of 0°, 90° on one or both molecules, respectively. Finally if one molecule is rotated around $\hat{u}_{\perp}$, the dimer is found in a T-shaped (TS) configuration.

Further subclasses can be created by exploiting the relative position of the side chains of the two dimer molecules. If the symbol " ⊂ " is used to sketch the HAB molecule, three subclasses can be determined depending on when the alkyl chains point in the same direction (⊂ ⊂), toward each other (⊂ ⊃), or in opposite directions (⊃ ⊂).

The PES sampling has been performed by computing the FRM energy of $N_{geom} \sim 2000$ geometries (see eq 5).

From a preliminary analysis of the resulting FRM database, the configurations with the more attractive interactions are the ⊃ ⊂ ones, where the steric chain repulsion effect does not prevent the rings and the NNO group to reach their most favorable positions with respect of the same groups of the other molecule. For the same reason, less attractive energies are found for the ⊂ ⊂ and ⊂ ⊃ arrangements, in this order.

In the latter, the steric effect dominates and the interaction energy wells are far less deep and shifted to higher values of the displacement coordinate. For example, for the p-FF case, reported in Figure 5, ⊃ ⊂ have a maximum interaction energy of about −50 kJ/mol, ⊂ ⊂ reach −20 kJ/mol at most, and ⊂ ⊃ only are −5 kJ/mol. The situation is similar for the a-FF geometry: here the ⊃ ⊂ is even more the favorite since the oxygens point in opposite directions and the maximum interaction energy can reach −57 kJ/mol.

Among all classes, the most favored geometries are p-FF, a-FF, and x-FF configurations, where the aromatic rings come closer to each other and the interaction π energy significantly contributes to the total interaction energy.[22,36] It may be worth noting that in the x-FF case, the aliphatic chains are not superimposed in any geometry with the result that their repulsion effect is not as important as in the previous configurations. This means that ⊂ ⊂, ⊂ ⊃, and ⊃ ⊂ geometries have similar situations, with −40 kJ/mol maximum interaction energies in all cases. The interaction energies for the SS configurations are less attractive: the a-SS geometries do not reach −30 kJ/mol and p-SS ones arrive at −18 kJ/mol, at most. Furthermore, p-SF does not overstep the −20 kJ/mol, showing energies between the FF case and the SS one. Also the x-SF configurations are less energetic than the x-FF (−25 kJ/mol and −40 kJ/mol, respectively). Finally TS arrangements present maximum interaction energy of about −12 kJ/mol. As far as the contact distance is concerned, it can be noted that, again, there are three different kinds of disposition: p-FF, a-FF, and x-FF configurations have a contact distance of about 3 Å, p-SS and a-SS of 6 Å, and TS above 8 Å. From this first picture, the anisotropic nature of the HAB molecule shows up clearly. This is a well-known feature of many calamitic LC, as for instance the nCB series:[47] the contact distance and the well depths of all interaction curves strongly depend on the monomer relative orientations.

All the sampled energies have been fitted by means of functional (5) with the model described in Figure 1. Equivalent sites have been given the same σ, ε, and q values.

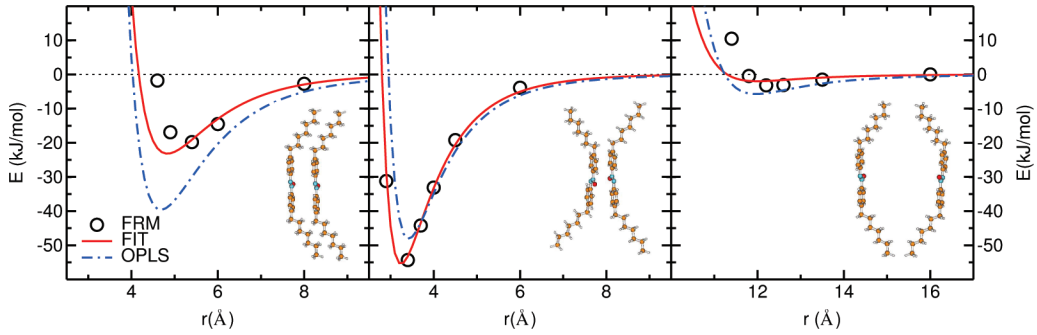


Figure 5. HAB computed and fitted energy curves for selected geometrical arrangements: p-FF ⊂ ⊂, p-FF ⊃ ⊂, and p-FF ⊂ ⊃. The points represent the FRM energies and the solid lines the fitted energies. OPLS predictions are also reported with dashed blue lines for comparison.

Table 6. HAB: Optimized Intermolecular Parameters

site	ε (kJ/mol)	σ (Å)	q (*e*)	site	ε (kJ/mol)	σ (Å)	q (*e*)
$C_{c1(*)}$	0.148	4.15	0.272	C_{n4}	3.402	2.69	−0.047
C_1	1.878	2.79	−0.367	C_{n3}	0.664	3.10	−0.126
C_2	0.145	3.52	−0.043	C_{n2}	0.066	4.20	−0.066
C_3	1.076	3.38	−0.043	C_{n1}	1.652	2.00	−0.050
C_4	2.632	2.00	0.098	H_{n3}	0.082	2.00	0.093
H_2	0.023	2.48	0.082	H_{n2}	0.053	2.00	0.119
H_3	0.074	2.35	0.048	C_{c1}	2.590	3.44	−0.009
N	1.135	3.05	−0.374	C_{c2}	0.246	4.10	0.000
N_O	0.020	3.82	0.729	C_{c3}	0.020	3.70	0.000
O	0.180	3.03	−0.428				

No other constraints were imposed but for σ and q of the methylene (C_{c2}) groups of the chains. These sites have been assumed chargeless as in previuos applications,[17,35] since this allows a straightfoward extension of the FF to higher homologues (see the nCB series[48]). As to the $\sigma_{C_{c2}}$ parameter, a first fitting yielded a standard deviation of 2.22 kJ/mol, with a value (4.4 Å) significantly larger than that obtained for butane (3.905 Å[49]) or 5OCB (3.76 Å[17]). With this set of parameters, a preliminary MD run, performed in the NPT ensemble on an isotropic system of 64 HAB molecule at 1 atm and 370 K gave a density underestimated by ∼4% with respect to the experimental value. However, another fit with $\sigma_{C_{c2}}$ reduced to 4.1 Å, and a slightly larger standard deviation (2.48 kJ/mol) yielded a density within 1% of the experimental data.

The final optimized parameter set that corresponds to this fit can be seen in Table 6.

In Figures 5 and 6, the energies obtained with the FRM are compared with those predicted by the OPLS[21,49] empirical force field, widely employed in simulations of the liquid phase of many smaller molecules. The OPLS-FF describes the FRM energy surface with a standard deviation of 9.56 kJ/mol. This can be seen as a remarkable achievement, and a proof of the good transferability of the OPLS parameters. However, the extreme sensitivity of the properties of liquid crystalline materials to even minor changes of the molecular structures and interactions supports the need for a more accurate representation of the FRM PES, as attained by our fitting procedure, which leads to the standard deviation of 2.48 kJ/mol. (See also Figures 5 and 6 for a visual estimate of the agreement at a few selected configurations.) The need for an accurate description of the PES is also apparent if we consider that the FRM approach is able to reproduce the interaction energies of HAB dimers with great accuracy, as Figure 6 shows. In each panel of this figure we include a point (labeled DIMER) that corresponds to the true *ab initio* interaction energy of the dimer in that configuration, i.e. the value obtained without decomposing the molecule as in the FRM scheme. The results of Figure 6 prove two things: i) that the FRM gives an excellent agreement with the true value of the interaction energy (e.g., −17.1 vs −16.9 kJ/mol with the FRM for the p-FF arrangement) and ii) that the fit we employ faithfully describes the PES we sampled.

As a test of the predictive capability of the ABD-FF, the energy curves of unsampled geometries have been calculated both with the model function and using FRM. The results are reported in Figure 7, together with the OPLS predictions. In the top panel an energy curve for a β rotation is shown. One HAB molecule was also translated by 2 Å along $\hat{u}_\perp$, 3.5 Å along $\hat{R}$ and rotated by $\alpha = 180°$. Here, both curves are in good agreement with the FRM points.

The central panel shows an interaction energy curve, obtained by displacing one molecule along the $u_{||}$ direction, with a shift of 2.5 Å along $\hat{u}_\perp$, 4.0 Å along $\hat{R}$, and a rotation of $\alpha = 40°$ and $\beta = 50°$. In this case the fitted curve traces out the FRM point positions better than the OPLS. Finally, in the third panel, a similar curve is reported for a a-FF ⊃ ⊂ type, again obtained by displacing the second molecule along $u_{||}$ after a translation of 3.4 Å along $\hat{R}$ and a 180° rotation of both α and β. Even if the OPLS prediction is not so far from the FRM points, the fitted curves are significantly closer to the FRM data.

3.4. Preliminary MD Results. MD simulations have been carried out on condensed phases of HAB in the NPT and NVE ensembles. The first type of conditions were used when searching for the transition temperatures, while the latter have been adopted for the calculation of dynamic properties. In all cases, we have used the largest number of molecules we could afford with a cubic box, to prevent any artificial ordering of the sample, induced e.g. by an elongated box.

210 HAB molecules have been considered a sensible starting point, mainly for the simulation of the isotropic phase. However, a larger system of 600 molecules (corresponding to 22200 interaction sites) has been adopted to study the ordered phases and to evaluate the system size effect on the results.

The isotropic phase has been obtained starting from a 6 × 7 × 5 cubic lattice disposition where the HAB centers of mass were placed on the nodes, with the long molecular axis

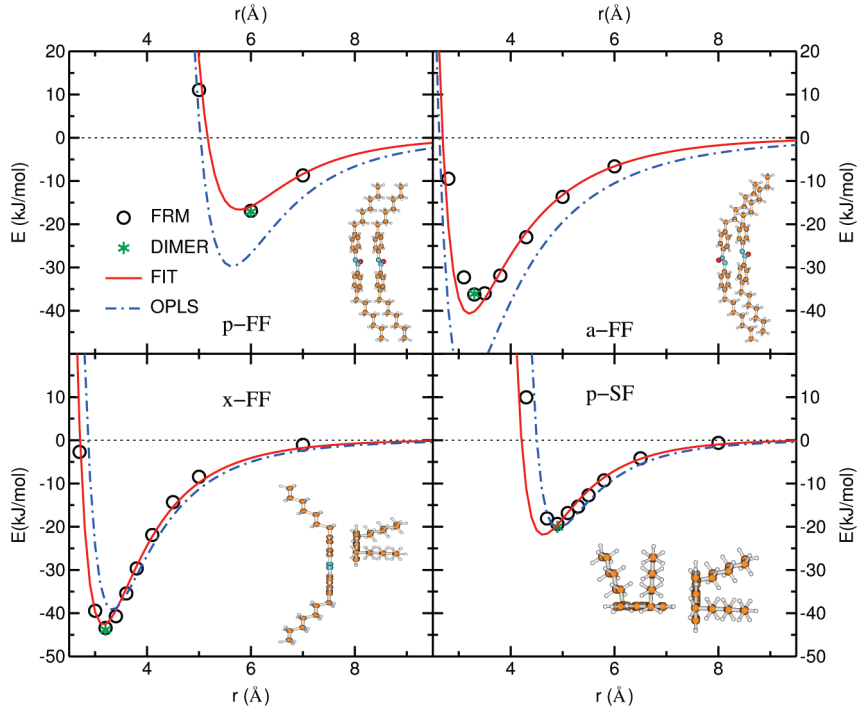


Figure 6. HAB computed and fitted energy curves for selected geometrical arrangements: p-FF $\subset\subset$, a-FF $\subset\subset$, x-FF $\supset\subset$, and p-SF $\supset\subset$. In the first two geometries, a shift of -2 Å was applied along $\hat{u}_\perp$. The points represent the FRM energies and the solid lines the fitted energies. The green point marked 'DIMER' in each panel shows the interaction energy of the dimer, computed *ab initio* for the *whole* molecules (no FRM, see text). OPLS predictions are also reported with dashed blue lines for comparison.

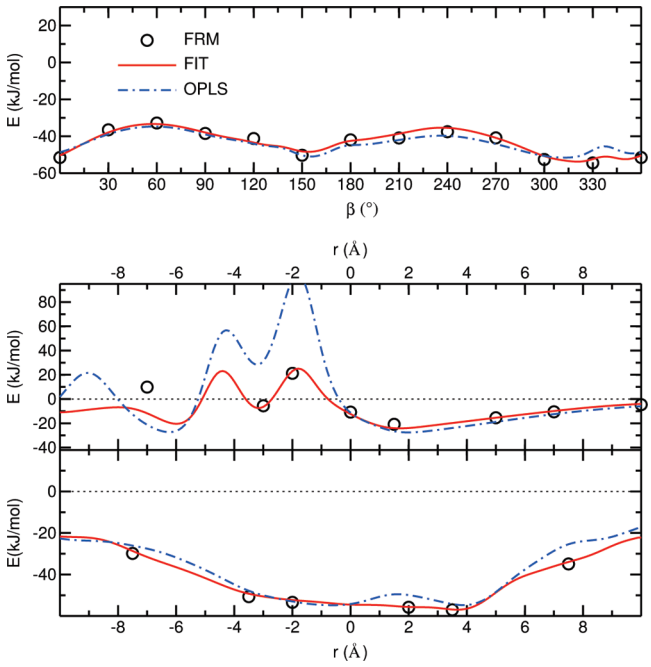


Figure 7. HAB interaction energy predicted by the fitting function (solid line) and control FRM values (open circles). The OPLS prediction is also shown (dashed line).

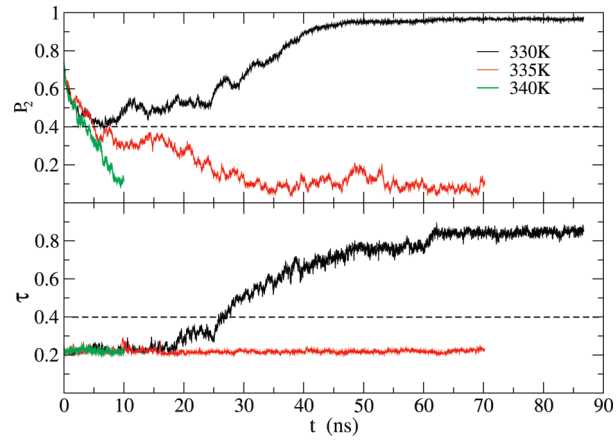


Figure 8. Time evolution of orientational, P_2, and positional, τ, order parameter at three temperatures for the system with 600 molecules. Horizontal dashed lines represent the conventional orientational and positional order threshold.

aligned along one of the box axes ($\hat{z}$). The system has been equilibrated in the NPT ensemble at 400 K for 5 ns and then cooled at 380 K and in successive steps at 370 K, 350 K, 330 K, and 300 K. Given that the experimental[50] nematic−isotropic transition takes place at 342 K, the temperatures relevant for the true isotropic phase are 350 and 370 K.

The system of 600 molecules was obtained from a lattice structure in a cubic box at low density which was later compressed to the isotropic liquid density.

After equilibration, an orienting field was applied for 0.5 ns under NPT conditions ($T = 340$ K). The field was able to produce a P_2 value of 0.74 with no positional order. Later, the field was switched off, and this metastable state was used as a common starting configuration for three runs at 340, 335, and 330 K.

The evolution of the orientational and positional order parameter was then followed for some tens ns, with the results shown in Figure 8. These data show clearly the

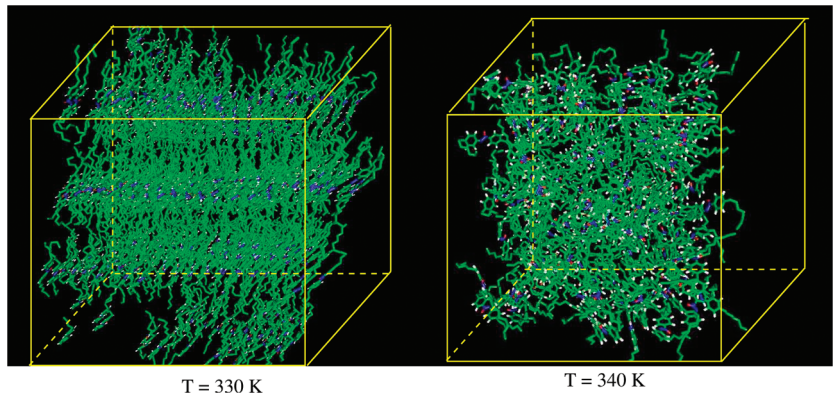


Figure 9. Snapshots of the smectic (left, $T = 330$ K, 600 molecules) and isotropic (right, $T = 340$ K, 210 molecules) phases of the simulated HAB systems.

Table 7. Experimental and Simulated Results of Density and Translational Diffusion Coefficient in the Isotropic Phase[a]

T (K)	ρ_{exp}[b] (g/cm^3)	ρ_{MD}^{NPT} (g/cm^3)	T_{NVE} (K)	ρ_{MD}^{NVE} (g/cm^3)	D_{exp}[b] (10^{-10} m^2/s)	D_{MD} (10^{-10} m^2/s)
300	1.002	–	302.0	0.972	0.27	0.23 ± 0.01
330	0.964	–	330.4	0.947	0.80	1.04 ± 0.04
350	0.939	0.930 ± 0.004	353.0	0.931	1.71	1.82 ± 0.06
370	0.915	0.914 ± 0.004	373.5	0.913	3.10	3.05 ± 0.01
380	0.902	0.906 ± 0.003	–	–	–	–
400	0.877	0.889 ± 0.003	–	–	–	–

[a] At $T = 300$ K and $T = 330$ K ρ_{exp} means the density of the supercooled isotropic phase, *i.e.* that extrapolated from experimental densities in the isotropic range, while D has been obtained from a geometric average of the longitudinal and transverse diffusion coefficients. [b] Reference 25.

existence of a smectic phase at 330 K and an isotropic phase at 335 and 340 K. This might indicate that our model does not lead to the nematic phase of HAB. However, it must be stressed that reproducing by atomistic simulation a phase that spans a range of temperature of just ≃10 K is a formidable challenge.[15]

As to the smectic phase, our analysis finds an interlayer spacing of ≃22.5 Å, significantly smaller than the experimental value of 28.9 Å,[51] that corresponds to a tilt angle of ≃18° and identifies a smectic A phase. As a consequence it is likely that our system is not arranged in a smectic A phase. However, it is certainly encouraging to observe a spontaneous positional reordering at a temperature just above the experimental smectic-nematic transition temperature (327 K).[50] The layered structure of the smectic phase is apparent in the left panel of Figure 9, while the right panel shows a snapshot of the isotropic phase.

In the case of the isotropic phase, the MD density values at constant atmospheric pressure are in very good agreement with the experimental data,[50] with errors always below 1% (see Table 7). It is worth noticing that a good reproduction of density is essential, not only for a correct description of the structure of the system but also for an accurate evaluation of dynamic properties, e.g. diffusion, which is extremely sensitive to the density in these materials.

Translational diffusion has been evaluated from trajectories obtained in NVE runs, at four different temperatures in the isotropic phase. The same data have also been used for a collective dynamical property, namely shear viscosity.

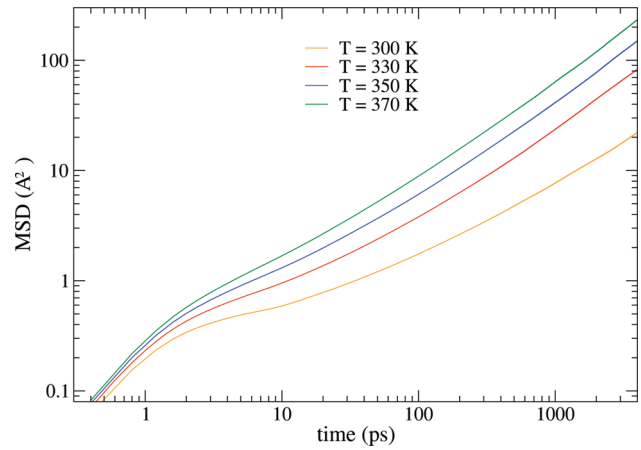


Figure 10. Time dependence of the MSD in the normal and supercooled isotropic phase (log−log scale).

As far as diffusion is concerned, for all trajectories the MSDs were computed according to eq 8 using a 4 ns correlation time window, and the curves are reported in Figure 10.

The computed translational diffusion coefficient is reported in Table 7 together with its experimental counterpart,[25] which is reproduced to a good extent. The Arrhenius treatment of the temperature dependence of D leads to an activation energy of 33.4 KJ/mol, also in good agreement with the experimental data of 31.9 KJ/mol.[25] Upon supercooling the isotropic system down to 330 and 300 K, a subdifusive behavior becomes apparent, with the β relaxation regime plateau between 1 and 10 ps.[52] This is the same time window already found in a previous work on supercooled isotropic mesogens,[53,54] so it may be considered a fairly general feature of this kind of materials.

Shear viscosity, η_s, for the isotropic phase has been investigated, too. This is a collective property that can only be averaged on successive time origins: as a consequence, it is affected by larger statistical uncertainty, and much longer simulations are needed to obtain reliable values for the long time limit of the function $C_\sigma(t)$ of eq 9. To some extent, the situation can be improved fitting the curves with a double exponential function, as done elsewhere.[24] In Figure 11 the integral, whose infinite time limit yields η_s, is reported vs time, together with the fitting curves, where a fitting window of 0−400 ps has been used at all temperatures.

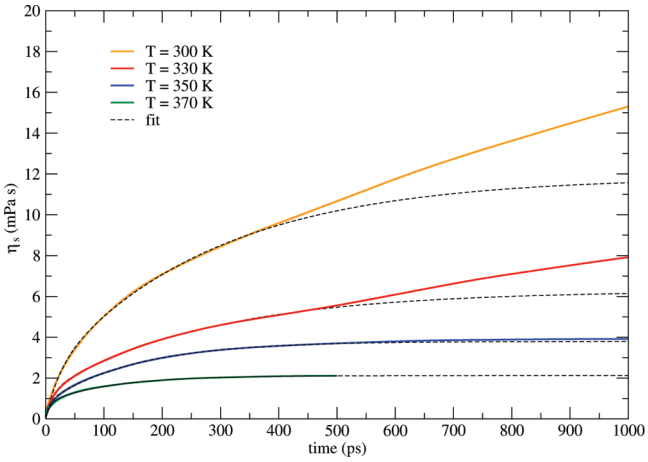


Figure 11. Calculation of HAB shear viscosity at different temperatures for the isotropic phase, according to eq 9. The exponential fits are reported in dashed lines.

Table 8. Temperature Dependence of Shear Viscosity in the Normal and Supercooled Isotropic Phase

T (K)	η_s^{MD} (mPa s)
300.2	11.9
330.4	6.3
353.0	3.8
373.5	2.1

Unfortunately, experimental data are not available for HAB viscosity, as far as we know, so we can try to estimate our MD results by comparison with the corresponding experimental data for the nCB series. $T = 353$ K corresponds roughly to a reduced temperature of 1.05, where HAB shows (see Table 8) $\eta_s = 3.8$ mPa s, while experimental values for the nCB series oscillate between 10 and 20 mPa s (see ref 24 and references therein). Assuming[55] an inverse relationship between D and η_s, the experimental D value for HAB (1.71), compared to that of nCB series (0.60−1.0 (10^{-10} m^2/s)[24]), would yield a viscosity value in the range 5−10 mPa s, slightly larger than our computed value. However, only experimental data can definitely assess the accuracy of the MD results.

4. Conclusions

The FRM approach to force fields has been applied to HAB in the direct scheme of implementation. Compared to the indirect scheme adopted in the first application of FRM to mesogens (the series of nCB[23,24]), the direct one turned out to provide more accurate results of key quantities for the description of condensed phase behavior, e.g. density and diffusion coefficient of 5OCB.[17] In the present study of HAB, the density of the isotropic phase is reproduced within 1% of the experimental value, and also a basic probe of dynamics, i.e. the diffusion coefficient D, agrees quite satisfactorily with the measured data. The success obtained with density and D supports some confidence that the MD results of shear viscosity also may be close to that of the real system, which are not available, to our knowledge.

In addition to the isotropic phase, that appears faithfully modeled, we have obtained a smectic phase at 330 K. It is rewarding that this smectic phase develops with a spontaneous positional reordering from a system with a P_2 as low as 0.4 and essentially no positional order. This locates the transition temperature to the isotropic phase between 330 and 340 K, i.e. within 10 K of the measured value of 342 K. However, from the data collected so far, the positionally ordered phase seems a SmC, instead of the SmA formed by the real system. More importantly, we have been unable to obtain a nematic phase with the present parametrization of the force field.

The incorrect smectic obtained and the missing nematic phase indicate that some changes are to made on our model. We are currently focusing on two main issues. The first improvement would entail abandoning the hybrid model adopted so far (the hydrogens of the alkyl chains are fused to the carbon they are linked to). At a significant increase of the number of interaction sites and hence computational time, a truly full atomic model should be able to better match the *ab initio* PES, maybe resorting to the exp-6 potential instead of the less flexible LJ. In turn, this would reduce distorsions that might affect some dimer configurations and propagate into an incorrect modeling of the phase diagram of the system, given the well-known sensitivity of these materials to apparently minor variations of the molecular structure and interactions.

The second point to address regards the selection of configurations whose interaction energy is to be *ab initio* calculated. We plan to combine chemical intuition with short MD runs, from which significant dimer and trimer configurations could be extracted. This way, the extent of nonadditivity could be evaluated and the database of dimer arrangements increased in a more physically driven approach.

References

(1) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon: Oxford, 1987.

(2) Frenkel, D.; Smith, B. *Understanding Molecular Simulations;* Academic Press: San Diego, 1996.

(3) *Advances in the Computer Simulations of Liquid Crystals NATO ASI series*; Pasini, P., Zannoni, C., Eds.; Kluwer: Dordrecht, 2000.

(4) Kremer, K. *Macromol. Chem. Phys.* **2003**, *204*, 257.

(5) *Computer Simulations of Liquid Crystals and Polymers NATO ASI series*; Pasini, P., Zannoni, C., Zumer, S., Eds.; Kluwer: Dordrecht, 2005.

(6) Care, C. M.; Cleaver, D. J. *Rep. Prog. Phys.* **2005**, *68*, 2665.

(7) Amovilli, A.; Cacelli, I.; Cinacchi, G.; De Gaetani, L.; Prampolini, G.; Tani, A. *Theor. Chim. Acc.* **2007**, *117*, 885.

(8) Maple, J. R.; Dinur, U.; Hagler, A. T. *Proc. Natl. Acad. Sci. U. S. A.* **1988**, *85*, 5350.

(9) Dasgupta, S.; Yamasaki, T.; Goddard, W. A., III *J. Chem. Phys.* **1996**, *104*, 2898.

(10) Palmo, K.; Mannfors, B.; Mirkin, N. G.; Krimm, S. *Biopolymers* **2003**, *68*, 383.

(11) Maple, J. R.; Hwang, M.-J.; Stockfish, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T. *J. Comput. Chem.* **1994**, *15*, 162.

(12) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490.

(13) Yang, L.; Tan, C.; Hsieh, M.; Wang, J.; Duan, Y.; Cieplak, P.; Caldwell, J.; Kollmann, P. A.; Luo, R. *J. Phys. Chem. B* **2006**, *110*, 13166.

(14) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157.

(15) Bockmann, M.; Peter, C.; Delle Site, L.; Doltsinis, N.; Kremer, K.; Marx, D. *J. Chem. Theory Comput.* **2007**, *3*, 1789.

(16) Cacelli, I.; Prampolini, G. JOYCE is free software, available under the terms of the GNU License at http://tglc.dcci.unipi.it/ Pisa (Italy) 2007.

(17) Cacelli, I.; Lami, C.; Prampolini, G. *J. Comput. Chem.* **2009**, *30*, 366.

(18) Wang, L.; Sadus, R. *J. Chem. Phys.* **2006**, *125*, 074503.

(19) Tiberio, G.; Muccioli, L.; Berardi, R.; Zannoni, C. *ChemPhysChem* **2009**, *10*, 125.

(20) Cacelli, I.; Cinacchi, G.; Prampolini, G.; Tani, A. *J. Am. Chem. Soc.* **2004**, *126*, 14278.

(21) Jorgensen, W. L.; Severance, D. L. *J. Am. Chem. Soc.* **1990**, *112*, 4768.

(22) Amovilli, C.; Cacelli, I.; Campanile, S.; Prampolini, G. *J. Chem. Phys.* **2002**, *117*, 3003.

(23) Cacelli, I.; Prampolini, G.; Tani, A. *J. Phys. Chem. B* **2005**, *109*, 3531.

(24) Cifelli, M.; De Gaetani, L.; Prampolini, G.; Tani, A. *J. Phys. Chem. B* **2008**, *112*, 9777.

(25) Cifelli, M.; Cinacchi, G.; De Gaetani, L. *J. Chem. Phys.* **2006**, *125*, 164912.

(26) 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.

(27) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(28) Cacelli, I.; Prampolini, G. *J. Chem. Theory Comput.* **2007**, *3*, 1803.

(29) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.

(30) Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Phys. Chem.* **1996**, *100*, 18790.

(31) Hobza, P.; Zahradník, R.; Müller-Dethlefs, K. *Collect. Czech. Chem. Commun.* **2006**, *71*, 443.

(32) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 104.

(33) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2006**, *110*, 10656.

(34) Cacelli, I.; Cinacchi, G.; Geloni, C.; Prampolini, G.; Tani, A. *Mol. Cryst. Liq. Cryst.* **2003**, *395*, 171.

(35) Bizzarri, M.; Cacelli, I.; Prampolini, G.; Tani, A. *J. Phys. Chem. A* **2004**, *108*, 10336.

(36) Prampolini, G. *J. Chem. Theory Comput.* **2006**, *2*, 556.

(37) Paschen, D.; Geiger, A. *MOSCITO 3.9*; Department of Physical Chemistry, University of Dortmund: 2000.

(38) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *55*, 3336.

(39) Darden, T. A.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.

(40) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, A.; Lee, H.; Pedersen, L. *J. Chem. Phys.* **1995**, *103*, 8577.

(41) Berendsen, H. J. C.; Postma, J. P. M.; van Gusteren, W. F.; Di Nola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.

(42) Bates, M. A.; Luckhurst, G. R. *J. Chem. Phys.* **1999**, *110*, 7087.

(43) Hansen, J. P.; McDonald, I. R. *Theory of Simple Liquids*; Academic Press: New York, 1986.

(44) Cinacchi, G.; Prampolini, G. *J. Phys. Chem. A* **2003**, *107*, 5228.

(45) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2000**, *122*, 3746.

(46) Cimoli, A. *Tesi di Laurea*; Università di Pisa, Pisa (Italy), 2007.

(47) Cacelli, I.; De Gaetani, L.; Prampolini, G.; Tani, A. *Mol. Cryst. Liq. Cryst.* **2007**, *465*, 175.

(48) Cacelli, I.; De Gaetani, L.; Prampolini, G.; Tani, A. *J. Phys. Chem. B* **2007**, *111*, 2130.

(49) Jorgensen, W.; Laird, E.; Nguyen, T.; Tirado-Rives, J. *J. Comput. Chem.* **1993**, *14*, 206.

(50) Jagadeesh, B.; Prabhakar, A.; Rao, M. H. V. R.; Murty, C. V. S.; Pisipati, V. G. K. M.; Kunwar, A. C.; Bowers, C. R. *J. Phys. Chem. B* **2004**, *108*, 11272.

(51) Pape, E. *Mol. Cryst. Liq. Cryst.* **1984**, *102*, 271.

(52) Götze, W.; Sjögren, L. *Rep. Prog. Phys.* **1992**, *55*, 241.

(53) De Gaetani, L.; Prampolini, G.; Tani, A. *J. Phys. Chem. B* **2007**, *111*, 7473.

(54) De Gaetani, L.; Prampolini, G.; Tani, A. *J. Chem. Phys.* **2008**, *128*, 194501.

(55) Hansen, J. P.; McDonald, I. R. *Theory of Simple Liquids*; Academic Press: New York, 1986.

CT900002P

JCTC Journal of Chemical Theory and Computation

Algorithm for Generating Defective Graphene Sheets

David R. Nutt*,† and Hilary Weller‡

Department of Chemistry, University of Reading, PO Box 224, Whiteknights, Reading, RG6 6AD, U.K., and National Centre for Atmospheric Science—Climate, Department of Meteorology, University of Reading, PO Box 243, Earley Gate, Reading, RG6 6BB, U.K.

Received March 10, 2009

Abstract: An algorithm is presented for the generation of molecular models of defective graphene fragments, containing a majority of 6-membered rings with a small number of 5- and 7-membered rings as defects. The structures are generated from an initial random array of points in 2D space, which are then subject to Delaunay triangulation. The dual of the triangulation forms a Voronoi tessellation of polygons with a range of ring sizes. An iterative cycle of refinement, involving deletion and addition of points followed by further triangulation, is performed until the user-defined criteria for the number of defects are met. The array of points and connectivities are then converted to a molecular structure and subject to geometry optimization using a standard molecular modeling package to generate final atomic coordinates. On the basis of molecular mechanics with minimization, this automated method can generate structures, which conform to user-supplied criteria and avoid the potential bias associated with the manual building of structures. One application of the algorithm is the generation of structures for the evaluation of the reactivity of different defect sites. Ab initio electronic structure calculations on a representative structure indicate preferential fluorination close to 5-ring defects.

Introduction

There is currently a growing interest in carbon structures based on 5-, 6-, and 7-membered rings.[1] The archetypal example is C_{60}, in which twelve pentagonal rings are distributed among twenty hexagonal rings in a football-like structure, such that no pentagon is adjacent to another pentagon.[2] The subsequent discovery of carbon nanotubes,[3] nanohorns,[4] and the associated nanoparticle side-products[5] provide a wide-range of contrasting structures and topologies, all of which are likely to contain a certain number of defects.[6]

It is well-known that nonhexagonal rings introduce curvature into an otherwise-planar graphitic sheet.[7] Introduction of an n-gon into a fragment of a graphene, where $n < 5$, leads to the formation of cones. When $n = 5$, fullerene-type structures emerge, and for $n > 6$, the negative curvature leads to saddle-like surfaces. As n is increased further (up to $n = 24$), calculations predict complex but stable distorted structures.[8]

Because of the stability of nonhexagonal rings in carbon structures, a number of groups have started to investigate whether there is evidence for polygonal defects in a number of carbon forms including nongraphitising (microporous) carbon,[9] glassy carbon,[10,11] and carbon black.[12] In many cases, it appears that incorporation of polygonal defects can help explain some of their many characteristics, such as low density, microporosity, and hardness.[1] Recent transmission electron microscopy (TEM) studies have demonstrated exceptional resolution, confirming the presence of 5-membered rings in samples of activated carbon[13] and providing atomic-level insight into the edges of graphene layers.[14]

Theoretical investigation of such structures can also yield useful insight. These can involve calculations of energy pathways involved in reconstructions, such as the Stone—Wales rearrangement in C_{60}[15] or the reconstruction of graphene edges.[16] Calculation of experimental observables, such as scanning tunnelling microscopy (STM) images, provides a

* To whom correspondence should be addressed. Telephone: +44-118-3786346. Fax: +44-118-3786331. E-mail: d.nutt@reading.ac.uk.

† Department of Chemistry.

‡ National Centre for Atmospheric Science—Climate.

10.1021/ct900113f CCC: $40.75
Published on Web 06/23/2009

direct comparison between experiment and theory and helps in the interpretation of experimental data.[17] Semiempirical or ab initio electronic structure calculations can provide information about reactive sites available for functionalization[18] or vibrational modes.[19]

However, the first stage of any theoretical investigation is the generation of a suitable structure. For small systems with a single defect, this is straightforward. It is possible to construct one *n*-gonal defect in a graphene sheet by joining *n* 60° segments of graphene together at a point.[8,17] Many such structures can be made "by hand", although this removes any randomness from the resulting structure and may lead to bias in the features created. For this reason, it is good to be able to automate the generation of suitable structures. Since the systematic determination of structures rapidly becomes a very large task,[20] molecular dynamics (MD)[21] or reverse Monte Carlo (RMC) methods are often used.[9] Experimentally, the growth of graphene on metal surfaces appears to occur by the addition of carbon cluster attachment.[22] Modeling of this process would offer an alternative method for generating graphene structures.

In this paper, a novel algorithm is presented for the automated generation of small graphene fragments containing small numbers of 5- and 7-ring defects. This algorithm is based on a Delaunay triangulation of a 2D array of points, followed by subsequent refinement. This new algorithm allows the user to define acceptable ratios of ring-sizes, based on experimental data, for example, but is otherwise designed around an initial random distribution of points, removing the potential for bias. Once a structure has been generated, energy minimization using classical molecular mechanics generates atomic coordinates, which can then be used in ab initio electronic structure calculations to identify sites for functionalization via fluorination[23,24] or electrochemical methods.[25]

Since the details and rationalization of the fluorination products of fullerenes and carbon nanotubes are still under debate,[23,24,26,27] the energetics of fluorination at isolated defects, such as those present in the graphene structures generated in the present work, can provide useful insight into the preferred regiochemistry.

Methods

Generation of an Initial Network. An initial distribution of points in 2-dimensional space was produced using a Sobol quasirandom sequence.[28,29] Quasirandom sequences have the property that they cover space "more uniformly" than true random sequences. As a result, they are often used in the numerical evaluation of high-dimensional integrals and global optimisation problems. Triangulation of these points, inserting additional Steiner points to avoid formation of triangles with internal angles smaller than 20°, was performed using the program Triangle.[30] The dual of the triangulation produces a Voronoi tessellation with a distribution of ring sizes. This process is illustrated in Figure 1.

The remaining task is therefore the conversion of the Voronoi tessellation from an abstract network of points to a structure with a connectivity typical of a molecular system,

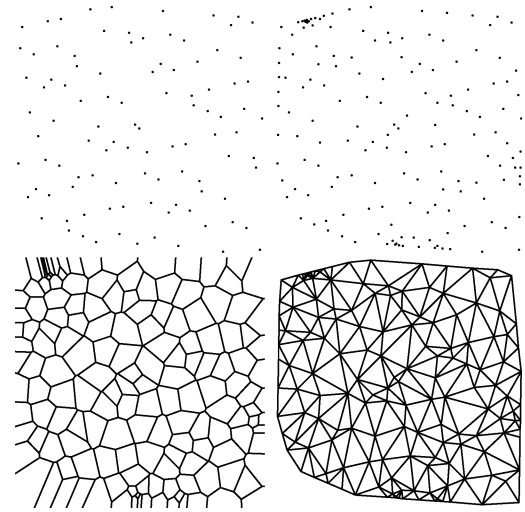


Figure 1. Processes during the generation of graphene-type networks. Clockwise from top-left: initial random points, initial points with additional Steiner points to increase the quality of the triangulation, Delaunay triangulation of the points, corresponding Voronoi tessellation.

in which the vertices are atoms and the edges represent chemical bonds. In the case of interest here, graphene, chemical knowledge dictates that the Voronoi tessellation must fulfill the following rules (Voronoi Rules):

(1) The network is limited to 5, 6, and 7-membered rings, with ratios as determined from experiments.
(2) Vertices must have a maximum connectivity of three (the connectivity is lower at the edges of a finite sheet).

The dual relationship between the Voronoi tessellation and the Delaunay triangulation leads to the following rule for the triangulation (Triangulation Rule):

(1) Vertices must have a connectivity of 5, 6, or 7, unless the point lies on the boundary, in which case the connectivity may be smaller.

Voronoi and Triangulation Rule 1 can be addressed by deleting or adding points to the triangulation in a well-defined way until the network complies (described below). Voronoi Rule 2 does not require an additional Triangulation Rule, since the triangulation is conforming Delaunay, meaning that all Voronoi vertices will automatically have a connectivity of three (except at the edge of the tessellation).

Modifying the Network to Comply with Chemical Knowledge. In the following, the connectivity in the triangulation is represented by c. Any point with $c < 5$ and not on the boundary of the triangulation was removed from the network. All points (x,y) with $c > 7$ were divided into two points (x_1,y_1) and (x_2,y_2), where $q_1 = q + \varepsilon_q$ and $q_2 = q - \varepsilon_q$, where q represents x or y and ε_q is a random number, small with respect to the interpoint distances.

All points with $c = 6$ were retained. Points with $c = 5,7$ were deleted or divided, respectively, according to a Monte Carlo procedure, with the acceptance and rejection probabilities chosen as free parameters. A number of additional points were also added to the convex hull of the triangulation, also in a Monte Carlo-type procedure, in order to reduce the number of unphysically long edges. The number of 5-, 6-, and 7-membered rings was determined and the current

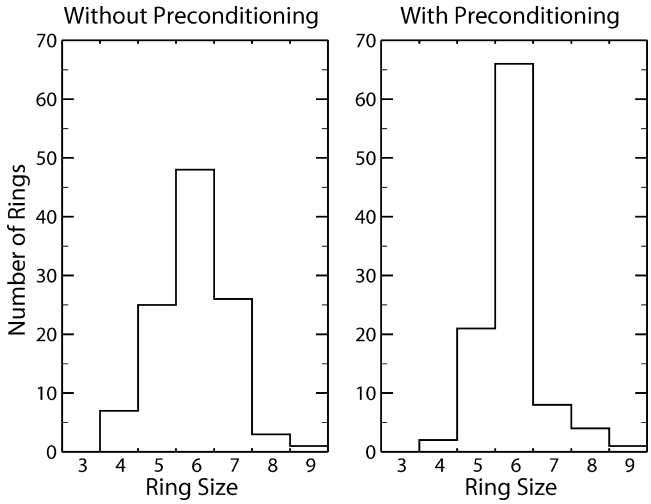


Figure 2. Distribution of ring sizes for a set of points without (left) and with (right) preconditioning.

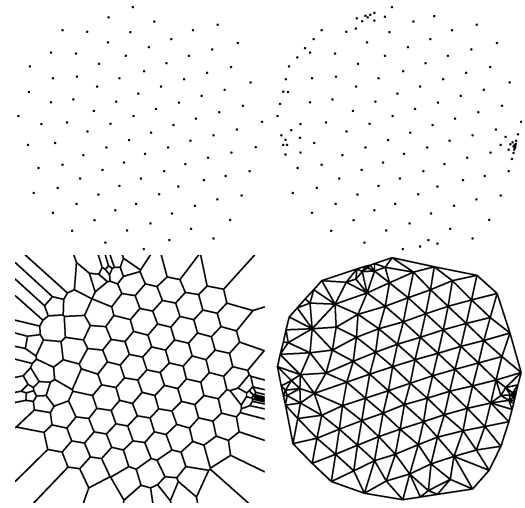


Figure 3. Processes during the generation of graphene sheets, with preconditioning. Clockwise from top-left: initial random points following relaxation, initial points with additional Steiner points to increase the quality of the triangulation, Delaunay triangulation of the points, corresponding Voronoi tessellation.

pentagon/hexagon and heptagon/hexagon ratios calculated. These are then compared with the target ratios determined from chemical knowledge of graphene systems, typically pentagon/hexagon = 3/100 and heptagon/hexagon = 1/100,[31] and compliance with the Triangulation Rule is checked. If all target ratios and network criteria are fulfilled, the network is accepted. Otherwise, further cycles of triangulation are performed (with no further addition of Steiner points) until convergence is found.

Geometry Optimization. Once a satisfactory network of points has been created, this can be converted into a chemical structure with atoms and bonds. In the present work, the structure was imported into the Charmm program,[32] each atom was assigned parameters corresponding to an aromatic carbon atom in the Charmm22 parameter set,[33] and the structure was minimized using Steepest Descent or Newton−Raphson methods to a gradient tolerance of 1×10^{-7} kcal/mol/Å.

Preconditioning of the Triangulation Points. Convergence of the above algorithm is typically slow. This is because no relaxation of the triangulation points is allowed during the refinement, leading to a broad distribution of connectivities and therefore a broad range of ring sizes in the Voronoi tessellation. Refinement is therefore slower, since a larger number of unsuitable rings must be eliminated before convergence is reached. This can be alleviated by allowing relaxation.

A close-packed network of points in a plane produces a perfect triangular network, the dual of which is a network of hexagons. Relaxation of points in a plane will move toward close-packing, giving a narrower distribution of connectivities in the network with a sharp peak at a connectivity of 6, as shown in Figure 2.

Each cycle of refinement (adding/deleting points) was therefore followed by relaxation. Each point was modeled as a lone aromatic carbon atom with the same parameters as above, and energy minimization was performed using the steepest descent algorithm in the Charmm program. Only a small number of minimization steps were performed, to remove as many unsuitable connectivities as possible through local rearrangements, while retaining a certain number of 5- and 7-connective defects, as required. Further minimization effectively leads to the removal of defects from the interior of the lattice through annealing. For comparison with the initial process (Figure 1), a corresponding set of figures showing the process with preconditioning is shown in Figure 3. It is clear that the topology of the final Voronoi tessellation is much closer to that of graphene.

Electronic Structure Calculations. Ab initio electronic structure calculations were performed on a small, defective graphene sheet to determine the most favorable sites for functionalization of the graphene sheet by fluorination. Hydrogen atoms were added around the edge of the sheet to satisfy bonding and ensure a closed-shell singlet ground state. Full geometry optimizations were performed for all fluorination sites on the mainly convex side of the sheet (an additional hydrogen atom was added to the periphery of the sheet to maintain the singlet electronic state), using the Gamess-UK electronic structure program[34] at the HF/3-21G* level. The energies of the optimized structures were then compared to determine the thermodynamically most-favorable fluorination site. Because many finite graphene-type structures have been shown to possess a spin-polarized ground state, even with a perfect hexagonal lattice and edge passivation,[35−38] tests were carried out to ensure that the ground state of the molecular system was neither magnetic nor metallic. Geometry optimizations using both spin-restricted and unrestricted Hartee Fock formalisms were performed for the original graphene structure and for the lowest energy fluorinated structure. In each case, the calculations yielded identical structures and energies. Visual inspection of the frontier molecular orbitals (HOMO−1, HOMO, LUMO, LUMO+1) revealed no differences for α and β electrons in the UHF calcula-

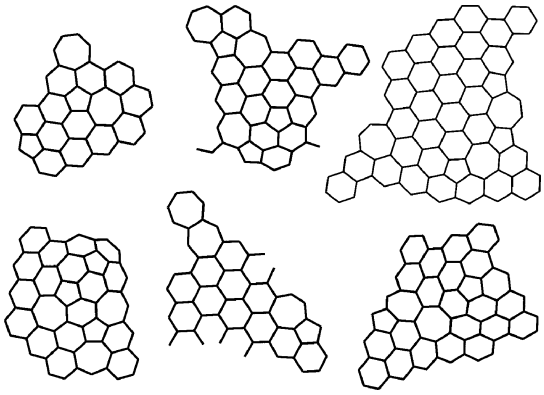


Figure 4. A selection of small defective graphene sheets generated using the new algorithm.

tions. As a result, the spin-restricted Hartree−Fock method was deemed to be appropriate.

Results

Small Structures. For a target ratio of ring sizes (pentagon/hexagon, heptagon/hexagon) of 0.2−0.3, small systems can be generated, with tens of rings. These can be generated with no preconditioning, starting from initial arrays of ∼100 points. A selection of energy-minimized structures are shown in Figure 4. Such structures are suitable for use in ab initio electronic structure calculations to probe the reactivity at various sites relevant to functionalization (see below).

Larger Structures. For smaller target ratios, such as pentagon/hexagon = 0.03 and heptagon/hexagon = 0.01, larger systems with hundreds of rings must be constructed for defects to be observed. This creates certain difficulties in automated production of suitable networks. One particular feature of the Voronoi tessellations produced by the above algorithm is that the majority of pentagonal defects are found around the edges of the network, especially at the armchair edges. This can be observed in Figures 1 and 3. This is an artifact of the calculation of Voronoi tessellation for finite systems. If these edges are included in the final network, not only will the number of pentagonal rings be much higher than required by the target ratio, but the system is also not likely to be representative of extended graphene sheets. However, they are likely to be of relevance for edge effects in graphene, where reconstructions are found to occur.[16]

To avoid introduction of these artifacts, structures were cut from the central region of a tessellation. As the relaxation of the sheet and growth around the edges (because of the addition of extra vertices along long edges) was found to occur in an isotropic manner (the circular nature of the relaxed network is obvious in Figure 3), a circle with a radius of approximately half the maximum radius was cut from the large network. Although it is possible to cut other geometric or random shapes from an extended sheet with no changes to the algorithm, we restrict ourselves to circles for simplicity.

Two representative structures obtained with target ratios of 0.03 (pentagon/hexagon) and 0.01 (heptagon/hexagon) are shown in Figure 5.

The structures generated by the algorithm are governed by a number of parameters, including the probability of dividing and deleting a vertex. Increasing these probabilities leads to an increase or reduction in the number of atoms in the final structure, respectively, as illustrated in Figure 6. The parameters used in the current work are provided in Table 1.

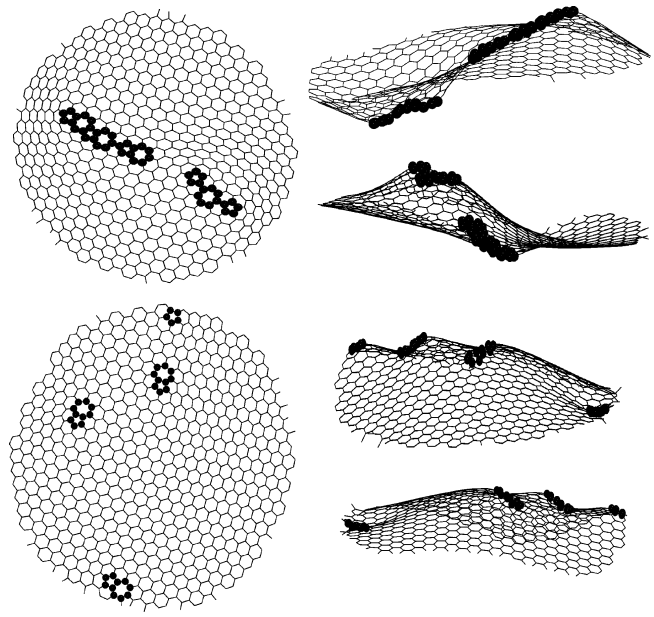


Figure 5. Two defective graphene sheets generated using the new algorithm. Side views of the sheet (revealing curvature) are placed adjacent to the planar view. Atoms forming defect rings are shown in a ball and stick representation.

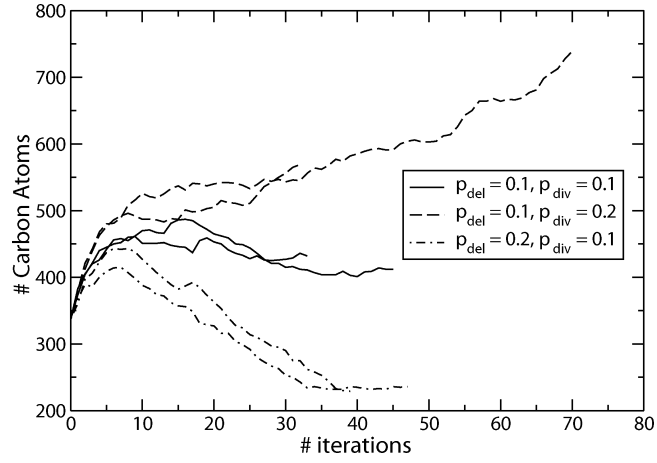


Figure 6. Effect of the probability for the deletion/division of points on the total number of carbon atoms in the structures created for six individual calculations, all of which converged to a valid structure.

Table 1. Free Parameters Used in the Refinement of a 625-Vertex Data Set

probability of deleting a vertex	0.1
probability of dividing a vertex	0.1
probability of adding a vertex to a boundary edge	0.5
min. length of a boundary edge before creation of a midpoint (Å)	10
initial number of minimization steps	500
subsequent number of minimization steps	50

Distribution of Rings. In the structures illustrated in Figure 5, it is noteworthy that 5-ring and 7-ring defects tend

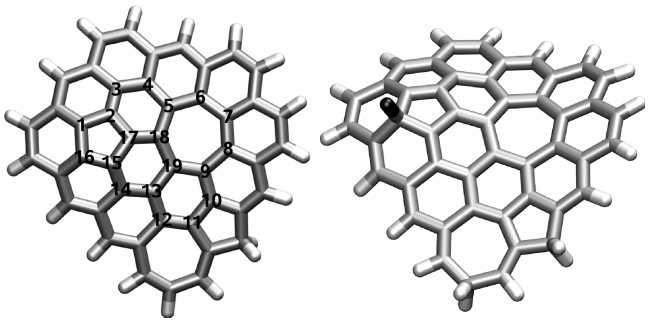


Figure 7. Left: The initial graphene structure showing the numbering of possible fluorination sites. Right: Structure of the most stable fluorinated structure.

to be found close to each other. This is due, in part, to the preconditioning process, in which defects are annealed out to obtain the required defect ratios, resulting in regions of perfect hexagonal ordering with isolated groups of defects. This could be avoided by introducing a more complex preconditioning process, involving a full three-dimensional treatment of the mesh. However, additional experimental data on the proximity of 5- and 7-ring defects would be required before the inclusion of further complexity can be justified.

Identifying the Most Favorable Sites for Fluorination Using Electronic Structure Calculations. The preferred sites for fluorination of a representative graphene structure were determined by ab initio electronic structure calculations. The initial structure contained 2 seven-membered rings, 2 five-membered rings, and 12 six-membered rings, with a total of 47 carbon atoms and is shown in Figure 7. Eighteen hydrogen atoms were added around the edge of the sheet to satisfy bonding and ensure a closed-shell singlet ground state. The relative energies and the local topology of the 19 fluorination isomers investigated are shown in Table 2, with the numbering of the fluorination sites and the lowest energy structure as illustrated in Figure 7. HOMO−LUMO gaps varied between 2 and 6 eV, indicating that the structures are not metallic. It can be seen that most of the low-energy structures are those in which the fluorination occurs at or close to the internal 5-membered ring. This can be rationalized by the positive curvature which already exists at a 5-ring defect. Only a small distortion of the carbon framework is then required to add an additional bonded neighbor and form a tetrahedral carbon atom. Many of the high-energy sites are those around the internal 7-membered ring, where the negative curvature makes it unfavorable to add a fourth bonded neighbor. However, the details are not so clear-cut. Significant distortions of the 7-membered ring appear to be tolerated at some positions making fluorination at sites 18 and 19 more favorable than would perhaps be expected. It is likely that this is caused by the proximity of the 5-membered ring

Table 2. Relative Energies of a Fluorinated Graphene Sheet with Defects

site	E_{rel}/kJ mol^{-1}	local topology[a]
16	0.0	5−6−6
2	87.9	5−6−6
14	114.1	6−6−6
18	154.3	6−6−7
19	196.4	6−6−7
17	200.2	5−6−6
10	213.2	5−6−6
15	216.4	5−6−6
6	232.4	6−6−7
12	234.3	6−6−7
1	238.7	5−5−6
11	249.2	5−6−7
13	267.7	6−6−6
8	269.1	6−6−7
4	282.8	6−6−6
3	287.9	6−6−6
9	309.9	6−6−7
5	319.0	6−6−7
7	333.6	6−6−7

[a] The local topology is defined in terms of the sizes of the three rings that share a common vertex at the position in question. For example, a vertex that is shared between three 6-membered rings (and hence expected to be in a locally planar environment) is labeled "6−6−6". The local topology gives an indication of the expected curvature (zero, positive, negative) of the graphene fragment at the position of interest.

Conclusions

A rapid and robust algorithm has been developed to facilitate the automated generation of graphene sheets containing a majority of 6-membered rings, with a small, user-specified, number of 5- and 7-ring defects.

In comparison with existing methods,[9,21] this algorithm allows the generation of small fragments of graphene (tens to thousands of atoms) according to user-defined criteria within a few minutes on a desktop PC. There is no requirement for periodic boundary conditions or computationally expensive simulations (which can be of the order of days[9]). After initial generation of set of random coordinates, which in the present work requires no special procedures to fulfill artificial periodicity or chemical bonding requirements, the central iterative procedure uses deterministic methodology (triangulation followed by minimization). Chemical information, in the form of an interatomic potential, is only added for the minimization step (which represents the most expensive part of the procedure). Chemical bonding is determined solely by the results of triangulation (which also automatically ensures the correction coordination number), meaning that no special procedures need to be incorporated in order to control bond making and breaking processes.[9,21] As a result, the algorithm presented here provides a simple and efficient solution to the generation of small, defective graphene sheets.

Once generated, these structures can be used in electronic structure calculations to identify favorable sites for functionalization by fluorination. The results obtained indicate that fluorination will occur preferentially at 5-ring defects, although the overall energetics depend on the details of the local topology. A further use of the structures generated could be in the simulation of STM images, providing a further tool for the interpretation and understanding of experimental observations.

Acknowledgment. D.R.N. is grateful to Peter Harris for introducing him to this interesting problem and for helpful discussions.

References

(1) Harris, P. J. F. *Crit. Rev. Solid State Mater. Sci.* **2005**, *30*, 235–253.

(2) Kroto, H. W.; Heath, J. R.; O'Brien, S. C.; Curl, R. F.; Smalley, R. E. *Nature* **1985**, *318*, 162–163.

(3) Iijima, S. *Nature* **1991**, *354*, 56–58.

(4) Iijima, S.; Yudasaka, M.; Yamada, R.; Bandow, S.; Suenaga, K.; Kokai, F.; Takahashi, K. *Chem. Phys. Lett.* **1999**, *309*, 165–170.

(5) Bandow, S.; Kokai, F.; Takahashi, K.; Yudasaka, M.; Qin, L. C.; Iijima, S. *Chem. Phys. Lett.* **2000**, *321*, 514–519.

(6) Charlier, J. C. *Acc. Chem. Res.* **2002**, *35*, 1063–1069.

(7) Iijima, S.; Ichihashi, T.; Ando, Y. *Nature* **1992**, *356*, 776–778.

(8) Ihara, S.; Itoh, S.; Akagi, K.; Tamura, R.; Tsukada, M. *Phys. Rev. B* **1996**, *54*, 14713–14719.

(9) Kumar, A.; Lobo, R. F.; Wagner, N. J. *Carbon* **2005**, *43*, 3099–3111.

(10) Harris, P. J. F. *Phil. Mag.* **2004**, *A84*, 3159–3167.

(11) Gogotsi, Y.; Libera, J. A.; Kalasknikov, N.; Yoshimura, M. *Science* **2000**, *290*, 317–320.

(12) Goel, A.; Hebgen, P.; Vander Sande, J. B.; Howard, J. B. *Carbon* **2002**, *40*, 177–182.

(13) Harris, P. J. F.; Liu, Z.; Suenaga, K. *J. Phys.: Condens. Matter* **2008**, *20*, 362201.

(14) Liu, Z.; Suenaga, K.; Harris, P. J. F.; Iijima, S. *Phys. Rev. Lett.* **2009**, *102*, 015501.

(15) Stone, A. J.; Wales, D. J. *Chem. Phys. Lett.* **1986**, *128*, 501–503.

(16) Koskinen, P.; Malola, S.; Häkkinen, H. *Phys. Rev. Lett.* **2008**, *101*, 115502.

(17) Kobayashi, K. *Phys. Rev. B* **2000**, *61*, 8496–8500.

(18) Ormsby, J. L.; King, B. T. *J. Org. Chem.* **2007**, *72*, 4035–4038.

(19) Malola, S.; Häkkinen, H.; Koskinen, P. *Phys. Rev. B* **2008**, *77*, 155412.

(20) Tošić, R.; Mašulović, D.; Stojmenović, I.; Brunvoll, J.; Cyvin, B. N.; Cyvin, S. J. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 181–187.

(21) Shi, Y. *J. Chem. Phys.* **2008**, *128*, 234707.

(22) Loginova, E.; Bartelt, N. C.; Feibelman, P. J.; McCarty, K. F. *New J. Phys.* **2008**, *10*, 093026.

(23) Mickelson, E. T.; Huffman, C. B.; Rinzler, A. G.; Smalley, R. E.; Hauge, R. H.; Margrave, J. L. *Chem. Phys. Lett.* **1998**, *296*, 188–194.

(24) Claves, D.; Rossignol, J. *Chem. Phys. Lett.* **2009**, *468*, 231–233.

(25) Ghanem, M. A.; Chrétien, J.-M.; Pinczewska, A.; Kilburn, J. D.; Bartlett, P. N. *J. Mater. Chem.* **2008**, *18*, 4917–4927.

(26) Selig, H.; Lifshitz, C.; Peres, T.; Fischer, J. E.; McGhie, A. R.; Romanow, W. J.; McCauley, J. P.; Smith, A. B., III *J. Am. Chem. Soc.* **1991**, *113*, 5475–5476.

(27) Jia, J. F.; Wu, H. S.; Xu, X. H.; Zhang, X. M.; Jia, H. H. *J. Am. Chem. Soc.* **2008**, *130*, 3985–3988.

(28) Sobol, I. M. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 86–112.

(29) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in Fortran*, 2nd ed.; Cambridge University Press: Cambridge, UK, 1992; pp 299−305.

(30) Shewchuk, J. R. Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator. In *Applied Computational Geometry: Towards Geometric Engineering*; Lin, M. C., Manocha, D. , Eds.; Springer-Verlag: Berlin, Germany, 1996; Vol. 1148, pp 203−222.

(31) Harris, P. J. F. Personal communication, 2008.

(32) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(33) MacKerell, A. D., Jr. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(34) Guest, M. F.; Bush, I. J.; van Dam, H. J. J.; Sherwood, P.; Thomas, J. M. H.; van Lenthe, J. H.; Havenith, R. W. A.; Kendrick, J. *Mol. Phys.* **2005**, *103*, 719–747.

(35) Ezawa, M. *Phys. Rev. B* **2007**, *76*, 245415.

(36) Fernández-Rossier, J.; Palacios, J. J. *Phys. Rev. Lett.* **2007**, *99*, 177204.

(37) Kudin, K. N. *ACS Nano* **2008**, *2*, 516–522.

(38) Hod, O.; Scuseria, G. E. *ACS Nano* **2008**, *2*, 2243–2249.

CT900113F

JCTC Journal of Chemical Theory and Computation

Pressure Annealing as a Complement to Temperature Annealing To Find Low-Energy Structures of Oligomeric Molecules

Christopher Adam Hixson and Ralph A. Wheeler*

Department of Chemistry and Biochemistry, University of Oklahoma, 620 Parrington Oval, Room 208, Norman, Oklahoma 73019

Received October 27, 2008

Abstract: Finding the lowest-energy geometry of a molecule or collection of molecules is a fundamental challenge of modern computational chemistry and is closely related to the more general problem of optimizing a function. Temperature annealing, popularly called simulated annealing, is a powerful and commonly used technique, but it is not well suited to conformational sampling of long, oligomeric molecules. A method is presented herein that incorporates pressure as an optimization parameter to complement temperature annealing, and several tests of its effectiveness are described. Bayesian statistical analysis shows that pressure−temperature annealing confers no advantage in control simulations of Lennard-Jones particles, but it yields lower-energy structures than pure temperature annealing with significant credibility for two model polyethers, monoglyme ($CH_3OCH_2CH_2OCH_3$) and tetraglyme [$CH_3(OCH_2CH_2)_4OCH_3$].

1. Introduction

Polymers or long-chain molecules designed to model polymers are generally difficult candidates for computational study because their multiple conformations make it difficult to find low-energy structures. This manifests as a glassy potential energy surface that complicates application of computational methods[1] typically employed in studying the structural and thermodynamic properties of these systems such as molecular dynamics (MD) or Monte Carlo (MC) methods.[2] Generally, MD and MC computations are supplemented by techniques such as simulated (temperature) annealing to improve the quality of the results.[3−9] This is especially true in structural studies, particularly when searching for low-energy states of such systems.[6−8] Alternative methods are available and are commonly used,[10,11] as no one class of optimization methods is suited to all problems. A technique is presented here that expands temperature annealing by using pressure as an additional control parameter. This extra parameter is tested to assess its utility in the special case of frustrated, glassy systems. First, we review optimization techniques employed in the study of complex systems, such as those studied here. Then, we explain the advantages of the temperature annealing approach and show how the addition of pressure as a parameter modifies the method. Finally, we present statistical analyses that demonstrate the value of our proposed approach.

1.1. Popular Molecular-Dynamics-Based Optimization Methods. One of the most important pieces of information useful to understanding chemical phenomena is structure. Understanding the spatial arrangement of the atoms contained in a chemical system might seem rather basic, but it is necessary to provide insights into the thermodynamics, reactivity, and spectroscopic properties of the system. Of course, determining the structure of chemical systems can be accomplished experimentally using X-ray crystallography,[12] NMR spectroscopy,[13] and other methods.[14] Quite commonly now, structural models are proposed based on computational studies. Using the quantum approach,[15] structures of small or even medium-sized molecules are routinely determined to great accuracy. For larger systems, including condensed-phase structures, accuracy is traded for computational efficiency, and molecular mechanics methods are employed.[16] Molecular mechanics methods take, for example, the AMBER force field[17] and use optimization methods traditionally used in mathematics, such as the conjugate gradient algorithm,[18] to find minima of the energy

* Corresponding author e-mail: rawheeler@ou.edu.

10.1021/ct800451c CCC: $40.75
Published on Web 05/22/2009

function. This is a useful approach, but it has limited utility for a variety of reasons, including limiting the search to finding local minima "near" the starting structure. More commonly used is molecular dynamics or Monte Carlo,[2] coupled with some optimization algorithm. Some popular algorithms, including temperature annealing (usually called simulated annealing),[4] a popular algorithm fundamentally important to this contribution, are discussed below.

1.1.1. Locally Enhanced Sampling. Locally enhanced sampling (LES)[19] is a technique that was first proposed in 1990 (the original idea was taken from another method)[20] and was initially designed to improve the search for diffusion paths of small ligands inside a protein matrix. It has since become a moderately popular optimization method, useful when the structure of a small part of the system is needed in relation to the remainder. This small part of the system is copied several times. None of the copies directly interacts with the others, but instead, each one interacts directly with the remainder of the system, generally referred to as the bath. Atoms in the bath interact normally with each other, but they interact with the average of each of the copied systems. It is often claimed that this algorithm allows the interaction between the copied parts of the system and the bath to be smoothed. Specifically, it has been reported that any barriers resulting from this interaction would be decreased proportionally to the number of times the smaller part of the system was copied.[21] Individual copies frequently gain energy during the simulation, allowing for a sort of "tunneling" behavior. Thus, each copy can be given a greater allotment of energy than in traditional molecular dynamics, and so can better explore the potential energy surface, until the energy is redistributed to the other systems via the bath interaction.[22]

The method has been used to explore potential energy surfaces in several situations.[11,21,23−26] The trajectory mapping application used in the original article is an important example, and there have been other applications of LES to finding low-energy structures.[11,21,23−26] Although not as commonly used as the method described in the next section, LES is still a relatively popular method and has been implemented in a variety of molecular dynamics packages in common use.

1.1.2. Replica Exchange. The replica-exchange molecular dynamics simulation method is quite popular. First proposed in 1999,[10] the method has a relatively straightforward implementation. The entire system under study is replicated a number of times. Each replica is independent and is held at a different temperature, with the set of replicas spanning a range of temperatures. The systems are allowed to evolve using molecular dynamics for a set number of time steps, and then, the temperatures of two systems are exchanged if they pass a Metropolis test. Because the probability of acceptance is low if the temperatures vary too much, only exchanges between systems of neighboring temperatures are typically attempted.

Generally described as finding minima on the free energy surface, replica exchange has been used to study a variety of systems, including polypeptides, proteins, and polymers.[10,27−32] The method has shown great utility in mapping low-energy structures of peptide systems, although many other examples can be found in the literature. This method owes a large part of its existence to temperature annealing, as the methods are quite similar in approach, although temperature annealing has been in use for a much longer time.

1.2. Temperature Annealing. The idea that a state having a given energy is populated with a calculable relative probability is one of the most fundamental ideas in statistical physics and is immortalized (at least for the canonical ensemble) in the Boltzmann distribution. Given this link between such a readily obtainable quantity (the energy) and relative population, it should come as no surprise that an optimization strategy, temperature annealing, would be formed from it. Simply put, temperature annealing[4] is a process in which an ensemble of states of a system is generated corresponding to high-energy conditions (by raising the temperature) and the system is allowed to evolve toward a lower-energy state. As cooler states are generated, the system tends to settle into areas of low energy, because, at lower temperatures, these areas are more likely to be populated. Raising the temperature initially widens the search area by increasing the volume of phase space available to be populated because the systems have a greater probability of surmounting barriers separating regions of phase space. Then, lowering the temperature traps the system in wells with a probability that depends on their energy and phase-space volume.

One surprising aspect of temperature annealing is that, despite its specific link to statistical physics, it has shown its worth in a variety of fields as a general-purpose optimization technique.[4] Such general use can be contemplated because a wide variety of problems exist that allow a cost function analogous to the energy to be defined. In one famous example, the traveling salesman problem, the object is to find the quickest path connecting two points along some complicated linkage structure sharing features found in road maps. The energy is represented by a cost function generally chosen to be a function of the number of nodes the salesman must cross. This function is optimized, generally, using a Monte Carlo procedure where each newly generated state is compared with the previous one and either accepted or rejected according to the Metropolis criterion.[2] The temperature is a fictitious parameter, but by increasing it, the algorithm accepts trial moves with a greater frequency, thereby allowing states of higher "energy" to be visited. Lowering the temperature biases the simulation into lower-energy states. This behavior is exactly analogous to the role of temperature in condensed-phase systems. Similar methods have been used to solve problems in electrical engineering to design circuits and in signal processing to process images and sounds.[4] Finally, the most traditional area of application is in finding energy minima in condensed systems either by using the Metropolis/Monte Carlo method outlined above or by using molecular dynamics as the ensemble-generation engine.[2]

1.3. Role of Pressure. Complementing temperature annealing with pressure annealing to locate low-energy geometries more quickly has the potential of improving the quality of temperature annealing searches. As described above, one

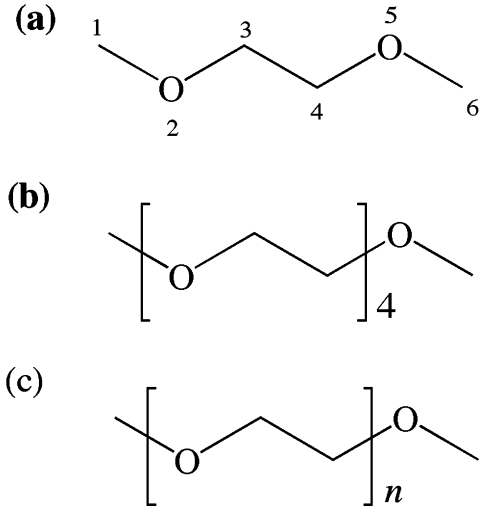


Figure 1. Structures of the oligomeric models used in this work: (a) monoglyme and (b) tetraglyme. These structures are oligomeric models of (c) the polymer polyethylene oxide.

role of temperature is to help define the phase-space area to search in a temperature annealing optimization. If the system is located in an area of phase space separated cleanly from another by a large barrier, the separated area will only rarely be visited within the calculation. This situation requires a particularly long simulation to obtain an accurate result. As the temperature is raised ever higher, the structure becomes more likely to surmount such a barrier; however, raising the temperature can become counterproductive by causing the simulation to become unstable or increasing the simulation time. Consequently, making more sensible changes to other simulation variables such as pressure is an avenue worth exploring. Adjusting the pressure indirectly affects the optimization because the pressure determines the volume (and thus the density) of the system, which, in turn, affects the amount of phase space available to be explored. It increases the fraction of explorable phase space by a different approach than using temperature alone. The effects of such changes are explored below.

1.4. Objectives of This Work. Several systems are tested here to illustrate the benefits that using pressure as an optimization control parameter can add to a temperature annealing optimization strategy. First, the method is tested on a system composed of Lennard-Jones particles.[33] This system is relatively simple and is used here are a control. Because the glassy nature of their energy surfaces makes conformational trapping a serious issue,[1] simulations of polymers and polymer models are expected to benefit greatly from pressure annealing. Thus, we also test two models for the polymer polyethylene oxide (PEO)[34−36] called monoglyme and tetraglyme (see Figure 1). PEO is interesting for several applications, including use as the matrix for polymer-ion batteries.[37,38] Monoglyme is an oligomer containing one unit of the PEO polymer, tetraglyme contains four units, and both of these systems have previously been studied.[34−36]

2. Statistical Mechanical Basis of Pressure Annealing

Temperature annealing is among the most frequently used optimization methods in chemistry for a variety of reasons. First, the method makes a great deal of intuitive sense and can be explained almost entirely using basic appeals to logic. On the other hand, temperature annealing has been formally proven to be an optimization technique and is on a firm mathematical basis.[5,39] It has even been shown that, when a proper cooling schedule is used, the method is guaranteed to find the global minimum of the system.[39] Although, in practice, the conditions required to realize the promise of temperature annealing are impossible to achieve completely, the fact that the method could, in principle, attain success is quite appealing. Perhaps this combination of being easily understood, effective, and firmly rooted in theory explains why temperature annealing is one of the most popular optimization methods in use, not only in chemistry, but also in a wide variety of other fields. What follows is an abridged summary of temperature annealing, to substantiate the pressure−temperature annealing procedure proposed in this work.

2.1. Temperature Annealing. A simple, qualitative way of explaining why temperature annealing works starts by noting that low-temperature states corresponding to low energies should be populated with greater probability than higher-energy states, whereas at higher temperatures, the populations are less strictly related to the energy of the state. This is due to the flattening of the probability distribution observed at higher temperatures. In the high-temperature phase of the optimization, the system wanders and explores phase space with relatively lax restrictions and freezes into the states of higher probability at lower temperatures. It has been proven that repeated application of heating−cooling cycles is guaranteed eventually to find the lowest-energy states.[5,39] It is also instructive to note that temperature annealing finds the minimum of the energy E by manipulating the parameter β in the expression

$$Q(N, V, T) = \frac{1}{C}\int_{-\infty}^{\infty} \mathrm{e}^{-\beta E}\Omega(N, V, E)\,\mathrm{d}E$$

where T is the temperature (β is the inverse of the product of the temperature and the Boltzmann constant), V is the volume, C is a normalization constant, and N is the number of particles. Q is the canonical partition function, and Ω is the microcanonical partition function. In this framework, simulated annealing can be interpreted as finding the lowest energy for a system containing N particles in a volume V.

2.2. Pressure Annealing. To understand how pressure can be used as an optimization variable, consider an expression similar to the canonical ensemble partition function written above, but for a related ensemble. Because the pressure divided by the temperature is conjugate to the volume,[40] we know that if βp (where p is the pressure of the system) is fixed, thus allowing the volume to fluctuate, we can write a partition function for that ensemble as

$$M(N, \beta p, E) = \frac{1}{C}\int_{0}^{\infty} \mathrm{e}^{-\beta pV}\Omega(N, V, E)\,\mathrm{d}V$$

All of the variables are the same as described in the previous section, and M is the partition function for the *NPE* ensemble. This expression for the partition function is reasonable given that $\Omega \sim V^N$ is weighted by $e^{-\beta pV}$ as V gets large. It is also apparent that, as the pressure increases, the volume decreases. Conversely, decreasing the volume increases pressure. When the pressure vanishes, the volume becomes unbounded. By manipulating βp as the temperature was manipulated in temperature annealing, the function V is minimized. This can be interpreted as finding the smallest volume able to contain N particles with an energy E and implies that pressure annealing alone is unlikely to prove effective as an energy optimization strategy.

2.3. Benefit to Including Pressure in Simulated Annealing Optimizations. In the isothermal−isobaric ensemble (*NPT*), pressure and temperature are fixed in the partition function. Using this knowledge, we constructed an optimization strategy to take advantage of pressure as a parameter alongside temperature in a pressure−temperature simulated annealing strategy. From the analysis in the previous two sections, it is clear that using pressure as an optimization parameter does not act to optimize the energy; however, the pressure does control features of a simulation that can be exploited to aid in simulated annealing. If the pressure is fixed at a small value, the density of the system decreases as the volume of the system increases. For some systems, this is not particularly helpful for optimization, but for bulky systems where geometry changes are hindered at the density of interest, increasing the available volume can be quite helpful. By allowing the molecules to separate, intramolecular energies can be minimized more easily. Then, by increasing the pressure, the molecules can be compressed efficiently into a low-energy state, optimizing the volume using the procedure explained in the previous section. Using these principles, pressure−temperature simulated annealing strategies can be designed. The strategy presented here consists of four steps. The first step allows an expansion to occur from the initial state by fixing the temperature and pressure at low values. After the expansion, the volume is fixed, and traditional temperature annealing is performed on the expanded system. Third, after temperature annealing is completed, the pressure is fixed to a large value, and the temperature remains fixed at the final annealing temperature. Finally, after the compression, the system's energy is minimized. The totality of the method allows two complications of temperature annealing to be addressed. First, in condensed-phase systems of molecules with hindered rotations, inducing intramolecular conformational changes can be difficult. This can also lead to difficulty in changing the way the system packs.

Understanding how lowered density might benefit a simulation is not difficult. Imagine the simple two-dimensional system depicted in Figure 2. This system consists of a string of eight beads connected to each other in a coil by springs of length r. The beads do not interact except through the springs and can be considered hard spheres. The entire coil is contained in a box with sides of length L. If an optimization algorithm seeks to characterize the low-energy states of this system, moving from the counterclockwise coil depicted in the figure to the clockwise coil would be an important transition. The height of the lowest barrier between these two states determines the difficulty of making the transition. If the length of the box is $8r$ or greater, no barrier to the transition exists because the system can align as a

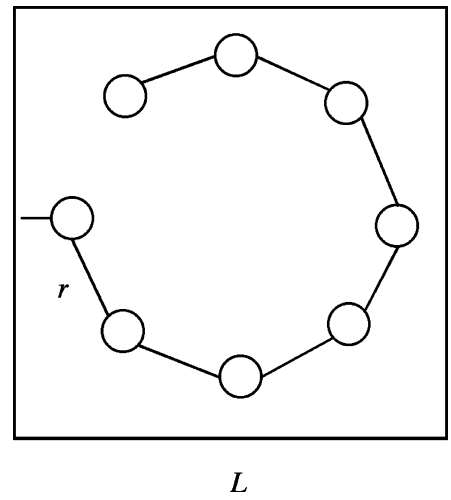


Figure 2. Simple two-dimensional model representing a coiled molecule. The molecule exists inside a box whose walls are treated as large potential energy barriers, each side of which has length L. The molecule is composed of beads connected by springs, whose equilibrium length is r. When the length of the side of the box is greater than $7r$, the molecule can change coil orientation without needing to surmount any barrier, illustrating that increasing the available volume can make conformational searches easier.

Table 1. Final Energies (kcal mol^{-1}) Determined by Temperature Annealing and Pressure−Temperature Annealing Algorithms for Each of the 50 Lennard-Jones Systems Studied[a]

system	temperature annealing	pressure−temperature annealing	system	temperature annealing	pressure−temperature annealing
1	−953.54	−959.72	26	−955.89	−954.60
2	−955.53	−958.15	27	−954.38	−957.08
3	−951.82	−941.17	28	−957.22	−957.02
4	−963.88	−958.96	29	−955.10	−964.38
5	−959.86	−957.37	30	−953.17	−954.55
6	−955.72	−962.54	31	−953.00	−954.04
7	−958.82	−960.63	32	−953.45	−953.38
8	−958.23	−952.44	33	−942.91	−955.55
9	−957.33	−957.85	34	−956.65	−949.70
10	−952.84	−955.26	35	−955.83	−955.21
11	−960.52	−958.35	36	−957.34	−959.68
12	−944.61	−951.79	37	−960.03	−959.50
13	−955.15	−957.99	38	−956.19	−948.28
14	−960.71	−947.03	39	−957.45	−960.63
15	−956.07	−951.94	40	−961.74	−956.73
16	−955.63	−960.72	41	−952.40	−962.08
17	−960.70	−953.16	42	−953.11	−961.18
18	−958.62	−956.65	43	−958.97	−960.29
19	−956.99	−957.29	44	−959.52	−961.30
20	−959.05	−953.37	45	−959.73	−959.84
21	−959.97	−956.79	46	−952.60	−958.94
22	−957.59	−955.11	47	−945.38	−958.56
23	−957.30	−954.24	48	−953.92	−963.10
24	−960.38	−945.11	49	−953.77	−956.70
25	−958.37	−952.55	50	−958.79	−955.04

[a] Each pair of simulations started from the same randomly generated initial structure. Neither method seems to be superior to the other, as the lowest energy was found with equal likelihood by the two methods.

Table 2. Final Energies (kcal mol^{-1}) Determined by Temperature Annealing and Pressure−Temperature Annealing Algorithms for Each of the 25 Monoglyme Systems[a]

system	temperature annealing	pressure−temperature annealing
1	−734.8209	−741.6008
2	−722.3525	−741.6008
3	−721.0294	−741.6008
4	−735.0073	−741.6008
5	−737.5804	−741.6008
6	−729.8086	−741.6008
7	−728.0307	−741.6008
8	−741.7752	−741.6008
9	−727.7178	−741.6008
10	−730.5577	−741.6008
11	−729.5520	−736.1320
12	−729.5520	−737.8802
13	−729.5520	−738.9512
14	−729.5520	−740.6107
15	−729.5520	−725.5618
16	−729.5520	−734.7025
17	−729.5520	−717.1497
18	−729.5520	−732.1472
19	−729.5520	−733.4708
20	−729.5520	−741.2944
21	−744.3963	−715.6028
22	−737.9864	−751.5224
23	−719.8457	−742.0861
24	−734.7466	−737.3870
25	−728.8936	−733.7409

[a] Each pair of simulations started from the same randomly generated initial structure. The pressure−temperature annealing algorithm preferentially gave the lowest energy approximately 80% of the time.

straight line while maintaining each of the springs at its equilibrium length. If the box is smaller than the radius of the semicircle, the system cannot make the transition to the low-energy state. Systems complicated enough to be of chemical interest must account not only for this constraint, but also for different packing arrangements. Both are dealt with by the pressure annealing procedure. Choosing the details of pressure annealing simulations with these facts in mind are important to a successful simulation.

3. Model Systems and Procedure

Several systems were used to compare pressure annealing with temperature annealing. These systems range from a relatively simple model Lennard-Jones system to oligomeric models of polyethylene oxide. In each case, two sets of simulations were performed. First, traditional temperature annealing was performed. In these simulations, the system was elevated in temperature starting from 25 to 1025 K over the course of a simulation on the nanosecond time scale. After the temperature annealing simulation, the system was further minimized using a conjugate gradient method. The minimized structure and energy were reported. Second, pressure−temperature annealing was performed. These simulations can be divided into three distinct steps. The first step involves heating while allowing the system to expand slowly. The second phase allows the system to expand dramatically while maintaining the hottest temperature of the initial annealing cycle. Finally, the system is compressed to its original volume while being cooled. The system is then subjected to the same minimization procedure as in the simulated annealing simulation, and the final minimized structure and energy are reported.

Table 3. Final Energies (kcal mol^{-1}) Determined by Temperature Annealing and Pressure−Temperature Annealing Algorithms for Each of the 25 Tetraglyme Systems[a]

system	temperature annealing	pressure−temperature annealing
1	−105.0748	−103.9599
2	−98.6349	−99.8905
3	−100.5817	−107.2664
4	−105.0766	−105.9834
5	−106.7726	−98.3929
6	−102.2849	−101.1396
7	−114.2732	−102.1396
8	−107.0713	−107.6627
9	−105.9728	−107.8258
10	−98.4871	−100.2481
11	−81.9171	−91.4357
12	−118.1657	−107.4743
13	−91.2899	−98.9836
14	−96.3824	−124.9147
15	−83.3597	−110.0198
16	−100.6137	−116.6695
17	−110.3598	−101.1773
18	−87.3115	−108.4409
19	−101.0185	−113.5207
20	−103.8985	−109.7915
21	−94.2803	−108.7402
22	−89.5132	−94.8436
23	−108.8482	−100.1544
24	−87.5746	−96.8615
25	−97.6049	−94.4017

[a] Each pair of simulations started from the same randomly generated initial structure. The pressure−temperature annealing algorithm preferentially gave the lowest energy approximately 70% of the time.

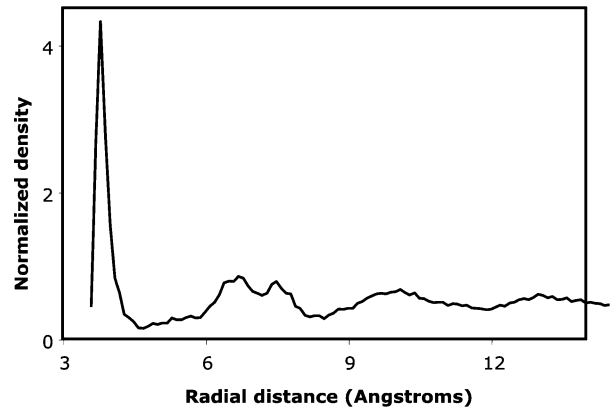


Figure 3. Radial distribution function formed from the data taken from the lowest-energy structure found in the Lennard-Jones simulations. It is representative of our results and compares well with previously obtained results.

3.1. Particles with Only Lennard-Jones Interactions. This model system contained particles of argon with a mass of 40 amu. Here, the force between the particles is the (12,6) Lennard-Jones model for van der Waals forces, written as

$$V(r) = 4\varepsilon\left(\frac{\sigma^{12}}{r^{12}} - \frac{\sigma^{6}}{r^{6}}\right)$$

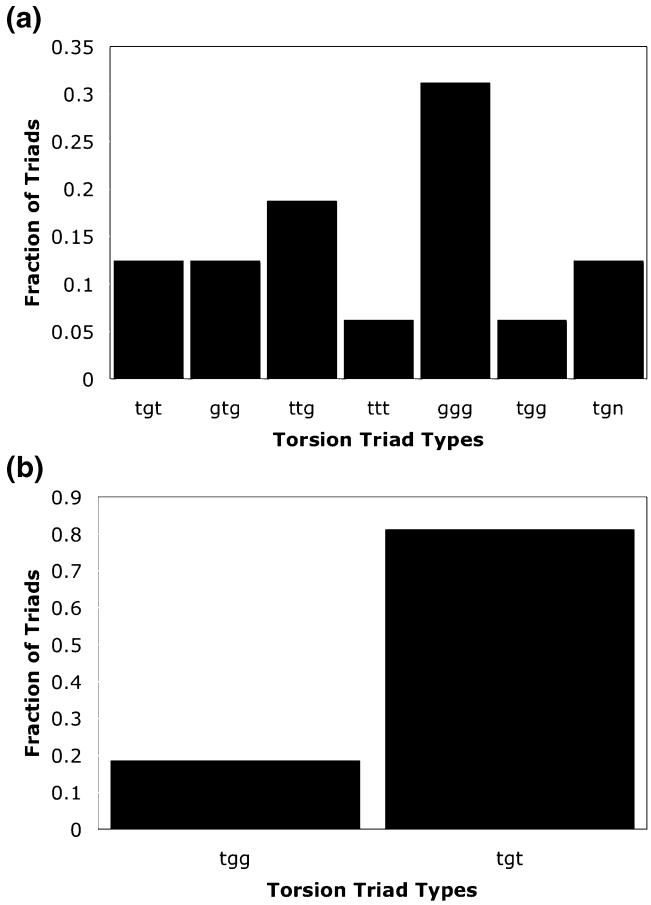


Figure 4. Pair of histograms detailing the population of triad types found in a representative monoglyme system (a) before and (b) after the simulation was performed. In the initial state, a wide variety of triad combinations existed in the system. After the optimization procedure, only two showed appreciable populations, both of which represented low-strain structures as determined by previous work.

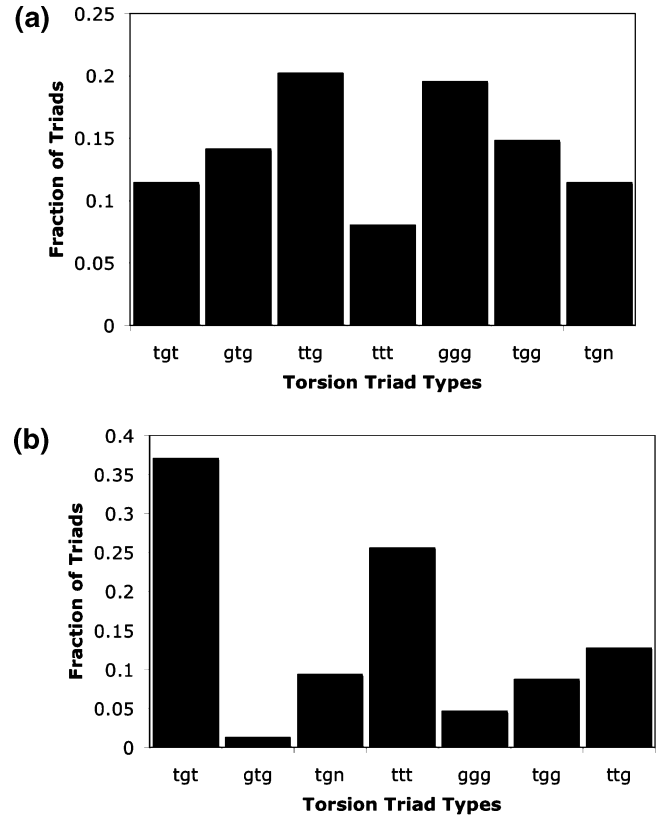


Figure 5. Pair of histograms detailing the population of triad types found in a representative tetraglyme system (a) before and (b) after the simulation was performed. In the initial state, a wide variety of triad combinations existed in the system. After the optimization procedure, the populations shifted. Triad combinations shown to be common in tetraglyme systems increased in population, whereas those rarely represented in previously published studies decreased.

The parameters for argon were taken from the literature,[41] and the simulation was performed using the NAMD programs[42] with a time step of 1 fs and a nonbonded cutoff of 12 Å. Each simulation involved 540 particles randomly placed inside the box and then minimized for 5000 conjugate gradient minimization steps. For the temperature annealing calculations, the volume of a cubic box was fixed at 30 Å per side. The temperature ranged from 25 to 1025 K in 100 K increments with 50-ps simulations at each temperature. After the highest temperature had been reached, the simulation was run for 200 ps and then cooled by reversing the heating schedule. After the system had returned to 25 K, a final equilibration was performed for 50 ps. The final structure and energy were obtained after a 5000-step conjugate gradient minimization. The pressure−temperature annealing simulations started from the same configuration as the corresponding simulated annealing run. They differed in the following ways: During the heating phase, the pressure was fixed at 0.1 bar, and the system was allowed to expand. During the first 50 ps, the system was held at 1025 K, and the pressure remained at 0.1 bar. During the next 100 ps at 1025 K, the system was allowed to expand dramatically, as the pressure was reduced to 0.025 bar. By the end of this period, the box generally expanded to approximately 40 Å on each side. During the final 50 ps, the system was held at 1025 K. During the rest of the cooling phase, the pressure was increased to as much as 1000 bar, until the system returned to its original volume, and then the volume was fixed. The final 50-ps equilibration and 5000-step minimization were then performed, and the final energy and structure were recorded.

3.2. Monoglyme. Monoglyme (Figure 1) is the smallest oligomeric analogue of the polymer polyethylene oxide (PEO), containing a single ethylene oxide repeat unit, capped with methyl and methoxy groups on either end. The simulations were performed using NAMD and a force field previously derived for PEO analogues.[34] Each of the simulations was performed after 120 molecules had been placed in a random orientation and then placed randomly in the box ensuring that no two atoms were closer than 1.9 Å apart. Again, the traditional temperature annealing runs were performed at a fixed volume in a cubic box having sides of length 26.25 Å. The procedure was exactly as described above for argon, with the only differences noted here. During the heating phase of the pressure−temperature annealing procedure, the pressure was set for 0.05 bar, but it was lowered to 0.025 bar to expand the system. The final box size was approximately 110 Å. The system was compressed during the cooling phase with a pressure of as much as 400

bar until the original volume was restored, although as the volume neared its original value, the pressure was slowly reduced to avoid overcompression.

3.3. Tetraglyme. Tetraglyme (Figure 1) is an oligomeric analogue for polyethylene oxide (PEO) containing four repeat units. The simulations were performed using NAMD and the same force field as used for monoglyme and intended for PEO analogues.[34] Each of the simulations was performed after 50 molecules had been placed in a random orientation and then placed randomly in the box ensuring that no two atoms were closer than 1.9 Å. Again, the traditional temperature annealing runs were performed at a fixed volume in a cubic box having sides of length 26.5 Å. The procedure was exactly as described above, with the only differences described here. During the heating phase of the pressure−temperature annealing procedure, the pressure was set to 0.1 bar, but it was lowered to 0.025 bar to expand the system. The final box size was approximately 100 Å. The system was compressed during the cooling phase with a pressure of 500 bar until the original volume was restored, although as the volume neared its original value, the pressure was slowly reduced to avoid overcompression.

3.4. Procedure. For each of the systems described above, a number of random structures were generated. For the Lennard-Jones system, 50 random configurations were generated. For the polymer models, 25 configurations were generated. For each of these random structures, both temperature annealing and pressure annealing simulations were performed, and the final energies and structures were recorded.

4. Results and Discussion

The data collected in the simulations described in the previous section demonstrate that the use of pressure as a coordinate in optimization simulations significantly affects the energies of the structures obtained. For 52% of the Lennard-Jones simulations, 84% of the monoglyme simulations, and 68% of the tetraglyme simulations, the energy of the state produced by the pressure annealing procedure was lower than that produced from the same starting structure but by following the traditional temperature annealing procedure. Each of the final states was characterized by a variety of structural measurements including radial distribution functions, radii of gyration, mean square end-to-end differences, and characteristic ratios. Torsion triad analysis was also performed for the polymer model systems. Each of the structures produced structural data consistent with published results.[34−36] Following is a more detailed look at the results, along with a discussion of their significance.

4.1. Energy Results. The final energies from each simulation started from a random structure are listed in Tables 1−3. They reveal two interesting features. First, of the 50 simulations performed on the Lennard-Jones system, only 52% of the random structures yielded a lower energy during the pressure annealing simulations. This indicates that pressure does not have a significant effect on the results of the simulation, as the results are randomly distributed within the two simulation techniques. For the other two types of simulations, the effect is more pronounced. For the monoglyme simulations, the pressure annealing method generated a lower energy 84% of the time, and for the tetraglyme simulations, pressure annealing generated a lower energy 68% of the time. Whereas the Lennard-Jones system shows no apparent advantage of pressure annealing, the bulkier models adopt structures of noticeably lower energies when the pressure annealing method is used.

4.1.1. Bayesian Statistical Analysis. To answer more precisely the question of whether simulations using pressure plus temperature annealing give lower-energy structures than simulations using temperature annealing alone, we appeal to Bayesian statistics. For similar questions that have a binary success/fail answer, the binomial distribution has been used. This discrete function

$$f(\theta;N,k) = \binom{N}{k}\theta^k(1 - \theta)^{N-k}$$

defines the probability for k successes to occur in a data set of N trials, where the probability of success is θ for each individual trial.[43] In the classic example of a coin toss, we can define a "success" as the coin landing on heads, and we generally believe $\theta = 0.5$. In our present work, we can define a success as a trial where the pressure annealing simulation provides a lower energy than the corresponding temperature annealing simulation. In our work, however, we are interested in estimating the value of the unknown probability θ. This value indicates the fraction of times we would expect pressure annealing to outperform temperature annealing given a random starting structure.

An estimation of θ can be readily achieved using Bayesian statistical analysis.[44] Even more usefully, Bayesian analysis allows the "credibility" that θ lies within a given interval to be estimated. This is accomplished by application of Bayes' theorem. This theorem links two complimentary conditional probabilities together

$$P(\theta|D) \propto P(D|\theta)\,P(\theta)$$

In the language of Bayesian statistical analysis, $P(\theta)$ is called the prior distribution; $P(D|\theta)$ is the conditional probability that our data, D, should be observed given any probability θ; and $P(\theta|D)$ is called the posterior distribution and represents our revised beliefs about the problem given both our prior opinion and our observed data. Thus, we can estimate the conditional probability, $P(\theta|D)$, if both $P(\theta)$ and $P(D|\theta)$ are known or can at least be plausibly represented.

We use the previously introduced binomial distribution to represent the conditional probability of our data (given θ). We choose the beta distribution[43]

$$g(\theta;\alpha,\beta) = \left(\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\right)\theta^{\alpha-1}(1 - \theta)^{\beta-1}$$

to represent our prior distribution. The beta distribution is a continuous-variable analogue of the binomial distribution (as α and β can take real values) and is chosen because it is a conjugate prior distribution. A prior distribution is conjugate when the chosen prior and conditional probabilities combine to give a posterior distribution with the same form as the prior distribution.

We analyze our simulations using four distinct choices for the prior distribution. In the first case, an "unbiased" distribution, we select $\alpha = 1$ and $\beta = 1$. This choice means that any choice of θ has equal probability. In the second case, an "even" distribution, we select $\alpha = 25$ and $\beta = 25$. This choice gives almost zero probabilities for values of θ far from 0.5, but relatively high probabilities in that region. The other two choices give high probabilities for choices of θ higher than 0.5 (the "over" distribution) and for choices of θ lower than 0.5 (the "under" distribution). These correspond to functions that select $\alpha = 25$ and $\beta = 5$ and $\alpha = 5$ and $\beta = 25$, respectively.

Thus we model our posterior distribution with the following functional form

$$P(\theta|D) = Cf(\theta;N,k)\, g(\theta;\alpha,\beta) = \left(\frac{\Gamma(N+\alpha+\beta-1)}{\Gamma(\alpha+k)\,\Gamma(N-k+\beta)}\right)\theta^{k+\alpha-1}(1-\theta)^{N-k+\beta-1}$$

and we can determine the credibility that $\theta > 0.5$ (which is expected if the pressure annealing technique is a better alternative than temperature annealing) by finding the value of the integral

$$I = \int_{0.5}^{1} P(\theta|D)\, \mathrm{d}\theta$$

The results of this analysis are given in Table 4.

The results are consistent with our analysis in the previous section. The control test involving the Lennard-Jones simulations indicates a minimal advantage to using pressure annealing, whereas both the monoglyme and tetraglyme test cases indicate a particularly clear advantage to pressure annealing. Even when analyzed using a prior distribution that is strongly biased against the favorability of pressure annealing, the credibility for the interval $0.5 < \theta < 1$ is still 34% for monoglyme.

4.1.2. Implications of Statistical Analysis. The results of the previous section support the hypothesis that the expansion of the system under low pressure, followed by compression, allows the system to pack more efficiently. The point particles represented in the Lennard-Jones simulations have little packing complexity, whereas each of the polymer models has not only intermolecular packing concerns but also intramolecular packing to take into account. Expansion frees the system to perform intramolecular reorganization, and compression helps find an optimal intermolecular arrangement. Based on these results, using pressure as an optimization parameter seems to benefit the search for low-energy structures of hindered systems.

4.2. Structural Results. It is important to verify that each of the states generated correspond to relevant structures that have been previously observed. Therefore, each of the states generated by these methods was characterized in a variety of ways and checked against previous results. For the Lennard-Jones simulations, pair radial distributions functions were generated and compared with those of similar systems. For the polymer models, radii of gyration and mean square end-to-end distances were calculated and compared. These values allowed calculation of the system's characteristic ratio, an additional check. Finally, the torsion angles along the backbone of the polymer models were analyzed using a technique called torsional triad analysis, which can be further compared with previous results. In general, the states generated during the course of this work match previous results well and imply that calculated average structures are similar to those reported previously.[34−36]

4.2.1. Radial Distribution Functions. The radial distribution function counts the density of atom pairs within a certain distance window from each other, relative to the density of a bulk fluid. The radial distribution function can be represented as

$$g(r) = \frac{V}{N^2}\left\langle \sum_i \sum_{j\neq i} \delta(r - r_{ij}) \right\rangle^2$$

where r is the distance between atom pairs, V is the volume, N is the number of particles in the system, and r_{ij} is the distance between a specific atom pair. In Figure 3, the pair-radial distribution function representing the lowest-energy Lennard-Jones system found in this work is shown. This radial distribution function is typical of fluids with a relatively high degree of order. It is zero until approximately 3.9 Å (the distance of closest approach allowed by the interatomic potential function). The high initial peak results from nearest-neighbor contacts at a distance approximately equal to the sum of Lennard-Jones radii for the particles. More distant, less distinct peaks have a similar interpretation. This figure is typical of the results generated in this work for these systems and is a good comparison with previous work performed on like systems.[41]

4.2.2. Radius of Gyration, Mean Square End-to-End Distance, and Characteristic Ratio. The remaining tests in this section were used to characterize the structures of the two oligomeric systems studied here. The three measurements described in this section are interrelated and measure bulk properties of the molecules under study; specifically, they describe the arrangement of the atoms in relation to the center of mass or the ends of the molecule. The first measurement, the radius of gyration, is defined as[45]

$$S^2 = \frac{1}{N}\sum_{k=1}^{N} \|\mathbf{r}_k - r_{\mathrm{mean}}\|^2$$

Table 4. Results of the Bayesian Statistical Analysis To Assess the Viability of Pressure−Temperature Annealing To Complement Temperature Annealing When Searching for Low-Energy Structures of Constrained Systems[a]

	Lennard-Jones systems ($N = 50$, $k = 26$)	monoglyme systems ($N = 25$, $k = 21$)	tetraglyme systems ($N = 25$, $k = 17$)
unbiased: $\alpha = 1$, $\beta = 1$	0.610	0.9997	0.962
even: $\alpha = 25$, $\beta = 25$	0.580	0.976	0.852
over: $\alpha = 25$, $\beta = 5$	0.994	0.99999999	0.99997
under: $\alpha = 5$, $\beta = 25$	0.021	0.342	0.067

[a] Contents of the table indicate the credibility that $\theta > 0.5$.

Table 5. Radius of Gyration ($\langle S^2 \rangle$), Mean Square End-to-End Distance ($\langle R^2 \rangle$), and Characteristic Ratio (C_n) for Each of the Runs Involving Monoglyme[a]

	temperature annealing			pressure−temperature annealing		
system	$\langle S^2 \rangle$ (Å^2)	$\langle R^2 \rangle$ (Å^2)	C_n	$\langle S^2 \rangle$ (Å^2)	$\langle R^2 \rangle$ (Å^2)	C_n
1	12.64294	29.95462	4.131672	12.63144	30.05138	4.145018
2	12.642	29.25339	4.03495	12.63144	30.05138	4.145018
3	12.68264	29.04504	4.006212	12.63144	30.05138	4.145018
4	12.62215	29.71162	4.098154	12.63144	30.05138	4.145018
5	12.65295	29.33647	4.04641	12.63144	30.05138	4.145018
6	12.67993	29.25582	4.035286	12.63144	30.05138	4.145018
7	12.64122	29.19647	4.027099	12.63144	30.05138	4.145018
8	12.6627	29.50217	4.069265	12.63144	30.05138	4.145018
9	12.73361	29.27499	4.03793	12.63144	30.05138	4.145018
10	12.64571	29.5682	4.078372	12.63144	30.05138	4.145018
11	12.64617	29.42254	4.058281	12.65642	29.56905	4.07849
12	12.64617	29.42254	4.058281	12.69759	29.81467	4.112368
13	12.64617	29.42254	4.058281	12.67364	29.93132	4.128458
14	12.64617	29.42254	4.058281	12.66215	29.8647	4.119269
15	12.64617	29.42254	4.058281	12.64929	29.47143	4.065025
16	12.64617	29.42254	4.058281	12.63597	29.97904	4.13504
17	12.64617	29.42254	4.058281	12.69603	29.2091	4.028841
18	12.64617	29.42254	4.058281	12.72194	29.59777	4.082451
19	12.64617	29.42254	4.058281	12.67191	29.86696	4.119581
20	12.64617	29.42254	4.058281	12.64283	29.80467	4.110989
21	12.68304	29.33573	4.046308	12.67301	29.23696	4.032684
22	12.65052	29.78779	4.108661	12.68439	29.82505	4.1138
23	12.62334	29.23872	4.032927	12.72182	30.34298	4.185239
24	12.66012	30.01462	4.139948	12.6111	30.06649	4.147102
25	12.65458	29.26337	4.036327	12.61566	29.70253	4.096901

[a] Structural data compiled from the final structures from each run for both the temperature annealing and pressure−temperature annealing simulations.

Table 6. Radius of Gyration ($\langle S^2 \rangle$), Mean Square End-to-End Distance ($\langle R^2 \rangle$), and Characteristic Ratio (C_n) for Each of the Runs Involving Tetraglyme[a]

	temperature annealing			pressure−temperature annealing		
system	$\langle S^2 \rangle$ (Å^2)	$\langle R^2 \rangle$ (Å^2)	C_n	$\langle S^2 \rangle$ (Å^2)	$\langle R^2 \rangle$ (Å^2)	C_n
1	12.95637	124.1425	4.280776	12.76893	129.1815	4.454534
2	12.88302	118.7149	4.093617	12.89379	130.1666	4.488503
3	12.90877	132.4862	4.56849	12.898	129.0914	4.451428
4	12.86335	129.6162	4.469524	12.976	133.3446	4.59809
5	12.92801	124.0264	4.276772	12.94306	123.1806	4.247607
6	12.95901	119.7428	4.129062	12.9213	138.1344	4.763255
7	12.86378	130.5432	4.50149	12.90611	128.7697	4.440334
8	12.82505	135.4648	4.6712	12.91239	132.9224	4.583531
9	12.90405	132.8103	4.579666	12.84688	125.0199	4.311031
10	12.90094	116.7909	4.027272	12.91853	128.7571	4.4399
11	12.93324	124.7005	4.300017	12.86183	128.4394	4.428945
12	12.79128	133.3978	4.599924	12.93111	134.0236	4.621503
13	12.89667	117.9688	4.06789	13.00522	138.3031	4.769072
14	12.87578	119.3006	4.113814	12.86455	125.1847	4.316714
15	12.9339	123.0498	4.243097	12.81507	139.8334	4.821841
16	12.9282	128.7497	4.439645	12.84843	134.7138	4.645303
17	12.84452	135.5878	4.675441	12.95821	140.6227	4.849059
18	12.98611	130.0819	4.485583	12.9139	136.3271	4.700934
19	12.87325	129.7357	4.473645	12.86061	141.1341	4.866693
20	12.96661	131.3619	4.529721	12.96339	148.0451	5.105003
21	12.92065	131.6316	4.539021	12.83847	131.696	4.541241
22	12.95	118.5199	4.086893	12.88816	134.5525	4.639741
23	12.85471	127.7803	4.406217	12.94454	125.419	4.324793
24	12.92728	134.0905	4.62381	13.00921	137.0178	4.724752
25	12.90026	110.0911	3.796245	12.95532	129.1935	4.454948

[a] Structural data compiled from the final structures from each run for both the temperature annealing and pressure−temperature annealing simulations.

and measures the extent to which the polymer's mass deviates from its center of gravity. It gives some indication as to how tightly the polymer arranges itself. Of course, the chemical and physical nature of the polymer determines this value, which varies according to the physical conditions in which the polymer exists. However, previous studies showed that, for tetraglyme, the value is approximately 20 Å^2.[35] Published data for monoglyme are unavailable. The mean square end-to-end distance, R^2, of the polymer is quite easily defined. It is simply the distance from one end of the polymer

to the other, squared,[45] and then averaged over time and the number of polymer molecules in the sample. This gives a good measure of whether the polymer is stretched (or coiled) or compact. Again, this measurement depends directly on the polymer in question and the physical conditions in which it exists. For tetraglyme, a value of approximately 140 Å has been reported. Finally, the characteristic ratio[45] is an important measurement, because it relates the radius of gyration in a unitless form that is easier to compare between oligomers of different lengths. Additionally, according to theoretical calculations, the value has significance in determining the amount of flexibility the polymer has. For a freely jointed polymer of infinite length, the characteristic ratio should be 1. In the freely rotating chain model, the value can be used to determine the bond angle.[45] The characteristic ratio is defined as

$$C_n = \frac{R^2}{nl^2}$$

where n is the number of backbone bonds in the oligomer and l is an average bond distance in the polymer system (approximately 1.5 Å for the glymes, based on published geometries in ref 33, for example). C_n for tetraglyme has been reported to be 4.9.[34,35] Tables 5 and 6 list the radii of gyration, mean squared end-to-end distances, and characteristic ratios for the oligomer simulations performed here. All calculated values for the mean squared end-to-end distances for tetraglyme are within 30% of the previously reported ensemble average and values obtained using pressure annealing are considerably closer, within 9%. In addition, most calculated values of the characteristic ratio are within a few tenths of an angstrom, with a maximum deviation of −1.1 Å (22%) from the published, ensemble-average value for temperature annealing and 13% (0.65 Å) for pressure annealing. The radii of gyration are uniformly lower than previously reported, and perhaps indicate a difference due to temperature effects or the fact that these values represent a single minimized structure, whereas the published values are taken from an ensemble average at 300 K.

4.2.3. Torsional Triad Population. This measurement[34−36] is specific to oligomers or polymers, as it can be defined only for a given backbone atom sequence of three bonds. Triad analysis for PEO requires measuring a series of three torsion angles each time they occur, classifying the conformations as trans (labeled t, encompassing angles from 120° to 240°), gauche (labeled g, encompassing angles from 0° to 120°), or gauche minus (labeled n, encompassing angles from 240° to 360°). The triads are characterized by three letters. The combination ttt, for example, indicates that all three angles in the triad are trans. Finally, the number of times each combination occurs is counted and used to generate the figures described below. Figure 1 shows that monoglyme contains only a single conformational triad defined by torsional angles around bonds labeled 2−3, 3−4, and 4−5. Tetraglyme, on the other hand, contains four triads in each molecule. Figure 4 shows how the distribution of triads changed between the random initial state and the final state for the monoglyme simulation that produced the lowest energy. Initially, the triad distribution was random, containing a wide variety of triad combinations. After the optimization was performed, only two types of triads predominated, corresponding to low-energy states of each individual strand. The tgt triad dominated, as expected because it is known to represent the dominant triad in glyme systems.[34−36] A similar situation occurred in the tetraglyme simulation that yielded the lowest energy, depicted in Figure 5. Again, the initial state contained a variety of traid combinations, but after the simulation was performed, more dominant triads emerged. Two triads in particular, tgt and ttt, each became more populated, which is consistent with previously reported results.[34−36] Also, states such as gtg that were only minimally populated in previous works dropped in population considerably.[34−36]

5. Conclusions

We have shown that pressure annealing (pressure−temperature annealing) can help to find low-energy structures of systems complicated by steric hindrance. We hypothesize that this is due to the effect that lowered density (due to the lowered pressure) has on a bulky molecule's ability to rearrange itself. Pressure annealing was tested here by allowing such conditions to be realized. Additionally, the final compression phase helps to pack the well-folded molecules onto each other. We showed that bulky molecules gave a lower energy nearly 70% of the time (and as high as 80% in monoglyme) during pressure−temperature annealing simulations when compared to conventional simulated annealing. In contrast, for the simpler Lennard-Jones model, pressure−temperature annealing performed better only 52% of the time and thus gave no appreciable benefits. The conventional simulated annealing simulations and the pressure annealing simulations executed identical numbers of MD and minimization steps and so took essentially equal amounts of time to complete, which demonstrates a further advantage for the pressure annealing simulations.

For the systems studied, the radial distribution functions, radii of gyration, mean squared end-to-end distances, characteristic ratios, and torsion triad populations generated by the method compare well with published results. Pressure annealing shows promise and should be further studied through application to other systems. In particular, applications of the technique to studying the low-energy conformations of longer oligomers, peptides, and small proteins might be pursued. This method, especially in simulations with explicit solvent, would seem to be a very promising approach.

Finally, different heating and cooling schedules can have a significant effect on the results of temperature annealing studies, and an extensive literature exists discussing various possible heating/cooling schedules. Although we selected one particular cooling (and expansion/compression) schedule simply to illustrate the utility of pressure annealing, many others exist, and their effects should be investigated systematically in subsequent work. Another avenue that deserves further exploration includes incorporating the ideas of this work into the replica-exchange framework. Exchanging replicas at different pressures as well as different tempera-

tures could yield a robust optimization technique, particularly in systems containing explicit solvent.

Acknowledgment. We are grateful for supercomputer time from the NSF/NRAC, through Award MCA96-N019, and from the Oklahoma Supercomputing Center for Education and Research (OSCER). Mr. Kevin Raymond first used this method for equilibrating simulations of long oligomers. We acknowledge helpful conversations with Prof. Kieran Mullen during the writing of this article.

References

(1) Malandro, D. L.; Lacks, D. J. Volume dependance of potential energy landscapes in glasses. *J. Chem. Phys.* **1997**, *107*, 5804–5810.

(2) Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids*; Clarendon Press: Oxford, U.K., 1987.

(3) Tsallis, C.; Stariolo, D. A. Generalized simulated annealing. *Physica A* **1996**, *233*, 395–406.

(4) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671–680.

(5) Cantoni, O. In *Lecture Notes in Mathematics/Seminaire de Probabilities*; Azema, J., Emery, M., Ledoux, M., Yor, M., Eds.; Springer-Verlag: Berlin, 1999; Vol. 33, pp 69−119.

(6) Nilges, M.; Gronenborn, A. M.; Brunger, A. T.; Clore, G. M. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Protein Eng.* **1988**, *2*, 27–38.

(7) Hohl, D.; Jones, R. O.; Car, R.; Parrinello, M. Structure of sulfur clusters using simulated annealing: S2 to S13. *J. Chem. Phys.* **1988**, *89*, 6823–6835.

(8) Wilson, S. R.; Cui, W. L. Applications of simulated annealing to peptides. *Biopolymers* **1990**, *29*, 225–35.

(9) Cohn, H.; Fielding, M. Simulated annealing: Searching for an optimal temperature schedule. *SIAM J. Optim.* **1999**, *9*, 779–802.

(10) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(11) Koehl, P.; Delarue, M. Mean-field minimization methods for biological macromolecules. *Curr. Opin. Struct. Biol.* **1996**, *6*, 222–6.

(12) Ladd, M. F. C.; Palmer, R. A. *Structure Determination by X-Ray Crystallography*; Springer: New York, 1994.

(13) Cui, F. Distance-based NMR Structure Determination and Refinement. Ph.D. Thesis, Iowa State University, Ames, IA, 2006.

(14) Townes, C. H.; Schawlow, A. L. *Microwave Spectroscopy*; Dover: Mineola, NY, 1975.

(15) Pulay, P.; Fogarasi, G. Geometry optimization in redundant internal coordinates. *J. Chem. Phys.* **1992**, *96*, 2856.

(16) Allinger, N. L.; Yuh, Y. H.; Lii, J.-H. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 1. *J. Am. Chem. Soc.* **1989**, *111*, 8551.

(17) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(18) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes: The Art of Scientific Computing*; Cambridge University Press: New York, 1986.

(19) Elber, R.; Karplus, M. Enhanced sampling in molecular dynamics: Use of the time-dependent Hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. *J. Am. Chem. Soc.* **1990**, *112*, 9161–75.

(20) Gerber, R. B.; Buch, V.; Ratner, M. A. Time-Dependent Self-Consistent Field Approximation for Intramolecular Energy Transfer. 1. Formulation and Application to Dissociation of van der Waals Molecules. *J. Chem. Phys.* **1982**, *77*, 3022–3030.

(21) Simmerling, C.; Miller, J. L.; Kollman, P. A. Combined Locally Enhanced Sampling and Particle Mesh Ewald as a Strategy To Locate the Experimental Structure of a Nonhelical Nucleic Acid. *J. Am. Chem. Soc.* **1998**, *120*, 7149–7155.

(22) Hixson, C. A.; Chen, J.; Huang, Z. N.; Wheeler, R. A. New perspectives on multiple-copy, mean-field molecular dynamics methods. *J. Mol. Graph. Model.* **2004**, *22*, 349–357.

(23) Simmerling, C.; Lee, M. R.; Ortiz, A. R.; Kolinski, A.; Skolnick, J.; Kollman, P. A. Combining MONSSTER and LES/PME to Predict Protein Structure from Amino Acid Sequence: Application to the Small Protein CMTI-1. *J. Am. Chem. Soc.* **2000**, *122*, 8392–8402.

(24) Roitberg, A.; Elber, R. Modeling side chains in peptides and proteins: Application of the locally enhanced sampling and the simulated annealing methods to find minimum energy conformations. *J. Chem. Phys.* **1991**, *95*, 9277–87.

(25) Zheng, Q.; Rosenfeld, R.; DeLisi, C.; Kyle, D. J. Multiple copy sampling in protein loop modeling: Computational efficiency and sensitivity to dihedral angle perturbations. *Protein Sci.* **1994**, *3*, 493–506.

(26) Stultz, C. M.; Karplus, M. On the potential surface of the locally enhanced sampling approximation. *J. Chem. Phys.* **1998**, *109*, 8809–8815.

(27) 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*, 1094–1107.

(28) Larios, E.; Pitera, J. W.; Swope, W. C.; Gruebele, M. Correlation of early orientational order of engineered λ_{6-85} structure with kinetics and thermodynamics. *Chem. Phys.* **2006**, *323*, 45–53.

(29) Gnanakaran, S.; Nussinov, R.; Garcia, A. E. Atomic-level description of amyloid β-dimer formation. *J. Am. Chem. Soc.* **2006**, *128*, 2158–2159.

(30) Furlan, S.; La Penna, G.; Perico, A.; Cesaro, A. Conformational Dynamics of Hyaluronan Oligomers in Solution. 3. Molecular Dynamics from Monte Carlo Replica Exchange Simulations and Mode Coupling Diffusion Theory. *Macromolecules* **2004**, *37*, 6197–6209.

(31) Sikorski, A. Properties of Star-Branched Polymer Chains. Application of the Replica Exchange Monte Carlo Method. *Macromolecules* **2002**, *35*, 7132–7137.

(32) Yamada, Y.; Udea, Y.; Kataoka, Y. Replica exchange Monte Carlo simulations for folding of di-block polyampholyte. *J. Comput. Chem. Jpn.* **2005**, *4*, 127–130.

(33) Liboff, R. L. In *Kinetic Theory: Classical, Quantum, and Relativistic Descriptions*. Wiley: New York, 1998.

(34) Dong, H.; Hyun, J.-K.; Durham, C.; Wheeler, R. A. Molecular dynamics simulations and structural comparisons of amorphous polyethylene oxide and polyethylenimine models. *Polymers* **2001**, *42*, 7809–7817.

(35) Dong, H.; Hyun, J.-K.; Rhodes, C. P.; Frech, R.; Wheeler, R. A. Molecular Dynamics Simulations and Vibrational Spectroscopic Studies of Local Structure in Tetraglyme: Sodium Triflate Solutions. *J. Phys. Chem. B* **2002**, *106*, 4878–4885.

(36) Smith, G. D.; Yoon, D. Y.; Jaffe, R. L.; Colby, R. H.; Krishnamoorti, R.; Fetters, L. J. Conformations and Structures of Polyoxyethlyene melts from molecular dynamics simulations and small-angle neutron scattering experiments. *Macromolecules* **1996**, *29*, 3462–3469.

(37) Dias, F. B.; Lambertus, P.; Veldhuis, J. B. J. Trends in polymer electrolytes for secondary lithium batteries. *J. Power Sources* **2000**, *88*, 169–191.

(38) Tarascon, J. M.; Gozdz, A. S.; Schmutz, C.; Shokochi, F.; Warren, P. C. Performance of Bellcore's plastic rechargable Li-ion batteries. *Solid State Ionics* **1996**, *86*−*88*, 49–54.

(39) Hajek, B. Cooling schedules for optimal annealing. *Math. Oper. Res.* **1988**, *13*, 311–329.

(40) McQuarrie, D. A. *Statistical Mechanics*; University Science Books: Sausalito, CA, 2000.

(41) Hixson, C. A.; Wheeler, R. A. Practical multiple-copy methods for sampling classical statistical mechanical ensembles. *Chem. Phys. Lett.* **2004**, *386*, 330–335.

(42) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(43) Stuart, A.; Ord, K. *Kendall's Advanced Theory of Statistics*, 6th ed.; Oxford University Press: New York, 1994; Vol. 1.

(44) Lee, P. M. *Bayesian Statistics: An Introduction*; Oxford University Press: New York, 2004.

(45) Flory, P. J. *Statistical Mechanics of Chain Molecules*; Interscience Publishers: New York, 1969.

CT800451C

JCTC Journal of Chemical Theory and Computation

Quantum Chemical Benchmark Studies of the Electronic Properties of the Green Fluorescent Protein Chromophore. 1. Electronically Excited and Ionized States of the Anionic Chromophore in the Gas Phase

Evgeny Epifanovsky,*,† Igor Polyakov,‡ Bella Grigorenko,‡ Alexander Nemukhin,‡,§ and Anna I. Krylov*,†

Department of Chemistry, University of Southern California, Los Angeles, California 90089, Department of Chemistry, M.V. Lomonosov Moscow State University, Moscow 119991, Russia, and Institute of Biochemical Physics, Russian Academy of Sciences, Moscow 119334, Russia

Received March 25, 2009

Abstract: We present the results of quantum chemical calculations of the electronic properties of the anionic form of the green fluorescent protein chromophore in the gas phase. The vertical detachment energy of the chromophore is found to be 2.4−2.5 eV, which is below the strongly absorbing $\pi\pi^*$ state at 2.6 eV. The vertical excitation of the lowest triplet state is around 1.9 eV, which is below the photodetachment continuum. Thus, the lowest bright singlet state is a resonance state embedded in the photodetachment continuum, whereas the lowest triplet state is a regular bound state. Based on our estimation of the vertical detachment energy, we attribute a minor feature in the action spectrum as due to the photodetachment transition. The benchmark results for the bright $\pi\pi^*$ state demonstrated that the scaled opposite-spin method yields vertical excitation within 0.1 eV (20 nm) from the experimental maximum at 2.59 eV (479 nm). We also report estimations of the vertical excitation energy obtained with the equation-of-motion coupled cluster with the singles and doubles method, a multireference perturbation theory corrected approach MRMP2 as well as the time-dependent density functional theory with range-separated functionals. Expanding the basis set with diffuse functions lowers the $\pi\pi^*$ vertical excitation energy by 0.1 eV at the same time revealing a continuum of "ionized" states, which embeds the bright $\pi\pi^*$ transition.

1. Introduction

Unique electronic properties of the green fluorescent protein (GFP) whose natural function is to convert blue light to green light have motivated a number of experimental and theoretical studies and have been exploited in numerous practical applications.[1−3] Due to their fundamental and practical importance, studies of the structure and properties of photoreceptor proteins and their denatured chromophores constitute an important field of modern research. Moreover, GFP can be considered as a model for other fluorogenic unsymmetric methine dyes[4−9] and is of interest to organic photovoltaic materials. For example, the fluorescent protein motif has already inspired the creation of new organic phototovoltaic sensitizers[10] and other optoelectronic materials.[11]

From the theoretical perspective, characterization of the electronic structure of isolated chromophores is the first step toward understanding their photochemical and photobiological properties in realistic environments. Modeling isolated species involves calculating the molecular parameters of the chromophores in the gas phase and in solution using quantum

* Corresponding authors. E-mail: epifanov@usc.edu (E.E.) and krylov@usc.edu (A.I.K.).

† University of Southern California.

‡ M.V. Lomonosov Moscow State University.

§ Russian Academy of Sciences.

10.1021/ct900143j CCC: $40.75
Published on Web 06/25/2009

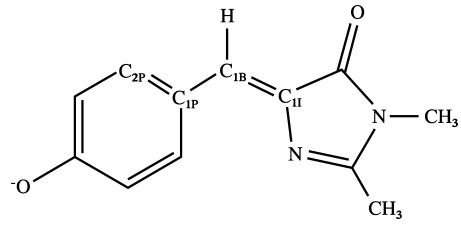


Figure 1. Chemical structure and atomic labels of the anionic form of the model GFP chromophore HBDI in the *cis*-conformation.

chemistry methods. This series of two papers focuses on accurate calculations of the properties of biological chromophores with ab initio methods using the model GFP chromophore, 4′-hydroxybenzylidene-2,3-dimethylimidazolinone (HBDI) anion, as a benchmark system (Figure 1). In this paper, we present the vertical excitation and electron detachment energies and discuss the electronic properties of the excited and detached states. In a subsequent paper,[12] we discuss the *cis*−*trans* isomerization of the HBDI anion in the ground electronic state.

Earlier experimental studies have characterized the absorption of HBDI in the native protein environment.[13] The spectrum of the wild-type GFP has two broad absorption bands at 396 nm (3.13 eV) and 476 nm (2.60 eV) assigned to the neutral and anionic forms of the chromophore, respectively. The spectra in aqueous solution[14,15] reveal strong pH sensitivity: the absorption maximum at neutral pH is at 370 nm (3.35 eV), whereas at pH = 13 and pH = 1, it is shifted to 426 nm (2.91 eV) and 396 nm (3.13 eV), respectively. The absorption of denatured wild-type proteins exhibits a similar pH-dependence.[16] The shifts were attributed to different protonation and deprotonation forms, as well as a strong interaction with water. The latter is consistent with a large change (about 7 D) of the dipole moment upon excitation determined in the Stark effect measurements taken in a buffered (pH = 6.5) glycerol solution at 77 K.[17] Therefore, the protonation and deprotonation of HBDI and the solvent effects, which strongly affect its electronic properties, needed to be accounted for in theoretical models. However, it is important to characterize the chromophore in the gas phase first in order to quantify the solvent effect separately.

Recently, a gas-phase action spectrum of anionic HBDI (as well as other protonated forms), using photodestruction spectroscopy of mass-selected ions injected into an electrostatic ion storage ring, was reported providing an important reference for theory.[14,15] The spectrum shows an absorption band centered at 2.59 eV (479 nm), which extends from 2.4−2.8 eV (440−520 nm) as well as a minor peak around 2.3 eV (540 nm). The authors emphasized a striking similarity between the absorption bands in the gas phase and the protein and suggested that the protein environment shields the chromophore from water and that the absorption in the protein is an intrinsic property of HBDI. While the role of the protein still needs to be investigated, this measurement facilitates more direct comparison between the gas-phase calculations of vertical excitation energies and the experimental absorption for benchmarking theoretical methods.

The large absorption band in the gas-phase spectrum of the HBDI anion has been assigned as the $\pi\pi^*$ transition, however, the nature of the minor feature at 2.3 eV has not been discussed.

A variety of electronic structure techniques ranging from simple semiempirical approximations to high-level ab initio methods have been applied to simulate the properties of the *cis*-anionic form of HBDI.[18−24] Selected representative results are summarized in Table 1.

These studies have identified the absorbing state of the HBDI anion as the S_1 state derived from a HOMO−LUMO excitation of the $\pi\pi^*$ character. LUMO, however, is a valence π^*-like orbital only in relatively small basis sets; including diffuse functions increases the number of molecular orbitals between the HOMO and the lowest π^*-like orbital. Although the bright state retains its $\pi\pi^*$ character, it is not, strictly speaking, a HOMO−LUMO transition in a realistic basis set.

The first theoretical studies on the chromophore, dating back to the late 1990s,[18,19] employed semiempirical methods based on the neglect-of-differential overlap approximation. They placed the $\pi\pi^*$ state at 2.77 eV or 0.18 eV above the experimental absorption maximum. Ab initio calculations using configuration interaction with single excitations (CIS) in a small basis set[20] grossly overestimated the excitation energy. These and other results from Table 1 reveal that accurate calculations of the vertical excitation energy for the bright $\pi\pi^*$ transition are challenging for quantum chemistry.

Time-dependent density functional theory (TD-DFT), which can be applied to very large systems, gave rise to high expectations in the field of photochemistry. However, the notorious self-interaction error[25−27] results in an unphysical description of charge-transfer (CT) states,[28] which are common in large molecules. In addition to artificially low excitation energies of real CT states, spurious CT states appear, spoiling the description of other states. The number of these false states increases steeply with the system size.[29] In the case of fluorescent protein chromophores, TD-DFT has been reported to perform quite modestly, as discussed in ref 23. In that study, which examined various functionals, the best agreement with the experiment was obtained for the BP86 functional.[30,31] This value, 2.94 eV (421 nm), which is listed in Table 1, is still 0.35 eV (60 nm) away from the experiment. Overall, TD-DFT excitation energies range from 2.94 to 3.23 eV,[23] the popular B3LYP functional[32,33] yields 3.05 eV.

CASPT2, the complete active space self-consistent field (CASSCF) method with second-order perturbation theory (PT2) correction,[34] has been applied to model photochemical properties of organic chromophores in various media.[22,35−44] The CASPT2/6-31G(d) result[22] from Table 1 agrees fairly well with the experimental value; however, the CAS distribution of 12 electrons over 11 orbitals is a truncation of the entire π-orbital active space, which requires the CASSCF(16/14) wave function for the GFP chromophore. Bravaya et al.[24] performed calculations of the vertical $\pi\pi^*$ transition energies for various forms of HBDI using a very expensive and highly correlated approach, i.e., state-averaged CASSCF(16/14) wave functions constructed in the full π-orbital active space augmented by perturbative corrections:

Table 1. Selected Theoretical Estimates of the S_0-S_1 Vertical Excitation Energy (ΔE) and the Corresponding Wavelengths for the Gas-Phase GFP Chromophore

method to compute excitation energy	method to optimize ground-state geometry	ΔE (eV)	wavelength (nm)	ref
INDO-CI	PM3	2.77	448	ref 18[a]
CIS/6-31G(d)	RHF/6-31G(d)	4.37	284	ref 20
TD-DFT(BP86)/6-31++G(d,p)	B3LYP/6-31++G(d,p)	2.94	421	ref 23[b]
CASPT2/6-31G(d)	CASSCF(12/11)/6-31G(d)	2.67	465	ref 22
SAC-CI/DZV	B3LYP/6-31G(d)	2.22	558	ref 21
MRMP2 based on SA-CASSCF (16/14)/(aug)-cc-pVDZ	PBE0/aug-cc-pVDZ	2.47	501	ref 24
aug-MCQDPT2 based on SA-CASSCF(16/14)/(aug)-cc-pVDZ	PBE0/aug-cc-pVDZ	2.54	489	ref 24
experiment		2.59	479	ref 14, 15

[a] A close value (444 nm) was obtained in the later semiempirical calculations of the NDDO type (ref 19). [b] For an overview of previous TD-DFT calculations using different functionals and basis sets, see Table 1 of ref 23; only the value closest to the experimental excitation energy is presented here.

multireference second-order Møller−Plesset perturbation theory (MRMP2)[45] and an extended version of the multiconfigurational quasidegenerate perturbation theory (aug-MCQDPT2)[46−48] (using ground-state equilibrium geometries optimized using DFT with the PBE0 functional[49,50] and the (aug)-cc-pVDZ basis set:[51] aug-cc-pVDZ on oxygen atoms and cc-pVDZ on all other atoms). The MRMP2 and aug-MCQDPT2 results are within 0.12 and only 0.05 eV from the experimental value, respectively (Table 1). These data suggest that one can compute the positions of absorption bands of biological chromophores with an accuracy of 10−20 nm (less than 0.1 eV) by applying perturbatively corrected CASSCF-based approaches. These techniques, however, are computationally demanding, and their execution requires advanced skills and extreme care, as the application of the method involves: (i) a careful selection of a large number of active space orbitals in fairly large basis sets; (ii) converging the state-averaged CASSCF solutions corresponding to the $\pi\pi^*$ transition, especially in realistic basis sets; and (iii) a careful and often ambiguous treatment of perturbative corrections to the reference CASSCF solutions. Moreover, gradient calculations are only available for bare CASSCF wave functions. Thus, it is desirable to find a more robust approach of a comparable accuracy for wider applications in modeling the properties of biological chromophores.

Contrarily to the bright singlet $\pi\pi^*$ state, little is known about the triplet states of the GFP chromophore. The fluorescence properties of GFP suggest that the intersystem crossing is not efficient, at least in the chromophore in the native protein environment. In the gas-phase photocycle, however, triplet states can play an important role. For example, possible population trapping in the triplet has been suggested as an explanation for observed millisecond-long lifetimes of the photoexcited ions in the ion storage ring experiments.[52] Spin-forbidden relaxation channels have also been considered in the studies of GFP mutants.[53,54]

In this work we discuss the character of the bright $\pi\pi^*$ state of the HBDI anion and continue to benchmark different perturbation theory corrected multiconfigurational approaches. We also apply the equation-of-motion coupled-cluster method with single and double excitations (EOM-CCSD)[55−60] as well as TD-DFT with the range-separated functionals, BNL[61] and ωPB97X.[62] We also characterize the lowest $\pi\pi^*$ triplet and report the vertical electron detachment energy (VDE). Based on the calculated VDE, we assign the minor feature as due to the photodetachment transition. This has important implications on the character of the bright state: the $\pi\pi^*$ transition is a resonance state embedded in the ionization continuum. The triplet state, however, lies below VDE. As a resonance state, the $\pi\pi^*$ singlet has a finite lifetime and can undergo autoionization due to coupling to the ionization continuum. Contrary to that, the triplet may have a much longer lifetime. Thus, population trapping in the triplet state in the gas-phase photocycle seems to be required to explain millisecond kinetics of the fragments yield in the ion storage ring experiments.[52] Moreover, the resonance nature of the $\pi\pi^*$ state in the anionic GFP might be responsible for very different kinetics of the photofragment yield of the anionic and protonated GFP.[52]

2. Computational Methods

The equilibrium geometries were optimized by DFT with the PBE0 variant[50] of the Perdew−Burke−Ernzerhof (PBE) hybrid functional[49] and by CASSCF(14/12). The cc-pVDZ basis set[51] was used in both calculations. After observing noticeable differences in vertical excitation energies computed using these two geometries, we reoptimized the equilibrium structure using MP2 with cc-pVTZ,[51] which yields very accurate structures for well-behaved closed-shell molecules.[63] MP2 calculations employed the resolution-of-the-identity (RI) technique. The CASSCF and PBE0 structures are C_1, whereas the RI-MP2 optimization produced a C_s minimum. The Cartesian coordinates of the optimized structures are given in the Supporting Materials.

Vertical excitation energies were computed by MRMP2, TD-DFT with the BNL and ωPB97X functionals, CIS, scaled opposite-spin CIS with perturbative doubles (SOS-CIS(D)), and EOM-CCSD for excitation energies (EOM-EE-CCSD). The VDE was computed at the CASSCF geometry as the energy of the Hartree−Fock HOMO (Koopmans' theorem) by EOM-CCSD for ionization potentials (EOM-IP-CCSD) and by using the BNL HOMO energy (as described below, this is equivalent to computing VDE as the difference between the total BNL energies of the anion and the neutral radical). The ωPB97X and B3LYP Koopmans' theorem and ΔE values are also given for comparison.

Each method is outlined below, and computational details are given in the Results Section. MRMP2 calculations were carried out with the PC GAMESS version[64] of the GAMESS(US) quantum chemistry package.[65] CIS, SOS-MP2, SOS-CIS(D), EOM-CCSD, and BNL calculations were performed with Q-Chem.[66]

2.1. Multireference Møller−Plesset Perturbation Theory (MRMP2). The MRMP2 model[45] is a special single-state case of multiconfigurational quasidegenerate second-order perturbation theory (MCQDPT2)[46] implemented in GAMESS(US)[65] and PC GAMESS.[64] The zeroth-order (reference) wave functions are state-averaged CASSCF wave functions (Ψ_α^{CAS}) for the target state α. The unperturbed Hamiltonian operator is a sum of one electron Fock-like operators in which the occupation numbers are replaced by the diagonal elements of the CASSCF density matrix. Beyond zeroth-order multiconfigurational functions, the complementary eigenfunctions of the complete active space configuration interaction Hamiltonian as well as the wave functions generated by one- and two-electron excitations from the reference functions (S-space), are considered via the perturbation theory. The second-order corrections to the energy are given by

$$E_\alpha^{(2)} = \sum_j \frac{\langle \Psi_\alpha^{CAS} | V \Phi_j \rangle \langle \Phi_j V | \Psi_\alpha^{CAS} \rangle}{E_\alpha^0 - E_j^{(0)}} \quad (1)$$

where functions Φ_j belong to the S-space of uncontracted determinants. In a similar approach, CASPT2,[34] the S-space consists of the contracted configurations. Both MRMP2 and CASPT2, diagonalize-then-perturb methods, are widely used to calculate the excitation energies of organic chromophores. A comprehensive benchmark study[67] reports excellent agreement between zeroth-order-corrected CASPT2[68] and an accurate approximation to the coupled-cluster with singles, doubles, and triples method (CC3).[69,70] Similar accuracy is observed in the recent computational studies employing the MRMP2 methodology.[46−48]

2.2. Scaled Opposite-Spin MP2 and Scaled Opposite-Spin CIS(D). Spin-component-scaled MP2 (SCS-MP2) is a semiempirical approach based on scaling different spin contributions to the MP2 correction as proposed by Grimme:[71]

$$E_c \approx E_{c,\mathrm{SCS-MP2}} = p_{ss} E_{ss}^{(2)} + p_{os} E_{os}^{(2)} \quad (2)$$

The parallel-spin component $E_{ss}^{(2)}$ and the antiparallel-spin component $E_{os}^{(2)}$ are scaled to correct for their unbalanced contributions to the MP2 correlation energy. Empirically found optimal values of the coefficients are $p_{ss} = 1/3$ and $p_{os} = 6/5$. SCS-MP2 provides an improved description for many systems in which the ground state has a single reference character.

As the opposite-spin contribution is the major one, Jung et al.[72] further simplified the model by dropping the same-spin term altogether and scaling the opposite-spin contribution up. Along with the RI technique, SOS-MP2 offers a significant improvement in computational performance compared to the original MP2. The scaling coefficient − the only empirical parameters in the method − can be optimized for a wide variety of systems.[73]

In the same manner as SOS-MP2 is introduced for correcting the ground-state energy, SOS-CIS(D) is designed for excitation energies.[74] The computational scaling of SOS-CIS(D) is only $\mathcal{O}(N^4)$, which is a significant improvement over the $\mathcal{O}(N^5)$ scaling of CIS(D). The accuracy of SOS-CIS(D) is very similar to that of CIS(D) for valence states, whereas the performance for the Rydberg states is improved. Based on a set of over 40 various excited states in over 20 organic molecules, the mean signed error in the SOS-CIS(D) vertical excitation energy is 0.02 eV for valence states and −0.08 eV for Rydberg transitions. Limitations of SOS-MP2 and SOS-CIS(D) are the same as MP2 and CIS(D), respectively. For example, these methods fail when the ground-state wave function acquires a significant multiconfigurational character, as at a *cis-trans* isomerization transition state, and for excited states with doubly excited character. Open-shell (e.g., doublet) states can also cause difficulties due to spin contamination.

2.3. Long-Range-Corrected Density Functionals. In long-range-corrected functionals, a range-separated representation of the Coulomb operator[75,76] is used to mitigate the effects of the self-interaction error. The contribution from the short-range part is described by a local functional, whereas the long-range part is described using the exact Hartree−Fock exchange. The separation depends on a parameter γ. In the BNL approach,[61] γ is optimized for each system using Koopmans-like arguments: γ is adjusted such that the HOMO energy equals the difference between the total BNL energies of the $N1$ and N-electron systems. Initial benchmarks[61,77] demonstrated an encouraging performance for excited states and even such challenging systems as ionized dimers. In ωPB97X,[62] γ and other parameters are optimized using standard training sets. Benchmark results have demonstrated consistently improved performance relative to uncorrected functionals.

2.4. Equation-of-Motion Coupled-Cluster Methods for Excitation Energies and Ionization Potentials. In EOM-CC,[55−60,78−80] the Hamiltonian is similarity transformed:

$$\bar{H} \equiv \exp(-T) H \exp(T) \quad (3)$$

using the cluster operator T obtained from coupled-cluster equations for the ground state:

$$\langle \Phi_\mu | \bar{H} - E^{CC} | \Phi_0 \rangle = 0 \quad (4)$$

where $|\Phi_\mu\rangle$ denotes all μ-tuply excited determinants with respect to the Hartree−Fock reference $|\Phi_0\rangle$. The solution for the m-th excited state is found from

$$\langle \Phi_\mu | \bar{H} - E_m^{EOM} | R_m \Phi_0 \rangle = 0 \quad (5)$$

where R_m is a linear excitation operator, the form of which depends on the target states. For example, in equation-of-motion for excitation energies (EOM-EE) R_m is

$$R_m = r_{m,0} + \sum_{ia} r_{m,i}^a a^\dagger i + \frac{1}{4} \sum_{ijab} r_{m,ij}^{ab} a^\dagger b^\dagger j i + \ldots \quad (6)$$

whereas for ionized states, R_m is not particle conserving

$$R_m = \sum_i r_{m,i} i + \frac{1}{4} \sum_{ija} r^a_{m,ij} a^\dagger j i + \ldots \quad (7)$$

The EOM wave function of the m-th state is given by

$$|\Psi_m\rangle = R_m \exp(T)|\Phi_0\rangle \quad (8)$$

In EOM-CCSD, the excitation operator R_m is truncated after the two-body term (i.e., two-hole two-particle for EE and two-hole one-particle for IP), and the similarity transformed Hamiltonian is diagonalized in the basis of all singly and doubly excited determinants.

The EOM-CCSD error bars are 0.1−0.3 eV for electronic states dominated by a single excitation. Including triples reduces the error to 0.01−0.02 eV.[81] In a recent benchmark study, Schreiber et al.[67] reported EOM-CCSD mean absolute and maximum errors of 0.12 and 0.23 eV, respectively. A recent study of uracil[82] demonstrated that even for well-behaved molecules inclusion of triple excitations and extending the basis set beyond augmented double-ζ can affect vertical excitations by as much as 0.3 eV.

3. Results and Discussion

3.1. Molecular Orbital Framework. Figure 2 shows two resonance structures of the HBDI anion. The interaction between these two structures results in charge delocalization between the two oxygen atoms and in scrambling CC bond orders, as discussed for example in ref 42. Changes in bond orders in the bridge region due to the resonance are believed to be important in thermal isomerization,[83] as discussed in detail in the companion paper.[12]

The natural bond orbital (NBO)[84,85] charges on the phenoxy or imidazolin oxygens (computed using the RI-MP2 densities) are −0.65 and −0.66, respectively. The almost equal values of the charges suggest significant contributions from both resonance structures. As the result of the resonance interaction, the C_{1P}−C_{1B} bond gains double-bond character, whereas the order of C_{1B}−C_{1I} bond is reduced, and the bridge moiety acquires allylic character. It is interesting to compare the respective bond lengths with the values for the double and single CC bonds between sp^2 hybridized carbons, e.g., in a planar and twisted ethylene (see ref 86 for explanation regarding the choice of the reference structures). Our best estimates of the C_{1P}−C_{1B} and C_{1B}−C_{1I} bond lengths (RI-MP2/cc-pVTZ) are 1.394 and 1.378 Å, respectively. The lengths of the CC bond in ethylene is 1.333 Å at the planar geometry (where the formal bond order is 2) and 1.470 Å[87] at the twisted configuration, where the π bond is completely broken.[86] Thus, assuming a linear relationship between the bond order and bond length, one can assign 55 and 67% of a double-bond character to the C_{1P}−C_{1B} and C_{1B}−C_{1I} bonds, respectively. The NBO analysis[84,85] assigns 1.8 electrons to a slightly asymmetric allylic three-center bond. The relative contributions of the C_{1P} and C_{1I} carbons are 38 and 62%, respectively, in a semiquantitative agreement with the bond orders derived from the bond lengths.

The resonance interaction is also reflected by the shape of MOs. Figure 3 shows two Hartree−Fock orbitals involved in the bright $\pi\pi^*$ and the photodetachment transitions (HOMO and valence LUMO). These orbitals, which are traditionally referred to as π and π^*, have quite complicated shapes and are delocalized over the entire molecule. Their character in the bridge region can be explained by considering two interacting π orbitals, as shown in Figure 3. The HOMO can be described as an out-of-phase combination of two localized π bonds, whereas the large electron density on C_{1B} in the LUMO can be derived from the in-phase combination. Of course, due to the delocalized character of the orbitals, this picture is just an approximation, but it allows one to see the origin of the large oscillator strengths and changes in charge distribution in the excited state and also provides a useful framework for explaining the character of the transition state along the *cis-trans* isomerization coordinate.[12]

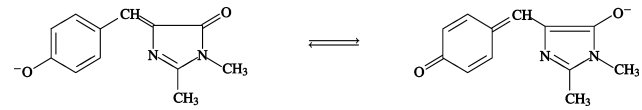


Figure 2. Two resonance structures of the HBDI anion.

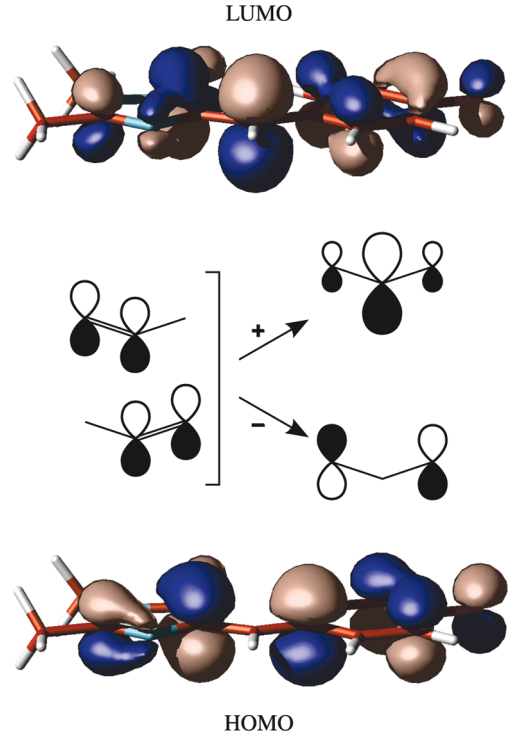


Figure 3. Two molecular orbitals of the HBDI anion giving rise to the $\pi\pi^*$ state. The character of the orbitals can be explained by considering the linear combination of two localized π-bonding orbitals.

3.2. Vertical Electron Detachment Energy of the HDBI Anion. Since the HBDI anion is a closed-shell system, it is stable in the gas phase and has a relatively large VDE. However, as shown below, it does not support bound electronically excited singlet states, and the lowest valence excitation is embedded in an ionization continuum. Such resonance states are very common in molecular anions[88] and play an important role in dissociative electron attachment processes.[89,90] Thus, the broad character of the experimental action spectrum[14,15] is at least partially due to the broadening of the resonance-like $\pi\pi^*$ state by its interaction with the ionization continuum. Finite lifetime and autoionization

decay of this state should be taken into account when considering photoinduced dynamics and lifetime of the HBDI anion in the gas phase. In the condensed phase, solvent may stabilize the anion such that its excited states become bound. However, one photon photodetachment channel may still be relevant for the anionic forms, especially when the production of solvated electrons is considered.[91] The resonance character of the $\pi\pi^*$ state also has significant consequences in the electronic structure calculations[59,88] of excited states, as described below.

The simplest estimate of VDE obtained by applying Koopmans' theorem is 2.56−2.93 eV depending on the basis set, the largest value obtained with the 6-311(2+,+)G(2df,2pd) basis. The EOM method for ionization energies (EOM-IP), which includes electronic correlation, can provide more reliable estimates of VDE. However, due to the size of the system, we are limited to relatively modest basis sets. In the 6-31G* basis, Koopmans' VDE is 2.56 eV, while EOM-IP-CCSD yields 2.05 eV. The larger 6-31+G* basis increases the energies to 2.91 and 2.48 eV, respectively. Thus, including correlation reduces VDE by 0.4−0.5 eV and assuming the effect is consistent throughout basis sets, we estimate that the target EOM-IP-CCSD/6-311(2+,+)-G(2df,2pd) energy is 2.4−2.5 eV.

B3LYP/cc-pVDZ energy difference calculations yield 2.46 eV, which is considerably larger than the respective Koopmans value of 0.92 eV. The ωPB97X/cc-pVDZ energy difference value is 2.39 eV, whereas the respective Koopmans IE is much higher (2.83 eV).

By construction, the BNL energy difference VDE is equal to the respective HOMO energy. Recent benchmarks[92] demonstrated that BNL produces very accurate ionization energies. VDE calculated with BNL in the small cc-pVDZ basis is 1.99 eV ($\gamma = 0.250$), and it increases to 2.52 eV in the 6-311G(2+,+)G(d,p) basis for which $\gamma = 0.228$. Additional sets of diffuse orbitals or more extensive polarization do not affect this value, e.g., VDE calculated with BNL/6-311G(3+,2+)G(2df,2pd) is 2.53 eV.

The discrepancies between the Koopmans and ΔE values have important implications for the excited-state calculations, as the former value defines the onset of the ionization continuum in CIS/TD-DFT calculations (see Appendix). Thus, with B3LYP, the continuum begins 1.54 eV below its own VDE, whereas the situation with ωPB97X is reverse, i.e., the continuum states appear 0.44 eV above the respective VDE. BNL, by constriction, is internally consistent, and the continuum states in TD-DFT calculations appear exactly at the respective VDE.

Thus, our estimate of VDE is 2.4−2.5 eV, within 0.1 eV from the maximum of the weak absorption feature at 2.3 eV. The remaining discrepancy between the two values might be due to the uncertainties in equilibrium geometries or possible vibrational excitation of the molecules in the experiment.

Although correlation has significant effect on VDE, the ionized state has Koopmans-like character, i.e., the leading EOM-IP amplitude corresponds to ionization from the HOMO (see Figure 3) and equals 0.96. Thus, the Hartree−Fock HOMO is a good approximation to a correlated Dyson orbital.[93]

3.3. Vertical Excitation Energy and Electronic Properties of the Singlet and Triplet $\pi\pi^*$ Transitions of the HBDI Anion. *3.3.1. Singlet $\pi\pi^*$ State: Benchmark Results.* Motivated by the discrepancies in previous theoretical estimates of the $\pi\pi^*$ excitation energy of the HBDI anion (Table 1), we set out to benchmark other electronic structure methods with the purpose of identifying a rigorous, yet fairly inexpensive, quantum chemistry approach that can be employed in condensed-phase applications. The experimental maximum of absorption is at 2.59 eV (479 nm), and the band's full width at half-maximum (fwhm) is 0.25 eV (45 nm).[14,15] Assuming that the absorption maximum corresponds to the vertical transition of the lowest-energy isomer (which is not entirely clear, as the temperature of the ions in the ring is unknown), one would like the computed wavelength to fall within 2.47−2.72 eV (456−502 nm). However, due to the resonance nature of the $\pi\pi^*$ state, calculating the excitation energies and oscillator strengths, as well as comparing them with the experimental spectrum, are not as straightforward as in the case of the excited states lying below the electron detachment energy. In the following, we will use the term "detached states" to identify the electronic states that compose the continuum instead of the usual "ionized states" as the initial species is anionic.

The ground-state equilibrium geometry was optimized with PBE0/cc-pVDZ, CASSCF/cc-pVDZ, and RI-MP2/cc-pVTZ (see Computational Methods section). Although the differences between the geometries are small (maximum bond length deviation is 0.03 Å and angles agree within 2 degrees), the $\pi\pi^*$ excitation energy computed using wave function-based methods differs by about 0.1 eV. This effect of geometry is consistent with previous calculations by Olsen[94] for a similar system (*p*-hydroxybenzylidene-imidazolin-5-one, HBI). Of the three structures, the RI-MP2/cc-pVTZ is the most accurate.[63] Since most of the changes in electron density occur in the bridge region, it is interesting to compare the $C_{1P}-C_{1B}$ and $C_{1B}-C_{1I}$ bond lengths computed by different methods. The RI-MP2 values are 1.394 and 1.378 Å, which is very close to the PBE0 values of 1.404 and 1.385 Å. Due to the absence of dynamical correlation, CASSCF exaggerates the bond alternation giving 1.406 and 1.397 Å.

Calculations in small basis sets create discrete states from the continuum and artificially exclude the detached states from the picture. Assuming that the character of the resonance state of interest is well described, such calculations may provide a fairly good estimate of the position of this state in the continuum, however, it is difficult to predict how expanding the basis set will affect the resonance state.[59,88] Moreover, one should anticipate broadening of the resonance state due to the interaction with the detached states. Increasing the size of the one-electron basis brings the continuum states down. When the basis set is large enough to accommodate a detached electron, the lowest excited state will correspond to the detached state (this fact has been exploited in pilot implementations of EOM-IP-CCSD based on the EOM-EE-CCSD code by adding a very diffuse orbital to the basis to describe the ionized electron).[95,96] It can be shown formally (see Appendix) that, in the case of CIS and

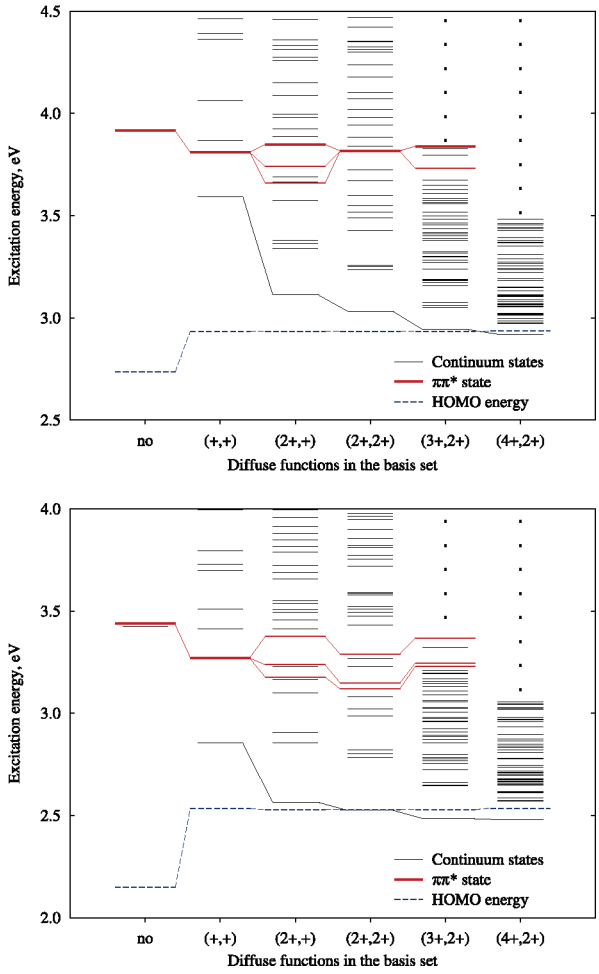


Figure 4. Effect of increasing the number of diffuse functions in the basis set on the density of states and the convergence of the lowest excited and the bright $\pi\pi^*$ state. The calculations were performed with CIS (top) and TD-DFT/BNL (bottom). The basis set was varied from 6-311G(2pd,2df) to 6-311-(4+,2+)G(2pd,2df).

TD-DFT, the energy of the lowest of such CIS-IP states equals the HOMO energy, that is, the Koopmans theorem can be proven by considering the configuration interaction of all singly detached determinants.

When the one-electron basis set is expanded with diffuse functions, the density of states rapidly increases with both CIS and TD-DFT (BNL), as illustrated in Figure 4. At the same time, the lowest excited state converges to Koopmans' VDE. Both methods split the oscillator strength of the bright $\pi\pi^*$ transition among multiple states, but BNL does that to a larger degree, which results in three to four states with close oscillator strengths (Table 2). That may be due to the remaining self-interaction error in BNL calculations.

Seemingly in contradiction with the proof given in the Appendix, the lowest excited CIS and TD-DFT states shown in Figure 2 fall below the Koopmans estimate in large basis sets due to additional stabilization by the interaction with other CIS determinants that vanish once the detached electron is infinitely far from the core. Indeed, if a separate large noninteracting orbital is used instead of diffuse functions, the lowest excitation energy in such a system is exactly equal to the HOMO energy.

Table 2. Interaction of the Bright $\pi\pi^*$ State with the Electron Detached Continuum as Reflected by the Diminishing Oscillator Strength (f_L) in CIS and TD-DFT/BNL Calculations[a]

basis set	number	energy, eV	f_L
	CIS		
6-311G(2df,2pd)	1	3.92	1.58
6-311(+,+)G(2df,2pd)	2	3.81	1.50
6-311(2+,+)G(2df,2pd)	6	3.66	0.22
	9	3.74	0.14
	10	3.85	1.14
6-311(2+,2+)G(2df,2pd)	9	3.60	0.08
	12	3.82	1.35
	17	4.02	0.06
6-311(3+,2+)G(2df,2pd)	37	3.73	0.43
	40	3.84	1.00
	TD-DFT (BNL)		
6-311G(2df,2pd)	2	3.44	1.47
6-311(+,+)G(2df,2pd)	3	3.27	1.38
6-311(2+,+)G(2df,2pd)	7	3.18	0.54
	9	3.24	0.31
	10	3.38	0.52
6-311(2+,2+)G(2df,2pd)	8	3.12	0.15
	9	3.15	0.15
	12	3.29	0.39
	13	3.29	0.63
6-311(3+,2+)G(2df,2pd)	35	3.23	0.47
	37	3.24	0.39
	39	3.37	0.15
	40	3.37	0.15

[a] Also see Fig 4.

As the density of low-lying states that approximate the continuum increases and the oscillator strength gets redistributed, it becomes increasingly more difficult to compute or even identify the resonance state. By following the evolution of the states with the basis set increase, the limiting value of the resonance state can be extrapolated using the stabilization method by Taylor,[97] as has been done in the calculations of resonance electron attached states.[59,88] From the data in Table 2, we can estimate that the excitation energy converges to 3.8 eV with CIS and to 3.3 eV with TD-DFT/BNL. Overall, adding sets of diffuse functions results in lowering the excitation energy of the bright state by 0.1 eV.

The CIS and TD-DFT/BNL calculations demonstrate that the resonance state is embedded in an electron detached continuum, and the broad character of the experimental spectrum can be due to the interaction with the continuum. Thus, small basis set calculations of the vertical excitations energies of the $\pi\pi^*$ state can only provide a rough estimate of the position of the absorption maximum.

Introducing a noniterative doubles correction via SOS-CIS(D) significantly lowers the energy of the $\pi\pi^*$ transition, while the detached state energies are affected less. We could not obtain SOS-CIS(D) results in heavily augmented basis sets due to the lack of the corresponding auxiliary bases for the RI procedure, which becomes unstable upon adding a second set of the diffuse functions. The magnitude of correction is 0.9−1.2 eV for the $\pi\pi^*$ state and 0.1−0.3 eV for the detached states. MRMP2 based on sa-CASSCF(14/12) in the modest cc-pVDZ basis yields 2.51−2.61 eV for the $\pi\pi^*$ excitation energy. SOS-CIS(D)/cc-pVDZ gives 2.71−2.81 eV, which is consistent with the more rigorous MRMP2 estimates.

Table 3. Vertical $\pi\pi^*$ Electronic Excitation Energies (ΔE, in eV), Corresponding Wavelengths (nm), and Oscillator Strength (f_L) for the Gas-Phase GFP Chromophore (Figure *1*)[d]

	Ground-state geometry optimized with								
	PBE0/cc-pVDZ			CASSCF(14/12)/cc-pVDZ			RI-MP2/cc-pVTZ		
Method	ΔE	f_L	nm	ΔE	f_L	nm	ΔE	f_L	nm
MRMP2 based on sa-CASSCF(14/12)/cc-pVDZ	2.52[b]	–	491[b]	2.61[c]	–	476[c]	–	–	–
EOM-CCSD/6-31G(d)	3.08	1.26	402	3.16	1.27	392	3.12	1.27	398
EOM-CCSD/6-31+G(d)	2.97	1.24	418	3.04	1.25	408	3.00	1.25	413
SOS-CIS(D)/cc-pVDZ	2.71	1.59[a]	457	2.81	1.61[a]	441	2.75	1.60[a]	451
SOS-CIS(D)/aug-cc-pVDZ	2.57	1.45[a]	482	2.67	1.45[a]	464	2.61	1.45[a]	475
SOS-CIS(D)/cc-pVTZ	2.58	1.54[a]	480	2.68	1.56[a]	463	2.62	1.54[a]	473
SOS-CIS(D)/aug-cc-pVTZ	2.58	1.37[a]	480	2.90	1.04[a]	427	2.72	1.23[a]	456
TD-DFT/BNL/cc-pVDZ	3.44	1.51	360	3.50	1.51	354	3.59	1.55	346
TD-DFT/BNL/6-311(+,+)G(2df,2pd)	3.22	1.38	385	3.27	1.38	379	3.24	1.38	383
TD-DFT/ωPB97x/cc-pVDZ	3.52	1.52	352	3.59	1.53	346	3.55	1.53	349
TD-DFT/ωPB97x/6-311(+,+)G(2df,2pd)	3.38	1.45	367	3.44	1.46	360	3.40	1.45	364

[a] Oscillator strength calculated with CIS. [b] At the equilibrium geometry computed with PBE0/(aug)-cc-pVDZ (diffuse functions only on oxygen atoms). [c] At the equilibrium geometry computed with CASSCF(12/11)/(aug)-cc-pVDZ (diffuse functions only on oxygen atoms). [d] The reference experimental value is 2.59 eV (479 nm) (ref 14, 15).

At the CASSCF optimized geometry, EOM-CCSD/6-31G* yields an excitation energy of 3.16 eV and an oscillator strength $f_L = 1.27$. Including only one set of diffuse functions lowers the energy to 3.04 eV ($f_L = 1.25$). The change in the oscillator strength is consistent with the observed drop of the EOM amplitude corresponding to the $\pi\pi^*$ transition from 0.85 to 0.74. Benchmarks on the electronically excited states of closed-shell molecules have shown that the combined effect of increasing the basis set in EOM-CCSD calculations and the triples correction can be as large as 0.3−0.4 eV.

The computational cost of MRMP2 and EOM-CCSD does not allow us to use these methods to fully explore the basis set effect on the $\pi\pi^*$ excitation energy, e.g., via stabilization method. We anticipate a noticeable effect of improving the basis set on the energy of the $\pi\pi^*$ state. For example, with SOS-CIS(D), the excitation energy with the cc-pVTZ basis is lower by 0.13 eV relative to cc-pVDZ. The EOM-CCSD excitation energy also drops by as much as 0.2 eV upon expanding the basis set from 6-31G* to 6-31+G*. The oscillator strength calculated with CIS is about 10% lower when basis sets with diffuse functions are used compared to those without, demonstrating the beginning of the interactions with the continuum.

The $\pi\pi^*$ excitation energy obtained with MRMP2 based on state-averaged CASSCF(16/14), which spans the entire π-electron active space (Table 1), and a more compact (14/12) active space (Table 3) agree equally well with the experimental result: 2.47 eV (501 nm) and 2.52−2.61 eV (476−491 nm), respectively, versus 2.59 eV (479 nm). The most expensive aug-MCQDPT2/CASSCF(16/14) approach (Table 1) does not perform noticeably different than MRMP2/CASSCF(14/12). Therefore, one can rely on the fairly practical MRMP2 approach based on a reduced active space in the CASSCF wave function. Overall, the effect of contracting the active space is less than 0.1 eV, which is smaller than the uncertainty due to the equilibrium geometry and anticipated effects of extending the basis set beyond the double-ζ level.

In view of complexities associated with performing multireference perturbation theory and underlying CASSCF calculations, we find the results of the inexpensive SOS-CIS(D) method very encouraging: at the PBE0-optimized geometry, the SOS-CIS(D)/cc-pVTZ $\pi\pi^*$ excitation energy of 2.58 eV is within 0.01 eV from the experimental absorption maximum and agrees very well with MRMP2 calculations.

The EOM-CCSD/6-31+G* excitation energy is 0.38 eV above the experimental maximum, which is outside the EOM-CCSD error bars. Analysis of the EOM amplitudes confirms a singly excited character of the $\pi\pi^*$ state, thus, with an adequate basis set, EOM-CCSD error should not exceed 0.3 eV. In addition to anticipated basis set effects and interactions with the continuum states, other factors, such as discrepancies due to equilibrium geometry as well as the uncertainty in the experimental value (fwhm of the experimental absorption is 0.25 eV), make it difficult to arrive at a definite conclusion. More extensive calculations using larger basis sets (and ideally using stabilization graphs) and estimates of triples corrections (to account for dynamical correlation) are required to assess the performance of EOM-CCSD for this molecule. Basis set and triples effects alone have been shown to account for as much as 0.3 eV in excitation energies of $\pi\pi^*$ character.[82]

Finally, we present TD-DFT results computed with the range-separated functionals. Using the BNL functional[61] with the small cc-pVDZ basis set, the $\pi\pi^*$ state appears the second lowest, which is consistent with the respective VDE. At the CASSCF geometry, its excitation energy is 3.50 eV, and the oscillator strength is $f_L = 1.51$. The 6-311(+,+)G(2df,2pd) basis lowers the excitation energy to 3.27 eV, and the oscillator strength to 1.38. ωPB97X[62] gives similar values for the excitation energy and the

Table 4. Vertical Triplet $\pi\pi^*$ Electronic Excitation Energies (ΔE, in eV) of the HBDI Anion in the Gas Phase

basis set	CIS	SOS-CIS(D)
cc-pVDZ	2.03	1.91
aug-cc-pVDZ	2.02	1.88
cc-pVTZ	2.03	1.86

oscillator strengths, i.e., 3.59 ($f_L = 1.51$) and 3.44 eV ($f_L = 1.46$) with cc-pVDZ and 6-311(+,+)G(2df,2pd), respectively. Because the Koopmans continuum with ωPB97X begins at higher energies (see Section 3.2, Vertical Electron Detachment Energy of the HDBI Anion), the detached states do not appear below the excited states in these calculations.

Overall, it is unrealistic to expect accuracy better than 0.1 eV (20 nm) from computational protocols applicable to a molecule of this size even for nonresonance states, and the observed discrepancies between different methods confirm that. Moreover, when assessing the accuracy of computed values, one should keep in mind the finite width of the experimental absorption band. Thus, more calculations are necessary to provide a converged theoretical estimate, especially stabilization analysis. For practical applications, however, it is important that all the reliable theoretical methods agree with each other in that the origin of the intensity in the resonance state is due to $\pi\pi^*$ excitation. SOS-CIS(D) offers an inexpensive alternative to more rigorous multireference methods if single excitations are dominant in the wave function.

3.3.2. Changes in Electronic Density in the Singlet $\pi\pi^*$ State. As one may expect from the molecular orbital character (Figure 3) and the large oscillator strength, electronic excitation results in a significant redistribution of electronic density. A convenient measure of charge distribution is the permanent dipole moment. Although in a charged system it is origin dependent, the difference between the two dipole moments, $\Delta\mu = \mu_{gr} - \mu_{ex}$, is not. At the CIS level of theory, the value of $|\Delta\mu|$ is 0.6 D, and its direction is in the molecular plain pointing toward the bridge carbon. This value can be compared with the experimentally measured $\Delta\mu$, derived from Stark effect measurements in a buffered (pH = 6.5) glycerol solution at 77 K.[17] This work also reports the angle between $\Delta\mu$ and $\Delta\mu_{tr}$. Strikingly, the experimental value is 10 times larger than the computed one. Since $\Delta\mu$ is related to the changes in orbital occupations upon excitation, it is dominated by contributions from the leading excitation amplitudes and should be reproduced fairly accurately at the CIS level. Thus, large discrepancy is likely to be due to the solvent effect. Indeed, polar solvents result in the increased dipole moment of the solute. For example, the dipole moment of water in bulk water is about 30% larger than in the gas phase. More polar charge distribution in the ground state in solvent is clearly seen from the respective NBO charges (see Table 2 in ref 12). Thus, for the difference of dipole moments of two states, one may anticipate an enhanced effect.

3.3.3. Triplet $\pi\pi^*$ State. The vertical excitation energies of the lowest triplet state at the RI-MP2/cc-pVTZ geometry are summarized in Table 4. The analysis of the wave function confirms that the triplet is derived from the transitions between the same orbitals as the singlet (Figure 3). As expected, all methods consistently place the triplet considerably below the singlet. The variations between the methods are smaller for the triplet state. Our best value (SOS-CIS(D)/cc-pVTZ) is 1.86 eV. The 0.76 eV gap between the singlet and triplet does not suggest efficient intersystem crossing at this geometry. The triplet state is 0.3−0.4 eV below VDE and is, therefore, a bound electronic state. Thus, much longer lifetime is expected for this state (as compared to the singlet), not only because the radiationless relaxation to the ground state is a spin-forbidden process, but also because the autoionization channel is absent.

4. Conclusions

In this work we exploit modern quantum chemical methods for calculations of the electronic properties of the GFP chromophore and compare the results to the gas-phase absorption spectrum obtained by photodestruction spectroscopy in the ion storage ring.[14,15,52]

The experimental action spectrum of the denatured gas-phase anionic GFP chromophore features a broad line (2.4−2.8 eV) with a maximum at 2.59 eV and a minor feature at 2.3 eV. Wave function-based and DFT calculations estimate a VDE of 2.4−2.5 eV. Thus, we assign the minor peak as due to the photodetachment transition. Based on our estimate of VDE, the absorption band at 2.6 eV corresponds to the transition to the resonance state embedded in an electron detached continuum, and the broad character of the spectrum is at least partially due to the interaction with the continuum states.

The resonance nature of the $\pi\pi^*$ state suggests a finite lifetime, and that autoionization channel should be considered when modeling the anionic GFP photocycle. The triplet state is found to be well below the photodetachment threshold (vertical excitation energy 1.86 eV). Thus, the two states are expected to have very different lifetimes, which makes the suggested[52] population trapping in the triplet state even more essential for explaining slow fragmentation kinetics. The resonance nature of the $\pi\pi^*$ state in the anionic GFP might be responsible for very different behavior of the photofragment yield of the anionic and protonated GFP,[52] however, more detailed electronic structure calculations are required in order to suggest a viable mechanism. An important question is how photoinduced isomerization and other structural changes[42,98] affect the relative-states energies.

All wave function-based and TD-DFT methods agree on the nature of the transition lending the intensity to the resonance state, which is a bright $\pi\pi^*$ transition (HOMO−LUMO in a small basis set), however, quantitative agreement is more difficult to achieve. Most importantly, small basis set calculations discretize the ionization continuum, and the results of such calculations provide only a crude estimate of the energy of the resonance state. In order to account for basis set effects, the stabilization analysis can be used; however, in view of the large size of the GFP chromophore molecule, we were able to only conduct it with the CIS and BNL methods.

Nevertheless, it is instructive to compare vertical excitation energies of the $\pi\pi^*$ state computed with different methods

in a moderate basis set and with the experimental band maximum. While the CIS/aug-cc-pVTZ excitation energy is more than 1 eV off, perturbative inclusion of double excitations by SOS-CIS(D) yields a value which is within 0.1 eV from the experimental band maximum. The EOM-CCSD values computed in the modest 6-31+G* basis are within 0.38 eV from the experimental absorption maximum. An analysis of the EOM-CCSD wave function confirms the dominant one electron character of the $\pi\pi^*$ state, however, perturbative inclusion of triple excitations and a larger basis set are required for a converged (with respect to the level of theory) EOM-CC value. Based on previous studies, a proper account of dynamical correlation by including triple excitations and increasing basis set can change the vertical excitation energy by as much as 0.3 eV. Basis set effects evaluated using inexpensive SOS-CIS(D) calculations affect vertical excitation energies by 0.14 eV upon the transition from cc-pVDZ to aug-cc-VTZ.

Additional uncertainties arise from the ground-state geometry. For example, different choices of the ground-state geometry (optimized with CASSCF, DFT, and MP2) introduce an uncertainty of 0.1 eV in the vertical excitation energies. A relatively strong dependence of the excitation energy on the structure suggests additional broadening of the absorption band due to vibrational excitations of the chromophore.

The best MRMP2 estimate is within 0.07 eV from the experimental band maximum, however, the inclusion of basis set effects (using the SOS-CIS(D)/aug-cc-pVTZ estimate) increases the difference to 0.21 eV. Overall, the observed variations demonstrate that it is unrealistic to expect an accuracy better than 0.1 eV from computational protocols applicable to a molecule of this size even for nonresonance excited states.

BNL/cc-pVDZ vertical excitation energy of the bright $\pi\pi^*$ transition is above MRMP2/cc-pVDZ value by 0.9 eV. Since the self-interaction error is considerably reduced in BNL, the photodetachment continuum is likely not contaminated by spurious low-lying charge-transfer states ubiquitous in TD-DFT calculations.

Acknowledgment. We thank Alex Granovsky and Dr. Ksenia Bravaya for their generous help and valuable discussions of this work. This work is supported by the joint grant from the U.S. Civilian Research and Development Foundation (project RUC1-2914-MO-07) and the Russian Foundation for Basic Research (08-03-91104-AFGIR). I.P., B.G., and A.N. thank the SKIF-GRID program and SKIF-Siberia for providing computational resources. E.E. and A.I.K. thank Prof. Roi Baer and Dr. Esti Livshitz for their help with performing BNL calculations. E.E. and A.I.K. acknowledge the iOpenShell Center for Computational Studies of Electronic Structure and Spectroscopy of Open-Shell and Electronically Excited Species (http://iopenshell.usc.edu) supported by the National Science Foundation through the CRIF:CRF CHE-0625419 + 0624602 + 0625237 grant as well as through the CHE-0616271 grant (A.I.K.). A.I.K. is grateful to the Institute of Mathematics and its Applications in Minnesota for its hospitality and productive environment during her stay at IMA as a visiting professor. We also wish to thank the reviewer of the manuscript for his/her thoughtful and insightful comments.

Supporting Information Available: Optimized geometry of the HBDI anion, CASSCF natural orbital figures, and leading EOM-CCSD amplitudes additional information. This material is available free of charge via the Internet at http://pubs.acs.org.

Appendix

The electronic Hamiltonian in the second quantization form is

$$H = \sum_{pq} \langle p|h|q\rangle p^\dagger q + \frac{1}{2}\sum_{pqrs} \langle pq|rs\rangle p^\dagger q^\dagger sr \qquad (9)$$

where h is the core Hamiltonian operator, $\langle pq|rs\rangle$ denotes two electron integrals, and the sums run over all the MOs.

The choice of the reference determinant $|\Phi_0\rangle$ defines the separation of the orbital space into the occupied and virtual subspaces. Let $\{i,j,...\}$ be the subspace of all occupied and $\{a,b,...\}$ be the subspace of all virtual orbitals in $|\Phi_0\rangle$. Indexes $p,q,...$ denote all orbitals, occupied or virtual.

Let $|\Phi_0\rangle$ be the solution to the Hartree−Fock equations with an energy value:

$$E_0 = \langle\Phi_0|H|\Phi_0\rangle = \sum_i \langle i|h|i\rangle + \frac{1}{2}\sum_{ij}\langle ij||ij\rangle \qquad (10)$$

$|\Phi_n\rangle$ is the determinant derived by removing an electron from orbital n of $|\Phi_0\rangle$: $|\Phi_n\rangle = n|\Phi_0\rangle$, $\langle\Phi_m| = \langle\Phi_0|m^\dagger$. The Hamiltonian matrix element between two ionized determinants is

$$\langle\Phi_m|H|\Phi_n\rangle = \sum_{pq}\langle\Phi_0|m^\dagger p^\dagger qn|\Phi_0\rangle\langle p|h|q\rangle + \frac{1}{2}\sum_{pqrs}\langle\Phi_0|m^\dagger p^\dagger q^\dagger srn|\Phi_0\rangle\langle pq|rs\rangle \qquad (11)$$

Using the anticommutation relation $pq^\dagger + q^\dagger p = \delta_{pq}$, it is not difficult to show that

$$\sum_{pq}\langle\Phi_0|m^\dagger p^\dagger qn|\Phi_0\rangle = \sum_i \langle i|h|i\rangle\delta_{mn} - \langle n|h|m\rangle \qquad (12)$$

$$\frac{1}{2}\sum_{pqrs}\langle\Phi_0|m^\dagger p^\dagger q^\dagger srn|\Phi_0\rangle\langle pq|rs\rangle = \frac{1}{2}\sum_{ij}\langle ij||ij\rangle\delta_{mn} - \sum_j\langle nj||mj\rangle \qquad (13)$$

$$\langle\Phi_m|H|\Phi_n\rangle = \delta_{mn}\left(\sum_i\langle i|h|i\rangle + \frac{1}{2}\sum_{ij}\langle ij||ij\rangle\right) - \langle n|h|m\rangle - \sum_j\langle nj||mj\rangle = \delta_{mn}E_0 - \langle n|f|m\rangle \qquad (14)$$

Since the Fock matrix f is diagonal in the basis of Hartree−Fock orbitals, the matrix element becomes

$$\langle\Phi_m|H|\Phi_n\rangle = (E_0 - \varepsilon_n)\delta_{mn} \qquad (15)$$

where $\varepsilon_n = \langle n|h|n\rangle + \sum_j \langle nj||nj\rangle$ is a diagonal Fock matrix element and the energy of the n-th Hartee−Fock orbital.

Therefore, the Hamiltonian is diagonal in the basis of the electron detached or ionized determinants. The excitation energies of such states are equal to the respective orbital energies. Thus, configuration interaction of singly detached state functions is equivalent to Koopmans' theorem. This provides a useful diagnostic for TD-DFT: spurious states will appear in large bases at the onset of Koopmans continuum, i.e., HOMO energy.

References

(1) Tsien, R. Y. *Annu. Rev. Biochem.* **1998**, *67*, 509.

(2) Zimmer, M. *Chem. Rev.* **2002**, *102*, 759.

(3) Stepanenko, O. V.; Verkhusha, V. V.; Kuznetsova, I. M.; Uversky, V. N.; Turoverov, K. K. *Curr. Protein Pept. Sci.* **2008**, *9*, 338.

(4) Chen, R. F. *Arch. Biochem. Biophys.* **1977**, *179*, 672.

(5) Steiner, R. F.; Albaugh, S.; Nenortas, E.; Norris, L. *Biopolymers* **1997**, *32*, 73.

(6) Babendure, J. R.; Adams, S. R.; Tsien, R. Y. *J. Am. Chem. Soc.* **2003**, *125*, 14716.

(7) Silva, G. L.; Ediz, V.; Yaron, D.; Armitage, B. A. *J. Am. Chem. Soc.* **2007**, *129*, 5710.

(8) Constantin, T. P.; Silva, G. L.; Robertson, K. L.; Hamilton, T. P.; Fague, K.; Waggoner, A. S.; Armitage, B. A. *Org. Lett.* **2008**, *10*, 1561.

(9) Özhalici-Ünal, H.; Pow, C. L.; Marks, S. A.; Jesper, L. D.; Silva, G. L.; Shank, N. I.; Jones, E. W.; Burnette, J. M.; Berget, P. B.; Armitage, B. A. *J. Am. Chem. Soc.* **2008**, *130*, 1260.

(10) Jain, V.; Rajbongshi, B. K.; Mallajosyula, A. T.; Bhattacharjya, G.; Iyer, S. K.; Ramanathan, G. *Sol. Energy Mater. Sol. Cells* **2008**, *92*, 1043.

(11) You, Y. J.; He, Y. K.; Borrows, P. E.; Forrest, S. R.; Petasis, N. A.; Thompson, M. E. *Adv. Mater.* **2000**, *12*, 1678.

(12) Polyakov, I.; Epifanovsky, E.; Grigorenko, B. L.; Krylov, A. I.; Nemukhin, A. V. *J. Chem. Theory Comput.* 2009; DOI, 10.1021/ct9001448.

(13) Heim, R.; Prasher, D. C.; Tsien, R. Y. *Proc. Nat. Acad. Sci. U.S.A.* **1994**, *91*, 12501.

(14) Nielsen, S. B.; Lapierre, A.; Andersen, J. U.; Pedersen, U. V.; Tomita, S.; Andersen, L. H. *Phys. Rev. Lett.* **2001**, *87*, 228102.

(15) Andersen, L. H.; Lappierre, A.; Nielsen, S. B.; Nielsen, I. B.; Pedersen, S. U.; Pedersen, U. V.; Tomita, S. *Eur. Phys. J. D* **2002**, *20*, 597.

(16) Ward, W. W.; Cody, C. W.; Hart, R. C.; Cormier, M. J. *Photochem. Photobiol.* **1977**, *31*, 611.

(17) Chattoraj, M.; King, B. A.; Bublitz, G. U.; Boxer, S. G. *Proc. Nat. Acad. Sci. U.S.A.* **1996**, *93*, 8362.

(18) Voityuk, A. A.; Michel-Beyerle, M.-E.; Rösch, N. *Chem. Phys. Lett.* **1997**, *272*, 162.

(19) Weber, W.; Helms, V.; McCammon, J. A.; Langhoff, P. W. *Proc. Nat. Acad. Sci. U.S.A.* **1999**, *96*, 6177.

(20) Helms, V.; Winstead, C.; Langhoff, P. W. *J. Molec. Struct. (Theochem)* **2000**, *506*, 179.

(21) Das, A. K.; Hasegawa, J.-Y.; Miyahara, T.; Ehara, M.; Nakatsuji, H. *J. Comput. Chem.* **2003**, *24*, 1421.

(22) Martin, M. E.; Negri, F.; Olivucci, M. *J. Am. Chem. Soc.* **2004**, *126*, 5452.

(23) Nemukhin, A. V.; Topol, I. A.; Burt, S. K. *J. Chem. Theory Comput.* **2006**, *2*, 292.

(24) Bravaya, K. B.; Bochenkova, A. V.; Granovsky, A. A.; Nemukhin, A. V. *Russ. J. Phys. Chem. B* **2008**, *2*, 671.

(25) Zhang, Y.; Yang, W. *J. Chem. Phys.* **1998**, *109*, 2604.

(26) Polo, V.; Kraka, E.; Cremer, D. *Mol. Phys.* **2002**, *100*, 1771.

(27) Lundber, M.; Siegbahn, P. E. M. *J. Chem. Phys.* **2005**, *122*, 224103.

(28) Dreuw, A.; Weisman, J. L.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 2943.

(29) Lange, A. W.; Rohrdanz, M. A.; Herbert, J. M. *J. Phys. Chem. B* **2008**, *112*, 6304.

(30) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.

(31) Perdew, J. P. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *33*, 8822.

(32) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

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

(34) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. *J. Phys. Chem.* **1990**, *94*, 5483.

(35) Molina, V.; Merchán, M. *Proc. Nat. Acad. Sci. U.S.A.* **2001**, *98*, 4299.

(36) Andruniów, T.; Ferré, N.; Olivucci, M. *Proc. Nat. Acad. Sci. U.S.A.* **2004**, *101*, 17908.

(37) Murakami, A.; Kobayashi, T.; Goldberg, A.; Nakamura, S. *J. Chem. Phys.* **2004**, *120*, 1245.

(38) Cembran, A.; Bernardi, F.; Olivucci, M.; Garavelli, M. *Proc. Nat. Acad. Sci. U.S.A.* **2005**, *102*, 6255.

(39) Coto, P. B.; Strambi, A.; Ferré, N.; Olivucci, M. *Proc. Nat. Acad. Sci. U.S.A.* **2006**, *103*, 17154.

(40) Frutos, L. M.; Andruniów, T.; Santoro, F.; Ferré, N.; Olivucci, M. *Proc. Nat. Acad. Sci. U.S.A.* **2007**, *104*, 7764.

(41) Schreiber, M.; Barbatti, M.; Zilberg, S.; Lischka, H.; González, L. *J. Phys. Chem. A* **2007**, *111*, 238.

(42) Olsen, S.; Smith, S. C. *J. Am. Chem. Soc.* **2008**, *130*, 8677.

(43) Strambi, A.; Coto, P. B.; Frutos, L. M.; Ferré, N.; Olivucci, M. *J. Am. Chem. Soc.* **2008**, *130*, 3382.

(44) Tokmachev, A. M.; Boggio-Pasqua, M.; Bearpark, M. J.; Robb, M. A. *J. Phys. Chem. A* **2008**, *112*, 10881.

(45) Hirao, K. *Chem. Phys. Lett.* **1992**, *190*, 374.

(46) Nakano, H. *J. Chem. Phys.* **1993**, *99*, 7983.

(47) Nemukhin, A. V.; Bochenkova, A. V.; Bravaya, K. B.; Granovsky, A. A. *Proc. SPIE - Int. Soc. Opt. Eng.* **2007**, *6449*, 64490N.

(48) Bravaya, K.; Bochenkova, A.; Granovsky, A.; Nemukhin, A. *J. Am. Chem. Soc.* **2007**, *129*, 13035.

(49) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *77*, 3865.

(50) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.

(51) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(52) Andersen, L. H.; Bluhme, H.; Boyé, S.; Jørgensen, T. J. D.; Krogh, H.; Nielsen, I. B.; Nielsen, S. B.; Svendsen, A. *Phys. Chem. Chem. Phys.* **2004**, *6*, 2617.

(53) Greemers, T. M. H.; Lock, A. J.; Subramaniam, V.; Jovin, T. M.; Völker, S. *Proc. Nat. Acad. Sci. U.S.A.* **2000**, *97*, 2974.

(54) Jung, G.; Bräuchle, C.; Zumbusch, A. *J. Chem. Phys.* **2001**, *114*, 3149.

(55) Rowe, D. J. *Rev. Mod. Phys.* **1968**, *40*, 153.

(56) Emrich, K. *Nucl. Phys. A* **1981**, *351*, 379.

(57) Sinha, D.; Mukhopadhyay, D.; Mukherjee, D. *Chem. Phys. Lett.* **1986**, *129*, 369.

(58) Stanton, J. F.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 7029.

(59) Simons, J. Equation of motion (EOM) methods for computing electron affinities. In *Encyclopedia of Computational Chemistry*; J. Wiley & Sons: New York, 2004.

(60) Krylov, A. I. *Annu. Rev. Phys. Chem.* **2008**, *59*, 433.

(61) Livshits, E.; Baer, R. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2932.

(62) Chai, J.-D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106.

(63) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular electronic structure theory*, Wiley & Sons: Chichester, England, 2000.

(64) Granovsky, A. *PC GAMESS*. http://classic.chem.msu.su/gran/gamess/index.html (accessed April 27, 2009).

(65) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Mastunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.

(66) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neil, D. P.; Distasio, R. A., Jr.; Lochan, R. C.; Wang, T.; Beran, G. J. O.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Van Voorhis, T.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Bird, E. F. C.; Daschel, H.; Doerksen, R. J.; Drew, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C.-P.; Kedziora, G. S.; Khalliulin, R. Z.; Klunziger, P.; Lee, A. M.; Liang, W. Z.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock III, H. L.; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Herhe, W. J.; Schaefer III, H. F.; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172.

(67) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110.

(68) Ghigo, G.; Roos, B. O.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **2004**, *396*, 142.

(69) Christiansen, O.; Koch, H.; Jorgensen, P. *J. Chem. Phys.* **1995**, *103*, 7429.

(70) Koch, H.; Christiansen, O.; Jørgensen, P.; de Meras, A. M. S.; Helgaker, T. *J. Chem. Phys.* **1997**, *106*, 1808.

(71) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095.

(72) Jung, Y.; Lochan, R. C.; Dutoi, A. D.; Head-Gordon, M. *J. Chem. Phys.* **2004**, *121*, 9793.

(73) DiStacio, R. A., Jr.; Head-Gordon, M. *Mol. Phys.* **2007**, *105*, 1073.

(74) Rhee, Y. M.; Head-Gordon, M. *J. Phys. Chem. A* **2007**, *111*, 5314.

(75) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540.

(76) Baer, R.; Neuhauser, D. *Phys. Rev. Lett.* **2005**, *94*, 043002.

(77) Stein, T.; Kronik, L.; Baer, R. *J. Am. Chem. Soc.* **2009**, *131*, 2818–2820.

(78) Pal, S.; Rittby, M.; Bartlett, R. J.; Sinha, D.; Mukherjee, D. *J. Chem. Phys.* **1988**, *88*, 4357.

(79) Stanton, J. F.; Gauss, J. *J. Chem. Phys.* **1994**, *101*, 8938.

(80) Levchenko, S. V.; Krylov, A. I. *J. Chem. Phys.* **2004**, *120*, 175.

(81) Larsen, H.; Hald, K.; Olsen, J.; Jørgensen, P. *J. Chem. Phys.* **2001**, *115*, 3015.

(82) Epifanovsky, E.; Kowalski, K.; Fan, P.-D.; Valiev, M.; Matsika, S.; Krylov, A. I. *J. Phys. Chem. A* **2008**, *112*, 9983.

(83) Hager, B.; Schwarzinger, B.; Falk, H. *Monatsh. Chem.* **2006**, *137*, 163.

(84) Weinhold, F.; Landis, C. R. *Chem. Edu.: Res. Pract. Eur.* **2001**, *2*, 91.

(85) Glendening, E. D.; Badenhoop, J. K.; Reed, A. E.; Carpenter, J. E.; Bohmann, J. A.; Morales, C. M.; Weinhold, F. *NBO 5.0*; Theoretical Chemistry Institute, University of Wisconsin: Madison, WI, 2001.

(86) Slipchenko, L. V.; Krylov, A. I. *J. Chem. Phys.* **2003**, *118*, 6874.

(87) Krylov, A. I.; Sherrill, C. D. *J. Chem. Phys.* **2002**, *116*, 3194.

(88) Simons, J. *J. Phys. Chem. A* **2008**, *112*, 6401.

(89) Simons, J. *Acc. Chem. Res.* **2006**, *39*, 772.

(90) Sobczyk, M. Simons, J. *J. Phys. Chem. B* **2006**, *110*, 7519.

(91) Solntsev, K. M.; Poizat, O.; Dong, J.; Rehault, J.; Lou, Y.; Burda, C.; Tolbert, L. M. *J. Phys. Chem. B* **2008**, *112*, 2700.

(92) Baer, R.; Krylov, A. I. in preparation.

(93) Oana, M.; Krylov, A. I. *J. Chem. Phys.* **2007**, *127*, 234106.

(94) Olsen, S. Ph. D. Thesis, The University of Illinois at Urbana-Champaign, Urbana, IL, 2004.

(95) Stanton, J. F.; Gauss, J. *J. Chem. Phys.* **1999**, *111*, 8785.

(96) Pieniazek, P. A.; Arnstein, S. A.; Bradforth, S. E.; Krylov, A. I.; Sherrill, C. D. *J. Chem. Phys.* **2007**, *127*, 164110.

(97) Hazi, A. U.; Taylor, H. S. *Phys. Rev. A: At., Mol., Opt. Phys.* **1970**, *1*, 1109.

(98) Dong, J.; Abulwerdi, F.; Baldridge, A.; Kowalik, J.; Solntsev, K. M.; Tolbert, L. M. *J. Am. Chem. Soc.* **2008**, *130*, 14096.

CT900143J

JCTC Journal of Chemical Theory and Computation

Quantum Chemical Benchmark Studies of the Electronic Properties of the Green Fluorescent Protein Chromophore: 2. *Cis*−*Trans* Isomerization in Water

Igor Polyakov,[†] Evgeny Epifanovsky,*[,‡] Bella Grigorenko,[†] Anna I. Krylov,*[,‡] and Alexander Nemukhin[†,§]

Department of Chemistry, University of Southern California, Los Angeles, California 90089, Department of Chemistry, M.V. Lomonosov Moscow State University, Moscow 119991, Russia, and Institute of Biochemical Physics, Russian Academy of Sciences, Moscow 119334, Russia

Received March 25, 2009

Abstract: We present quantum chemical calculations of the properties of the anionic form of the green fluorescent protein (GFP) chromophore that can be directly compared to the results of experimental measurements: the *cis*−*trans* isomerization energy profile in water. Calculations of the *cis*−*trans* chromophore isomerization pathway in the gas phase and in water reveal a problematic behavior of density functional theory and scaled opposite-spin-MP2 due to the multiconfigurational character of the wave function at twisted geometries. The solvent effects treated with the continuum solvation models, as well as with the water cluster model, are found to be important and can reduce the activation energy by more than 10 kcal/mol. Strong solvent effects are explained by the change in charge localization patterns along the isomerization coordinate. At the equilibrium, the negative charge is almost equally delocalized between the phenyl and imidazolin rings due to the interaction of two resonance structures, whereas at the transition state the charge is localized on the imidazolin moiety. Our best estimate of the barrier obtained in cluster calculations employing the effective fragment potential-based quantum mechanics/molecular mechanics method with the complete active space self-consistent field description of the chromophore augmented by perturbation theory correction and the TIP3P water model is 14.8 kcal/mol, which is in excellent agreement with the experimental value of 15.4 kcal/mol. This result helps to resolve previously reported disagreements between experimental measurements and theoretical estimates.

1. Introduction

The properties of the green fluorescent protein (GFP), which converts blue light to green light, have inspired numerous experimental and theoretical studies as well as many important applications (see ref 1–3 and references therein). This paper is the second in a series[4] that focuses on accurate calculations of the properties of biological chromophores with ab initio methods using the model GFP chromophore, 4′-hydroxybenzylidene-2,3-dimethylimidazolinone (HBDI) anion, as a benchmark system (Figure 1).

The *cis*−*trans* isomerization (or *Z*/*E* diastereomerization) of the photoswitchable fluorescent chromophores plays an essential role in their photophysical properties. For example, their functionality is believed to be driven by photoinduced *cis*−*trans* isomerization inside a protein matrix.[5] Kindling and blinking phenomena, as well as loss of fluorescence yield of bare chromophores in solution, are also related to this process. In a broader context, the photophysics of GFP is similar to that of other fluorogenic unsymmetric methine

* Corresponding authors. E-mail: epifanov@usc.edu (E.E.) and krylov@usc.edu (A.I.K.).

[†] M.V. Lomonosov Moscow State University.

[‡] University of Southern California.

[§] Russian Academy of Sciences.

10.1021/ct9001448 CCC: $40.75
Published on Web 06/25/2009

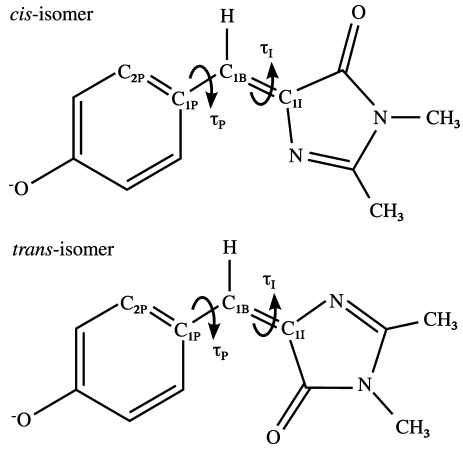


Figure 1. Chemical structure and atomic labels of the anionic form of the model GFP chromophore HBDI in the *cis* (top) and *trans* (bottom) conformations.

dyes[6–11] and is of interest with connection to organic photovoltaic materials.

The majority of the experimental and computational studies of this process focused on the excited-state dynamics. However, the ground-state electronic potential-energy surface (PES) also needs to be considered for two main reasons. First, after photoisomerization has completed, the photoswitchable protein returns to its initial state to prepare for the next cycle. The recovery apparently takes place on the ground-state PES. Second, the details of the ground-state isomerization can elucidate the chromophore's rearrangements along the more complicated excited-state route. Despite previous studies of the *cis*−*trans* photoisomerization of GFP-like chromophores[12–17] using quantum chemistry modeling,[18–24] many questions remain unanswered.

The process of *cis*−*trans* isomerization of the HBDI anion in aqueous solution was investigated experimentally by He et al.[12] who estimated the free-energy differences and the activation energies using the NMR technique. The authors stressed that the activation energy of 15.4 kcal/mol derived from their measurements in aqueous solution is in distinct disagreement with the results of calculations,[18,19] which estimated that the barrier is above 21 kcal/mol. The relatively low value of the barrier (as compared to other isomerization reactions involving exocyclic double bonds) has also been emphasized in subsequent studies of the isomerization and several explanations have been suggested.[13,17] For example, thermal isomerization studies of model GFP-like compounds[13] suggested that different substituent groups may have a significant effect on the activation energy by changing the interaction between two resonance structures of the chromophore. Tolbert and co-workers considered mechanisms involving changes in the chemical structure of the chromophore, e.g., addition/elimination pathway.[17] No ab initio calculations have been reported so far to resolve this disagreement between experimental measurements and theoretical estimates and to explain the low value of the barrier.

2. Computational Methods

The equilibrium geometries were optimized by density functional theory (DFT) with the PBE0 variant[25] of the Perdew−Burke−Ernzerhof (PBE) hybrid functional[26] and by complete active space self-consistent field (CASSCF)(14/12). The cc-pVDZ basis set[27] was used in both calculations. The *cis*−*trans* isomerization pathways of the chromophore were studied with the DFT and CASSCF methods. We estimated the solvent effects by using continuum solvation models as well as by explicit treatment of water molecules in a quantum mechanics/molecular mechanics (QM/MM) scheme. The Cartesian coordinates of optimized structures along the isomerization pathway are given in the Supporting Information.

The methods are outlined below and in the first paper in this series,[4] and the computational details are given in the Results and Discussion section. Multireference second-order perturbation theory (MRMP2) and dielectric polarizable continuum model (D-PCM)[33] calculations were carried out with the PC GAMESS version[28] of the GAMESS(US) quantum chemistry package.[29] Scaled opposite-spin-MP2 (SOS-MP2) calculations were performed with Q-Chem.[30] GAMESS(US)[29] was employed for C-PCM and surface and volume polarization for electrostatic (SVPE) approach[35] computations. The QM/MM implementation is based on PC GAMESS[28] and the Tinker molecular mechanics package.[31]

2.1. Continuum Solvation Models. To simulate solvent effects on the chromophore's *cis*−*trans* isomerization energy profile in an aqueous solution, we employ three versions of the continuum solvation model:[32] D-PCM, C-PCM, and SVPE. In the simplest approach, the D-PCM,[33] the water solvent is treated as a continuous unstructured dielectric with a dielectric constant of 78.39. C-PCM[34] is a version of PCM that takes into account certain corrections in the boundary conditions of the electrostatic problem in accord with the conductor-like screening model. The SVPE approach[35] provides an improved description of the volume polarization contributions. This effect can be significant for reaction barriers, especially for ionic solutes.[35]

2.2. Effective Fragment Potential-Based QM/MM Method. Solvent effects can be described explicitly in a combined QM/MM approach based on the effective fragment potential (EFP) model.[36,37] In this scheme, solvent molecules (water in our case) are represented by effective fragments in the MM-part. The fragments affect the Hamiltonian of the QM-part (HBDI anion) by their electrostatic potentials expanded up to octupole terms. The parameters of these one-electron electrostatic potentials, as well as contributions from interactions between polarizable effective fragments and the QM-region, are computed in preliminary ab initio calculations of the electronic densities of individual fragments. The exchange−repulsion potentials, which are combined with the electrostatic and polarizability terms, are obtained from preliminary ab initio calculations as well.

The original EFP approach[36] treats interactions between solvent molecules as EFP−EFP interactions. Studies of chemical reactions in aqueous solution[38,39] have shown that replacing the EFP−EFP terms by the empirically calibrated TIP3P potential yields a faster computational scheme for large water clusters around the solute.

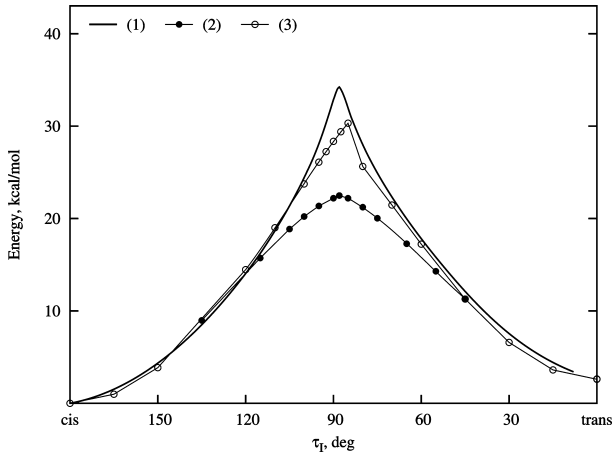


Figure 2. Computed energy profiles of *cis*−*trans* ground-state isomerization of the HBDI anion in the gas phase: (1) IRC calculated with DFT(PBE0)/6-31+G(d,p); (2) MEP calculated with CASSCF(12/11)/cc-pVDZ; and (3) MEP calculated with SOS-MP2/cc-pVDZ.

3. Results and Discussion

3.1. *Cis*−*Trans* Isomerization Pathway in the Gas Phase. On the ground-state PES, conversion between the lower-energy *cis*-isomer and the higher-energy *trans*-isomer takes place when the two rings are rotated along the $C_{1I}-C_{1B}$ bond as shown in Figure 1. The dihedral angle $N-C_{1I}-C_{1B}-H$, which defines the reaction coordinate, is denoted as τ_I.

The shape of the isomerization pathway can be characterized by: (i) the activation energy E_a defined as the relative energy of the transition state (TS) with respect to the *cis*-isomer; and (ii) the energy difference E_{tc} between the *trans*- and *cis*-forms. We began with gas-phase DFT calculations with the PBE0 functional and the 6-31+G(d,p) basis set and located stationary points that correspond to the *cis*-, *trans*- and TS structures. The latter is characterized by a single imaginary frequency of $725i$ cm^{-1}. The intrinsic reaction coordinate (IRC) profile was computed by starting steepest descent pathways from the TS point in both directions along the Hessian eigenvector that corresponds to the imaginary frequency (shown in Supporting Information).

The results of this calculation are discouraging in two aspects: the cusp-like shape of the profile (Figure 2), which is consistent with the large value of the imaginary frequency at the saddle point, and the value of the activation energy E_a(DFT) = 34.5 kcal/mol, which is more than twice as high as the experimental estimate of 15.4 kcal/mol in aqueous solution.[12] The computed energy difference between the *trans*- and *cis*- structures E_{tc}(DFT) = 2.3 kcal/mol is in excellent agreement with the experimental estimate in solution,[12] most likely due to error cancellation.

In the CASSCF energy profile (Figure 2), the stationary points were fully optimized with CASSCF(12/11)/cc-pVDZ. The points in between representing the minimum-energy path (MEP) were computed by varying the value of the dihedral angle τ_I and minimizing the energy by relaxing all other degrees of freedom. Although this curve does not represent the true IRC path, it is expected to be a good approximation to it. Along with a smoother curvature in the vicinity of the saddle point (imaginary frequency $77i$ cm^{-1}, see Supporting Information), this profile yields a reasonable value of the *trans-cis* energy difference E_{tc}(CASSCF) = 3.5 kcal/mol and a much lower (and closer to the experimental estimate) value of the activation energy E_a(CASSCF) = 22.5 kcal/mol.

When calculating the energy profile in the CASSCF(12/11) approximation with a fairly large active space, we performed careful selection of the orbitals in order to avoid dubious solutions of the variational problem. The finally optimized orbitals and the corresponding occupation numbers at selected points along the energy graph are presented in the Supporting Information. To better understand changes in the electronic structure along the isomerization pathway, we also computed the energy profile with the smallest active space CASSCF(2/2). Selection of active orbitals in this approach was performed on the base of previously optimized orbitals at the TS point. Then the descent in both directions toward minimum-energy structures was easy to accomplish. As expected, the corresponding value of the activation energy, 25 kcal/mol, was slightly larger than that computed with CASSCF(12/11).

As discussed in more detail below (and illustrated by the population analysis presented in Supporting Information), there is an increase in charge localization at the TS relative to that the minimum-energy points.

The transition state located with SOS-MP2/cc-pVDZ is characterized by an imaginary frequency of $188i$ cm^{-1}. Starting from that point, the MEP was taken to the *cis*- and *trans*-configurations (Figure 2). Both DFT and SOS-MP2 exhibit a cusp at the transition state, which is a manifestation of the multireference character of the ground-state wave function. The composition of the CASSCF wave function in the region discussed below confirms that.

Table 1 presents computed equilibrium geometry parameters in the bridging region: the $C_{1P}-C_{1B}$ and $C_{1I}-C_{1B}$ bond lengths and the τ_P ($C_{2P}-C_{1P}-C_{1B}-H$) and τ_I ($N-C_{1I}-C_{1B}-H$) dihedral angles (Figure 1). The parameters optimized at the PBE0/6-31+G(d,p), CASSCF(12/11)/cc-pVDZ, and SOS-MP2/cc-pVDZ levels are in agreement with those obtained by Olsen and Smith[23] with SA3-CAS(4/3)/DZP, which stands for CASSCF with the (4/3) active space averaged over three states computed with the DZP basis set. Overall, the geometry parameters at the stationary points are rather insensitive to the level of theory: DFT, large active space state-specific CASSCF, small active space state-averaged CASSCF, and SOS-MP2 yield bond lengths that agree within 0.02 Å and angles that agree within 2°.

At both *cis*- and *trans*-equilibrium points, the molecule is essentially planar except that the methyl groups naturally have out-of-plane atoms. Analysis of the structures reveals that the difference between $C_{1P}-C_{1B}$ and $C_{1B}-C_{1I}$ is less than expected for the structure shown in Figure 1. This can be explained by considering two resonance structures of the HBDI anion, see figure 2 from ref 4. The interaction between these two structures results in charge delocalization between two oxygen atoms and in scrambling CC bond orders as discussed, for example, in ref 23. The $C_{1P}-C_{1B}$ bond acquires a double-bond character, whereas the order of $C_{1B}-C_{1I}$ bond is reduced. Natural bond orbital (NBO)[40,41] charges and bond

Table 1. Geometry Parameters at the Stationary Points for the Gas-Phase *Cis*−*Trans* Chromophore Isomerization[a]

structure	method	C_{1P}−C_{1B}	C_{1B}−C_{1I}	τ_P	τ_I
cis-isomer	PBE0/6-31+G(d,p)	1.404	1.385	0	180
	CASSCF(12/11)/cc-pVDZ	1.406	1.383	0	180
	SOS-MP2/cc-pVDZ	1.415	1.397	0	180
	SA3-CAS(4,3)/DZP[b]	1.408	1.382	0	180
transition state	PBE0/6-31+G(d,p)	1.365	1.458	0.3	88.2
	CASSCF(12/11)/cc-pVDZ	1.362	1.477	1.0	86.4
	SOS-MP2/cc-pVDZ	1.374	1.476	0.0	87.6
trans-isomer	PBE0/6-31+G(d,p)	1.403	1.392	0	0
	CASSCF(12/11)/cc-pVDZ	1.402	1.395	0	0
	SOS-MP2/cc-pVDZ	1.414	1.405	0	0
	SA3-CAS(4/3)/DZP[b]	1.407	1.390	0	0

[a] Distances in Å and angles in degrees. [b] Ref 23.

Table 2. Cumulative Natural Charges on the Fragments of the HBDI molecule: the Phenyl and Dimethylimidazolin Rings and the CH bridge Calculated with CASSCF(12/11) and DFT (PBE0 Functional)

	CASSCF(12/11)/cc-pVDZ			PBE0/6-31+G(d,p)		
	cis	TS	*trans*	*cis*	TS	*trans*
			Gas Phase			
phenyl	−0.61	−0.20	−0.50	−0.59	−0.30	−0.57
bridge	0.13	0.13	0.11	0.10	0.20	0.09
imidazolin	−0.52	−0.93	−0.61	−0.51	−0.90	−0.52
			Solution[a]			
phenyl	−0.85	−0.09	−0.91	−0.63	−0.19	−0.70
bridge	0.14	0.08	0.11	0.13	0.16	0.13
imidazolin	−0.29	−0.99	−0.20	−0.50	−0.97	−0.43

[a] Calculated with QM/MM: EFP for the solvent−QM part interaction and TIP3P for the water−water interaction.

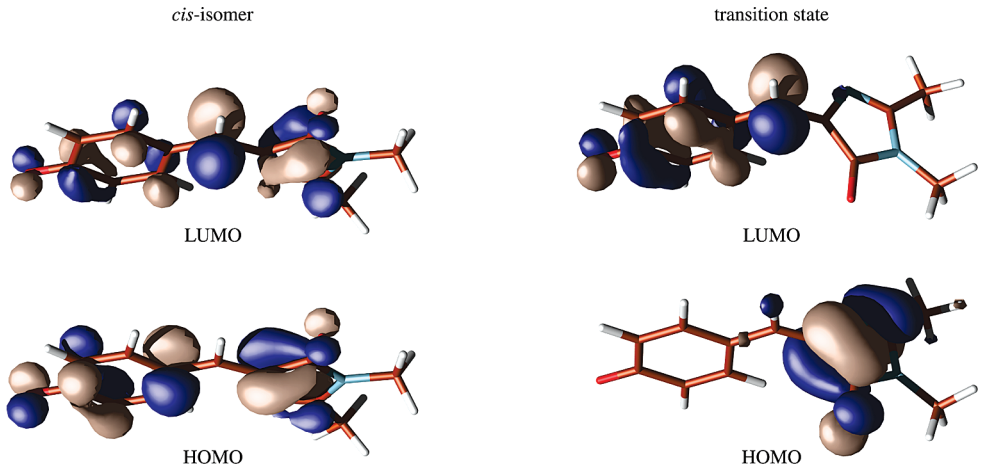


Figure 3. Frontier valence molecular orbitals of the HBDI anion in the *cis*-form (left) and at the transition state of the *cis*−*trans* isomerization path (right).

orders (see ref 4) are consistent with computed bond lengths. Overall, CASSCF slightly exaggerates bond alternation relative to DFT (or MP2, see ref 4) in favor of the canonical structure (Figure 1).

At the TS, the two rings are nearly perpendicular to each other ($\tau_I = 87-88°$), which disturbs the π system and breaks the resonance interaction. The bond alternation pattern is reversed such that C_{1P}−C_{1B} becomes shorter than C_{1B}−C_{1I}, suggesting that one of the two resonance structures becomes dominant. All the methods agree on the magnitude of the change relative to the equilibrium structures: C_{1P}−C_{1B}, which is longer at the *cis*- and *trans*-geometries, becomes shorter by about 0.04 Å at the TS, whereas the C_{1B}−C_{1I} bond is about 0.07 Å longer at the saddle point. NBO charges (Table 2) and the molecular orbital picture (Figure 3) reveal almost complete charge localization on the imidazolin ring.

Despite the similarity of the computed geometry parameters and charge distributions at the minima and the TS computed with DFT and CASSCF, respective activation energies differ by more than 10 kcal/mol (34.5 versus 22.5 kcal/mol). Note that the corresponding *cis*−*trans* energy differences are very close: 2.3 and 3.5 kcal/mol, respectively. A close inspection of the charges from Table 1 reveals slightly more polar charge distribution for the PBE0 density: the positive charge on the bridge moiety is 0.19 versus 0.13 at the CASSCF level. Because of the large energy penalty

Table 3. *Cis-Trans* Energy Difference ΔE_{tc} and the Energy Barrier E_a for the *Cis*−*Trans* Isomerization of the HBDI Anion Calculated at Various Levels of Theory for the Chromophore Molecule and Solvent

method		ΔE_{tc}, kcal/mol			E_a, kcal/mol		
chromophore	solvent	gas phase	solution	shift	gas phase	solution	shift
PBE0/6-31+G(d,p)		2.3			34.5		
	D-PCM		2.1	−0.2		33.5	−1.0
	C-PCM		2.3	+0.0		34.0	−0.5
	SVPE		2.1	−0.2		24.6	−9.9
	QM/MM[a]		5.0	+2.7		26.0	−8.5
CASSCF(12/11)/cc-pVDZ		3.5			22.5		
	SVPE		2.6	−0.9		9.9	−12.6
	QM/MM[a]		2.1	−1.4		11.1	−11.4
MRMP2/cc-pVDZ		3.7			26.2		

[a] Calculated with QM/MM: EFP for the solvent−QM part interaction and TIP3P for the water−water interaction.

due to charge separation in the gas phase, small differences in ionicity may produce a large effect. Another notable difference is larger asymmetry in oxygen charges: at the CASSCF level the imidazolin oxygen is by 0.22*e* more negative than that of the phenyl oxygen, whereas at the DFT level this difference is reduced to 0.15. This suggests a larger contribution of the second resonance structure in the CASSCF wave function, which can also contribute to the energy difference.

More ionic character of the PBE0 density and the cusp-like shape of the profile (Figure 2) are due to the multiconfigurational character of the wave function at the TS, which is not adequately described by DFT (or MP2). The CASSCF amplitudes show almost equal weights of the two dominant configurations, $(HOMO)^2$ and $(HOMO)^1(LUMO)^1$ at the TS. The HOMO and LUMO at the TS differ considerably from those at the equilibrium geometry (shown in figure 3 of ref 4), they become localized on the imidazolin and phenyl rings, respectively. Thus, $(HOMO)^2$ and $(LUMO)^2$ correspond to the two charge-localized configurations, and their interaction results in less ionic electron distribution. The ionicity is also reduced by $(HOMO)^1(LUMO)^1$.

A stability analysis of the Hartree−Fock wave function at the TS shows a RHF−UHF (restricted and unrestricted Hartree-Fock) instability with a negative eigenvalue of −0.037. The DFT/PBE0 solution, however, proves to be stable, which means that using the symmetry broken unrestricted solution to achieve better description of the barrier will not be useful in this case. Therefore, a multiconfigurational approach is necessary not only for describing excited-state isomerization of the GFP-like chromophores but also for modeling isomerization in the ground state.

Although the CASSCF wave function is capable of capturing the multiconfigurational character of the wave function, it needs to be augmented by dynamical correlation to provide accurate energy differences. We included dynamical correlation correction via MRMP2 for the E_{tc} and E_a energies (at the CASSCF geometries). This yields an activation energy of $E_a = 26.2$ kcal/mol, which is 3.7 kcal/mol higher than that of the CASSCF result.

The CASSCF results represent an improvement over DFT; however, there is still a considerable discrepancy between the theoretical (22−26 kcal/mol) and experimental (15.4 kcal/mol) values for E_a, as noted by the authors of experimental studies.[12,13,17] Below we demonstrate that this discrepancy is resolved when solvent effects are taken into account.

3.2. *Cis*−*Trans* Isomerization Pathway in Aqueous Solution: Continuum Solvation Models. Continuum solvation models[32] provide a reasonable starting point in modeling ground-state isomerization in aqueous solution. At first, two versions of the polarized continuum model (PCM), D-PCM[33] and C-PCM,[34] were applied to optimize the equilibrium geometry parameters of the *cis*- and *trans*-isomers and the TS configuration and to compute the relative energies ΔE_{tc} and E_a at the PBE0/6-31+G(d,p) level. Both models produced similar results: $\Delta E_{tc} = 2.1$ (D-PCM) and 2.3 kcal/mol (C-PCM); $E_a = 33.5$ (D-PCM) and 34.0 kcal/mol (C-PCM). The geometry parameters and the energies are close to the gas-phase values obtained with the same DFT model. These results are fairly stable with respect to the basis set: upon expanding it to 6-311++G(2d,p), E_a and ΔE_{tc} change by less than 1.5 and 0.7 kcal/mol, respectively. Table 3 summarizes solvent effects on E_{tc} and E_a computed using different approaches.

Next, we considered a new version of the continuum solvation model, SVPE,[35] which has recently been implemented in GAMESS(US),[29] for the single point calculations of relative energies ΔE_{tc} and E_a at the gas-phase geometry parameters. At the PBE0/6-31+G(d,p) level, we obtained a considerable reduction of the activation energy compared to the PCM results: $E_a = 24.6$ kcal/mol, while the energy $\Delta E_{tc} = 2.1$ kcal/mol was almost the same as in the PCM model. According to a comment by Chipman,[35] the improvements introduced in SVPE do affect the reaction barrier estimated with the continuum solvation models, especially for charged substrates. The large solvent effect on E_a (9.1 kcal/mol reduction relative to the gas-phase value) is consistent with the more ionic character of the TS (see Table 2).

Finally, we combined the SVPE model with the adequate description of the electronic structure of the solute and carried out the calculations of relative energies with the CASSCF(12/11)/cc-pVDZ description for the chromophore. The quantities, $\Delta E_{tc} = 2.6$ and $E_a = 9.9$ kcal/mol, can now be directly compared to the experimental free-energy difference of 2.3 and 15.4 kcal/mol.[12] The reduction of about 15 kcal/mol in E_a when going from DFT to CASSCF results within the SVPE solvation model is consistent with the corresponding reduction (13 kcal/mol) in the gas-phase calculations.

Moreover, DFT and CASSCF agree in the magnitude of the reduction of E_a due to solvent (9.1 and 12.6 kcal/mol, respectively). Thus, the absolute value of E_a is overestimated by DFT due to the multiconfigurational character of the electronic structure at the TS geometry, which is correctly captured by CASSCF.

3.3. *Cis*−*Trans* Isomerization Pathway in Aqueous Solution: Explicit Solvent Molecules. The relative energies ΔE_{tc} and E_a for the HBDI anion inside a cluster of water molecules were computed using a QM/MM technique. To build the starting point, the chromophore molecule at the gas-phase TS geometry was placed in a sphere of 200 water molecules using the VMD computer program.[42] After running short molecular dynamics trajectories at various temperatures, the energy was minimized using molecular mechanics with the CHARMM force field[43] by keeping the solute species frozen. Then the solvent outside the first solvation shell was removed with 49 water molecules completely covering the chromophore remaining in the system. This saddle point structure was reoptimized for all geometric degrees of freedom by QM/MM using DFT with the PBE0 functional and the 6-31+G(d,p) basis for the QM-part (chromophore), EFP[36] for QM-solvent interactions, and the empirical TIP3P potential for water−water interactions. The steepest descent pathways taken in both directions from the optimized TS lead to the *cis*- and *trans*-isomers of the trapped chromophore (Figure 4). The energies obtained with this method, ΔE_{tc} = 5.0 kcal/mol and E_a = 26.0 kcal/mol, are consistent with the DFT/SVPE calculations (Table 3).

The energies of the *cis*-, *trans*- and TS structures were also calculated using CASSCF(12/11) for in the QM-part. The interactions between the chromophore molecule and solvent and between the water molecules were handled with EFP and TIP3P, respectively. In line with the gas-phase and dielectric continuum model results, the activation energy is lower at the CASSCF level relative to those of DFT: ΔE_{tc} = 2.1 kcal/mol, E_a = 11.1 kcal/mol.

The relative energies can be further refined by including the dynamical correlation effects. In the gas phase, accounting for dynamical correlation via MRMP2 raises the CASSCF activation barrier by 3.7 kcal/mol. Applying that correction, we estimate the activation energy for the *cis*−*trans* isomerization of the GFP chromophore in aqueous solution to be 14.8 kcal/mol. The remaining discrepancy with the experimental value of 15.4 kcal/mol[12] can be partly attributed to the differences between the potential-energy and the free-energy barriers. The present QM/MM model with only 49 explicit water molecules is not sufficient for describing the statistical state in solution. This water solvation shell completely covers the chromophore molecule and accounts for principal environmental effects, but precise free-energy values, along the isomerization pathway, should be estimated by using more appropriate condensed-phase models. As shown, e.g., in ref 44, no considerable changes in conclusions are expected if the dielectric continuum model is applied on top of the model with explicit water molecules.

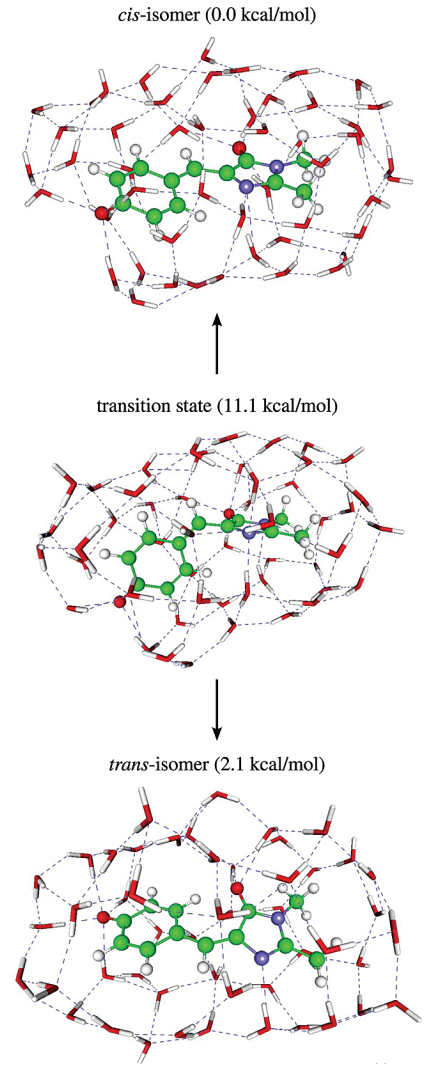


Figure 4. *Cis-trans* isomerization of the chromophore inside a cluster of water molecules. Relative energies of the stationary points are computed with QM(CASSCF(12/11)/cc-pVDZ)/EFP/MM(TIP3P).

4. Conclusion

Calculations of the *cis*−*trans* isomerization pathway require wave functions that are flexible enough to reflect the multiconfigurational character of the transition state. CASSCF gives a qualitatively correct curve, whereas DFT and SOS-MP2 fail at the twisted geometries. The gas-phase CASSCF value of the barrier height is 10 kcal/mol higher than that of the experimental value; however, the inclusion of solvent effects brings it down to 9.9−11.1 kcal/mol, which agrees well with 15.4 kcal/mol derived from experimental measurements. Including dynamical correlation correction yields 14.8 kcal/mol, which is within 0.8 kcal/mol from the experiment. Good agreement between SVPE and QM/MM calculations further supports the validity of our results. A large solvent effect on E_a is due to the more ionic character of the TS, where the negative charge is localized on imidazolin ring, which is in contrast to the equilibrium structure, where the negative charge is delocalized between the two rings. Our calculations help to resolve previous

disagreements between theory and experiment with respect to the GFP chromophore isomerization in an aqueous solution.

Acknowledgment. We thank Alex Granovsky for generous help and valuable discussions of this work. This work is supported by the joint grant from the U.S. Civilian Research and Development Foundation (project RUC1-2914-MO-07) and the Russian Foundation for Basic Research (08-03-91104-AFGIR). I.P., B.G., and A.N. thank the SKIF-GRID program and SKIF-Siberia for providing computational resources. E.E. and A.I.K. acknowledge the iOpenShell Center for Computational Studies of Electronic Structure and Spectroscopy of Open-Shell and Electronically Excited Species (http://iopenshell.usc.edu) supported by the National Science Foundation through the CRIF:CRF CHE-0625419 + 0624602 + 0625237 grant as well as through the CHE-0616271 grant (AIK). A.I.K. is grateful to the Institute of Mathematics and its Applications in Minnesota for its hospitality and productive environment during her stay as a visiting professor.

Supporting Information Available: Optimized geometry of the HBDI anion and the anion in water, the natural population analysis, the imaginary frequency mode at the TS, and the CASSCF natural orbitals figures. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Tsien, R. Y. *Annu. Rev. Biochem.* **1998**, *67*, 509.

(2) Zimmer, M. *Chem. Rev.* **2002**, *102*, 759.

(3) Stepanenko, O. V.; Verkhusha, V. V.; Kuznetsova, I. M.; Uversky, V. N.; Turoverov, K. K. *Curr. Protein Pept. Sci.* **2008**, *9*, 338.

(4) Epifanovsky, E.; Polyakov, I.; Grigorenko, B. L.; Nemukhin, A. V.; Krylov, A. I. *J. Chem. Theory Comput.* 2009; DOI: 10.1021/ct903187f.

(5) Henderson, J. N.; Remington, S. J. *Physiology* **2006**, *21*, 162.

(6) Chen, R. F. *Arch. Biochem. Biophys.* **1977**, *179*, 672.

(7) Steiner, R. F.; Albaugh, S.; Nenortas, E.; Norris, L. *Biopolymers* **1997**, *32*, 73.

(8) Babendure, J. R.; Adams, S. R.; Tsien, R. Y. *J. Am. Chem. Soc.* **2003**, *125*, 14716.

(9) Silva, G. L.; Ediz, V.; Yaron, D.; Armitage, B. A. *J. Am. Chem. Soc.* **2007**, *129*, 5710.

(10) Constantin, T. P.; Silva, G. L.; Robertson, K. L.; Hamilton, T. P.; Fague, K.; Waggoner, A. S.; Armitage, B. A. *Org. Lett.* **2008**, *10*, 1561.

(11) Özhalici-Ünal, H.; Pow, C. L.; Marks, S. A.; Jesper, L. D.; Silva, G. L.; Shank, N. I.; Jones, E. W.; Burnette, J. M.; Berget, P. B.; Armitage, B. A. *J. Am. Chem. Soc.* **2008**, *130*, 1260.

(12) He, X.; Bell, A. F.; Tonge, P. J. *FEBS Lett.* **2003**, *549*, 35.

(13) Hager, B.; Schwarzinger, B.; Falk, H. *Monatsh. Chem.* **2006**, *137*, 163.

(14) Loos, D. C.; Habuchi, S.; Flors, C.; Hotta, J.; Wiedenmann, J.; Nienhaus, G. U.; Hofkens, J. *J. Am. Chem. Soc.* **2006**, *128*, 6270.

(15) Yang, J.-S.; Huang, G.-J.; Liu, Yi.-H.; Peng, S.-M. *Chem. Commun.* **2008**, 1344.

(16) Nienhaus, K.; Nar, H.; Heiker, R.; Wiedenmann, J.; Nienhaus, G. U. *J. Am. Chem. Soc.* **2008**, *130*, 12578.

(17) Dong, J.; Abulwerdi, F.; Baldridge, A.; Kowalik, J.; Solntsev, K. M.; Tolbert, L. M. *J. Am. Chem. Soc.* **2008**, *130*, 14096.

(18) Voityuk, A. A.; Michel-Beyerle, M.-E.; Rösch, N. *Chem. Phys. Lett.* **1997**, *272*, 162.

(19) Weber, W.; Helms, V. J. A.; McCammon; Langhoff, P. W. *Proc. Nat. Acad. Sci. U.S.A.* **1999**, *96*, 6177.

(20) Levine, B. G.; Martinez, T. J. *Annu. Rev. Phys. Chem.* **2007**, *58*, 613.

(21) Schäfer, L. V.; Groenhof, G.; Klingen, A. R.; Ullmann, G. M.; Boggio-Pasqua, M.; Robb, M. A.; Grubmüller, H. *Angew. Chem., Int. Ed.* **2007**, *119*, 536.

(22) Schäfer, L. V.; Groenhof, G.; Boggio-Pasqua, M.; Robb, M. A.; Grubmüller, H. *PLoS Comput. Biol.* **2008**, *4*, e1000034.

(23) Olsen, S.; Smith, S. C. *J. Am. Chem. Soc.* **2008**, *130*, 8677.

(24) Andresen, M.; Wahl, M. C.; Stiel, A. C.; Gräter, F.; Schäfer, L. V.; Trowitzsch, S.; Weber, G.; Eggeling, C.; Grubmüller, H.; Hell, S. W.; Jakobs, S. *Proc. Nat. Acad. Sci. U.S.A.* **2005**, *102*, 13070.

(25) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.

(26) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *77*, 3865.

(27) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(28) Granovsky, A. *PC GAMESS*. http://classic.chem.msu.su/gran/gamess/index.html (accessed April 27, 2009).

(29) Schmidt, M. W.; Baldridge, K. K. J. A.; Boatz, S. T.; Elbert; Gordon, M. S.; J. H.; Jensen, S.; Koseki; Mastunaga, N.; Nguyen, K. A. S.; Su, T. L.; Windus; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.

(30) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neil, D. P.; Distasio, R. A. R. C., Jr.; Lochan, Wang, T.; Beran, G. J. O.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Van Voorhis, T.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Bird, E. F. C.; Daschel, H.; Doerksen, R. J.; Drew, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C.-P.; Kedziora, G. S.; Khalliulin, R. Z.; Klunziger, P.; Lee, A. M.; Liang, W. Z.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L.; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Herhe, W. J.; Schaefer, H. F.; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172.

(31) Ponder, J. W. *TINKER*−*Software Tools for Molecular Design*. http://dasher.wustl.edu/tinker/ (accessed May 7, 2009).

(32) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.

(33) Cossi, M.; Barone, V. *J. Chem. Phys.* **1998**, *109*, 6246.

(34) Barone, V.; Cossi, M. *J. Phys. Chem. A* **1998**, *102*, 1995.

(35) Chipman, D. M. *J. Chem. Phys.* **2006**, *124*, 224111.

(36) Gordon, M. S.; Freitag, M. A.; Bandyopadhyay, P.; Jensen, J. H.; Kairys, V.; Stevens, W. J. *J. Phys. Chem. A* **2001**, *105*, 293.

(37) Adamovic, I.; Freitag, M. A.; Gordon, M. S. *J. Chem. Phys.* **2003**, *118*, 6725.

(38) Nemukhin, A. V.; Grigorenko, B. L.; Topol, I. A.; Burt, S. K. *Phys. Chem. Chem. Phys.* **2004**, *6*, 1031.

(39) Grigorenko, B. L.; Rogov, A. V.; Nemukhin, A. V. *J. Phys. Chem. B* **2006**, *110*, 4407.

(40) Weinhold, F.; Landis, C. R. *Chem. Edu.: Res. Pract. Eur.* **2001**, *2*, 91.

(41) NBO 5.0. Glendening, E. D.; Badenhoop, J. K.; Reed, A. E.; Carpenter, J. E.; Bohmann, J. A.; Morales, C. M.; Weinhold, F. *NBO 5.0*; Theoretical Chemistry Institute, University of Wisconsin: Madison, WI, 2001.

(42) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33.

(43) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. S.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.

(44) Nemukhin, A. V.; Topol, I. A.; Burt, S. K. *J. Chem. Theory Comput.* **2006**, *2*, 292.

CT9001448

JCTC Journal of Chemical Theory and Computation

Catalytic Mechanism of Diaminopimelate Epimerase: A QM/MM Investigation

Marco Stenta,*,† Matteo Calvaresi,† Piero Altoè,† Domenico Spinelli,‡ Marco Garavelli,† Roberta Galeazzi,§ and Andrea Bottoni*,†

Dipartimento di Chimica "G. Ciamician", Università di Bologna, Via Selmi 2, 40126 Bologna, Italy, Dipartimento di Chimica Organica "A. Mangini", Università di Bologna, Via S. Giacomo 11, 40126 Bologna, Italy, and Dipartimento di Scienze e Tecnologie Chimiche, Università Politecnica delle Marche, via Brecce Bianche, 60131, Ancona, Italy

Received January 2, 2009

Abstract: A QM/MM investigation, based on a DFT(B3LYP)//Amber-ff99 potential, has been carried out to elucidate the mechanism of diaminopimelate epimerase. This enzyme catalyzes the reversible stereoconversion of one of the two stereocenters of diaminopimelate and represents a promising target for rational drug design aimed to develop new selective antibacterial therapeutic agents. The QM/MM computations show that the reaction proceeds through a highly asynchronous mechanism where the side-chain of a negatively charged Cys-73 (thiolate) deprotonates the α-carbon substrate. Simultaneously, the Cys-217 thiolic proton moves toward the same carbon atom on the opposite face, thus determining the configuration inversion. A fingerprint analysis provides a detailed description of the influence of the various residues surrounding the active site and clearly shows the electrostatic nature of the most important contributions to the catalysis.

I. Introduction

During the past decade hybrid Quantum Mechanics/Molecular Mechanics (QM/MM)[1−8] methods have been successfully used to investigate large molecular systems. The study of enzymatic reactivity[8] certainly represents a field where QM/MM methods have been most widely applied. These hybrid methods are particularly suitable to deal with this class of problems since the enzyme can be easily and almost "naturally" partitioned into two regions: one (described at the QM level) approximately corresponding to the active site and the other (described at the MM level) that includes the remaining part of the enzyme and, in some cases, the solvent molecules.

In the present paper we use a QM/MM approach to provide a complete and exhaustive analysis of an interesting enzymatic system (the diaminopimelate epimerase), and we present a general strategy for the study of similar problems involving enzymatic systems. In particular, we suggest general criteria to build a reliable model-system (using various dynamics techniques), and we show how, after having computed the potential energy surface (PES), we can use various tools to analyze the results and identify the key-factors that control the catalytic mechanism.

In section II we introduce the basic principles of the computational techniques employed in our study. These techniques concern the features of our QM/MM[1−8] approach and the type of analysis used to rationalize in details the enzyme catalytic effects. Our QM/MM code has been developed according to the hybrid approach described in a previous paper[12] and is included in the **COBRAMM**[9] suite of programs together with the analysis tools described in the following sections.

The catalytic mechanism of diaminopimelate (DAP) epimerase (E.C. 5.1.1.7)[10−12] is the case study discussed in

* Corresponding author phone: ++39-051-2099477; fax: ++39-051-2099456; e-mail: andrea.bottoni@unibo.it.

† Dipartimento di Chimica "G. Ciamician", Università di Bologna.

‡ Dipartimento di Chimica Organica "A. Mangini", Università di Bologna.

§ Università Politecnica delle Marche.

10.1021/ct900004x CCC: $40.75
Published on Web 05/27/2009

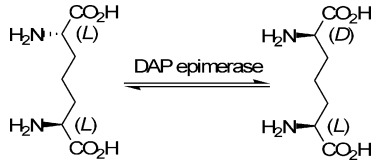


Figure 1. The reaction catalyzed by the enzyme diaminopimelate epimerase.

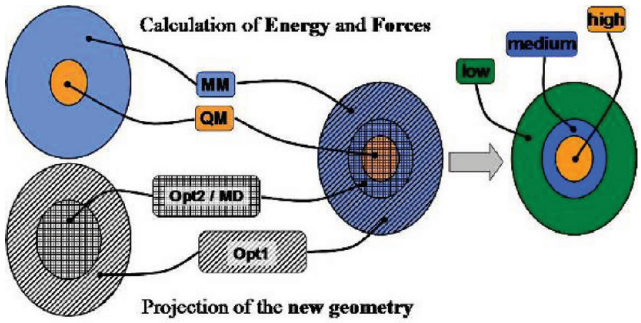


Figure 2. A schematic representation of the partition scheme of the whole system. The two main regions (**MM** and **QM**) are subdivided into three layers (**H**, **M**, and **L**) when computing energy and forces. "Microiterations" (**Opt1**) and "macroiterations" (**Opt2**) refer to the optimization process applied to the **L** and **H**+**M** regions, respectively.

section III. This enzyme belongs to the group of lyases which are capable of inverting the absolute configuration of a carbon atom in substrates containing one (racemases) or more stereocenters (epimerases).[13] While many enzymes belonging to this group (for instance alanine,[14,15] serine, threonine racemases) need the presence of a molecule of pyridoxal-5′-phosphate (PLP)[16] behaving as a cofactor, diaminopimelate epimerase,[10,11] and a few other members of the family, like glutamate[17−19] and proline[20−22] racemases, can exert their catalytic action without the participation of additional molecules. The enzymes belonging to this class of racemases and epimerases (i.e., those not requiring a cofactor), despite a substantial difference in the active site structure, which is due to the specificity for different substrates, share a pair of catalytically important cysteine residues. In particular, diaminopimelate epimerase, that catalyzes the epimerization of L,L- to D,L-*meso*-diaminopimelate (see Figure 1), without assistance of PLP, operates via a "two base" mechanism involving one active-site cysteine thiolate (Cys-73) acting as a base that deprotonates the α-carbon of the substrate and a second cysteine thiol group (Cys-217) that protonates the opposite side *via* a general-acid catalysis.

To stress the importance of understanding the catalytic mechanism of DAP epimerase and its potential effects on pharmacological research, a short discussion on the involvement of this enzyme in some important metabolic pathways can be helpful. It is well-known that bacteria, plants, and fungi metabolize aspartic acid to produce four different amino acids (lysine, threonine, methionine, and isoleucine) through a sequence of reactions known as the "aspartate pathway".[23] These reactions produce several important metabolic intermediates such as diaminopimelic acid, an essential component of the bacterial cell wall biosynthesis.[24] Members of the animal kingdom do not possess this pathway and must therefore acquire these essential amino acids through their diet. Since the enzymes[23] involved in this pathway are not present in animals, inhibitors of them are promising targets for the development of novel antibiotics, herbicides, and fungicides. The recent emergence of bacterial resistance to currently available antibiotics has determined a renewed interest in the search for novel antibacterial compounds. Since DAP epimerase is the key enzyme of the diaminopimelic acid/lysine branch of the aspartate pathway, it represents an optimal target for developing new selective drugs capable of blocking the synthesis of Gram-positive bacterial cell walls (by interrupting the lysine synthesis) or able to interfere with the building of the peptidoglycan layer of Gram-negative and mycobacterial cell walls (by stopping the D,L-diaminopimelate[25] synthesis) without interfering with host cell metabolism.

Thus, detailed information on this enzymatic mechanism (concerted or stepwise process, role of the various residues in the vicinity of the active site dyad (Cys-73 and Cys-217), electrostatic interactions controlling the substrate selectivity) is of primary importance and opens the way to important applications in drug design. In principle, a new class of antibiotics could originate from the discovery of specific inhibitors of the diaminopimelate epimerase.

In the present paper we have chosen to investigate the *Haemophilus influenzae* DAP epimerase because, for this enzyme, kinetics[10] and crystallographic[11] studies are available in literature. The kinetic[10] and the structural X-ray data[11] have been very helpful in detecting the two catalytic cysteine residues (Cys-73 and Cys-217), which are responsible for the observed pH dependency of the reaction velocity (V/K profiles in ref 10). These are characterized by p*K* values of 6.7 and 8.5 and must be unprotonated and protonated, respectively, to allow the L,L- to D,L-*meso*-diaminopimelate conversion (forward reaction). The protonation state must be inverted for the reverse reaction (D,L-*meso*- to L,L-diaminopimelate). The primary deuterium isotope effects suggest that substrate epimerization together with the double proton transfer strongly affect the reaction rate (rate-determining step). However, the experiments indicate that another slow step, which follows the product dissociation, could be partially rate determining. This step, which corresponds to the back-proton transfer involving, after product release, the two catalytic Cys residues, restores the original protonation state of the active site for subsequent turnovers. On the basis of this experimental evidence the experimental k_{cat} value (128 s^{-1}) for the forward (L,L → D,L) reaction and the corresponding barrier of 15.6 kcal mol^{-1} (obtained from k_{cat} by applying the Eyring equivalence[26]) can be associated, in principle, either to the substrate epimerization step or the back-proton transfer occurring in the substrate-free enzyme. Thus, this experimental value should be considered as an upper limit for the activation energy of the epimerization step.

II. Computational Details

II.A. Setting-up the System. To build a reliable model-system of the diaminopimelate (DAP) epimerase, we used the recently obtained crystal structures[11] of this enzyme (from

Haemophilus influenzae) binding two isomers of the irreversible inhibitor aziridino-DAP,[27] which mimics the natural substrate [PDB[28] codes: 2GKE(L,L-AziDAP), resolution 1.35 Å and 2GKJ (D,L-AziDAP), resolution 1.70 Å]. The methylene carbon of the aziridine ring of the two diastereomeric inhibitors is covalently bonded to the sulfur atom of Cys-73 or Cys-217 after the nucleophilic attack of the sulfur on the aziridine ring that irreversibly inhibits the enzyme. The DAP epimerase backbones obtained from the two crystal structures show negligible differences.[11] The structural features of the covalent bond between the cysteine sulfur and the inhibitor (L,L-AziDAP or D,L-AziDAP) provides information to model the mechanism of approach of the thiol base during the α-deprotonation/protonation process of the DAP substrate. We decided to use the coordinates from the 2GKE PDB file to build the model-system since 2GKE has a better resolution than 2GKJ. We modified the L,L-AziDAP into the natural substrate L,L-DAP, and we retained only the A-conformer where multiple conformations of the amino acids were available in the crystal structure. The model-system obtained from these crystallographic data was protonated with the **H**++[29] software, using the default parameters available in that package. This code employs an automatic algorithm that computes pK_a values for the various ionizable groups in macromolecules and adds missing hydrogen atoms according to the specified pH value of the environment. The positions of the added hydrogen atoms are also optimized by this algorithm. The protonation state of all titratable residues was carefully checked by visual inspection of the **H**++ output structure. It followed that all Asp and Glu residues were unprotonated (negatively charged), while all Lys and Arg residues were protonated (positively charged). This result is consistent with the solvent exposure of the side chain of all these titratable residues, with two important exceptions: Glu-208 and Arg-209 which, however, interact with the zwitterionic substrate by their charged side chains. Furthermore, all His residues were found to be neutral with the proton on the ϵ nitrogen atom. The only exception was the double protonated (thus, positively charged) His-50 whose side chain is exposed to the solvent but is also interacting with the hydroxyl group of the near Tyr-98 residue through the δ hydrogen of the side chain imidazole ring. Since the accurate computation of the pK values of the active site residue was beyond the purposes of our research - and adequate experimental data were available - we did not calculate the pK value of the catalytic Cys residues. The protonation state of these two residues was decided on the basis of the kinetics studies,[10] (suggesting the presence of one unprotonated Cys side chain, as outlined in the Introduction) and by inspection of the crystallographic structure of the enzyme bound to a substrate mimic,[27] indicating that the thiolate moiety should correspond to Cys-73 to allow the reaction to take place in the forward direction (L,L → D,L). Identical results were obtained using the Propka2.0 software.[30]

The L,L-DAP and D,L-*meso*-DAP molecules were parametrized using the Generalized Amber Force Field (GAFF).[31] Partial atomic charges were assigned to atoms using the AM1-BCC method[32,33] as implemented in the *antechamber* module of the AMBER8.0 package.[34]

The initial model-system geometry was fully minimized at the MM level using the *sander* module of AMBER8.0. The minimization was carried out with the Amber Force Field (Amber-*ff99*)[35] until the root-mean-square deviation (rmsd) of the Cartesian elements of the gradient was less than 0.0001 kcal mol^{-1}. A full conjugate gradient minimization approach and the General Born (GB) model[36] to simulate the aqueous environment (as implemented in the *sander* module of the AMBER8.0 code) were used.

Finally we checked if the crystal structures of 2GKE and 2GKJ mimic precisely the natural binding mode of the substrates or if the covalent bond between the enzyme and the L,L-AziDAP and D,L-AziDAP inhibitors can perturb the natural binding mode of L,L-DAP and D,L-*meso*-DAP. To this purpose we carried out a conformational study of the binding mode of L,L-DAP and D,L-*meso*-DAP within the protein environment using different methods i.e. Cluster Analysis, Simulated Annealing and Docking.

Cluster Analysis. We carried out high temperature Molecular Dynamics starting from the optimum structure obtained for the complex. A region of 5 Å around the substrate was free to move during the MD simulation. The system was heated from 0 to 800 K in 100 ps, and then a trajectory of 2 ns was computed at constant temperature (800 K). The integration step of 2 fs was used in conjunction with the SHAKE algorithm[37] to constrain the stretching of bonds involving hydrogen atoms. The coordinates of the system were saved on a trajectory file every 2 ps, giving a total of 1000 structures. Solvation effects were taken into account using the GB model with a dielectric constant of 78.5. To determine the most populated conformations of L,L-DAP and D,L-*meso*-DAP within the protein binding pocket, we performed a Cluster Analysis on the different conformations visited by the two molecules during the simulation. To this purpose we used the MMTSB toolset,[38] and we clustered different conformations of the substrates on the basis of structural similarity; we carried out our analysis employing the *kclust* module with a fixed radius of 1.0 Å on the Cartesian coordinate rmsd computed for heavy atoms. Then, we determined the centroid of each cluster. For each cluster we chose the structure closest to the corresponding centroid as representative of the cluster itself (this structure is characterized by the smallest rmsd value with respect to the centroid).

Simulated Annealing. We performed 10 cycles of simulated annealing for the two complexes formed by the protein and the substrate molecules L,L-DAP and D,L-*meso*-DAP. We used the same parameters of the previous simulation, and we heated the system from 0 to 1000 K in 30 ps, holding at 1000 K for equilibration for 10 ps. Then, we cooled from 1000 to 0 K in 60 ps. The heat bath coupling for the system was tight for heating and equilibration (0.1 ps). The cooling phase was divided in three periods: during the first 48 ps the cooling was very slow (coupling of 5.0 ps); this was followed by a cooling phase of 6 ps (with coupling of 1.0 ps) and a last cooling phase of 6 ps (with a coupling changing from 0.1 to 0.05 ps). At the end of the simulated annealing

a complete minimization was carried out, and the final coordinates of the complexes were retained.

Docking. The previous model obtained from the PDB structure 2GKE was used in the docking calculations after having deleted the ligand from the cavity site. The orientation sampling of the substrate into the cavity was carried out using spheres calculated by the *sphgen* module of DOCK6 [39,40] within 10.0 Å of root-mean-square deviation (rmsd) from every atom of the minimized structure of the ligand. Partial atomic charges for the substrates were obtained with the AM1-BCC method. We carried out flexible ligand docking using the Anchor-and-Grow algorithm implemented in DOCK6 with the Grid-Based Score function as primary scoring function. The results were then rescored with the new Amber score as secondary scoring function allowing a minimization of the ligand and residues within 5 Å (the same mobile residues of the previous calculations) for 100 steps. The best 10 poses obtained in the ranking of L,L-DAP and D,L-*meso*-DAP were considered for subsequent QM/MM calculations.

II.B. QM/MM Details. The QM/MM potential[9] is based on a subtractive scheme[2,9,41,42] (see Figure 1). The boundary zone between the QM and MM regions is handled by means of a hydrogen atom link approach,[2] and a charge shifting scheme within the electrostatic embedding method[2] is adopted to avoid hyper-polarization of the QM wave function. A special and particularly important feature of our QM/MM approach[9] is the partition of the system into three layers. The innermost layer called "high" (**H**) is treated at the QM level, while the outermost one, named "low" (**L**), is treated at the MM level. The presence of an intermediate layer, denoted as "medium" (**M**), improves the efficiency of the geometry optimization procedure. It has been shown that the decoupling[43,44] of the QM an MM regions during the optimization process can significantly improve the geometry convergence (faster optimization) and the accuracy of the results.[43] This approach allows a full relaxation of the MM region ("microiteration" phase[44,45] carried out with a cheap and fast optimization algorithm, like "steepest descent", indicated as **Opt1** in Figure 1) at each optimization cycle of the QM region ("macroiteration", based on an accurate algorithm, like BFGS[46−49] and denoted as **Opt2**). It is valuable to notice that, in the "macroiteration" step, the new geometry projection task can be, alternatively, carried out by a molecular dynamics code (MD), to obtain a molecular dynamics simulation on the **HM** region (**H** layer + **M** layer). During "microiterations" the **HM** region is kept frozen, and its electrostatic potential is taken into account by means of atomic point charges coming from the MM force field for **M** atoms or from the QM calculations for **H** atoms (CHELPG[50] charges have been used in this work). This approach, as pointed out by many authors and demonstrated in a previous paper,[22] gives good results in terms of the simulation cost/efficiency ratio. The obtained results are also in good agreement with more expensive methods. A possible pitfall of this approach is the transition state search procedure when it involves the simultaneous rearrangement of MM and QM atoms. This can be solved by expanding the QM subregion, but this causes, of course, an increase of the computational cost. The introduction of the intermediate (or "buffer") layer (**M**) between the QM (**H**) and MM (**L**) subregions partially overcomes this problem, because the **M** layer is treated at the MM level but is optimized together with the **H** region. This strategy allows a detailed description of large molecular motions involving several tens of atoms with a minor increase in the computational cost.

The QM/MM potential adopted throughout this work is based on DFT[51,52]/B3LYP[53] calculations using the double-ζ DZVP[54] basis set for all atoms of the QM region, while the Amber-*ff99*[35] force field has been employed for the MM atoms (GAFF[31] parameters have been adopted for the DAP substrate). In the following discussion this potential will be referred to as DFT(B3LYP/DZVP)/Amber-*ff99* potential.

The nature of the critical points on the PES can be determined by means of QM/MM numerical frequency calculations on the whole enzyme. In these computations we change only the geometry of the **HM** region in the presence of the MM potential determined by the frozen **L** region. Thus, a complete numerical frequency run (denoted as "**fullfreq**" calculation) would require a total of $1 + 6N^{HM}$ QM/MM energy evaluations, N^{HM} being the number of atoms of the **HM** region. A different and stronger level of approximation for frequency computations (simply denoted as "**freq**" calculation) has been tested. Within this approximation we hypothesize that the small motions of an MM atom have only a tiny effect on the wave function. Under this assumption it becomes possible to save $6N^{M}$ QM computation (N^{M} being the number of M atoms) by summing the current MM energy value to the reference (initial) QM energy when an MM atom is moved. Then, a new QM computation is carried out only when an atom of the **H** region is moved. In this way only $1 + 6N^{H}$ wave function evaluations are required. Both these approximations have been tested, and the results obtained at the two levels show a good qualitative agreement.

All the QM/MM computations described in this paper have been carried out using the general-purpose package **COBRAMM**,[9] which interfaces many commercially available QM and MM codes as well as some analysis routines. For the present work we used the GAUSSIAN03[55] (C02 version) and AMBER8.0[34] packages to perform QM and MM calculations, respectively. In the geometry optimization (to locate minima and saddle-points) we applied for the **HM** region ("macroiteration") the BFGS optimization algorithm implemented in the Gaussian code.[45,56,57] For the **L** region ("microiteration") we used the "steepest descent" method from the *sander* tool of AMBER8.0.

II.C. Fingerprint Analysis. In our QM/MM scheme all contributions to energy within the **H** region are computed by means of single point computations on a molecular system (denoted as *model-H*) formed by the **H** region where properly placed hydrogen atoms saturate the dangling bonds at the QM-MM boundary (according to the hydrogen atom link scheme).[2,22] Bonding and nonbonding terms of the **M** and **L** regions are computed at the MM level. A special caution is required to take into account the QM-MM cross terms, and many recipes have been proposed in literature to handle this problem. We chose a general and rather popular approach

that describes all cross terms (i.e., van der Waals, bonding, bending, torsions), except the electrostatic ones, at the MM level. We adopted the electrostatic embedding scheme (EES) to describe the electrostatic contributions.[22] This consists in the computation of the QM wave function in the presence of the atomic point charges of the **M** and **L** layers. We assume that the polarization of the wave function determined by this EES computations accounts for the electrostatic cross-term interactions. Under this assumption it becomes easy to derive a procedure that splits the cross-terms energy contributions into single-residue contributions.

In this section we provide a short description of two general procedures, named Direct and Reverse Finger Print analysis (**DFP** and **RFP**, respectively), that allow to rank the electrostatic effects of the single residues and a third procedure (**vdWFP**) able to evaluate the van der Waals contributions. These analyses can provide semiquantitative information about the role of a given residue (or group of residues) in determining the relative stabilization/destabilization of two critical points. If, for instance, a transition state is compared to the nearest minimum, then we can rank the effects of each residue on the entity of the barrier for the corresponding process. It is worth remembering that, since the original publication by Karplus,[58] the **DFP** approach has been used by several others authors[59−66] (in many cases the term "charge perturbation method" has been used), and the method has been demonstrated to be a valuable approach to obtain precious information and new insight into enzyme catalysis.

To illustrate the specific features of our analysis, we consider two critical points A and B located on the QM/MM Potential Energy Surface. The overall electrostatic contribution can be easily computed as follows. QM calculations *in vacuo* on *model-H* (i.e., the QM region after hydrogen addition) provide for the two points the corresponding energy values E_0^A and E_0^B. QM calculations in the presence of all atomic point charges give the two energy values E_t^A and E_t^B. From these values, after subtraction of the corresponding charge self-energies (e_t^A and e_t^B), we obtain (eqs 1 and 2) the two quantities E_{QM}^A and E_{QM}^B (the charge self-energy corresponds to the coulomb term describing the interaction between point charges). E_{QM}^A and E_{QM}^Brepresent, for A and B, respectively, the sum of pure QM and electrostatic cross terms, as inserted in our QM/MM potential.

$$E_{QM}^A = E_t^A - e_t^A \tag{1}$$

$$E_{QM}^B = E_t^B - e_t^B \tag{2}$$

The net electrostatic effects of the MM regions on the QM wave function $E_{pol}^A$and E_{pol}^B can be estimated by means of eqs 3 and 4.

$$E_{pol}^A = E_{QM}^A - E_0^A \tag{3}$$

$$E_{pol}^B = E_{QM}^B - E_0^B \tag{4}$$

It is important to notice that the more negative these values are, the greater is the charge stabilization effect (more precisely, values smaller or greater than zero indicate that the charge polarization contribution is stabilizing or destabilizing, respectively).

A first comparison between A and B can be carried out using the terms computed with eqs 1−4. In particular, $\Delta E_0(A,B)$ and $\Delta E_{QM}(A,B)$ (eqs 5 and 6) represent the QM energy difference between A and B, in the absence and in the presence of the MM atomic point charges, respectively. Equation 7 outlines the connection between the differential stabilization effect of charges on A and B. The stability factor $S_{tot}(A,B)$ computed by eq 7 represents the magnitude of the point charge effect in promoting or discouraging the passage from A to B.

$$\Delta E_0(A,B) = E_0^B - E_0^A \tag{5}$$

$$\Delta E_{QM}(A,B) = E_{QM}^B - E_{QM}^A \tag{6}$$

$$S_{tot}(A,B) = \Delta E_{QM}(A,B) - \Delta E_0(A,B) = E_{QM}^B - E_{QM}^A - E_0^B + E_0^A = E_{pol}^B - E_{pol}^A \tag{7}$$

This analysis can provide important information on the influence of the protein environment (described at the MM level) on the rate of the reaction occurring within the small QM region. An estimate, even if qualitative, of the contribution coming from each single residue can make this approach particularly useful. If A and B are, for instance, a stable species (minimum) and the near transition state (saddle point), it becomes possible to detect the residues that most significantly affect the barrier height, thus playing the main catalytic effect.

We adopted two different decomposition schemes to rank the influence of the various enzyme residues on the rate of reaction A → B.

The first scheme, denoted here as Direct Finger Print (**DFP**) analysis, requires a series of single point QM calculations (SPc) on the QM region (i.e., *model-H*) for both structures A and B (using the optimized QM/MM geometry).

After the evaluation of all terms needed to compute $S_{tot}(A,B)$ (eqs 1 to 7), we perform N SPc's (where N is the total number of residues to analyze). In each calculation *model-H* is surrounded by the atomic point charges of the **i**$^{\text{th}}$ residue only (see Figure S1a in the Supporting Information), with the charges placed according to the atomic coordinates of the **i**$^{\text{th}}$ residue itself. This procedure provides NE_i^A and NE_i^B energy values together with the corresponding charge self-energy values e_i^A and e_i^B (this term represents the pure electrostatic contribution among the point charges of the **i**$^{\text{th}}$ residue only). We can easily compute the electrostatic (Coulomb) effect of the **i**$^{\text{th}}$ residue on the QM region ($E_{pol,i}^A$ and $E_{pol,i}^B$ from eqs 8 and 9) using eqs 3 and 4 (proposed for the total electrostatic effect). This effect is stabilizing or destabilizing if the corresponding value is lesser or greater than zero, respectively.

$$E_{pol,i}^A = (E_i^A - e_i^A) - E_0^A \tag{8}$$

$$E_{pol,i}^B = (E_i^B - e_i^B) - E_0^B \tag{9}$$

We obtain the stability parameter S_i (eq 10) for the **i**$^{\text{th}}$ residue by comparing the values of $E_{pol,i}^A$ and $E_{pol,i}^B$. If $S_i < 0$,

then the **i**th residue favors the transition from A to B. On the contrary, if $S_i > 0$ the **i**th residue slows down the process.

$$S_i = E^B_{pol,i} - E^A_{pol,i} \tag{10}$$

A further concern is important to establish the accuracy of our method in decomposing the total electrostatic effect into smaller components. An ideal decomposition scheme should provide an exact equivalence between the total electrostatic effects E^A_{pol} and E^B_{pol} and the terms $E^A_{polsumm}$ and $E^B_{polsumm}$ obtained by summing up all single contributions E^A_{pol} and E^B_{pol} values) as stated in eqs 11 and 12.

$$E^A_{polsumm} = \sum_{i=1}^{N} E^A_{pol,i} \tag{11}$$

$$E^B_{polsumm} = \sum_{i=1}^{N} E^B_{pol,i} \tag{12}$$

It is evident that, under the adopted approximations, the terms $E^A_{polsumm}$ and $E^B_{polsumm}$ are not perfectly equivalent to E^A_{pol} and E^B_{pol}. To estimate the error we introduce the following three error parameters.

$$X^A_{add,S} = E^A_{pol} - \sum_{i=1}^{N} E^A_{pol,i} \tag{13}$$

$$X^B_{add,S} = E^B_{pol} - \sum_{i=1}^{N} E^B_{pol,i} \tag{14}$$

$$X_{add,S}(A,B) = X^B_{add,S} - X^B_{add,S} \tag{15}$$

If the errors made on the two critical points A and B ($X_{add,S}{}^A$ and $X_{add,S}{}^B$) differ significantly, $X_{add,S}(A,B)$ is an important evaluation of the reliability of the adopted decomposition scheme.

An immediate outlook of the whole enzyme influence on the A → B transformation can be obtained by plotting each S_i value against the corresponding residue number. The relative magnitude of the stability parameter S_i can be used to rank the residues according to their importance in the catalytic process. The results of the DFP analysis can be affected by the basic approximations of our approach: the S_i factor is estimated by comparing the unperturbed *in vacuo* system (to obtain $E^A_{pol}$and E^B_{pol}) and the system perturbed by a single-residue (to obtain $E^A_{pol,i}$and $E^B_{pol,i}$). However, in principle, the latter situation can be rather different, in terms of wave function polarization, with respect to the QM region fully embedded into the MM point charge cloud (according to the electrostatic embedding scheme).

An opposite scheme, referred to here as Reverse Finger Print (**RFP**) analysis, can be used to improve the previous description. We compute again the E^A_{QM}, E^B_{QM}, E^A_0, and E^B_0 terms by means of single point calculations (SPc) on the *model-H* system in the presence and absence of the whole set of atomic point charges. We define from these calculations the two terms E^A_{dest} and E^B_{dest} (eqs 16 and 17).

$$E^A_{dest} = -E^A_{pol} = E^A_0 - E^A_{QM} \tag{16}$$

$$E^B_{dest} = -E^B_{pol} = E^B_0 - E^B_{QM} \tag{17}$$

We perform N SP computations (N is the total number of residues to analyze) where the system *model-H* is surrounded by all atomic point charges except those corresponding to the **i**th residue (see Figure S1b in the Supporting Information). Charges are placed, as previously described for DFP, according to the QM/MM optimized geometry. This procedure provides N $E^A_{h,i}$ and N $E^B_{h,i}$ energy values and the corresponding charge self-energies $e^A_{h,i}$ and $e^B_{h,i}$, which represent the pure electrostatic contributions of all point charges except those corresponding to the "hole" of the missed **i**th residue. We can now compute the electrostatic effects $E^A_{dest,i}$ and $E^B_{dest,i}$on the QM region, which are due to the absence of the **i**th residue (eqs 18 and 19). These terms represent a destabilization or stabilization if they are lesser or greater than zero, respectively.

$$E^A_{dest,i} = (E^A_{h,i} - e^A_i) - E^A_{QM} \tag{18}$$

$$E^B_{dest,i} = (E^B_{h,i} - e^B_i) - E^B_{QM} \tag{19}$$

A destabilization parameter D_i (eq 20) can be obtained for each residue **i**th. $D_i < 0$ or $D_i > 0$ indicate that the **i**th residue has the effect of reducing or enhancing the rate of the A → B transformation, respectively.

$$D_i = E^B_{dest,i} - E^A_{dest,i} \tag{20}$$

Again, as previously outlined for the **DFP** analysis, the summation over the single contributions E^A_{dest} and E^B_{dest} does not perfectly correspond to the total electrostatic effects $E^A_{dest,i}$ and $E^B_{dest,i}$. In other words eqs 21 and 22 do not hold.

$$E^A_{destsumm} = \sum_{i=1}^{N} E^A_{dest,i} \tag{21}$$

$$E^B_{destsumm} = \sum_{i=1}^{N} E^B_{dest,i} \tag{22}$$

The error can again be estimated by the following three error parameters (see eqs 23, 24, and 25).

$$X^A_{add,D} = E^A_{dest} - \sum_{i=1}^{N} E^A_{dest,i} \tag{23}$$

$$X^B_{add,D} = E^B_{dest} - \sum_{i=1}^{N} E^B_{dest,i} \tag{24}$$

$$X_{add,D}(A,B) = X^B_{add,D} - X^A_{add,D} \tag{25}$$

Thus, $X_{add,D}(A,B)$ is a good estimate of the reliability of the adopted decomposition procedure.

Plotting the D_i values as a function of a residue index provides information which is similar to that obtained from the S_i diagram and allows a ranking of the importance of each residue in favoring/disfavoring the A → B transformation. The two diagrams are only apparently different, since a residue which favors the process has a negative *S* factor but a positive *D* factor.

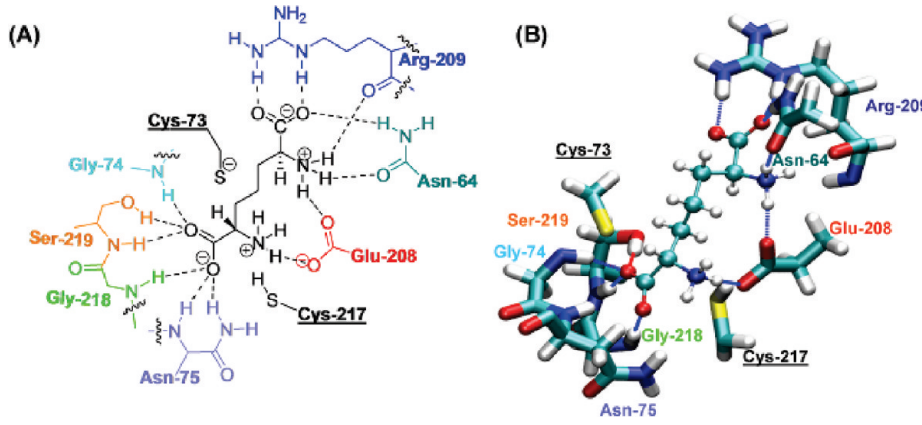


Figure 3. Two-dimensional (A) and three-dimensional representation of L,L-DAP within the active site. The most important residues involved in the H-bond network are shown (Asn-11 and Gln-44 are omitted for clarity).

The estimate of the van der Waals contributions (**vdWFP** analysis) is an additional useful tool to analyze the enzyme catalytic effect. This analysis is rather straightforward, being these contributions included in the QM/MM potential at the MM level. Moreover the single-residue contributions are additive for the MM force-field definition. We use a procedure similar to **DFP** and **RFP** analysis to obtain information on the role played by van der Waals interactions of each residue on the A → B transformation. We compute separately the van der Waals interaction energy between each residue and the **H** layer (note that the system here is not *model-H*). We obtain N (the total number of residues) energy values ($E^A_{vdW,i}$ and $E^B_{vdW,i}$) for both A and B. The lower $E^B_{vdW,i}$ is (relative to $E^A_{vdW,i}$), the greater is the stabilization for the critical point under examination. It is easy to compare the values obtained for the two structures and derive a stabilization factor W_i (eq 26). In this case the contribution of each component is perfectly additive (this follows from the definition of the adopted force field).

$$W_i = E^B_{vdW,i} - E^A_{vdW,i} \tag{26}$$

$$E^B_{vdW,i} > E^A_{vdW,i} \Rightarrow W_i > 0 \tag{27}$$

$$E^B_{vdW,i} < E^A_{vdW,i} \Rightarrow W_i < 0 \tag{28}$$

The **vdWFP** analysis has been performed by means of several calculations carried out with the *anal* module from the AMBER8.0 package. The **vdWFP** results can be easily represented using a plot of W_i versus i, similar to the diagrams obtained for **DFP** and **RFP**.

The three methods of analysis previously described have been implemented in the **COBRAMM** package. Useful discussions on this type of analysis can be found in previous works by Karplus and co-workers.[58−63,65] These papers have been very helpful and informative to develop our **DFP** and **RFP** approaches. These computational tools have been used here to obtain a per-residue analysis, but they can be easily exploited for a generic A → B transformation to investigate the effects arising from different groups of atoms or single atoms. Thus, in principle, we can easily provide a per-atom analysis, or alternatively, we can perform our analysis to understand the role played in the catalysis by secondary structures (α-helices or β-sheets or combinations of them) present in the enzyme under examination.

III. Results and Discussion

III.A. Structure of the Initial Complex. The conformational studies of the binding mode of L,L-DAP and D,L-*meso*-DAP using the three different methods described in section II.A point to enzyme−substrate complexes which are very similar in structure. The geometry of these complexes is also very close to the crystallographic structures of 2GKE and 2GKJ, that strictly mimic the natural binding mode of the substrates but are characterized by a covalent bond between the enzyme and L,L-AziDAP and D,L-AziDAP. In the following discussion we refer to the structure obtained in the full MM optimization of the final point provided by the simulated annealing procedure (no cutoff concerning long-range interactions was used in this procedure). This geometry has been used to construct the starting point of the QM/MM study.

The structural features of these complexes show why DAP epimerase can bind only DAP isomers characterized by configuration L at the distal ε-carbon. The binding pocket has an asymmetric arrangement of residues that can form (as a donor or acceptor) hydrogen bonds strictly suitable to bind the L isomer at the distal site. The carboxyl group forms a salt bridge with the positively charged side chain of Arg-209 and three H-bonds with the side chains of Asn-64 and Asn-190. At the same time the positively charged amino group is hydrogen bonded to the side chains of Asn-64 and Glu-208 and the carbonyl oxygen of Arg-209. These structures show that the substrates enter the active site as zwitterion.

When the L,L-DAP and D,L-*meso*-DAP are bound to the enzyme the α-carboxyl group is bonded with the amidic hydrogen of Gly-74, Asn-75, Gly-218, and Ser-219 and also with the side chain of Ser-219. The charged amino group forms hydrogen bonds with the side chains of the Asn-11, Gln-44, and Glu-208 residues. These interactions are schematically represented in Figure 3.

III.B. Model1: the Simplest Model System. The construction of the model system is a crucial point in QM/MM computations if we wish to obtain the most convenient cost/

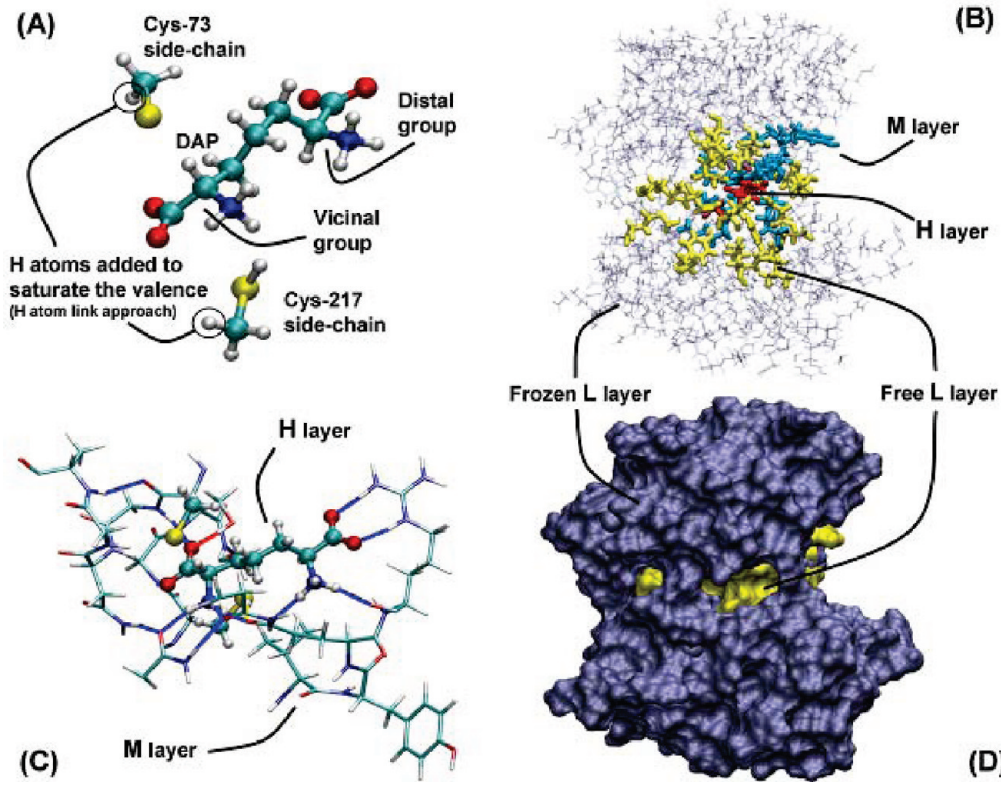


Figure 4. Partitioning scheme adopted for **Model1**. (A) QM subsystem. (B) **H** layer (red), **M** layer (blue), free **L** layer (yellow), frozen **L** layer (dark blue). (C) Hydrogen-bond network in the active site (**H** layer: ball and stick, **M** layer: stick). (D) Solvent exposed surface: free **L** layer (yellow), frozen **L** layer (dark blue).

efficiency ratio for the adopted hybrid potential. Thus, the atom selection to define the various layers (**H**, **M**, and **L** regions) is of primary importance. In the present case we have reduced the **H** layer to the smallest possible set of atoms (see Figure 4A). Since only two cysteine residues are directly involved in the enzyme-catalyzed stereoinversion process, we have included in the **H** layer the side chains of Cys-73 and Cys-217, after saturation of the dangling bonds with hydrogen atoms (atom link approach) and the entire diaminopimelate substrate. Some additional residues (see Table S1 in the Supporting Information and Figure 4B,C) surround the reacting core and are hypothesized to have an important effect on the catalytic process. These residues have been included in the **M** layer to improve the description of the system without increasing the computational demand. All remaining residues form the **L** region. No solvent effects (either explicitly or implicitly) have been taken into account in the PES computation. This can be considered a satisfactory approximation because the active site is a deep pocket far beneath the enzyme surface exposed to solvent. Moreover, a careful comparison of the crystallographic structures of L,L-AziDAP and D,L-AziDAP (the covalent complexes between enzyme and reactant-like and product-like inhibitor, respectively) shows that the substrate → product conversion does not cause important changes in the enzyme structure outside the active site. For this reason, to avoid unrealistic deformations of the structure, due to the lack of solvent, a few residues of the **L** layer in the vicinity of the **M** border were free to move during the "microiteration" steps of geometry optimization, while all other residues belonging to **L** have been kept "frozen" at the initial positions. The "free" residues are not exposed (or exposed only to a negligible extent) to the external enzyme surface (see Figure 4D). In the following discussion we will refer to this model system as **Model1**.

The investigation of the Potential Energy Surface (PES) for **Model1** has demonstrated the existence of two critical points **M1** and **M2** that describe the enzyme bound to the reactant (L,L-DAP) and product (D,L-*meso*-DAP) molecule, respectively. An opposite protonation state of the catalytic cysteine dyad features **M1** and **M2**. In particular, in **M1** Cys-217 is protonated and Cys-73 is unprotonated and also oriented in such a way to easily grab a proton from substrate. In the hypothesized reaction mechanism the side-chain of the negatively charged Cys-73 captures a proton from the carbon substrate, while the Cys-217 thiolic proton moves toward the same carbon atom on the opposite face. It is not evident from the experimental results if the stereoinversion of the carbon atom is a concerted or stepwise process. In the second case the PES should be characterized by an intermediate species between two transition states. Any attempt to locate the intermediate of the hypothetical stepwise process has failed, and we have located only one Transition State (**TSC**) where the two cysteine residues are almost completely protonated, being the Sγ-H distance 1.43 and 1.36 Å for Cys-73 and Cys-217, respectively. This finding accounts for a concerted but highly asynchronous process. In **TSC** the substrate is deprotonated and planar. This structure allows a delocalization of the partial negative charge over the extended π orbital system. A measure of charge

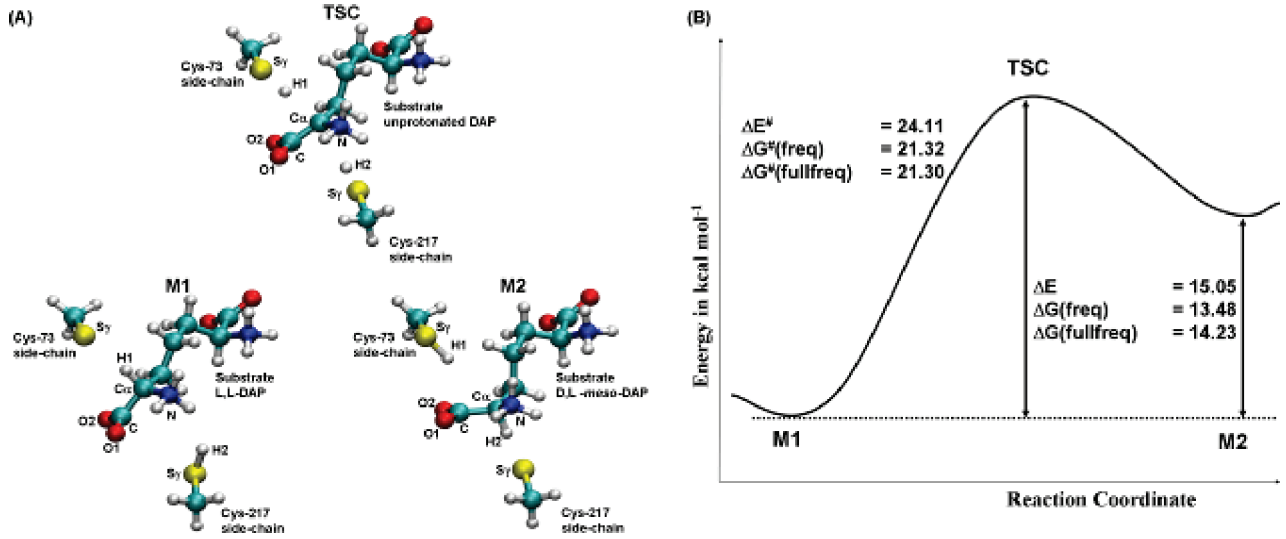


Figure 5. A) Schematic representation of the *model-H* subsystem for the three critical points **M1**, **TSC**, and **M2**. B) Reaction profile (E and G denote total and Gibbs free energy, respectively).

delocalization can be obtained from the comparative analysis of some relevant atomic distances and atomic point charges (see Table S2 in the Supporting Information, where CHELPG charges are reported). It is evident that the negative charge is mainly localized on the two oxygen atoms of the carboxyl group. Also, the carboxylic and -NH_3 groups become more negative and less positive, respectively, on passing from the minima to the transition state. On the contrary, the deprotonated α carbon atom does not show a significant charge variation. The charge delocalization is also proved by a bond order decrease of both carboxylic C−O bonds and a simultaneous bond order increase of the Cα−C bond.

The computed reaction profile for DAP epimerization is shown in Figure 5B (see also Table S3 in the Supporting Information). Energy (ΔE and ΔE^*) and Gibbs free energy (ΔG and ΔG^*) values relative to **M1** are given in the figure. Thermal energy corrections to compute Gibbs free energy have been obtained by numerical frequency calculations on the **HM** layer with a frozen **L** layer. Results obtained using either the approximated (**freq**) or complete (**fullfreq**) frequency computation approach are reported.

It is evident from the reaction profile that the product complex (Enzyme/D,L-*meso*-DAP) is less stable than the reactant complex (Enzyme/L,L-DAP) by 15.05 kcal mol^{-1}. This value becomes 13.48 and 14.23 kcal mol^{-1} when we consider the free energy computed with the **freq** and **fullfreq** procedure, respectively. The computed barrier for the stereoinversion is about 24.11 kcal mol^{-1}, while the corresponding free activation energies obtained with the **freq** and **fullfreq** procedure are almost identical (21.32 and 21.30 kcal mol^{-1}, respectively). The normal mode corresponding to the imaginary frequency obtained for **TSC** describes the protonation/deprotonation process of the planar substrate. The computation of numerical frequencies using the approximated procedure **freq** gives results comparable to those obtained by the **fullfreq** approach but with a significant saving of CPU time. Also, the shape of the transition vector is very similar in the two cases. Here the **H** and **M** layers are composed by 36 and 170 atoms, respectively. Thus, the **fullfreq** procedure requires 1+(36 + 170)*6=1+(206)*6=1237 wave function evaluations, while the **freq** procedure requires only 1+(36)*6= 217, which saves about the 80% of computation time.

Table 1. Effects (kcal mol^{-1}) of the Electrostatic Interactions on the **M1/M2** and **M1/TSC** Energy Difference, As Obtained for **Model1**

$\Delta E_{QM/MM}$**(M1,M2)**	15.05
$\Delta E_{QM/MM}$**(M1,TSC)**	24.11
ΔE_0**(M1,M2)**[a]	−1.22
ΔE_0**(M1,TSC)**[a]	20.21
ΔE_{qm}**(M1,M2)**[b]	16.50
ΔE_{qm}**(M1,TSC)**[b]	26.42

[a] See eq 5. [b] See eq 6.

We have carried out the fingerprint analysis for the three critical points **M1, M2**, and **TSC** to obtain a detailed description of the influence of the various residues surrounding the reacting core. This analysis should be also helpful to ascertain the reliability of the adopted model system. In Table 1 we have collected the values of the **M1/M2** and **M1/TSC** energy differences as obtained from QM calculations (single point) on the *model-H* system (obtained from the QM/MM optimized geometries) *in vacuo* and in the presence of the atomic point charges of the whole enzyme.

The two minima are almost isoenergetic *in vacuo*: ΔE_0**(M1,M2)** is only −1.22 kcal mol^{-1}. However, in the presence of the atomic point charges the energy difference (ΔE_{qm}**(M1,M2)**) significantly increases and becomes 16.50 kcal mol^{-1}, a value which is close to the QM/MM value of 15.05 kcal mol^{-1}. This finding demonstrates the importance on the **M1/M2** equilibrium of the electrostatic interactions due to the protein environment. In particular, the negatively charged Glu-208 side chain (described at the MM level) seems to play an important role in destabilizing **M2** with respect to **M1**. This is caused by the unfavorable electrostatic interaction, occurring in **M2**, with the negatively charged Cys-217 (see discussion below). A similar effect, even less significant, is evident for the **M1/TSC** pair. The computed barrier changes from 20.21 (ΔE_0**(M1,TSC)**) to 26.42 kcal

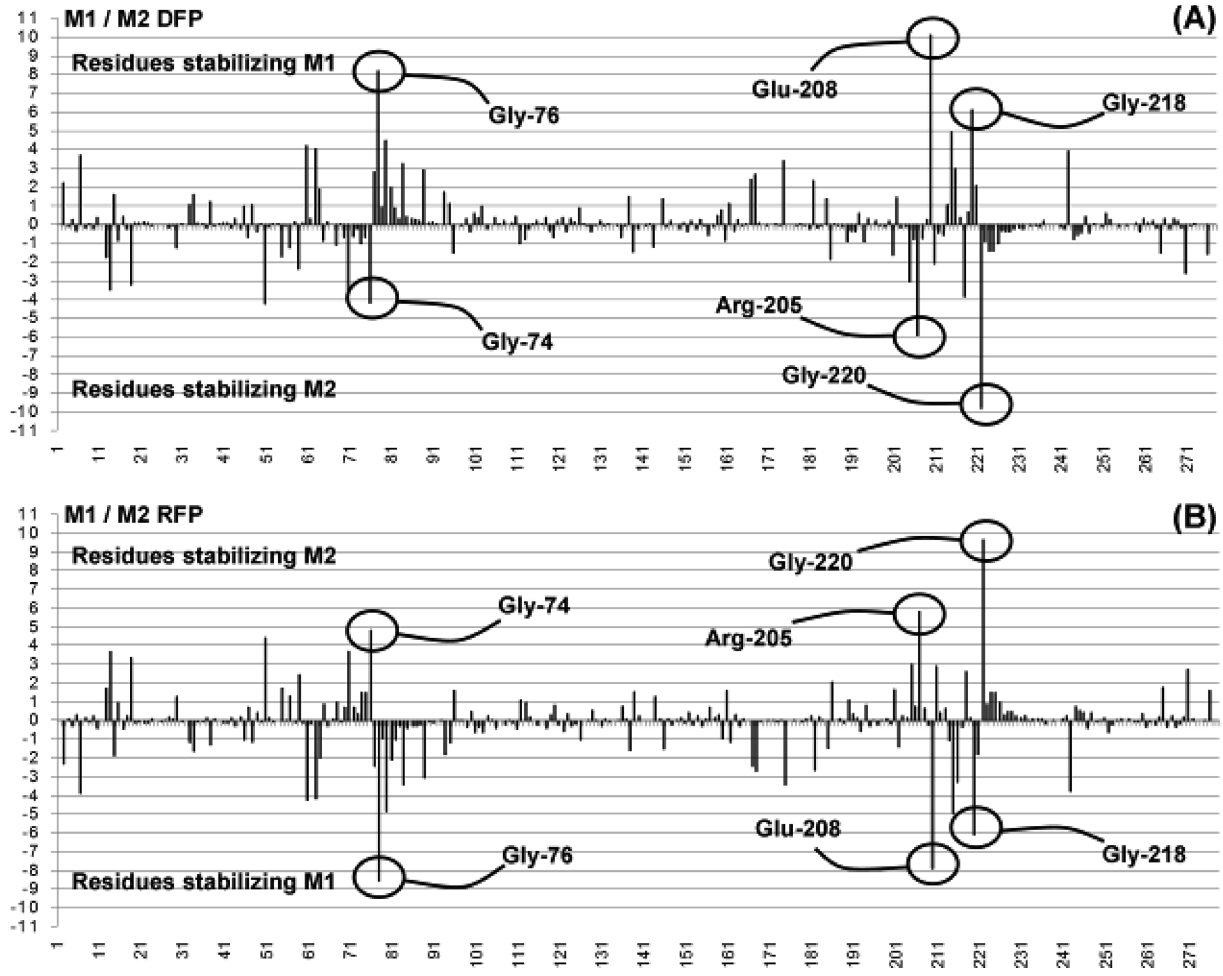


Figure 6. **DFP** and **RFP** diagrams obtained for the **M1/M2** pair in **Model1** (see Table S4 in the Supporting Information).

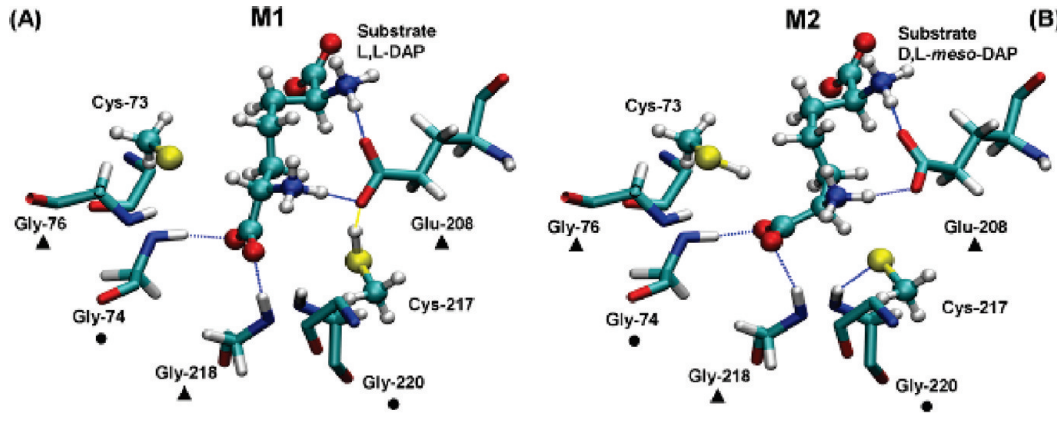


Figure 7. Schematic representation of the residues playing a key-role in the stabilization/destabilization of the **M1/M2** pair as found in **Model1**. (•) Residues stabilizing **M2** over **M1**; (▲) Residues stabilizing **M1** over **M2**.

mol^{-1} (ΔE_{qm}**(M1,TSC)**) when atomic point charges are added to the bare QM core.

In the Supporting Information we have reported the computed values of the various terms occurring in eqs 1 to 36 for **DFP** and **RFP** analysis. Here we discuss the most important stabilization (**S**) and destabilization (**D**) values obtained with the two approaches for the two pairs **M1/M2** and **M1/TSC**. **S** and **D** values for a few selected residues for the pair **M1/M2** are reported in Table S4 of the Supporting Information. The effects of the various residues are represented in the two diagrams of Figure 6.

It is evident from Figure 6 that Gly-76 plays a fundamental role in stabilizing **M1** with respect to **M2**. This stabilization (Figure 7) is due the hydrogen bond involving the Sγ atom of Cys-73. This interaction is stronger in **M1** than in **M2** because of the different protonation state of the sulfur atom: in **M2** this atom grabs the proton from the substrate and loses part of its negative charge (see Table S2 in the Supporting Information). This trend is evidenced by the (Gly-76)-NH---Sγ-(Cys-73) distance that increases from 2.34 Å to 2.39 Å on passing from **M1** to **M2**. A major role in stabilizing **M1** with respect to **M2** is also played by Glu-208. In **M1** this

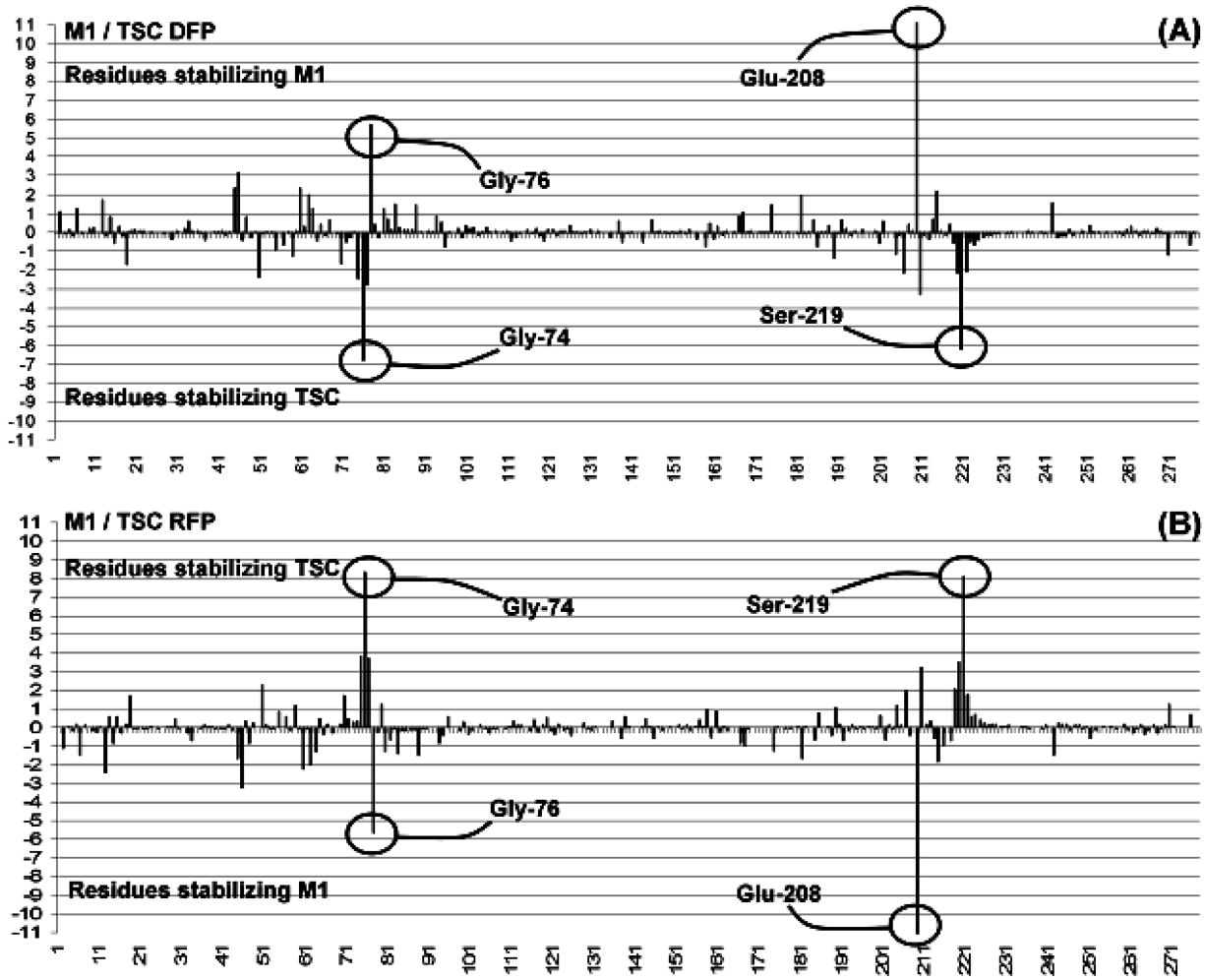


Figure 8. DFP and RFP diagrams for the **M1**/**TSC** pair in **Model1** (see Table S5 in the Supporting Information).

residue behaves as a triple hydrogen-bond acceptor: the unprotonated carboxylic side chain interacts with the two NH_3 substrate groups and with the thiolic hydrogen atom of Cys-217. Since in **M2** this hydrogen atom has been transferred to the substrate, only two hydrogen bonds remain active. Thus, the loss of one hydrogen bond and the negative charge on the Cys-217 Sγ atom explain the strong destabilization due to Glu-208 on **M2**. Gly-218 behaves as an H-bond donor toward the substrate carboxylic group adjacent to the α carbon undergoing the stereoinversion. The best structural arrangement is found in **M1** where the (Gly-218)-NH---$^{(-)}$OCO-(DAP) distance is 1.81 Å. This value increases to 2.09 Å in **M2** because of the stereo inversion, thus causing a decrease of stabilization in **M2**. On the contrary, some residues play an important role in stabilizing **M2** with respect to **M1.** For instance, the effect of Gly-74 is similar, but opposite, to that of Gly-218. It behaves as an H-bond donor toward the same substrate carboxylic group, but the interaction favors the D,L-*meso*-DAP (**M2**). This is confirmed by the (Gly-74)-NH---$^{(-)}$OCO-(DAP) distance which becomes shorter (from 1.77 to 1.70 Å) on passing from **M1** to **M2**. In a similar way Ala-216 stabilizes **M2** because of a more favorable structural arrangement of the NH_3 group of D,L-*meso*-DAP. The most important stabilizing effect for **M2** is due to Gly-220. This residue is a hydrogen-bond donor toward the Sγ atom of Cys-217: when this atom is not protonated (as in **M2**) this interaction becomes much stronger and **M2** is stabilized.

A list of selected residues (with the corresponding **S** and **D** values) that play an important role in the electrostatic catalysis by stabilizing or destabilizing **TSC** with respect to **M1** is reported in Table S5 of the Supporting Information. The effects of the various residues are indicated in the two diagrams of Figure 8, and the key interactions are shown in Figure 9. The effect of Gly-76 and Gly-74 is similar to that found for the **M1/M2** pair: these residues stabilize and destabilize **M1** with respect to **TSC**, respectively, but to a smaller extent. Glu-208 again stabilizes **M1**, because in **TSC** one of the three H-bonds (involving the substrate NH_3 group) is almost completely lacking. The substrate carboxylic group is a double hydrogen-bond acceptor toward Ser-219, and the better structural arrangement favoring these interactions is found in **TSC**. However, while the distance that features the H-bond between the carboxylic group and (Ser-219)-OH varies from 1.60 Å in **M1** to 1.53 Å in **TSC** and 1.61 Å in **M2**, the distance between the carboxylic group and (Ser-219)-NH is almost constant in **M1** and **TSC** (1.76 Å and 1.77 Å, respectively) and becomes 1.87 Å in **M2**. Thus, the latter hydrogen bond does not play a key role in stabilizing the transition state.

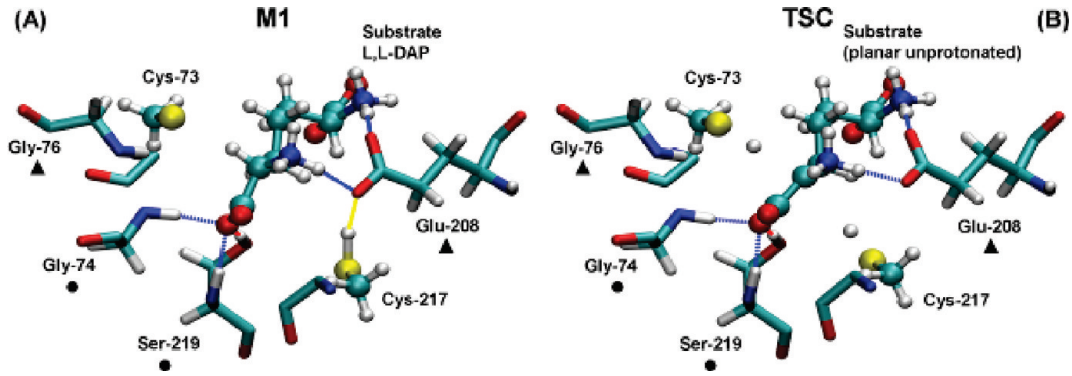


Figure 9. Schematic representation of the residues playing a key-role in the stabilization/destabilization of the **M1**/**TSC** pair as found in **Model1**. (•) Residues stabilizing **TSC** over **M1**; (▲) Residues stabilizing **M1** over **TSC**.

The analysis of the contributions of the various residues demonstrates the general agreement of the **DFP** and **RFP** analysis.

In Tables S6 and S7 of the Supporting Information we have collected the van der Waals contributions of a few selected residues. It is evident that these contributions are significantly less important than the electrostatic contributions (evidenced by **DFP** and **RFP** analysis) in determining the relative stabilization/destabilization of the **M1/TSC** and **M1/TSC** pairs. This suggests that, in the present case, it is reasonable to consider only the electrostatic contributions to rationalize the enzyme catalytic effect.

III.C. Model2: A More Accurate Model System. The **DFP** and **RFP** analysis on **Model1** has clearly shown the role of single residues and their importance in determining the shape of the potential surface. Since **DFP** and **RFP** can identify the residues exerting the strongest effects on the catalysis, they can help to improve the features of the model-system and obtain more reliable results.

In the previous section we have shown that Glu-208 has the most important effect in stabilizing **M1** with respect to **TSC** and **M2**. In **Model1** Glu-208 is completely described by an MM potential, while their interactions with the QM core are taken into account by electrostatic and van der Waals QM-MM cross terms. We focus here on the electrostatic effects, being that the van der Waals terms are much smaller. To check the reliability of **Model1** and to establish if the interactions between the residue Glu-208 and the substrate are correctly described, it is essential to build a different and more accurate model system, that we denote as **Model2** (see Figure S2 in the Supporting Information). The important feature of **Model2** is the inclusion of the side-chain of Glu-208 in the **H** layer (see Figure S3). In this way the negatively charged Glu-208 side-chain and its interactions with the substrate molecule and the Cys-217 residue is fully described at the QM level. We have determined for **Model2** the three critical points **M1**, **M2**, and **TSC** starting from the corresponding structures obtained for **Model1**, and we have carried out for each critical point **freq** and **fullfreq** computations and **DFP** and **RFP** analysis. Some important parameters (structural features and charge distribution) are reported in Table S8 in the Supporting Information.

The reaction energetics, as obtained for **Model2**, is reported in Figure S3 in the Supporting Information. The energy difference between reactants (Enz/L,L-DAP) and products (Enz/D,L-*meso*-DAP) decreases and becomes 11.74 kcal mol^{-1} (the corresponding free energy values computed with **freq** and **fullfreq** are 11.56 and 10.80 kcal mol^{-1}, respectively). The activation barrier only slightly changes and with respect to **Model1** becoming 25.31 kcal mol^{-1} (23.00 and 22.48 kcal mol^{-1} are the free energy values obtained from **freq** and **fullfreq** procedure, respectively). Again, the normal mode associated with the **TSC** imaginary frequency is very similar when computed with the two approaches (**freq** and **fullfreq**) and describes the protonation/deprotonation process of the planar substrate.

Thus, a comparison of the results obtained for the two model systems (Figures 5 and S3 and Tables S3 and S9 in the Supporting Information) shows that the barrier associated with the stereoinversion only slightly changes on passing from **Model1** (24.11 kcal mol^{-1}) to **Model2** (25.31 kcal mol^{-1}), while the inclusion of Glu-208 in the QM layer has a significant effect on the **M1/M2** relative energies (15.05 and 11.74 kcal mol^{-1} for **Model1** and **Model2**, respectively). The negligible variation of the barrier height suggests that the description of the Glu-208 residue is reliable enough at the MM level. On the contrary, since the inclusion of Glu-208 in the model-system has a significant effect on the **M1/M2** energy difference, a QM description of Glu-208 seems to be essential in the **M1**-**M2** comparison. This finding can be understood if we consider the particular geometrical arrangement of **M2** where the Glu-208 carboxylate is rather close to the negatively charged sulfur atom of Cys-217. The localized point charges placed on the carboxylate oxygen atoms in the MM treatment are evidently not able to correctly account for the interaction between the Cys-217 and Glu-208 residues in the **M2** structure.

A further understanding of the factors controlling the energetics of the process is provided by the values reported in Table 1 and 2 where we have collected the results of QM calculations *in vacuo* (ΔE_0) and in the presence of the enzyme atomic point charges (ΔE_{qm}) on the *model-H* system. Interestingly, for both **Model1** and **Model2** the **M1/TSC** and **M1/M2** energy differences significantly vary with respect to the QM/MM values when the "naked" *model-H* system is considered (see (ΔE_0(**M1**,**TSC**) and ΔE_0(**M1**,**M2**)). However, these terms become very similar to the QM/MM values after inclusion of the enzyme

Table 2. Effects (kcal mol^{-1}) of the Electrostatic Interactions on the **M1/M2** and **M1/TSC** Energy Difference, As Obtained for **Model2**

$\Delta E_{QM/MM}$(M1,M2)	11.74
$\Delta E_{QM/MM}$(M1,TSC)	25.31
ΔE_0(M1,M2)[a]	7.17
ΔE_0(M1,TSC)[a]	37.81
ΔE_{qm}(M1,M2)[b]	10.20
ΔE_{qm}(M1,TSC)[b]	24.45

[a] See eq 5. [b] See eq 6.

atomic point charges (compare ΔE_{qm}(**M1**,**TSC**) and ΔE_{qm}(**M1**,**M2**) to $\Delta E_{QM/MM}$(**M1**,**TSC**) and $\Delta E_{QM/MM}$(**M1**,**M2**), respectively). This clearly indicates that the most important contribution to the catalysis is electrostatic.[67]

The results of **DFP** and **RFP** analysis (see Tables S10 and S11 in the Supporting Information) are comparable to those obtained for **Model1**: the ranking order for the various residues is almost identical in the two cases, except, of course, for Glu-208 which has been included in the QM region (see Figures S4-S7 in the Supporting Information).

III.D. Model3: Toward a Reliable Description of the Chemical Transformation. We have shown in the previous sections that **DFP** and **RFP** are extremely useful in detecting the residues that play a key role in the catalysis. In both **Model1** and **Model2** these important residues have not been included in the QM region but belong to the MM shell, and, consequently, their interactions are described by the Amber potential. Thus, it is not surprising that the reaction barriers computed for **Model1** and **Model2** (21.3 and 22.5 kcal mol^{-1}, respectively) differ from the experimental value (15.6 kcal mol^{-1}, as reported in the Introduction). This is probably due to the failure of the MM force field to properly describe the changes (on passing from **M1** to **TSC**) of the charge distribution that features the hydrogen bond network involving the substrate and the above-mentioned residues. The following important point, which stems from the previous discussion, must be outlined. Even if small model systems such as **Model1** and **Model2** are not able to quantitatively reproduce the experimental data, they provide, when coupled to **DFP** and **RFP,** a valuable tool to build more reliable model systems, by making possible the inclusion in the QM region of the residues indicated as important for the catalysis. Following this approach we have defined the new model system **Model3** by adding to the QM shell all residues detected by the **DFP** and **RFP** analysis carried out for the three critical points **M1**, **M2**, and **TS**. In particular, we have considered Gly-74, Gly-76, the backbone of Asn-75, Gly-218, part of Ser-219 and Gly-220. We have carried out single point calculations on **Model3** at the geometry previously obtained for **Model2** (see section III.C). These new computations have been performed at two different levels: DFT(B3LYP/DZVP)/Amber-ff99 and DFT(B3LYP/TZVP)/Amber-ff99 levels to examine the effect of increasing the basis set accuracy.

The results confirm the capability of **DFP** and **RFP**, applied to small systems like **Model1** and **Model2**, of identifying the catalytically important residues. It is evident from the results reported in Table 3 that the inclusion of the new residues in the QM layer determines a significant improvement in the computed reaction barrier (18.7 kcal mol^{-1}), even if single point calculations, rather than full geometry optimizations, have been carried out. A further improvement of the computed barrier (16.6 kcal mol^{-1}) is observed when the more accurate basis set TZVP is employed.

Table 3. Reaction Energies ($\Delta E_{QM/MM}$(**M1,M2**)) and Activation Energies ($\Delta E_{QM/MM}$(**M1,TSC**)) Obtained for **Model3**[b]

	$\Delta E_{QM/MM}$(M1,M2)	$\Delta E_{QM/MM}$(M1,TSC)
Model3 (DZVP)	10.9	18.7
Model3 (TZVP)	10.2	16.6
experimental[a]		15.6

[a] The experimental activation energy corresponds to a free energy change ΔG, as obtained by applying the Eyring equation to the experimental k_{cat} value. [b] Values are in kcal mol^{-1}.

IV. Biological Insights into a Class of Enzymes: Reaction Mechanism and Catalysis

According to sequence and structural similarity, glutamate and aspartate racemases belong to one homologous family of enzymes,[11] while DAP epimerase and proline racemase are usually collected in a different group. To elucidate the general catalytic mechanism of this second family we investigated in a previous paper[22] the reaction surface of proline racemase (TcPRAC).[21] In particular we considered the enzyme found in the eukaryotic parasite *Trypanosoma cruzi*[68] because it represents a promising target for drug design against Chagas' disease[69] and can be considered a reliable model for prokaryotic proline racemases.[20,70,71] In that study we used the same computational approach of the present paper (i.e., a combination of a DFT(B3LYP/DZVP)/Amber-*ff99* potential[9] and fingerprint analysis on the critical points), and we described the mechanism of stereoinversion of the proline α carbon to afford the un-natural D-proline from the more abundant L-proline. We explained the enzyme catalysis in term of electrostatic stabilization of the transition state.

Even if TcPRAC and DAP epimerase belong to the same family, they show very poor similarities in the active site region (except for the conserved catalytically active cysteine pair). This strong structural dissimilarity is due to the need of accommodating in the reacting region two substrates which significantly differ in the shape and electronic properties. In both enzymes the substrate within the active site is present as a zwitterionic ion and is bound to the various residues by a tight network of hydrogen bonds. These interactions determine the right orientation of the substrate by anchoring the amino and the carboxylic groups and, in the case of DAP epimerase, the distal site of the molecule. In particular, the carboxylic group seems to be highly relevant to the catalytic process although not directly involved in the chemical reaction (breaking and forming of chemical bonds). To better understand this aspect, we must focus our attention on the nature of the transition state that describes for both

enzymes a concerted but highly asynchronous mechanism. This transition state corresponds to an almost fully deprotonated substrate where the α carbon (originally the sp^3 carbon) and the carboxylic group form an extended planar π system that delocalizes the partial negative charge left on the substrate after proton abstraction. This charge delocalization decreases the carbanionic character of the transition state (as suggested by charge analysis) and increases the negative charge on the carboxylic group (especially the two O atoms). The final effect on the transition state is an increase of the strength (and stabilization) of the hydrogen bonds between the carboxylic group and some residues in the vicinity of the cysteine pair.

The fingerprint analysis (DFP for TcPRAC[9] and both DFP and RFP for DAP epimerase) has evidenced the residues which are responsible for the electrostatic catalysis. These residues are hydrogen bond donor to the substrate carboxylic group. It is worth pointing out that, even if the stabilization of the transition state is determined by different residues in the two enzymes, the nature of the overall catalytic effect remains the same and can be mainly ascribed to specific electrostatic contributions of "pre-organized" active sites where some residues are in the optimum positions to emphasize the effect of the stabilizing hydrogen-bonds. This finding is in very good agreement with the conclusions reached by Warshel.[6,67,72−77]

The presented results, beside helping to obtain a general picture of the catalytic mechanism for an important class of enzymes, allow the identification of common features shared within the same enzyme family. This information should be highly useful in the future for the design and development of new drugs targeting this group of PLP-independent racemases/epimerases.

V. Conclusions

In this paper we have provided a detailed description of the reaction mechanism of the enzyme diaminopimelate (DAP) epimerase, a promising target for rational drug design aimed at developing new selective antibacterial therapeutic agents. This enzyme represents a model for the PLP-independent racemases/epimerases acting *via* a two-base mechanism involving a pair of cysteine residues (thiol/thiolate pair at neutral pH[11,27]).

We have used a QM/MM computational approach based on a DFT(B3LYP/DZVP)//Amber-*ff99* potential.[9] This approach is similar to that employed in a previous paper where we have investigated the mechanism of proline racemase (TcPRAC).[21] Two different model-systems have been investigated. In one case (**Model1**) the entire substrate (DAP molecule) and the side-chains of Cys-73 and Cys-217 (after saturation of the dangling bonds with hydrogen atoms) have been included in the **H** layer described at the QM level. The remaining part of the enzyme has been treated at the MM level (**M** and **L** layers). In the second model-system (**Model2**) the side-chain of Glu-208 has been included in the **H** region. Thus, in this case the negatively charged Glu-208 side-chain and its interactions with the substrate molecule and the Cys-217 residue have been completely described at the QM level. Both model-systems have provided the same mechanistic picture: the reaction proceeds through a highly asynchronous mechanism where the side-chain of the negatively charged Cys-73 (thiolate) captures a proton from the carbon substrate. Simultaneously, the Cys-217 thiolic proton moves toward the same carbon atom on the opposite face. In the transition state the substrate is essentially unprotonated and planar.

Direct and inverse fingerprint analysis (**DFP** and **RFP** analysis) on the three critical points **M1, M2**, and **TSC** for both **Model1** and **Model2**, have provided a detailed description of the influence of the various residues surrounding the active site and have clearly indicated that the most important contribution to the catalysis is electrostatic. DFP and FFP analysis carried out on **Model1** have pointed out that Glu-208 has the most important effect in stabilizing reactants (**M1**) with respect to transition state (**TSC**) and products (**M2**). The indication of the fingerprint analysis has suggested including this residue in the QM region of the model. The aim of this choice was to establish if the MM potential was reliable to describe the interactions of Glu208 with the substrate and Cys-217. A comparison of the energetics obtained for the two model-systems has shown that, while the stereoinversion barrier does not significantly change (24.11 and 25.31 kcal mol^{-1} for **Model1** and **Model2** respectively), the inclusion of Glu-208 in the QM layer has a stronger effect on the **M1**/**M2** relative energy, which is 15.05 kcal mol^{-1} in **Model1** and becomes 11.74 kcal mol^{-1} in **Model2**. This finding suggests that the MM description of the Glu-208 residue is reliable when comparing reactants (**M1**) and transition state (**TSC**), while a QM description of this residue seems to be essential in the **M1**-**M2** comparison.

Using the results of the fingerprint analysis on **Model2** we have built a larger (and more reliable) model system, **Model3**, where all important residues detected by **DFP** and **RFP** have been included in the QM region. Single point computations on **Model3**, using the **Model2** structures and the two basis sets DZVP and TZVP for the QM shell, have provided activation energies in good agreement with the experimental value of 15.6 kcal mol^{-1}: 18.7 and 16.6 kcal mol^{-1} at the DZVP and TZVP levels, respectively. These results confirm the validity of our approach and the possibility of using **DFP** and **RFP** analysis to identify the catalytically important residues and, thus, build reliable model systems.

Acknowledgment. We would like to thank Ministero dell'Università e della Ricerca (MIUR) (PRIN 2007 "Sintesi e Stereocontrollo di Molecole Organiche per lo Sviluppo di Metodologie Innovative di Interesse Applicativo") and CINECA computer center of Bologna for financial support.

Supporting Information Available: Computational details, additional results in the form of tables and figures in text format, molecular geometries in PDB format and parameters in Amber format for nonstandard residues in a ZIP archive. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Gao, J. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VHC Publishers: New York, 1995; Vol. 7, pp 119−185.

(2) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2006**, *117*, 185–199.

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

(4) Hillier, I. H. *J. Mol. Struct.* **1999**, *463*, 45–52.

(5) Sherwood, P. NIC series 2000, 3, 285−305. http://www.fz-juelich.de/nic-series/NIC-Series-e.html (accessed March 03, 2009).

(6) Klahn, M.; Braun-Sand, S.; Rosta, E.; Warshel, A. *J. Phys. Chem. B* **2005**, *109*, 15645–15650.

(7) Senn, H. M.; Thiel, W. In *Atomistic Approaches in Modern Biology: from Quantum Chemistry to Molecular Simulations*; 2007; Vol. 268, pp 173−290.

(8) Senn, H. M.; Thiel, W. *Curr. Opin. Chem. Biol.* **2007**, *11*, 182–187.

(9) Altoe, P.; Stenta, M.; Bottoni, A.; Garavelli, M. *Theor. Chem. Acc.* **2007**, *118*, 219–240.

(10) Koo, C. W.; Blanchard, J. S. *Biochemistry* **1999**, *38*, 4416–4422.

(11) Pillai, B.; Cherney, M. M.; Diaper, C. M.; Sutherland, A.; Blanchard, J. S.; Vederas, J. C.; James, M. N. G. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 8668–8673.

(12) Brunetti, L.; Galeazzi, R.; Orena, M.; Bottoni, A. *J. Mol. Graphics Modell.* **2008**, *26*, 1082–1090.

(13) Yoshimura, T.; Esak, N. *J. Biosci. Bioeng.* **2003**, *96*, 103–109.

(14) Major, D. T.; Gao, J. L. *J. Am. Chem. Soc.* **2006**, *128*, 16345–16357.

(15) Major, D. T.; Nam, K.; Gao, J. *J. Am. Chem. Soc.* **2006**, *128*, 8114–8115.

(16) Amadasi, A.; Bertoldi, M.; Contestabile, R.; Bettati, S.; Cellini, B.; Luigi di Salvo, M.; Borri-Voltattorni, C.; Bossa, F.; Mozzarelli, A. *Curr. Med. Chem.* **2007**, *14*, 1291–1324.

(17) Glavas, S.; Tanner, M. E. *Biochemistry* **2001**, *40*, 6199–6204.

(18) Puig, E.; Garcia-Viloca, M.; Gonzalez-Lafont, A.; Lluch, J. M.; Field, M. J. *J. Phys. Chem. B* **2007**, *111*, 2385–2397.

(19) Puig, E.; Garcia-Viloca, M.; Gonzalez-Lafont, A.; Lluch, J. M. *J. Phys. Chem. A* **2006**, *110*, 717–725.

(20) Rudnick, G.; Abeles, R. H. *Biochemistry* **1975**, *14*, 4515–4522.

(21) Buschiazzo, A.; Goytia, M.; Schaeffer, F.; Degrave, W.; Shepard, W.; Gregoire, C.; Chamond, N.; Cosson, A.; Berneman, A.; Coatnoan, N.; Alzari, P. M.; Minoprio, P. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 1705–1710.

(22) Stenta, M.; Calvaresi, M.; Altoe, P.; Spinelli, D.; Garavelli, M.; Bottoni, A. *J. Phys. Chem. B* **2008**, *112*, 1057–1059.

(23) Viola, R. E. *Acc. Chem. Res.* **2001**, *34*, 339–349.

(24) Born, T. L.; Blanchard, J. S. *Curr. Opin. Chem. Biol.* **1999**, *3*, 607–613.

(25) Work, E. *Nature* **1950**, *165*, 74–75.

(26) Eyring, H. *Chem. Rev.* **1935**, *17*, 65–77.

(27) Diaper, C. M.; Sutherland, A.; Pillai, B.; James, M. N. G.; Semchuk, P.; Blanchard, J. S.; Vederas, J. C. *Org. Biomol. Chem.* **2005**, *3*, 4402–4411.

(28) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(29) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. *Nucleic Acids Res.* **2005**, *33*, 368–371.

(30) Bas, D. C.; Rogers, D. M.; Jensen, J. H. *Proteins: Struct., Funct., Bioinf.* **2008**, *73*, 765–783.

(31) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(32) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132–146.

(33) Jakalian, A.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

(34) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

(35) Ponder, J. W.; Case, D. A.; Valerie, D. In *Advances in Protein Chemistry*; Academic Press: 2003; Vol. 66, pp 27−85.

(36) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

(37) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.

(38) Feig, M.; Karanicolas, J.; Brooks, C. L. *MMTSB Tool Set (2001)*; MMTSB NIH Research Resource, The Scripps Research Institute: 2001.

(39) Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 601–619.

(40) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. *J. Mol. Biol.* **1982**, *161*, 269–288.

(41) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357–19363.

(42) Vreven, T.; Byun, K. S.; Komaromi, I.; Dapprich, S.; Montgomery, J. A.; Morokuma, K.; Frisch, M. J. *J. Chem. Theory Comput.* **2006**, *2*, 815–826.

(43) Prat-Resina, X.; Gonzalez-Lafont, A.; Lluch, J. M. *J. Mol. Struct. - Theochem* **2003**, *632*, 297–307.

(44) Prat-Resina, X.; Bofill, J. M.; Gonzalez-Lafont, A.; Lluch, J. M. *Int. J. Quantum Chem.* **2004**, *98*, 367–377.

(45) Vreven, T.; Morokuma, K.; Farkas, 951 > O.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **2003**, *24*, 760–769.

(46) Broyden, C. G. *J. Inst. Math. Appl.* **1970**, *6*, 76–90.

(47) Fletcher, R. *Comput. J.* **1970**, *13*, 317–322.

(48) Goldfarb, D. *Math. Comput* **1970**, *24*, 23–26.

(49) Shanno, D. F. *Math. Comput* **1970**, *24*, 647–656.

(50) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361–373.

(51) Geerlings, P.; De Proft, F.; Langenaeker, W. *Chem. Rev.* **2003**, *103*, 1793–1873.

(52) Cohen, A. J.; Mori-Sanchez, P.; Yang, W. *Science* **2008**, *321*, 792–794.

(53) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

(54) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. *Can. J. Chem.* **1992**, *70*, 560–571.

(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.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2004 (http://www.gaussian.com/).

(56) Odon, F.; Schlegel, H. B. *J. Chem. Phys.* **1999**, *111*, 10806–10814.

(57) Schlegel, H. B. *J. Comput. Chem.* **2003**, *24*, 1514–1527.

(58) Bash, P. A.; Field, M. J.; Davenport, R. C.; Petsko, G. A.; Ringe, D.; Karplus, M. *Biochemistry* **1991**, *30*, 5826–5832.

(59) Cui, Q.; Karplus, M. *J. Am. Chem. Soc.* **2001**, *123*, 2284–2290.

(60) Cui, Q.; Elstner, M.; Karplus, M. *J. Phys. Chem. B* **2002**, *106*, 2721–2740.

(61) Cui, Q.; Karplus, M. *J. Phys. Chem. B* **2002**, *106*, 1768–1798.

(62) Cui, Q.; Karplus, M. *J. Am. Chem. Soc.* **2002**, *124*, 3093–3124.

(63) Zhang, X.; Harrison, D. H. T.; Cui, Q. *J. Am. Chem. Soc.* **2002**, *124*, 14871–14878.

(64) Ranaghan, K. E.; Ridder, L.; Szefczyk, B.; Sokalski, W. A.; Hermann, J. C.; Mulholland, A. J. *Org. Biomol. Chem.* **2004**, *2*, 968–980.

(65) Banerjee, A.; Yang, W.; Karplus, M.; Verdine, G. L. *Nature* **2005**, *434*, 612–618.

(66) Zhang, X.; Bruice, T. C. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 6148–6153.

(67) Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. M. *Chem. Rev.* **2006**, *106*, 3210–3235.

(68) Chamond, N.; Gregoire, C.; Coatnoan, N.; Rougeot, C.; Freitas, L. H.; da Silveira, J. F.; Degrave, W. M.; Minoprio, P. *J. Biol. Chem.* **2003**, *278*, 15484–15494.

(69) Chamond, N.; Goytia, M.; Coatnoan, N.; Barale, J. C.; Cosson, A.; Degrave, W. M.; Minoprio, P. *Mol. Microbiol.* **2005**, *58*, 46–60.

(70) Cardinale, G. J.; Abeles, R. H. *Biochemistry* **1968**, *7*, 3970–3978.

(71) Fisher, L. M.; Albery, W. J.; Knowles, J. R. *Biochemistry* **1986**, *25*, 2529–2537.

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

(73) Warshel, A. *J. Biol. Chem.* **1998**, *273*, 27035–27038.

(74) Warshel, A. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 425–443.

(75) Olsson, M. H. M.; Parson, W. W.; Warshel, A. *Chem. Rev. (Washington, DC, U. S.)* **2006**, *106*, 1737–1756.

(76) Rosta, E.; Klahn, M.; Warshel, A. *J. Phys. Chem. B* **2006**, *110*, 2934–2941.

(77) Warshel, A.; K, S. P.; Mitsunori, K.; Parson, W. W. *Biochim. Biophys. Acta* **2006**, *1764*, 1647–1676.

CT900004X

JCTC Journal of Chemical Theory and Computation

Large Protein Dynamics Described by Hierarchical-Component Mode Synthesis

Jae-In Kim, Sungsoo Na,* and Kilho Eom*

Department of Mechanical Engineering, Korea University, Seoul 136-701, Republic of Korea

Received January 15, 2009

Abstract: Protein dynamics has played a pivotal role in understanding the biological function of protein. For investigation of such dynamics, normal-mode analysis (NMA) has been broadly employed with atomistic model and/or coarse-grained models such as elastic network model (ENM). For large protein complexes, NMA with even ENM encounters the expensive computational process such as diagonalization of Hessian (stiffness) matrix. Here, we suggest the hierarchical-component mode synthesis (hCMS), which allows the fast computation of low-frequency normal modes related to conformational change. Specifically, a large protein structure is regarded as a combination of several structural units, for which the eigen-value problem is utilized for obtaining the frequencies and their normal modes for each structural unit, and consequently, such frequencies and normal modes are assembled with geometrical constraint for interface between structural units in order to find the low-frequency normal modes of a large protein complex. It is shown that hCMS is able to provide the normal modes with accuracy, quantitatively comparable to those of original NMA. This implies that hCMS may enable the computationally efficient analysis of large protein dynamics.

1. Introduction

Normal mode analysis (NMA) has enabled one to understand the protein dynamics, related to the biological function of protein, based on the low-frequency normal modes that are usually associated with the conformational change of protein.[1−3] The fundamental of NMA is to solve the eigen-value problem for diagonalization of Hessian (stiffness) matrix for protein structure.[4−6] Here, the stiffness matrix is computed based on the second-derivative of anharmonic potential field with respect to atomistic coordinates at equilibrium state, where potential is globally minimum. Complexity of potential field for protein atomistic structure leads to the computationally expensive process such as energy minimization (to find the equilibrium state) and calculation of stiffness matrix. This has led many research groups to develop the computationally efficient algorithm (or reduced model) to estimate the stiffness matrix and its related low-frequency normal modes computed from NMA with given stiffness matrix.

In a recent decade, there has been an attempt to develop the coarse-grained model for protein structure by reducing the degrees of freedom as well as simplifying the potential field. One of the successful, broadly accepted coarse-grained models is the Go model,[6−9] where only α carbon atoms are taken into account with a simplified potential field composed of a backbone covalent bond stretch and ver der Waal's interaction for native contact. The Go model has successfully predicted the protein dynamics such as conformational fluctuation dynamics[6] as well as protein unfolding mechanics.[7−10] Currently, the Go model can be regarded as a versatile model for the description of protein dynamics and/or mechanics. In a similar spirit, Tirion[11] first suggested a more simplified protein structural model, referred to as an elastic network model (ENM), in such a way that α carbon atoms are only prescribed by the harmonic potential field in the neighborhood. Despite its simplicity, ENM is able to reproduce the low-frequency normal modes and the thermal fluctuation behavior, quantitatively comparable to those estimated by experiments (X-ray crystallography or nuclear magnetic resonance) and/or atomistic simulation.[11] ENM has been broadly employed for gaining insight into the conformational

* Corresponding author e-mail: nass@korea.ac.kr (S.N.) and kilhoeom@korea.ac.kr.

10.1021/ct900027h CCC: $40.75
Published on Web 05/26/2009

transition upon ligand-binding. For instance, Bahar and co-workers[12−15] reported that conformational change of proteins is well described by a few low-frequency normal modes. Further, several research groups[16−18] developed the model for the description of conformational transition by employing the ENM with constraints for computing the incremental displacement based on normal modes for conformational change at a certain state. Karplus and co-workers[19] reported the plastic network model (PNM) (similarly, mixed network model by Hummer and co-workers[20]) based on ENM by mixing the two potential fields near two distinct equilibrium states to find the pathway for conformational change. Recently, ENM has been employed even for studying protein mechanics such as protein unfolding mechanics. These indicate that ENM becomes a universal model for understanding the protein dynamic and/or mechanics.

However, for a large protein complex, ENM may exhibit the computational inefficiency for studying protein dynamics based on NMA. In order to overcome the computational inefficiency in obtaining the low-frequency normal modes, there have been attempts to introduce the model reduction methods applicable to ENM. For instance, Cui and co-workers[1] have implemented the block normal mode (BNM) analysis to protein structure based on atomistic potential (see Chapter 4 in ref 1). In their work, the protein motion is described in the normal modes of blocks at the residue level as well as the diagonalization scheme of sparse matrix for blocks. Bahar and co-workers[21] have first suggested the coarse-grained elastic network model composed of nodes, much less than the total number of residues, which are connected by entropic springs. Here, they have used the empirical parameters (such as force constant and cutoff distance) to describe the coarse-grained elastic network model. Recently, Eom et al.[22,23] provided the model reduction method, referred to as model condensation, inspired by the skeletonization scheme provided by Rohklin and co-workers.[24] Further, Jernigan and co-workers[25,26] reported the rigid cluster model, which regards protein structure as a combination of rigid domains connected by harmonic springs. Sanejouand and co-workers[27] developed the rotational/translational block (RTB) model, which dictates the protein structure as the rigid blocks (containing more than one residue). Those methods show that the low-resolution structure described by a few degrees of freedom is sufficient to study the protein dynamics such as conformational fluctuation. Recently, Ma and co-worker[28] suggested the coarse-grained network model based on the RTB method. Nonetheless, the quality of low-frequency normal modes is generally degraded as the protein structure is further coarse-grained. This indicates that coarse-graining of protein structure may be sometimes inappropriate for studying the conformational change that is related to low-frequency normal modes.

In this work, we report the hierarchical component mode synthesis (hCMS) for quantitative study on the low-frequency normal modes of a large protein complex. Here, a protein structure is regarded as consisting of structural subdomains, where NMA is implemented, and then such normal-mode information for each subdomain is assembled based on geometric constraint. It is shown that hCMS is capable of fast computation on low-frequency normal modes, quantitatively comparable to those obtained by conventional NMA. This implies that hCMS may enable one to study the large protein dynamics with computational efficiency as well as accuracy.

2. Model

2.1. Normal Mode Analysis (NMA) and Elastic Network Model (ENM). NMA assumes that protein motion is described by harmonic motion near equilibrium state.[1,4,5] For a given potential field V for a protein structure, the protein motion is represented in the form of $\mathbf{M}(d^2\mathbf{x}/dt^2) + \mathbf{Kx} = \mathbf{0}$, where $\mathbf{M}$ is the mass matrix (typically, diagonal matrix) and $\mathbf{K}$ is the stiffness (Hessian) matrix given by $\mathbf{K} = \partial_{\mathbf{x}}\partial_{\mathbf{x}}V$, where $\partial_{\mathbf{x}}$ is the gradient with respect to coordinates $\mathbf{x}$. Let $\mathbf{x} = \mathbf{u}\exp[i\omega t]$ with natural frequency ω and its corresponding normal mode $\mathbf{u}$. Then, the protein motion is described by an eigen-value problem as follows: $\mathbf{Ku} = \omega^2\mathbf{Mu}$.

ENM describes the protein structure as the harmonic spring network such that residues within the neighborhood are connected by a harmonic spring with an identical force constant. The potential field V for ENM is in the form of[11,13]

$$V = \frac{\gamma}{2}\sum (R_{ij} - R_{ij}^0)^2 \cdot H(R_c - R_{ij}^0) \tag{1}$$

where R_{ij} is the distance between two residues i and j, R_c is the cutoff distance given as $R_c = \sim 10$ Å, γ is the force constant, $H(x)$ is the Heaviside unit-step function, and superscript 0 indicates the equilibrium state. The potential field V can be also represented in the form of $V = (1/2)\mathbf{v}^T\mathbf{Kv}$, where $\mathbf{v}$ is the displacement vector for all residues, a symbol T represents the transpose of a vector, and $\mathbf{K}$ is the stiffness matrix composed of 3×3 block matrices $\mathbf{K}_{ij}$ given by

$$\mathbf{K}_{ij} = -\left[\gamma H(R_c - R_{ij}^0)\frac{(\mathbf{R}_{ij}^0)^T\mathbf{R}_{ij}^0}{(R_{ij}^0)^2}\right](1 - \delta_{ij}) - \delta_{ij}\sum_{l\neq i}\mathbf{K}_{il} \tag{2}$$

Here, $\mathbf{R}_{ij} = \mathbf{R}_j - \mathbf{R}_i$ with $\mathbf{R}_i$ being a position vector for residue i, and δ_{ij} is the Kronecker delta defined as $\delta_{ij} = 1$ if $i = j$; otherwise $\delta_{ij} = 0$.

Statistical mechanics theory allows the computation of correlation matrix $\mathbf{S}$ representing the thermal fluctuation behavior[1,13,29,30]

$$\mathbf{S} = \langle \mathbf{v}^T\mathbf{v}\rangle = \sum_{p=7}^{3N}\frac{k_BT}{\omega_p^2}\mathbf{u}_p^T\mathbf{u}_p \tag{3}$$

where <A> represents the ensemble average (time average) of the quantity A, k_B is the Boltzmann's constant, T is the absolute temperature, and index p indicates the mode index. Here, it should be noted that six zero-normal modes corresponding to rigid body motions are excluded for computing the correlation matrix $\mathbf{S}$. The mean-square fluctuation for residue i is given by $\langle|\mathbf{R}_i - \mathbf{R}_i^0|^2\rangle = \mathbf{S}_{3(i-1)+1,\,3(i-1)+1} + \mathbf{S}_{3(i-1)+2,\,3(i-1)+2} + \mathbf{S}_{3(i-1)+3,\,3(i-1)+3}$.

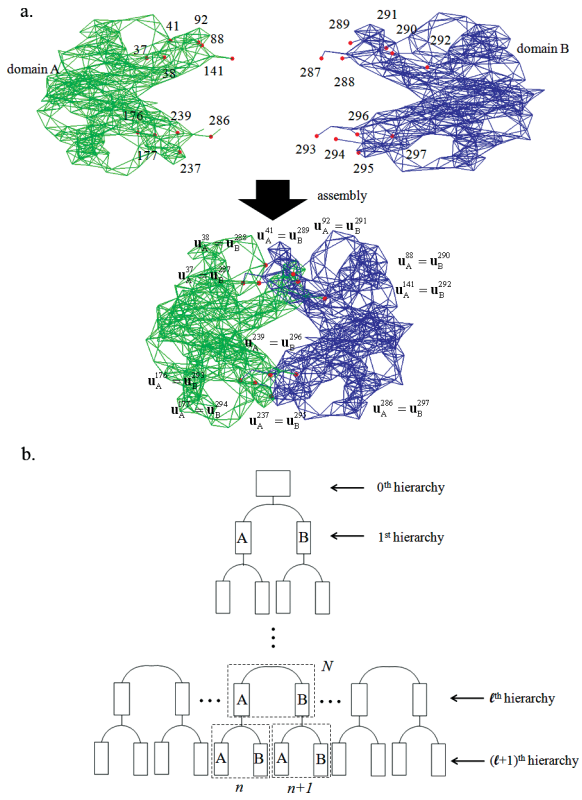


Figure 1. (a) Schematic illustration of component mode synthesis (CMS) applied to hemoglobin. Here, hemoglobin is decomposed into 2 subdomains (subdomains *A* and *B*). The red dotted points represent the nodal points (residues) belonging to interface between 2 subdomains. Here, for nodal points at interface, the geometric constraints are imposed such that displacement vectors of residues at interface are continuous. (b) Schematic illustration of hierarchical component mode synthesis (hCMS) which decomposes the protein structure into a hierarchy composed of several subdomains. The eigenvalue problem for the *n*-th subdomain or (*n*+1)-th subdomain in the (*l*+1)-th hierarchy is solved for obtaining the normal modes of the *N*-th subdomain in the *l*-th hierarchy. This process is repeated until one runs into the 0-th hierarchy. In this work, we perform the hierarchical decomposition of protein into subdomains equivalent to protein domains.

2.2. Component Mode Synthesis (CMS). For normal-mode analysis of a protein structure, we have employed the component mode synthesis (CMS), which has been broadly utilized in engineering mechanics.[31−33] For a clear description of CMS, we consider the protein structure (e.g., hemoglobin shown in Figure 1(a)) which is decomposed into 2 subdomains. The motion of subdomain *A* is constrained by the subdomain *B*, and vice versa, in such a way that the nodal points (residues) at interface between 2 subdomains are constrained under the continuity of a displacement field. In other words, the interactions between 2 subdomains are prescribed by constraints at interface between 2 subdomains. The correlation between motions of 2 subdomains with using geometric constraints is a generic computational scheme in CMS used in structural dynamics[31−33] rather than using the domain−domain interaction directly. In general, component mode synthesis is implemented such that the number of nodal points (residues) of a subdomain is much larger than that of nodal points belonging to an interface (relevant to geometric constraint). This indicates that block (subdomain) should be selected in such a way that the degrees of freedom related to geometric constraints (i.e., nodal points related to block−block interaction described by constraint) should be much less than that of block.

Now, for convenience, let us describe the motion of protein structure decomposed into 2 subdomains without applying the constraints at this moment. The constraints will be implemented later in the assembly process. The potential energy without constraints is given by

$$V' = \frac{1}{2}(\mathbf{u}_A^T \mathbf{K}_A \mathbf{u}_A + \mathbf{u}_B^T \mathbf{K}_B \mathbf{u}_B) \tag{4}$$

Here, prime indicates that constraints were not implemented at this moment. $\mathbf{K}_i$ and $\mathbf{u}_i$ represent the stiffness matrix and displacement field for subdomain i (where $i = A$ or B), respectively. In the similar manner, the kinetic energy without constraints is in the form of

$$T' = \frac{1}{2}(\dot{\mathbf{u}}_A^T \mathbf{M}_A \dot{\mathbf{u}}_A + \dot{\mathbf{u}}_B^T \mathbf{M}_B \dot{\mathbf{u}}_B) \tag{5}$$

where $\mathbf{M}_i$ indicates the mass matrix for subdomains i (where $i = A$ or B), and a symbol dot represents the time-derivative. Then, we introduce the linear transformation such that the displacement vector $\mathbf{u}_i$ (where $i = A$ or B) is represented in the form of $\mathbf{u}_i(\mathbf{x}, t) = \Phi_i(\mathbf{x}) \cdot \mathbf{v}_i(t)$, where $\Phi_i(\mathbf{x})$ is the matrix whose column vector is the eigenvector of the stiffness matrix $\mathbf{K}_i$. That is, $\Phi_i(\mathbf{x})$ satisfies the eigen-value problem such as $\mathbf{K}_i\Phi_i(\mathbf{x}) = \Phi_i(\mathbf{x})\Lambda_i$, where Λ_i is the diagonal matrix whose component is the eigen-value of $\mathbf{K}_i$. With transformation, the potential energy and the kinetic energy without constraints can be represented in the space spanned by normal modes of each subdomain.

$$V' = \frac{1}{2}\begin{bmatrix}\mathbf{v}_A^T & \mathbf{v}_B^T\end{bmatrix}\begin{bmatrix}\Lambda_A & 0\\ 0 & \Lambda_B\end{bmatrix}\begin{bmatrix}\mathbf{v}_A\\ \mathbf{v}_B\end{bmatrix} \equiv \frac{1}{2}\mathbf{v}^T\Lambda\mathbf{v} \tag{6.a}$$

$$T' = \frac{1}{2}\begin{bmatrix}\dot{\mathbf{v}}_A^T & \dot{\mathbf{v}}_B^T\end{bmatrix}\begin{bmatrix}\Phi_A^T\mathbf{M}_A\Phi_A & 0\\ 0 & \Phi_B^T\mathbf{M}_B\Phi_B\end{bmatrix}\begin{bmatrix}\dot{\mathbf{v}}_A\\ \dot{\mathbf{v}}_B\end{bmatrix} \equiv \frac{1}{2}\dot{\mathbf{v}}^T\mathbf{L}\dot{\mathbf{v}} \tag{6.b}$$

Here, the displacement vector represented in the normal mode space, $\mathbf{v}^T = [\mathbf{v}_A{}^T\ \mathbf{v}_B{}^T]$, has the degrees of freedom larger than the degrees of freedom of a protein, since the constraints are not imposed.

Now, in order to describe the motion of entire protein domains, we have to impose the geometric constraints as shown in Figure 1(a). For instance, nodal points 37 of the domain *A* colored green is the identical nodal point 287 of the domain *B* colored blue, so that the displacement field for such two nodal points should be continuous, i.e. $\mathbf{u}_A{}^{37} = \mathbf{u}_B{}^{287}$. In general, the geometric constraints for interface between two subdomains can be represented in the form of $\mathbf{Pv} = \mathbf{0}$. Since $\mathbf{v}$ has the redundancy because of redundant enumeration of nodal points at the interface belonging to subdomains *A* and *B*, a vector $\mathbf{v}$ can be decomposed into independent variable $\mathbf{w}(t)$ and dependent variable $\mathbf{z}(t)$. The constraint equation, i.e. $\mathbf{Pv} \equiv \mathbf{P}_1\mathbf{w}(t) + \mathbf{P}_2\mathbf{z}(t) = 0$, leads to the relation of

$$\mathbf{B} = \begin{bmatrix} \mathbf{I} \\ -\mathbf{P}_2^{-1}\mathbf{P}_1 \end{bmatrix} \tag{7}$$

where **B** is the constraint matrix. Then, with the application of constraints, the potential energy and the kinetic energy for a protein structure, respectively, become

$$V = \frac{1}{2}\mathbf{w}^T(\mathbf{B}^T\Lambda\mathbf{B})\mathbf{w} \equiv \frac{1}{2}\mathbf{w}^T\mathbf{D}\mathbf{w} \tag{8.a}$$

$$T = \frac{1}{2}\dot{\mathbf{w}}^T(\mathbf{B}^T\mathbf{L}\mathbf{B})\dot{\mathbf{w}} \equiv \frac{1}{2}\dot{\mathbf{w}}^T\mathbf{S}\mathbf{w} \tag{8.b}$$

Here, **D** and **S** are the stiffness matrix and mass matrix, respectively, for a protein structure, represented in the space spanned by the normal modes of subdomains. It should be noted that the potential energy V and the kinetic energy T given by eqs 8.a and 8.b, respectively, are the exact form for potential energy and kinetic energy for a protein structure composed of 2 subdomains, since the geometric constraints that describe the domain−domain interaction are imposed.

The normal-mode analysis of a protein structure can be, thus, represented by the following eigen-value problem such as $\mathbf{DU} = \mathbf{SU}\Omega$, where **U** is the modal matrix and Ω is the diagonal matrix whose component is the eigen-value of a protein structure. In order to describe the protein dynamics with respect to normal modes, the modal matrix **U** has to be transformed into matrix **Z**, whose column vectors represent the normal modes, such as $\mathbf{Z} = \Phi\mathbf{B}\mathbf{U}$, where Φ is the matrix given by $\Phi^T = [\Phi_A{}^T\ \Phi_B{}^T]$.

As stated above, the component mode synthesis for a protein with two domains is straightforward and easy to be implemented. However, a large protein complex has several rigid domains, so that we need to consider component mode synthesis with several subdomains. For such a large protein, we introduce the hierarchical component mode synthesis (hCMS), which adopts the component mode synthesis in a hierarchical manner. As shown in Figure 1(b), we divide the protein structure into subdomains in a hierarchical manner. Let us denote the index l representing the index of hierarchy and the index n to indicate the index of subdomain in the l-th hierarchy. The process for hCMS is given as follows:

(i) Define the stiffness matrix $\mathbf{K}_n^{(l+1)}$ and the mass matrix $\mathbf{M}_n^{(l+1)}$ for the n-th subdomain in the $(l+1)$-th hierarchy.

(ii) Solve the eigen-value problem with using $\mathbf{K}_n^{(l+1)}$ in order to obtain the normal modes of subdomains to construct the matrices such as $\Phi_{A,n}{}^{(l+1)}$ and $\Phi_{B,n}{}^{(l+1)}$.

(iii) Convert the stiffness matrix $\mathbf{K}_n^{(l+1)}$ using normal modes $\Phi_{A,n}{}^{(l+1)}$ and $\Phi_{B,n}{}^{(l+1)}$, that is, find the matrix $\mathbf{D}_n^{(l+1)}$. Similarly, transform the mass matrix $\mathbf{M}_n^{(l+1)}$ to obtain the matrix $\mathbf{S}_n^{(l+1)}$.

(iv) From the eigen-value problem $\mathbf{D}_n^{(l+1)}\mathbf{U}_n^{(l+1)} = \mathbf{S}_n^{(l+1)}\mathbf{U}_n^{(l+1)}\mathbf{\Omega}_n^{(l+1)}$, the normal mode $\mathbf{Z}_n{}^l$ for the n-th subdomain in the l-th hierarchy can be obtained.

(v) Repeat the process (i)−(iv) until the normal modes $\mathbf{Z}_n{}^l$ for every subdomain in the l-th hierarchy are obtained.

(vi) Set the normal modes $\mathbf{Z}_n{}^l$ as eigen-modes $\Phi_{A,N}{}^l$ for the subdomain A belonging to the N-th subdomain in the l-th hierarchy. Similarly, the normal modes $\mathbf{Z}_{n+1}{}^l$ is set to the eigen-modes $\Phi_{B,N}{}^l$.

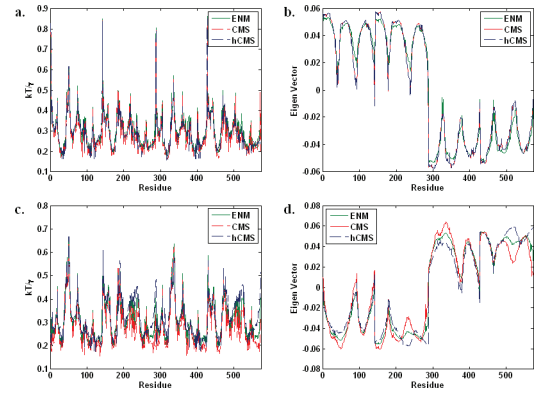


Figure 2. (a) Debye−Waller temperature factors obtained from ENM, CMS (consisting of two subdomains), and hCMS (composed of four subdomains) for hemoglobin in close form (pdb: 1a3n), (b) lowest-frequency normal mode (excluding the zero modes corresponding to rigid body motions) obtained from ENM, CMS, and hCMS for hemoglobin (pdb: 1a3n), (c) Debye−Waller temperature factor of hemoglobin in close form (pdb: 1bbb) described by ENM with cutoff radius of $R_c = 7$ Å, CMS based on such an ENM, and hCMS, and (d) lowest-frequency normal mode for hemoglobin (pdb: 1bbb) obtained from such an ENM, CMS, and hCMS.

(vii) Repeat the process (i)−(vi) until one runs into the 0-th hierarchy.

Here, it should be noted that we perform the hierarchical decomposition of a protein structure until each subdomain has the appropriate degrees of freedom (see Results and Discussion).

Results and Discussion

We considered the model proteins such as hemoglobin (in open and close forms), citrate synthase, and motor protein F_0-ATPase. These proteins have several subdomains, i.e. 2 subdomains for citrate synthase, 4 subdomains for hemoglobin, and 13 subdomains for F_0-ATPase, so that CMS (or hCMS) is applicable to understanding the dynamics of such proteins as well as low-frequency normal modes relevant to conformational change.

Conformational Dynamics of Hemoglobin. We consider the hemoglobin, which is a good model protein that is well described by NMA and ENM. Hemoglobin consists of 4 subdomains (chains) such as α_1, β_1, α_2, and β_2 chains. In our study, we take into account two types of CMS for a description of hemoglobin dynamics. Specifically, we first consider the CMS, which regards the protein structures as a combination of 2 subdomains. We also take into account hCMS to describe the hemoglobin structure as two subdomains, each of which is composed of α and β chains.

First, let us take into account the mean-square fluctuation (MSF) of hemoglobin, obtained by ENM, CMS (hemoglobin composed of 2 subdomains), and hCMS (hemoglobin consisting of 4 subdomains). At this moment, we employed the hemoglobin in close form (pdb: 1a3n), in which a ligand is bounded. Here, the force constant for ENM is given as $\gamma = 0.886$ kcal/mol·Å^2 with setting $R_c = 7$ Å by comparing the Debye−Waller factor (B-factor). Figure 2(a) shows the conformational fluctuation behavior predicted by ENM, CMS, and hCMS. This

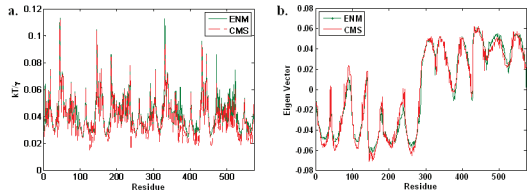


Figure 3. (a) Debye−Waller temperature factor for hemoglobin (pdb: 1bbb) described by ENM with cutoff radius of R_c = 12 Å, and CMS based on such an ENM, and (b) lowest-frequency normal modes computed from such an ENM and CMS.

indicates the robustness of CMS or hCMS for analyzing the conformational fluctuation behavior. More specifically, in order to validate the robustness of CMS or hCMS, we have also considered the lowest-frequency normal mode (with excluding the zero normal modes corresponding to rigid body motion), which is highly related to the conformational change of protein. As shown in Figure 2(b), normal modes obtained from ENM and CMS (or hCMS) are almost identical to each other with a correlation of $r > 0.99$. This indicates that low-frequency normal mode is well dictated by CMS (or hCMS). However, for a hemoglobin in open form (pdb: 1bbb) described by ENM with using $R_c = 7$ Å, the low-frequency normal mode (or B-factor) anticipated from CMS (or hCMS) is similar but not exactly identical to that of ENM (see Figure 2(c) and (d)). Even though the constraint condition is changed (i.e., different definition of atoms at interface), the quality of normal mode or B-factor computed from CMS (or hCMS) has not been improved (not shown). This may be attributed to a protein structure such that the structural feature of hemoglobin in an open form may not be well depicted by ENM with $R_c = 7$ Å. For validation of this conjecture, we describe the hemoglobin by using ENM with $R_c = 12$ Å and its related CMS (or hCMS). It is shown that the lowest-frequency normal modes computed from ENM and CMS (or hCMS) are almost identical to each other, based on protein topology dictated by $R_c = 12$ Å (see Figure 3(a) and (b)).

For further investigation of the quality of normal modes estimated from CMS (or hCMS), we have introduced the parameter, referred to as *overlap*,[34] defined as $\chi_{ij} = \mathbf{v}_i^{ENM} \cdot \mathbf{v}_j^{CMS}$, where $\mathbf{v}_i^{ENM}$ and $\mathbf{v}_j^{CMS}$ are the i-th and j-th normal modes obtained from ENM and CMS, respectively. The quantity χ_{ij} close to 1 indicates the high correlation (similarity) between the i-th normal mode obtained from ENM and the j-th normal mode computed from CMS, while such a quantity close to zero represents that two normal modes are rarely correlated. We have shown the density map of *overlap* for hemoglobin in both forms (see Figure 4). It is remarkable that low-frequency normal modes, which play a role in conformational change, obtained from CMS are highly correlated to those estimated from ENM. This suggests the robustness of CMS (or hCMS) in the prediction of low-frequency normal modes that are typically involved in conformational change. However, the correlation between high-frequency normal modes obtained from ENM and CMS is decreased. This provides that high-frequency normal modes, related to localized motion (e.g., fast motion of receptor due to ligand−receptor binding), may not be anticipated from CMS (or hCMS). Further, for quantifying

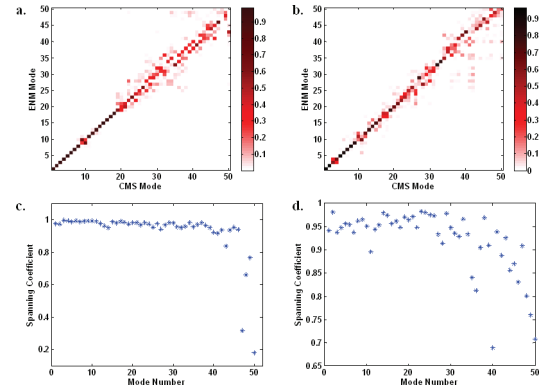


Figure 4. Overlap between normal modes obtained from ENM and CMS for hemoglobin in (a) close form (pdb: 1a3n) and (b) open form (pdb: 1bbb). Spanning coefficient between normal modes computed from ENM and CMS for hemoglobin in (c) close form and (d) open form. It is shown that low-frequency normal modes estimated from CMS are highly correlated with those from ENM.

the correlation between normal modes computed from ENM and CMS, we have adopted the quantity such as *spanning coefficient*,[34] defined as $\delta_i = \sum_{j=1}^{M} \chi_{ij}$. The spanning coefficient indicates that a normal mode of CMS can be spanned by M normal modes of ENM. Here, we set $M = 50$, which means that the spanning coefficient indicates how much a normal mode of CMS can be represented by 50 low-frequency normal modes of ENM. As shown in Figure 4(c) and (d), low-frequency normal modes (up to ∼40 normal modes) of CMS can be well dictated by the space spanned by 50 low-frequency normal modes of ENM. This implies that, for hemoglobin, low-frequency normal modes (up to ∼40 low-frequency modes) can be well predicted by CMS (or hCMS).

Conformational Dynamics of Model Proteins. We have considered several model proteins such as citrate synthase in open, and close forms, and F0-ATPase motor protein. We compared the B-factors of model proteins obtained from CMS (or hCMS) with those estimated from ENM. Similar to the case of hemoglobin, the B-factors are well reproduced by CMS (or hCMS), indicating the robustness of CMS for predicting the fluctuation behavior. For further validation, we have taken into account the lowest-frequency normal modes of model proteins obtained from ENM as well as CMS (or hCMS). It is shown that low-frequency normal modes anticipated from CMS (hCMS) are almost identical to those evaluated from ENM. With similar analyses on *overlap* and *spanning coefficient*, the low-frequency normal modes of model proteins related to their biological function (e.g., conformational change) can be well dictated by CMS (hCMS). For details, see the Supporting Information.

Effect of Constraints on Protein Dynamics Described by CMS. Constraint equation is essential in CMS (or hCMS) in order to convert the stiffness matrix (and/or the mass matrix) into that spanned by normal modes of subdomains. It is assumed that the boundary nodal points were selected in such a way that two residues belonging to two different subdomains are boundary nodal points at the interface between two subdomains if the distance between two such residues is within a certain distance, referred to as search

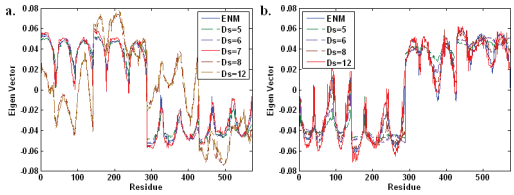


Figure 5. Lowest-frequency normal modes obtained from CMS with different types of constraint (i.e., different search distances D_S) for hemoglobin in (a) close form (pdb: 1a3n) and (b) open form (pdb: 1bbb). It is shown that the constraint should be selected such that $D_S \sim R_c$, indicating that overall stiffness should be maintained when constraint is determined.

distance D_S. If D_S is very small, then two subdomains will be less constrained so that the whole protein structure will be more flexible than it is. On the other hand, if D_S is very large, then two subdomains are so constrained that the protein structure is more rigid than it is. Figure 5 depicts the low-frequency normal modes of hemoglobin obtained from CMS with different constraints D_S. Here, the structure of hemoglobin in close form (pdb: 1a3n) is described by ENM with $R_c = 7$ Å, and D_S is varying from 5 Å to 12 Å. For $D_S = 5$ Å the number of constraints is 4, whereas for $D_S = 12$ Å the total number of constraints is 25. As shown in Figure 5(a), as long as D_S is in the range of 5 Å−7 Å, the lowest-frequency normal mode obtained from CMS is quantitatively comparable to that computed from ENM. In the case of hemoglobin in open form (pdb: 1bbb), we describe its structure as a Gaussian Network Model (GNM) with a cutoff distance of $R_c = 12$ Å. For such a case, the CMS with use of $D_S = 12$ Å provides the lowest-frequency normal mode quantitatively comparable to that estimated from GNM (Figure 5(b)), while the functional low-frequency normal mode cannot be dictated by the CMS with $D_S < 12$ Å. These two examples suggest that the search distance D_S for constraint should be chosen such that D_S is quantitatively comparable to the cutoff distance R_c used in ENM (or GNM). This may be attributed to the fact that, if $D_S \sim R_c$, the overall stiffness of a protein described by CMS is close to that dictated by the original structure. It is implied that the constraint should be selected as long as the constraint does not affect the overall stiffness of the protein structure responsible for conformational fluctuation dynamics.

Conformational Fluctuation Dictated by Normal Modes of CMS. The key of CMS is to transform the stiffness matrix **K** in the Cartesian coordinate into that represented in the space G spanned by normal modes of subdomains. This leads the matrix **S** (i.e., stiffness matrix represented in the space G) to be a diagonal matrix, and, consequently, it improves the computational efficiency to obtain the natural frequencies and their related normal modes for a protein. Here, we have further considered the space, in which **S** is represented, spanned by some normal modes (from lowest-frequency mode to a certain frequency normal mode) rather than all normal modes of subdomains. Here, such a space is denoted as G^*. This representation in G^* will enhance the computation of functional low-frequency normal modes of proteins as well as their conformational fluctuation. Figure 6 shows the Debye−Waller temperature factors of model proteins

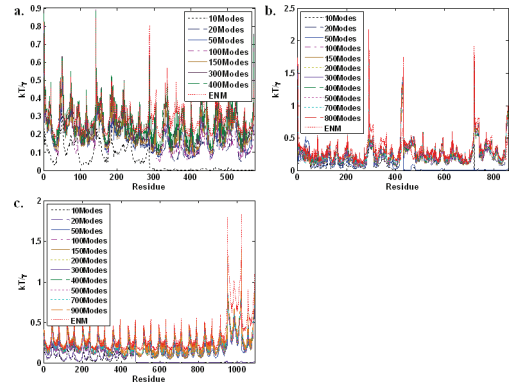


Figure 6. Debye−Waller temperature factors obtained from CMS in reduced space G* spanned by some normal modes (i.e., 10 ∼ O(N) normal modes, where N is the total degrees of freedom) for (a) hemoglobin (pdb: 1a3n), (b) citrate synthase (pdb: 5csc), and (c) F0-ATPase motor protein (pdb: 1c17). It is shown that, at least, more than 50 normal modes should be used for space G* in CMS.

obtained from both ENM and CMS which employs the different number of normal modes spanning the space G^* for **S**. It is shown that, at least, more than 50 normal modes should be utilized in CMS in order to have the physically meaningful conformational fluctuation information. In order to quantify the correlation of normal modes between ENM and CMS with the use of a different number of normal modes, we have introduced the correlation parameter r defined as[35]

$$r = \frac{\sum_{i=1}^{N}(B_i^{ENM} - \langle B^{ENM}\rangle)(B_i^{CMS} - \langle B^{CMS}\rangle)}{\sqrt{\sum_{i=1}^{N}(B_i^{ENM} - \langle B^{ENM}\rangle)^2 \sum_{j=1}^{N}(B_j^{CMS} - \langle B^{CMS}\rangle)^2}} \quad (9)$$

Here, B_i^{ENM} and B_i^{CMS} indicate the Debye−Waller temperature factor for residue i obtained from ENM and CMS, respectively, N is the total number of residues, and angle brackets < > represent the average given by $\langle B\rangle = (1/N)\sum_{j=1}^{N} B_j$. A value of the correlation parameter r close to 1 indicates that the B-factor obtained from CMS is highly correlated to that from ENM, while a value of r approaching 0 indicates the uncorrelation between B-factors obtained from ENM and CMS, and the value of r close to −1 shows the anticorrelation between B-factors computed from ENM and CMS. As shown in Figure 7, it is shown that, at least, more than 50 normal modes spanning G^* should be employed in CMS in order to gain the B-factors with a correlation of >80% compared to that obtained from ENM. It indicates that, if one utilizes the 50 normal modes spanning the space G^* in CMS, one is able to estimate the Debye−Waller temperature factor comparable to ENM. In other words, 50 normal modes employed in CMS are sufficient to represent the protein dynamics with computational efficiency.

Degree of Hierarchical Decomposition. We have performed the hierarchical component mode synthesis (hCMS) to protein structure until each hierarchical subdomain is identical to the protein domain. For instance, the hCMS has

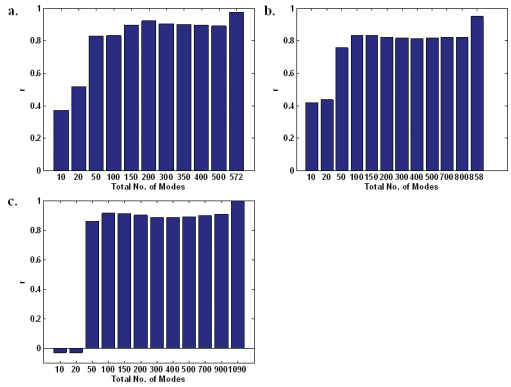


Figure 7. Correlation coefficient between Debye−Waller temperature factors computed from ENM and CMS with different reduced space G* for (a) hemoglobin (pdb: 1a3n), (b) citrate synthase (pdb: 5csc), and (c) F0-ATPase motor protein (pdb: 1c17). It is shown that 50 normal modes spanning space G* in CMS provides the Debye−Waller temperature factor quantitatively comparable to that computed from original structure (ENM) with a correlation of >80%.

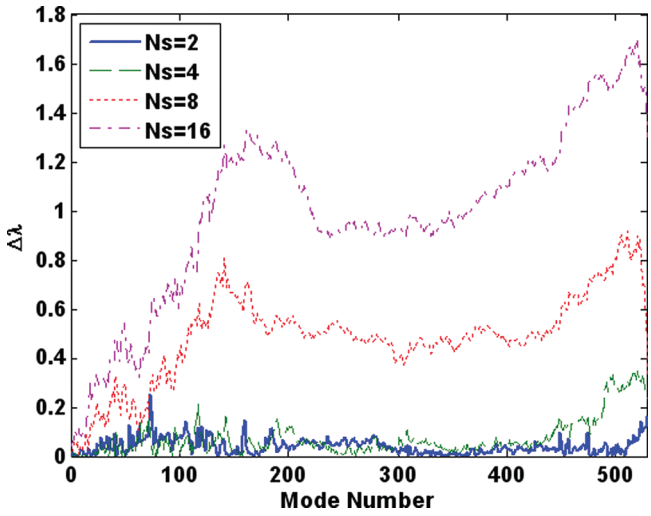


Figure 8. The difference between eigen-values, for hemoglobin, obtained from normal-mode analysis (NMA) and hierarchical component mode synthesis (hCMS) with different hierarchial decomposition. Here, $\Delta\lambda = |\ddot{\lambda}^{NMA} - \ddot{\lambda}^{hCMS}|$, where $\ddot{\lambda}^{NMA}$ and $\ddot{\lambda}^{hCMS}$ represent the eigen-value obtained from NMA and hCMS, respectively. When a hemoglobin is decomposed into 2 or 4 subdomains, the difference of eigen-values, $\Delta\lambda$, is insignificant. However, if a hemoglobin is decomposed into more subdomains, the difference of eigen-values, $\Delta\lambda$, becomes larger. This indicates that hCMS can be implemented until protein strucure is decomposed into protein domains.

been applied to hemoglobin such that the hemoglobin structure is decomposed to 4 subdomains. In order to investigate the available degree of hierarchy for hCMS, we have considered the hemoglobin (composed of 4 protein domains) with different hierarchies — (i) 2 subdomains, (ii) 4 subdomains, (iii) 8 subdomains, and (iv) 16 subdomains. Figure 8 shows the difference between eigen-values, for hemoglobin, obtained from NMA and hCMS with different hierarchies. It is shown that, as long as hemoglobin is decomposed into 2 or 4 subdomains, the difference between eigen-values obtained from NMA and hCMS is insignificant.

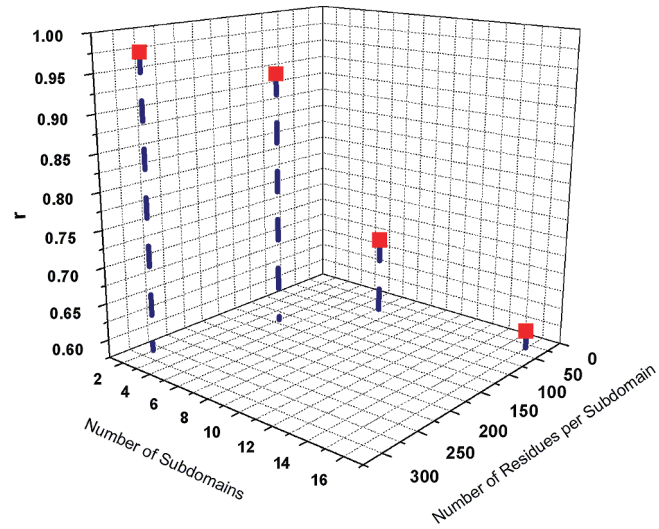


Figure 9. Correlation between thermal fluctations, for hemoglobin, obtained from normal-mode analysis (NMA) and hierarchical component mode synthesis (hCMS). It is shown that, as long as a hemoglobin is split into 2 or 4 subdomains, the fluctuation behavior (Debye−Waller B factor) obtained from hCMS is quantitatively comparable to that from NMA with correlation of >90%. However, if a hemoglobin is decomposed into many subdomains (e.g., 16 subdomains), the fluctuation behavior estimated from hCMS deviates from that predicted from NMA with correlation of <70%. This indicates that our hCMS has to be implemented until a protein structure is decomposed at protein domain level rather than residue level.

On the other hand, as hemoglobin is decomposed into smaller subdomains (i.e., smaller than protein domain), the eigen-values obtained from hCMS are deviated from those obtained from NMA. This indicates that hCMS with decomposition of the protein structure into many subdomains would not provide a good prediction of conformational fluctuation. Specifically, Figure 9 shows the correlation between B-factors, for hemoglobin, obtained from normal-mode analysis (NMA) and hCMS with given hierarchies. It is shown that the hCMS with 2 or 4 subdomains predicts the thermal fluctuation behavior with correlation of >90% to NMA. However, once hemoglobin is decomposed into more than 4 subdomains, then the correlation between B-factors obtained from NMA and hCMS is <80%. As the hemoglobin is decomposed into more subdomains (much larger than number of protein domains), the worse correlation between B-factors obtained from NMA and hCMS is obtained. For instance, if hemoglobin is divided into 16 subdomains (composed of ∼25 residues), then the correlation between B-factors obtained from NMA and hCMS is $r = \sim 60\%$. This indicates that our hCMS has to be implemented such that a protein structure can be decomposed into the subdomains at the protein domain level. This is attributed to the fact that, if protein is decomposed into many subdomains composed of a small number of residues, then the degrees of freedom related to geometric constraint is equivalent to degrees of freedom of subdomains, which leads to the overconstraint of the subdomain. This implies that our

Table 1. Computational Time for Estimating Normal Modes

	model protein		
	F0-ATPase (pdb:1c17)	citrate synthase (pdb:5csc)	citrate synthase (pdb: 6csc)
GNM	63.37 s	27.6 s	25.82 s
CMS ($N_s = 2$, $D_s = 5$ Å)	22.19 s	9.77 s	NA
CMS ($N_s = 2$, $D_s = 6$ Å)	22.76 s	9.93 s	10.33 s
CMS ($N_s = 2$, $D_s = 7$ Å)	23.73 s	10.09 s	10.61 s

hCMS can be implemented at the protein domain level rather than the residue level.

Computational Cost for hCMS. In order to compare the computational speed of hCMS with that of general NMA, we have measured the computation time to estimate the thermal fluctuation of model proteins, composed of >1000 residues, based on hCMS and NMA. As shown in Table 1, hCMS with different constraints exhibits the faster computation on low-frequency normal modes and conformational fluctuation than general NMA by a factor of ~2. This indicates that our hCMS enhances the computational time to estimate the conformational fluctuation by a factor of ~2 and that our hCMS can be applicable to iterative NMA of a large protein complex for understanding their conformational transition. The fast computation of fluctuation dynamics using CMS is attributed to fact that the representation of the stiffness matrix in the normal modes of subdomains enhances the computation in solving the eigen-value problem.[31−33]

Conclusion

We demonstrate the application of component mode synthesis (CMS) or hierarchical CMS (hCMS) for the conformational dynamics of a large protein complex. We have shown that hCMS enables the computationally efficient estimation of functional low-frequency normal modes, the Debye−Waller temperature factor, and correlated motion. The key of hCMS (or CMS) is to represent the stiffness matrix in the space G spanned by normal modes of subdomains. Moreover, it is shown that the reduced space G^* allows us to depict the large protein dynamics with enhanced computational efficacy. This computationally efficient hCMS may improve the computational estimation of conformational transition between two equilibrium states, which is usually computed from iterative normal-mode analysis with certain constraints.[16,17] In the long run, such hCMS will enable the understanding of the functional motion of large protein complexes as well as their energy landscape for conformational transition described by iterative normal-mode analysis.

Acknowledgment. This work was supported by KOSEF (Grant No. R11-2007-028-03002) and KRF (Grant No. KRF-2008-314-000012).

Supporting Information Available: Results for conformational dynamics of model proteins with hierarchical component mode synthesis. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Cui, Q.; Bahar, I. *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*; CRC Press: 2005.

(2) Bahar, I.; Rader, A. J. *Curr. Opin. Struct. Biol.* **2005**, *15*, 586–592.

(3) Tama, F.; Brooks, C. L. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 115–133.

(4) Brooks, B.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1983**, *80*, 6571–6575.

(5) Janezic, D.; Venable, R. M.; Brooks, B. R. *J. Comput. Chem.* **1995**, *16*, 1554–1566.

(6) Hayward, S.; Go, N. *Annu. Rev. Phys. Chem.* **1995**, *46*, 223–250.

(7) Cieplak, M.; Hoang, T. X.; Robbins, M. O. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 114–124.

(8) Cieplak, M.; Hoang, T. X.; Robbins, M. O. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 104–113.

(9) Sulkowska, J. I.; Cieplak, M. *Biophys. J.* **2008**, *95*, 3174–3191.

(10) Yoon, G.; Park, H.-J.; Na, S.; Eom, K. *J. Comput. Chem.* **2009**, *30*, 873–880.

(11) Tirion, M. M. *Phys. Rev. Lett.* **1996**, *77*, 1905–1908.

(12) Bahar, I.; Atilgan, A. R.; Demirel, M. C.; Erman, B. *Phys. Rev. Lett.* **1998**, *80*, 2733–2736.

(13) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. *Biophys. J.* **2001**, *80*, 505–515.

(14) Xu, C. Y.; Tobi, D.; Bahar, I. *J. Mol. Biol.* **2003**, *333*, 153–168.

(15) Tobi, D.; Bahar, I. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 18908–18913.

(16) Miyashita, O.; Onuchic, J. N.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12570–12575.

(17) Zheng, W. J.; Brooks, B. R. *Biophys. J.* **2005**, *88*, 3109–3117.

(18) Whitford, P. C.; Miyashita, O.; Levy, Y.; Onuchic, J. N. *J. Mol. Biol.* **2007**, *366*, 1661–1671.

(19) Maragakis, P.; Karplus, M. *J. Mol. Biol.* **2005**, *352*, 807–822.

(20) Zheng, W.; Brooks, B. R.; Hummer, G. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 43–57.

(21) Doruker, P.; Jernigan, R. L.; Bahar, I. *J. Comput. Chem.* **2002**, *23*, 119–127.

(22) Eom, K.; Ahn, J. H.; Baek, S. C.; Kim, J. I.; Na, S. *CMC: Comput. Mater. Continua* **2007**, *6*, 35–42.

(23) Eom, K.; Baek, S.-C.; Ahn, J.-H.; Na, S. *J. Comput. Chem.* **2007**, *28*, 1400–1410.

(24) Cheng, H.; Gimbutas, Z.; Martinsson, P. G.; Rokhlin, V.; SIAM, J. *Sci. Comput.* **2005**, *26*, 1389–1404.

(25) Kurkcuoglu, O.; Jernigan, R. L.; Doruker, P. *Polymer* **2004**, *45*, 649–657.

(26) Kim, M. K.; Jernigan, R. L.; Chirikjian, G. S. *Biophys. J.* **2005**, *89*, 43–55.

(27) Tama, F.; Gadea, F. X.; Marques, O.; Sanejouand, Y. H. *Proteins: Struct., Funct., Genet.* **2000**, *41*, 1–7.

(28) Lu, M.; Ma, J. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 15358–15363.

(29) Weiner, J. H. *Statistical mechanics of elasticity*; Dover Publication: 1983.

(30) Chandler, D. *Introduction to modern statistical mechanics*; Oxford University Press: 1987.

(31) Meirovitch, L. *Computational methods in structural dynamics*; SIJTHOFF & NOORDHOFF: Rockville, Maryland, USA, 1980.

(32) Bhat, R. B. *J. Sound Vib.* **1985**, *101*, 271–272.

(33) Sung, S. H.; Nefske, D. J. *AIAA J* **1986**, *24*, 1021–1026.

(34) Van Wynsberghe, A. W.; Cui, Q. *Biophys. J.* **2005**, *89*, 2939–2949.

(35) Kondrashov, D. A.; Cui, Q.; Phillips, G. N., Jr. *Biophys. J.* **2006**, *91*, 2760–2767.

CT900027H

JCTC Journal of Chemical Theory and Computation

Structure Prediction of Bis(amino acidato)copper(II) Complexes with a New Force Field for Molecular Modeling

Jasmina Sabolović*,† and Vjeran Gomzi‡

Institute for Medical Research and Occupational Health, Ksaverska cesta 2, P.O. Box 291, HR-10001 Zagreb, Croatia, and Ruđer Bošković Institute, Bijenička cesta 54, HR-10000 Zagreb, Croatia

Received January 12, 2009

Abstract: This article presents a new force field whose parameterization was based on experimental crystal data and quantum chemically obtained vacuum structures of a series of copper(II) complexes with aliphatic α-amino acids and their *N*-alkyl derivatives, along with the SPC/E water model. The ability of the new force field to reproduce and predict the structural properties of the copper(II) complexes in the gas phase, in simulated crystalline surroundings, and solvated in water is examined. Molecular dynamics (MD) simulations with the new force field yielded time-average structural coordinates of bis(glycinato)copper(II) [the only one of 25 modeled bis(amino acidato)copper(II) systems with published experimental structural data in aqueous solution at room temperature] within the experimental error values. The study of the cis−trans isomerization of bis(glycinato)copper(II) in aqueous medium at 300 K using the quantum chemical polarized continuum model revealed a small energy difference (5 kJ mol^{-1}) between the solvated cis and trans minima, in line with the MD energy estimations. The new force field proved promising in predicting the association of the complexes in aqueous solution and formation of a nucleus of crystallization.

Introduction

Copper, like other essential transition metals (iron, zinc, cobalt, manganese), is present in many biological fluids as the free ion or complexed in metalloproteins and low-molecular-weight complexes with peptides and amino acids.[1−3] In healthy organisms, physiological copper concentrations are maintained by a number of homeostatic mechanisms, such as absorption regulation, cellular uptake and efflux, intracellular transport, sequestration/storage, and copper excretion from the body.[4,5] Exposure to excess copper through an accident, occupational hazard, environmental contamination, or human genetic disorder (Menkes disease, occipital horn syndrome, or Wilson disease)[2,4] causes copper overload, disruptions to normal copper homeostasis, copper-induced oxidative damage, and toxic effects in organs.[5] From available data on human exposures worldwide, there is a greater risk of health effects from copper deficiency (which might also increase cellular susceptibility to oxidative damage)[5] than from excess copper intake.[6]

Copper is required as a cofactor for structural and catalytic activity in a number of enzymes (e.g., cytochrome *c* oxidase, lysyl oxidase, tyrosinase, superoxide dismutase).[3] An analysis of the types and frequencies of amino acid residues involved in the coordination of metal ions in metalloproteins that was performed on a set of structures extracted from the Protein Data Bank in October 2007 showed that copper preferred the coordination number 4 and that it was most often coordinated by histidine imidazole nitrogen atom, followed by cysteine sulfur atom.[7] L-Histidine was also identified as the predominant amino acid bound to copper(II) in bis-(L-histidinato)copper(II) (with imidazole nitrogen, amino nitrogen, and carboxylato oxygen donor atoms) and in mixed

* Corresponding author e-mail: jasmina.sabolovic@imi.hr; phone: +385 1 4673 188; fax: +385 1 4673 303.

† Institute for Medical Research and Occupational Health.

‡ Ruđer Bošković Institute.

10.1021/ct9000203 CCC: $40.75
Published on Web 06/02/2009

copper(II) complexes with other L-amino acids in human blood serum.[3]

Whereas metalloproteins are relatively easy to isolate, low-molecular-weight complexes form parts of multicomponent systems with different complexing species in solution. Most spectroscopic and electrochemical methods allow identification of the prevailing complexes and their stability constants in solutions, but they provide no structural information.[3,8−19] For instance, trans and cis isomers for a number of bis(amino acidato)copper(II) complexes were confirmed to exist in aqueous solution by the ^{14}N superhyperfine structure in electron paramagnetic resonance (EPR) spectra without details on their geometry.[16−19] Some structural data only for bis(glycinato)copper(II), $Cu(Gly)_2$, in aqueous solution were obtained by X-ray absorption spectroscopy,[20] whereas for other copper(II) chelates with aliphatic α-amino acids and their *N*-alkyl derivatives, the only experimentally available structures are those determined by X-ray and neutron diffraction studies.[21−43]

The structural properties can be predicted and reproduced using the molecular modeling methods, for example, molecular dynamics (MD) simulations. However, prerequisite for MD calculations is a reliable molecular mechanics (MM) force field.[44] To the best of our knowledge, such a force field has not yet been reported for low-molecular-weight transition metal complexes with amino acids.

The development of MM models and force fields for transition metal complexes has been complicated by the large number of transition metal elements; system-specific metal−ligand interactions; and the diversity of oxidation states, coordination numbers, and geometrical shapes around metal centers, together with the lack of experimental data for force field parameterization.[45−49] Both the development of quantum chemical methods and the expansion of computer power have contributed to an increase of quantum chemical data suitable for force field parameterization and/or validation of the MM results during past decade. Next, we mention several parameterization studies involving copper complexes.

The polarizable MM procedure SIBFA (sum of interactions between fragments ab initio computed) was used to treat copper(II) in complexes with HIV-1 protease inhibitors, N_1-(4-methyl-2-pyridyl)-2,3,6-trimethoxybenzamide, and N_2-(2-methoxybenzyl)-2-quinolinecarboxamide to study the structural and energetic aspects, as well as to compare the relative stability of the complexes.[50]

A ReaxFF reactive force field, in which the parameters were fitted to a substantial database of quantum chemical data (binding energies, ground-state systems, full reactive pathways), was developed for reactions involving carbon materials and transition metal atoms, including copper, usually employed in catalytic transformations.[51] The force field was applied for high-temperature MD simulations in the presence of metal atoms and carbon fragments, which demonstrated different catalytic abilities of the metals in the formation of small polycyclic structures serving as nucleation points for further nanostructure formation.

Given the need for an accurate and general force field for type 1 copper binding sites in the copper(II) form of blue copper proteins involved in electron transport, Comba and Remenyi[52,53] based the force field parameterization on potential energy curves [computed energy vs bond distance (valence angle) in the gas phase by density functional theory (DFT)] of the model compound $[Cu(imidazole)_2(SCH_3)S(CH_3)_2]^+$. DFT curves were used to fit corresponding MM curves by least-squares fitting and thus to develop the force field parameters.

The ligand-field MM (LFMM) model was also applied to model compounds for the oxidized type 1 copper center.[54,55] The parameters were developed on the basis of DFT geometries optimized in vacuo for the homoleptic model compounds $[Cu(SCH_3)_4]^{2-}$, $[Cu(S(CH_3)_2)_4]^{2+}$, $[Cu(imidazole)_4]^{2+}$, and $[Cu(O{=}CH_2)_4]^{2+}$; validated on the active-site model complex $[Cu(imidazole)_2(SCH_3)(S(CH_3)_2]^+$; and then applied for modeling of the active sites of 24 crystallographically characterized blue copper proteins surrounded with a layer of water molecules.[55] The development and applications of the LFMM model for transition metal complexes were summarized in a recent review.[56]

We have developed MM models and force fields for copper(II) complexes with aliphatic α-amino acids with the aim of reproducing and predicting the properties of the whole class of bis(amino acidato)copper(II) complexes (by subsequent inclusion of other amino acids in the force field parameterization) in different environments (vacuum, crystal, solution).[31,57−59] To achieve this aim, the environmental effects are calculated explicitly; that is, the same set of potential energy functions and their empirical parameters are used both for modeling the isolated systems and for simulating condensed phases. In the latter case, the effects of the surrounding environment of a molecule (such as crystal and solution) are calculated explicitly by including the environment in the calculations. The approach affects the force field parameterization based on available experimental crystal data and the results of quantum chemical studies.[31,58,59]

The experimental crystal and molecular structures of anhydrous and aqua copper(II) complexes with aliphatic α-amino acids and their *N*-alkyl derivatives[21−43] differ with respect to the copper(II) coordination polyhedron geometries and intermolecular interactions, which makes this class of complexes challenging for modeling. Copper(II), as a Jahn−Teller (or pseudo-Jahn−Teller) center,[60] adopts diverse coordination geometries such as irregular square-planar, distorted planar, flattened tetrahedral, elongated octahedral, distorted octahedral, and irregular or distorted square pyramidal in the crystal state.[59] In addition to the usual intermolecular van der Waals and hydrogen-bonding interactions, copper(II) can have weak coordinative bonding with an axially placed water oxygen atom and/or carbonyl oxygen atom from an adjacent molecule in the crystal lattice. The exceptions are truly four-coordinate trans copper(II) chelates with *N*,*N*-dialkylated valine,[21,24] isoleucine,[23] and alanine,[22,31] bonded only via van der Waals interactions. Any newly solved crystal and molecular structure of a copper(II) chelate with amino acids is very welcome because it enlarges the collection of knowledge about structural properties. Such an example is the recent X-ray crystal and molecular structure of anhydrous *trans*-bis(*N*,*N*-diethylglycinato)copper(II), $Cu(Et_2Gly)_2$, in which a relatively short copper-to-axial-

carbonyl-oxygen distance of 2.312 Å was found,[37] whereas in other anhydrous trans copper(II) amino acidates, this distance spans the range from 2.6 to 3.1 Å.[59]

Although many quantum chemical studies on model compounds of metal-binding sites in copper metalloproteins have been performed,[52,54,61,62] only a few have investigated bis complexes of copper(II) with amino acids.[58,59,63−65] Quantum chemical calculations of electronic and geometrical properties,[63−65] the gas-phase potential energy profile for the cis−trans isomerization reaction,[64] water molecules' bindings,[59,63,65] and reaction rate constants[64,65] have been done for $Cu(Gly)_2$. The quantum chemical calculations of equilibrium structures for three anhydrous copper(II) complexes with L-alanine, L-leucine, and L-*N*,*N*-dimethylvaline[58] and four aquabis copper(II) glycinato complexes[59] revealed that the experimental crystal and ab initio derived vacuum geometries differ. Their structure comparisons clarify the influence of the crystal environment on the copper(II) coordination geometry and overall complex geometry, and the modeling of the impact of the intermolecular interactions on the geometry changes represents an additional challenge.[31,37,38,58,59]

Although our most recent force field, FFW,[31,59] proved as reliable for modeling anhydrous and aquabis(amino acidato)copper(II) complexes in vacuo and in the crystal,[37,38,59] the empirical parameters of the nonbonded potential energy functions for the water molecule's oxygen and hydrogen atoms used in FFW yielded an incorrect water density (around 1140 kg m^{-3}) and a too-compact system with almost no diffusion of water molecules in the MD simulations. Specifically, the parameters yielded forces between the water molecules that were too attractive; this did not cause problems in hypothetical motionless equilibrium structures calculated by MM gas-phase and in-crystal simulations, but it did cause problems in MD calculations. Therefore, using the empirical parameters of the SPC/E water model,[66] the FFW force field was reparametrized in an exhaustive effort, and a set of force field empirical parameters equally applicable for vacuum, crystal, and aqueous solution was derived. This article presents the new force field and discusses its ability to reproduce and predict the properties of the copper(II) class of compounds in vacuo and in the crystal by MM calculations, as well as in aqueous solution by MD simulations. The MD results were verified by comparison with the quantum chemical results obtained within the polarized continuum model[67] (PCM) approximation. In addition, standard transition state theory was applied to PCM energies to gain insight into the cis−trans isomerization reaction of $Cu(Gly)_2$ in aqueous solution at 300 K.

Methods

MM Model and Calculations. The conformational (strain) potential energy was calculated from the following basic formula

$$V_{\text{strain}} = V(b) + V(\theta) + V(\varphi) + V(\chi) + V_{\text{LJ}} + V_{\text{Coulomb}} = \sum_{\text{bonds}} D_e(e^{-2\alpha(b-b_0)} - 2e^{-\alpha(b-b_0)} + 1) + \frac{1}{2}\sum_{\substack{\text{valence}\\ \text{angles}}} k_\theta(\theta - \theta_0)^2 + \frac{1}{2}\sum_{\substack{\text{torsion}\\ \text{angles}}} V_\varphi(1 \pm \cos n\varphi) + \frac{1}{2}\sum_{\substack{\text{out-of-plane}\\ \text{angles}}} k_\chi\chi^2 + \sum_{i<j}(A_iA_jr_{ij}^{-12} - B_iB_jr_{ij}^{-6}) + \sum_{l<m} q_lq_mr_{lm}^{-1} \quad (1)$$

Here, b, θ, φ, χ, and r are the bond length; the valence, torsion, and out-of-plane angles, and the nonbonded distance, respectively. D_e, α, and b_0 are empirical parameters for bond stretching (a Morse function); k_θ and θ_0 are empirical parameters for valence-angle bending; and k_χ is an empirical parameter for the out-of-plane deformational potential for the carboxyl groups. Torsional interactions are specified with V_φ and n (height and multiplicity of the torsional barrier, respectively). One torsion per bond was calculated. A and B are one-atom empirical parameters for the van der Waals interactions (a Lennard-Jones 12−6 function). q is a charge parameter. Intramolecular interactions between atoms separated by three or more bonds were considered nonbonded and were calculated with the Lennard-Jones (V_{LJ}) and electrostatic ($V_{\text{Coulomb,NB}}$) potentials, respectively.

The interactions within the copper(II) coordination sphere were modeled using the Morse potential between the copper and ligand donor atoms (two amino nitrogen and two carboxylato oxygen atoms); the repulsive electrostatic potential between the four donor atoms ($V_{\text{Coulomb,1−3}}$); and a torsion-like potential dependent on the "torsion" angle O−N−N−O, with two minima at 0° and 180° that correspond to the cis- and the trans-planar CuN_2O_2 configurations, respectively.[31] It is a model without any explicit valence-angle bending potential for the angles around copper.[31,57−59]

The interactions between the water molecule's oxygen and hydrogen atoms and the atoms of a bis(amino acidato)copper(II) complex were calculated with the Lennard-Jones 12−6 and electrostatic potentials regardless of the water molecule position around the copper(II) complex.[59]

All MM calculations were performed using the modified version[68] of the Lyngby version of the CFF program for conformational analysis.[69−71] The conformational potential energy was minimized for an isolated molecular system (in vacuo or a gas-phase approximation) and for a molecule surrounded by other molecules in the simulated crystal lattice (a condensed-phase approximation). The intermolecular atom−atom interactions were calculated using the same functional forms (Lennard-Jones 12−6 function and Coulombic potential) and empirical parameters as for the intramolecular nonbonded interactions. The crystal simulations were carried out using the Williams variant of the Ewald lattice summation method[72,73] with a spherical and abrupt cutoff limit of 14 Å, and convergence constants of 0.2 $Å^{-1}$, 0.2 $Å^{-1}$, and 0.0 for Coulomb, dispersion, and repulsion lattice summation terms, respectively. A detailed description of the crystal simulations is given elsewhere.[31,37,73]

In the Lyngby-CFF program, the input charge parameters are used for an assignment of fractional atomic charges.[71]

The assignment is done by a special charge redistribution algorithm, which keeps the total charge of the molecules neutral and distributes the charge values in a manner supposed to mimic ab initio results. The charge distribution routine was modified[31] to obtained the assigned fractional charges for the copper(II) chelates with amino acids close to the charges resulting from the natural population analysis[74] (NPA). The reason for selecting NPA charges as a guideline to the charge parameter values has been given elsewhere.[58,68] Our choice of the potential energy functions was based on available options in the Lyngby-CFF program, and the combination of the Lennard-Jones 12−6 and Coulombic potentials is the only one implemented for intermolecular atom−atom interactions in the crystal simulator of the program. Hence, the force field developed belongs to the class of nonpolarizable fixed-charge force fields.[75]

The empirical parameters of the potential energy functions were determined by combining trial-and-error guesses with the optimization algorithm, which was a variant of the general least-squares method (the Levenberg−Marquardt algorithm).[70,71]

MD Simulations. The simulations were performed using the program package Gromacs, version 3.2.1,[76,77] with the new MM force field FFWa-SPCE[78] (see MM Force Field Parameterization section) developed for copper(II) complexes with aliphatic amino acids. MD calculations were carried out for trans and cis isomers of $Cu(Gly)_2$ separately, and also for four $Cu(Gly)_2$ molecules. One $Cu(Gly)_2$ and four $Cu(Gly)_2$ molecules were solvated in rectangular boxes containing 2159 and 5457 water molecules, respectively, and equilibrated for 500 ps. Two nanoseconds of the productive MD phase was accomplished under constant temperature and pressure (298.15 K and 1 bar) using Berendsen T-coupling ($\tau_T = 0.1$ ps) and Berendsen p-coupling ($\tau_p = 0.5$ ps).[79] The time step was 1 fs. The water molecules' geometry was constrained by the SETTLE procedure,[80] whereas all $Cu(Gly)_2$ degrees of freedom were relaxed during the MD simulations. A cutoff limit of 1.5 nm was applied for the calculations of Coulomb and Lennard-Jones 12−6 interactions. The cutoff distance for the short-range neighbor list was set to 1.0 nm.

Quantum Chemical Calculations. Geometries and single-point energies of the stationary points of $Cu(Gly)_2$ [i.e., trans and cis minima, as well as a transition state (TS) structure] in the gas phase and in aqueous solution were investigated by the unrestricted B3LYP hybrid density functional method[81−84] using the LanL2DZ double-ξ basis set[85] to which an additional set of polarization functions[86] on heavy atoms except copper was added. In addition, diffuse functions[87] were added for the oxygen, nitrogen, and carbon atoms. Through the use of the aforementioned basis set, the effective core potentials of Hay and Wadt[88−90] were used to describe the shielding effects of the electrons in copper inner shells. The basis set used is denoted as B3LYP/LanL2DZ{D95v+(d)} throughout this article. All quantum chemical calculations were performed using the Gaussian 03 package program.[91] The choice of the B3LYP method and the basis set used was based on previous studies of the energies and geometries of copper(II)-glycinato systems.[59,63−65] They yielded values similar to those obtained by a higher-level theory method [G3(MP2)-B3] and thus indicated the qualitatively correct behavior of the lower-level method.[64]

The PCM of Tomasi and co-workers,[67] modified by Barone and co-workers,[92] was used to describe the effects of the aqueous medium in the self-consistent reaction field (SCRF) calculations. The environmental temperature was set to 300 K. The water solvent was specified by the dielectric constant of 78.39. The united-atom topological model was applied to solvent radii optimized for the PBE0/6-31G(d) level of theory.[93−95]

The initial atom positions for geometry optimizations of the cis and trans minima and TS structure of $Cu(Gly)_2$ were the Cartesian coordinates of the stationary points obtained from the quantum chemical gas-phase calculations.[64] The TS structure in aqueous solution at room temperature was successfully calculated only upon using a "good enough" initial guess and by applying the synchronous transit-guided quasi-Newton (STQN) method implemented in programs invoked using the QST3 keyword.[96,97] The optimized geometries of the stationary points were verified by frequency calculations to be those of the required optimization state (TS structure or energy minimum). The energies were calculated using the same level of the theory and basis set. Full geometry optimizations of the gas-phase and TS structures obtained in solution were performed to check the optimization process itself, as well as to verify whether the two isomer structures were accessible from the presumed TS structure.

Calculation of the Reaction Rate Constants. To evaluate the reaction rates of the cis−trans isomerization process in water solution (given that, to the best of our knowledge, there are no experimental data on these reaction rate constants) to the gas phase, we applied the equations from standard transition state theory

$$k_{\mathrm{cis}\rightarrow\mathrm{trans}} = \frac{k_B T}{h} \exp\left(-\frac{\Delta G_{\mathrm{TS}\rightarrow\mathrm{cis}}}{RT}\right) \quad (2a)$$

$$k_{\mathrm{trans}\rightarrow\mathrm{cis}} = \frac{k_B T}{h} \exp\left(-\frac{\Delta G_{\mathrm{TS}\rightarrow\mathrm{trans}}}{RT}\right) \quad (2b)$$

where $\Delta G_{\mathrm{TS}\rightarrow\mathrm{cis}}$ and $\Delta G_{\mathrm{TS}\rightarrow\mathrm{trans}}$ represent respective quantum chemically estimated Gibbs free energy differences between the TS structure and the cis and trans minima, h is Planck's constant, k_B is the Boltzmann constant, R is the gas constant, and the temperature (T) is 300 K. The thermal correction to the Gibbs free energy was calculated by adding the zero-point vibrational energy plus thermal rotational−vibrational free energy to the Gibbs free energy at temperature T obtained from the potential energy in the standard way.[91]

Results and Discussion

MM Force Field Parameterization. The force field FFW,[31,59] developed for anhydrous and aqua copper(II) chelates with aliphatic α-amino acids with *trans*- and *cis*-CuN_2O_2 coordination polyhedra, was taken as the initial empirical parameter set for parameter optimization. The empirical parameters of the SPC/E water model[66] were used

to model the interactions of the water molecule's oxygen and hydrogen atoms. As the SPC/E water model uses a harmonic approximation for the bond stretching potential, we fitted the D_e, α, and b_0 empirical parameters of the Morse potential for the O_W—H bond length (where O_W denotes the water oxygen atom) to fulfill the relation[69] $k_b = 2D_e\alpha^2$ (where k_b is the empirical parameter of the quadratic function for the O_W—H bond in the SPC/E model) and to yield average O_W—H bonds equal to 1.0 Å (Table 1). The empirical parameters k_θ and θ_0 for H—O_W—H valence-angle bending, *A* and *B* for the Lennard-Jones potential for O_W and H(O_W) (Table 1), and the partial charges [−0.8476*e* and 0.4238*e* for O_W and H(O_W), respectively] were taken to be the same as in SPC/E and were held constant during the force field parameterization.

The force field parameterization procedure was the same as for the FFW force field; specifically, it was based on the experimental crystal and molecular structures and the quantum chemical B3LYP minimum vacuum structures and NPA charges.[59]

The modeling included 14 anhydrous and 11 aqua copper(II) amino acid complexes with the same atom types. All of the molecules are electrically neutral. Eleven of them have *trans*- and four have *cis*-N_2O_2 coordination to the copper(II).

The empirical parameters were optimized with respect to the experimental data (bond lengths, valence and torsion angles, and unit cell dimensions) of the following seven anhydrous and six aqua copper(II) complexes: *trans*-bis(L-*N*,*N*-dimethylvalinato)copper(II)[21] [Cu(L-Me_2Val)$_2$], *trans*-bis(D,L-*N*,*N*-diethylalaninato)copper(II)[22] [Cu(D,L-Et_2Ala)$_2$], *trans*-bis(L-leucinato)copper(II)[26] [Cu(L-Leu)$_2$], *trans*-bis(D,L-2-aminobutyrato)copper(II)[26] [Cu(D,L-2-aBut)$_2$], *trans*-bis(L-2-aminobutyrato)copper(II)[27] [Cu(L-2-aBut)$_2$], *cis*-bis(D-alaninato)copper(II)[29] [Cu(D-Ala)$_2$], *trans*-Cu(Et_2Gly)$_2$,[37] *trans*-aquabis(L-*N*,*N*-dimethylvalinato)copper(II) hydrate[32] [Cu(L-Me_2Val)$_2$·2H_2O], *trans*-aquabis(L-*N*,*N*-diethylalaninato)copper(II)[33] [Cu(L-Et_2Ala)$_2$·H_2O], *trans*-bis(D,L-alaninato)copper(II) monohydrate[40] [Cu(D,L-Ala)$_2$·H_2O], *trans*-diaquabis(sarcosinato)copper(II)[39] [Cu(Sar)$_2$·2H_2O], *trans*-aquabis(L-*N*,*N*-dimethylalaninato)copper(II) hexahydrate[34] [Cu(L-Me_2Ala)$_2$·7H_2O], and *trans*-aquabis(*N*-*tert*-butyl-*N*-methylglycinato)-copper(II)[36] [Cu(tBuMeGly)$_2$·H_2O]. The empirical parameters were systematically tested on the set of 12 additional copper(II) amino acid complexes, by modeling the crystal and molecular structures of seven anhydrous and five aqua copper(II) complexes: *trans*-bis(D,L-*N*,*N*-dimethylvalinato)-copper(II)[24] [Cu(D,L-Me_2Val)$_2$], *trans*-bis(L-*N*,*N*-dimethylisoleucinato)copper(II)[23] [Cu(L-Me_2Ile)$_2$], *trans*-bis(L-*N*,*N*-dipropylalaninato)copper(II)[31] [Cu(L-Pr_2Ala)$_2$], *trans*-[25] and *cis*-[30] bis(L-alaninato)copper(II), [Cu(L-Ala)$_2$], *trans*-bis(1-aminocyclopentanecarboxylato)copper(II)[25] [Cu(1-Acpc)$_2$], *trans*-bis(α-aminoisobutyrato)copper(II)[28] [Cu(α-aiBut)$_2$], *trans*-aquabis(L-*N*,*N*-dimethylisoleucinato)copper(II)[35] [Cu(L-Me_2Ile)$_2$·H_2O], *trans*-aquabis(*N*,*N*-dimethylglycinato)copper(II) dihydrate[38] [Cu(Me_2Gly)$_2$·3H_2O], *trans*-aquabis(*N*,*N*-dimetylglycinato)copper(II)[38] [Cu(Me_2Gly)$_2$·H_2O], *cis*-bis(glycinato)copper(II) hydrate[41] [Cu(Gly)$_2$·H_2O], *cis*-aquabis(L-isoleucinato)copper(II)[42] [Cu(L-Ile)$_2$·H_2O]. To ensure the planarity of the copper(II) coordination geometry for anhydrous molecules in vacuo as obtained by quantum chemical calculations,[58] the parameters were optimized with

Table 1. Potential Energy Parameter Set[a,b]

Morse Potential Parameters

bond	D_e	α	b_0
Cu—N	219.4140	2.2000	1.9900
N—C	52.4710	3.6000	1.4650
C—C	62.3610	4.0700	1.5150
C—H	101.6000	1.8000	1.0900
O_w—H(O_w)	200.0000	1.4360	0.9800
N—H	93.0000	2.5000	1.0100
C_{pl}—$O_{carbonyl}$	54.6840	3.4800	1.2300
C—C_{pl}	114.1670	3.9500	1.5220
C_{pl}—O	59.1680	2.7000	1.2900
Cu—O	85.7680	2.8500	1.9010

Valence Angle Potential Parameters

valence angle	k_θ	θ_0
Cu—N—H	55.0000	1.9106
H(O_w)—O_w—H(O_w)	91.5390	1.9106
Cu—N—C	240.0000	1.9106
N—C—C	450.0000	1.9106
N—C—H	356.0000	1.9106
H—N—H	310.0000	1.9106
H—C—H	276.8110	1.9106
C—N—H	653.0000	1.9106
H—C—C	304.1030	1.9106
C—C—C	803.6020	1.9106
C—N—C	185.0000	1.9106
C_{pl}—C—H	471.1910	1.9106
N—C—C_{pl}	434.7020	1.9106
C—C—C_{pl}	235.0390	1.9106
O—C_{pl}—$O_{carbonyl}$	717.4690	2.1640
C—C_{pl}—$O_{carbonyl}$	383.8770	2.0940
C—C_{pl}—O	400.0000	2.0250
C_{pl}—O—Cu	1100.0000	2.0070

Torsion Potential Parameters

torsion angle	V_φ	*n*
C—C—C—N	9.5000	3.0000
C—C—N—Cu	6.0500	3.0000
Cu—O—C_{pl}—C	5.6000	−4.0000
O—C_{pl}—C—N	0.7000	12.0000
C_{pl}—O—Cu—N	4.3000	−4.0000
O—Cu—N—C	0.0900	12.0000
O—N—N—O	16.8000	−2.0000

Out-of-Plane Torsion Potential Parameter

out-of-plane torsion angle	k_χ
$O_{carbonyl}$(C_{pl}—C—O)	45.0000

Lennard-Jones Potential Parameters

atom	*A*	*B*
H	80.0000	3.5810
C	773.3199	9.7291
N	4451.0765	119.5843
O	1683.3682	73.8745
O_W	793.3470	25.0000
$O_{carbonyl}$	690.0000	40.3744
Cu	210.0000	4.3919
C_{pl}	310.0000	9.1000
H(O_w)	0.0000	0.0000

Charge Parameter

atom	*q*
O_W	−1.2698
H(O_W)	0.0010

[a] Uncommon symbols: C_{pl}, planar carbon atom; $O_{carbonyl}$, double-bonded oxygen atom, O_W, water oxygen. [b] Units are as follows: D_e, kcal mol^{-1}; α, Å^{-1}; b_0, Å; k_θ, kcal mol^{-1} rad^{-2}; θ_0, rad; V_φ, kcal mol^{-1}; *A*, (kcal mol^{-1} Å^{12})$^{1/2}$; *B*, (kcal mol^{-1} Å^6)$^{1/2}$; *q*, electron units (therefore, the expression for electrostatic interactions has to be multiplied by 332.091 kcal mol^{-1} Å in order to be in kcal mol^{-1}).

respect to the B3LYP valence angles around copper in three trans copper(II) chelates:[58] Cu(L-Me_2Val)$_2$, Cu(L-Ala)$_2$, and Cu(L-Leu)$_2$. They were then tested on the B3LYP calculated gas-phase structures of four aqua copper(II) glycinato complexes:[59] *cis*-Cu(Gly)$_2\cdot H_2O$, *trans*-Cu(Me_2Gly)$_2\cdot 3H_2O$, *trans*-Cu(Sar)$_2\cdot 2H_2O$, and *trans*-Cu(tBuMeGly)$_2\cdot H_2O$

The empirical parameters were reoptimized and fitted with the aim to obtain a potential energy parameter set that could yield the best possible reproduction of experimental crystal structures and in vacuo B3LYP structures, regardless of whether the modeled systems were anhydrous or included interactions with water molecules. A problem that might arise during the parameter fitting process is that several combinations of the potential energy parameter set might yield the same reproduction results. A solution to this problem was sought by restricting the empirical parameter hyperspace with several specific conditions that the force field had to fulfill to be considered reliable. We used the same seven conditions as before,[59] plus an additional one (denoted condition 8 below). Specifically, the conditions were (1) to yield both trans and cis amino acid chelation to copper(II); (2) in Cu(Sar)$_2\cdot 2H_2O$, to allow for the movement of $2H_2O$ from the axial position (as in the experimental crystal structure)[39] to the equatorial position (as in the equilibrium vacuum B3LYP structure)[59] accompanied by a conformational change[59] during in vacuo energy minimization; (3) to be able to simulate pronounced distortion from planarity of the chelate rings in aqua relative to anhydrous complexes; (4) to preserve the 2-fold crystallographic symmetry in the simulated crystal lattices of aquabis Cu(II) complexes with L-Me_2Ala, L-Et_2Ala, and tBuMeGly, with Cu(II) and O_W on a 2-fold axis; (5) to retain the irregular square-planar copper(II) coordination geometry in the complexes that have such coordination in the crystal state; (6) to retain in vacuo distorted planar copper(II) coordination geometry[58] for anhydrous complexes; (7) to keep the positions of the water molecules as close to their experimental crystal and B3LYP vacuum positions as possible; and (8) to be able to reproduce the span of experimental intermolecular Cu—$O_{carbonyl}$ distances from 2.3 to 3.1 Å.

The final force field, named FFWa-SPCE (Table 1), met all eight special requirements and still yielded the geometry reproduction of the anhydrous and aqua copper(II) amino acid complexes in vacuo and in crystal comparable to that obtained by the force field FFW (see sections below). The empirical parameters of the bond-stretching, valence-angle-bending, torsion, out-of-plane-deformation, and Lennard-Jones potentials of the FFWa-SPCE force field are listed in Table 1. The charge parameters and fractional charge values assigned according to these parameters by the Lyngby-CFF program are given elsewhere.[31,37,59] The charge parameters for O_W and H(O_W) in Table 1 yielded fractional charges of $-0.8476e$ and $0.4238e$, respectively, as assigned by the Lyngby-CFF program and were the same in all studied systems.

Efficacy of the New Force Field. To examine the efficacy of the new force field, the MM equilibrium geometries calculated in simulated crystalline surroundings and for isolated systems were compared with the experimental crystal data[21−42] and the ab initio B3LYP-derived vacuum geometries,[58,59] respectively. The FFWa-SPCE results were compared with the results yielded by the FFW force field,[37,38,59] to examine the simulation ability of the new force field relative to that of FFW. The comparisons between the two force fields' results also indicate the validity of the MM model used. The suitability of FFWa-SPCE for simulations and predictions in aqueous solution was examined for solvated Cu(Gly)$_2$, the only bis(amino acidato)copper(II) system for which, as mentioned in the Introduction, some experimental structural data in aqueous solution at room temperature were available from the literature.[20]

MM Simulations in Crystal. The experimental atomic coordinates and the unit cell lengths and angles were taken as the starting points for geometry optimization using the crystal simulator of the Lyngby-CFF program. Table 2 shows errors in reproducing the experimental internal coordinates and unit cell dimensions for each compound, total root-mean-square (rms) deviations between the experimental and theoretical values calculated for the anhydrous and aqua complexes separately, and the grand total rms values for all 25 compounds.

The FFWa-SPCE rms deviations between experimental and theoretical internal coordinates and unit cell dimensions are very much comparable to the values obtained with the FFW force field and discussed elsewhere.[37,38,59] In particular, using FFW, the total rms deviations for anhydrous complexes in the bond lengths, valence angles, torsion angles, and unit cell lengths are 0.018 Å, 2.2°, 4.6°, and 0.287 Å, respectively; for aqua complexes, the corresponding values are 0.016 Å, 2.3°, 6.1°, and 0.475 Å; and the grand total rms values for all 25 complexes are 0.017 Å, 2.2°, 5.3°, and 0.381 Å. The maximum difference between the experimental and theoretical unit cell angles is 9.6° by FFWa-SPCE and 8.7° by FFW [for Cu(D,L-Ala)$_2\cdot H_2O$]. Experimental unit cell volumes are reproduced from −5.4% to 7.0% by FFWa-SPCE and from −8.1% to 9.6% by FFW.

FFWa-SPCE is capable of simulating the flexibility (plasticity) of the copper(II) coordination (Table S1, Supporting Information). Table S1 (Supporting Information) presents the mean values and standard deviations for the experimental and FFWa-SPCE crystal coordination polyhedron Cu—N and Cu—O bond lengths and angles around copper for all studied copper(II) amino acidates. The differences between the theoretical and experimental means are commonly within the standard deviations given in Table S1 (Supporting Information). The shapes of the coordination polyhedra are generally well reproduced by FFWa-SPCE as well as by FFW.[59] FFWa-SPCE maintains the planarity of the copper(II) coordination polyhedron in the crystal (as well as in vacuo) for all molecules whose amino acid donor atoms form an irregular square-planar configuration around the copper(II) in the experimental crystal lattice.

The intermolecular axial copper—carbonyl-oxygen distances are experimentally in the range from 2.312 to 3.116 Å (Table S2, Supporting Information). The new force field manages to cover a greater span of the particular distances (from 2.376 to 2.877 Å; Table S2, Supporting Information) than the force field FFW (from 2.519 to 2.855 Å; Table S2,

Table 2. Comparison of Experimental and FFWa-SPCE Crystal Structures in Terms of the Root-Mean-Square Deviations (rms) in Internal Coordinates[a] and Unit Cell Constants (*a*, *b*, and *c*, in Å) and Differences, Δ, between Experimental and Theoretical Unit Cell Angles (α, β, and γ, in deg) and Volumes (*V*)

compound	rms(Δ*b*)	rms(Δθ)	rms(Δφ)	rms(Δ*a*,Δ*b*,Δ*c*)	Δα, Δβ, Δγ	100Δ*V*/V_{exp}
Cu(L-Me_2Val)$_2$	0.015	2.4	2.9	0.525	0.0, 0.0, 0.0	−1.4
trans-Cu(L-Ala)$_2$	0.017	2.7	4.1	0.207	0.0, 3.7, 0.0	−2.3
Cu(L-Leu)$_2$	0.016	2.2	4.2	0.263	0.0, −1.1, 0.0	−0.4
Cu(D,L-2-aBut)$_2$	0.015	1.7	4.2	0.171	0.0, −0.7, 0.0	−4.6
Cu(1-Acpc)$_2$	0.010	1.7	1.6	0.139	0.0, 4.6, 0.0	−2.4
Cu(D,L-Et_2Ala)$_2$	0.030	1.9	3.4	0.240	4.2, 1.3, −3.5	2.7
Cu(D,L-Me_2Val)$_2$	0.010	1.5	0.9	0.109	0.0, 1.1, 0.0	1.0
Cu(L-2-aBut)$_2$	0.017	2.2	4.9	0.254	0.0, 1.2, 0.0	−3.2
Cu(α-aiBut)$_2$	0.015	2.2	3.1	0.151	0.0, −0.7, 0.0	0.5
Cu(L-Me_2Ile)$_2$	0.020	2.7	4.7	0.479	0.0, 0.5, 0.0	−0.5
Cu(L-Pr_2Ala)$_2$	0.018	1.7	3.4	0.254	0.0, −0.2, 0.0	−5.4
Cu(D-Ala)$_2$	0.024	1.8	5.1	0.209	0.0, 0.0, 0.0	−4.4
cis-Cu(L-Ala)$_2$	0.009	1.9	5.7	0.207	0.0, 0.0, 0.0	−4.1
Cu(Et_2Gly)$_2$	0.018	2.5	3.6	0.041	0.0, 0.0, 0.0	−0.1
total	*0.018*	*2.1*	*3.8*	*0.264*		
Cu(L-Me_2Val)$_2$·2H_2O	0.010	1.7	3.5	0.816	1.7, 2.5, 4.1	2.6
Cu(Sar)$_2$·2H_2O	0.018	1.2	5.8	0.640	0.0, −2.7, 0.0	2.8
Cu(L-Et_2Ala)$_2$·H_2O	0.019	1.7	5.3	0.268	0.0, 0.0, 0.0	1.3
Cu(tBuMeGly)$_2$·H_2O	0.021	3.2	5.9	0.101	0.0, 0.0, 0.0	1.6
Cu(L-Me_2Ile)$_2$·H_2O	0.016	2.1	4.5	0.092	0.0, 0.0, 0.0	1.1
Cu(D,L-Ala)$_2$·H_2O	0.023	2.5	8.5	0.769	0.0, 9.6, 0.0	4.4
Cu(L-Me_2Ala)$_2$·7H_2O	0.008	2.0	1.7	0.524	0.0, 2.7, 0.0	7.0
Cu(Gly)$_2$·H_2O	0.018	1.3	14.2	0.473	0.0, 0.0, 0.0	3.8
Cu(L-Ile)$_2$·H_2O	0.013	3.2	6.7	0.123	0.0, 0.0, 0.0	−2.6
Cu(Me_2Gly)$_2$·3H_2O	0.011	2.1	3.3	0.179	0.0, 0.0, 0.0	6.0
Cu(Me_2Gly)$_2$·H_2O	0.012	2.2	2.8	0.197	0.0, 0.0, 0.0	1.1
total	*0.016*	*2.2*	*6.4*	*0.461*		
grand total	**0.017**	**2.2**	**5.0**	**0.364**		

[a] Internal coordinates: bond lengths, *b* (in Å), valence angles, θ (in deg), torsion angles φ (in deg). Hydrogen atoms were not taken into account.

Supporting Information), and hence, FFWa-SPCE copes quite well with the eighth special condition used in the force field parameterization. Both force fields overestimate the means of the copper—axial-water-oxygen-atom distances by 0.2 Å (Table S3, Supporting Information). This result is due to the force field parameterization procedure, that is, the best possible reproduction obtained by taking all special force field parameterization requirements for a reliable force field into the consideration.

Good overall reproduction of the experimental crystal and molecular structures confirms that FFWa-SPCE is reliable; specifically, it accurately reproduces the crystal lattice effects, and the van der Waals and hydrogen-bonding intermolecular interactions are properly modeled.

MM Simulations in Vacuo. A very good match obtained between vacuum quantum chemical B3LYP structures[58] and MM minimum geometries calculated by the two force fields, FFW and FFWa-SPCE, for three anhydrous copper(II) amino acid complexes is shown in Figure 1. FFWa-SPCE yielded very much the same reproduction of the B3LYP structures as FFW (discussed elsewhere[59]).

For the four copper(II) bis-glycinato systems containing from one to four water molecules (Figure 2), the match between the MM and B3LYP structures[59] is not as good as for the anhydrous complexes (as the aqua complexes are generally more difficult to model than the anhydrous complexes),[59] but it is still acceptable. Both force fields similarly reproduce molecular structures of the copper(II) glycinato complexes, although they yield different positions for the water molecules (Figure 2). The new force field overestimates the B3LYP Cu—O_W axial distance by 0.03 and 0.27 Å in *cis*-Cu(Gly)$_2$·H_2O and *trans*-Cu(Me_2Gly)$_2$·3H_2O, respectively. The B3LYP and FFWa-SPCE-calculated hydrogen-bond distances between O_W and carboxylato oxygen atoms also differ slightly, from −0.14 Å to 0.15 Å. The ab intio O···O_W and N···O_W hydrogen-bond distances amount, respectively, to 2.76 and 3.05 Å in *cis*-Cu(Gly)$_2$·H_2O and 2.72 and 2.89 Å in *trans*-Cu(Sar)$_2$·2H_2O.[59] FFWa-SPCE overestimates the corresponding distances by 0.12 and 0.35 Å in the cis complex and by 0.10 Å and 0.63 Å in the trans complex, still yielding values generally accepted to count as hydrogen bonds.

MD Simulations of Cu(Gly)$_2$ in Aqueous Solution at Room Temperature. MD simulations were performed using the new force field FFWa-SPCE. The out-of-plane deformation angle, χ (improper dihedral angle), is differently defined in the Lyngy-CFF program [as the angle between the plane defined by (C, C_{pl}, O) and $O_{carbonyl}$] than in Gromacs [as the angle between two planes defined by (C, C_{pl}, O) and by (C, $O_{carbonyl}$, O)]. To get the same equilibrium geometries in vacuo by the two programs, the empirical parameter of the

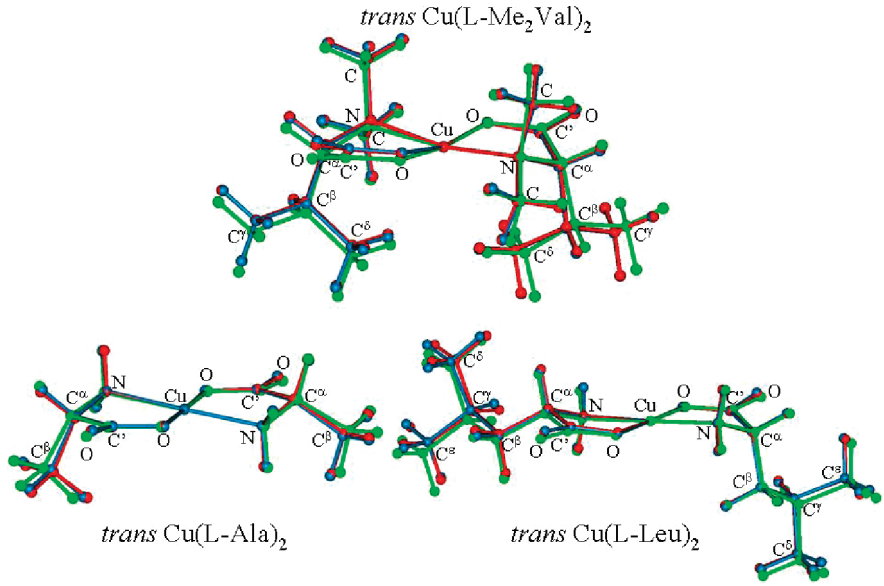


Figure 1. Superposition of in vacuo equilibrium geometries of three anhydrous copper(II) amino acid complexes: structures obtained using B3LYP (green), MM FFW (red), and FFWa-SPCE (blue).

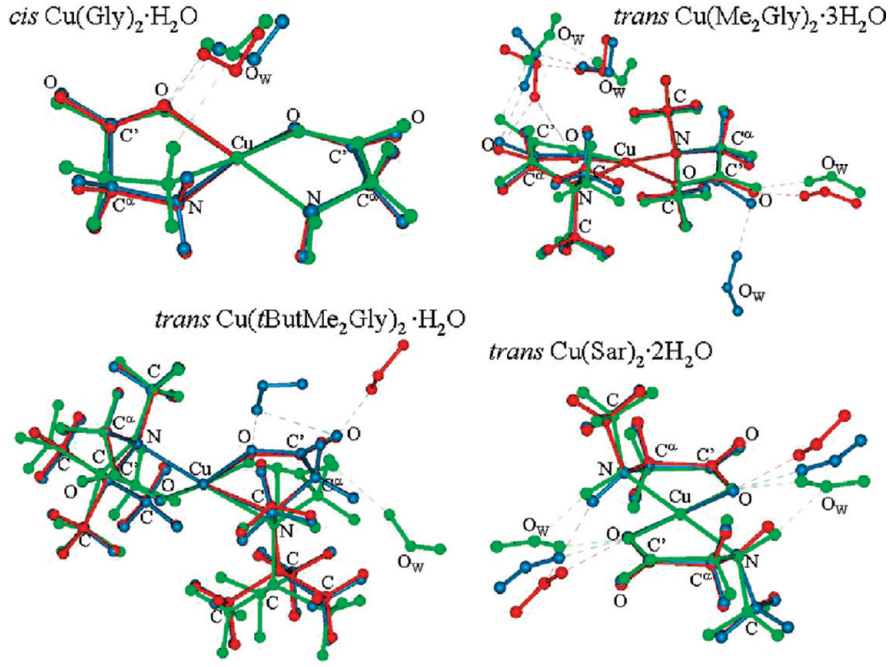


Figure 2. Superposition of the equilibrium geometries of four isolated aquabis glycinato copper(II) systems: structures obtained using B3LYP (green), MM FFW (red), and FFWa-SPCE (blue).

out-of-plane deformation potential $V(\chi)$ in eq 1 was adjusted to $k_\chi = 99.5$ kcal mol^{-1} rad^{-2} for use in Gromacs.

The ability of the FFWa-SPCE force field to predict structural properties in aqueous solution was examined by simulating solvated *cis*- and *trans*-Cu(Gly)$_2$ and by comparing the theoretical MD results with experimental data[20] obtained by X-ray absorption spectroscopy. Table 3 presents selected experimental structural coordinates (the technique could not distinguish one isomer from the other), along with the corresponding values of the average MD structures, as well as the means and standard deviations of the coordinate values obtained during 20 ns of MD simulations at room temperature, separately for the trans and cis isomers. Table 3 also includes the results obtained by quantum chemical PCM calculations at the B3LYP/LanL2DZ{D95v+(d)} level of theory. The MD copper—water-oxygen-atom distances in the first and second coordination spheres (or second hydration shell) in Table 3 were obtained from the Cu—O$_W$ radial distribution functions (Figure 3).

The deviations between the theoretical means and the experimental values of the bond distances and angles are within the theoretical and experimental error values (Table 3). This statement also applies to the valence angles around copper(II), although their values suggest the distortion of the copper(II) coordination geometry from planarity for the solvated complex during MD simulations. The distribution functions for the valence angles around copper(II) estimated from the MD simulation data revealed the most frequent

Table 3. Selected Structural Coordinates of $Cu(Gly)_2$ in Aqueous Solution at Room Temperature As Obtained from X-ray Absorption Studies[a] and Estimated from 20 ns of MD Simulations and by the PCM Method Separately for *trans*- and *cis*-$Cu(Gly)_2$[b]

			MD (FFWa-SPCE force field)					
	X-ray absorption[a]				average structure		PCM	
coordinate	value	variance	trans[c]	cis[c]	trans	cis	trans	cis
Cu—N	1.99	0.004	2.00 (1)	2.00 (2)	1.98	1.97	2.02	2.03
Cu—O	1.95	0.006	1.93 (2)	1.92 (2)	1.91	1.92	1.96	1.96
Cu···C^{α}	2.84	0.020	2.83 (3)	2.82 (3)	2.79	2.79	2.88	2.88
Cu···C′	2.79	0.003	2.73 (3)	2.73 (3)	2.69	2.71	2.79	2.79
C′═O	1.24	0.006	1.23 (2)	1.23 (2)	1.20	1.21	1.24	1.24
N—Cu—N			167 (7)		179.7		179.5	
N—Cu—O	179	36		169 (7)		175.3		177.4
O—Cu—O			167 (7)		179.7		179.8	
Cu—C′═O	168	36	161 (3)	161 (2)	163.0	163.2	163.0	162.9
Cu···$O_{W,ax}$	2.40 (6)	0.03 (1)	2.6	2.6				
Cu···$O_{W,s}$	3.3 (2)	0.06 (3)	3.5	3.4				

[a] Reference 20. [b] Bond distances are in Å, and bond angles are in deg. Bond and angle variances are in $Å^2$ and deg^2, respectively. Standard deviations are in parentheses. The $O_{W,ax}$ and $O_{W,s}$ denote the water oxygen atoms from the first and second coordination spheres, respectively. [c] Means and standard deviations calculated from the values attained in 20-ns trajectories.

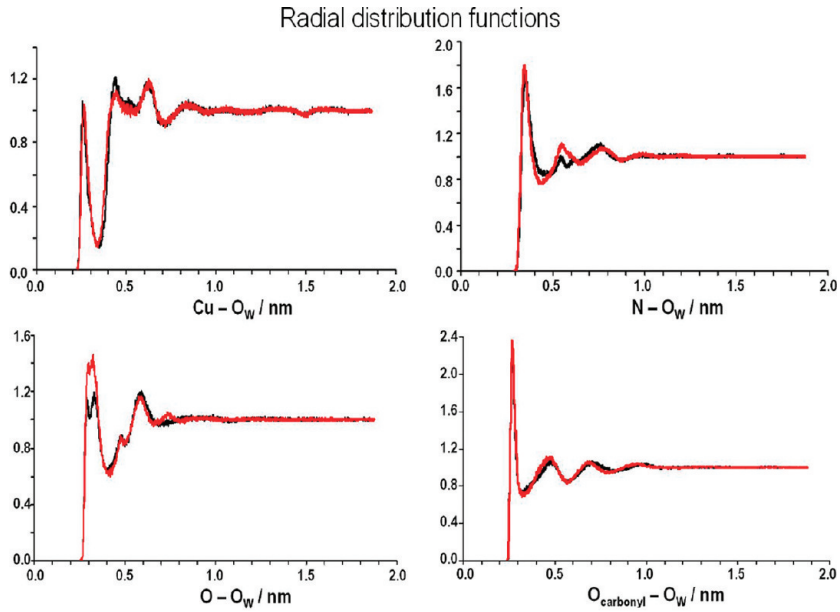


Figure 3. Radial distribution functions for solvated *trans*-$Cu(Gly)_2$ (black) and *cis*-$Cu(Gly)_2$ (red) determined during 20 ns of MD simulation at room temperature.

values to be 172° and 171° for the N—Cu—N and O—Cu—O angles, respectively, in *trans*-$Cu(Gly)_2$ and 173° for the N—Cu—O angles in *cis*-$Cu(Gly)_2$.

Although the average structures obtained from the MD simulations might not necessarily represent physically reasonable structures, their structural coordinates (Table 3) are in very good agreement with the experimental values (within experimental errors) and the PCM coordinates.

The *B* factor, which is estimated from the rms fluctuations of the atoms, was calculated and compared with available experimental values from the X-ray diffraction measurements of *cis*-$Cu(Gly)_2 \cdot H_2O$ at room temperature[41] and 173 K[43] (Figure 4). The calculated *B* factors, which indicate the atom motions in aqueous solution at room temperature, follow the general pattern of the experimental *B* factors for heavy atoms with a few exceptions, namely, the hydrogen-atom fluctuations are calculated as more pronounced in solution at 300 K under the influence of solvent—solute interactions than as measured in the 173 K crystal structure (Figure 4). Apparently, the MD simulations reproduced the molecular motion reasonably well.

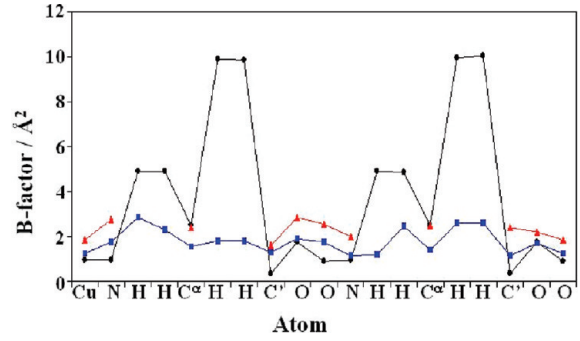


Figure 4. *B* factors for the atoms of *cis*-$Cu(Gly)_2$ calculated from the MD simulation (black) and from experimental X-ray crystal structures of *cis*-$Cu(Gly)_2 \cdot H_2O$ measured at room temperature (red) and at 173 K (blue) as reported in refs 41 and 43, respectively.

Table 4. MD Average Values of Energy Contributions and Corresponding rms Deviations (in Parentheses) Calculated from the Values Attained during 20 ns of MD Simulations at Room Temperature Separately for the *trans*- and *cis*-Cu(Gly)$_2$·2159H$_2$O Systems

	energy (kJ mol^{-1})			
	trans		cis	
potential energy	−101604.0	(244.9)	−101609.0	(246.4)
kinetic energy	16309.3	(148.0)	16311.1	(148.2)
total energy	−85294.8	(196.5)	−85298.2	(198.9)
	Cu(Gly)$_2$ Intramolecular Contributions			
$V(b)$	26.3	(8.4)	27.7	(8.7)
$V(\theta)$	37.1	(9.3)	34.9	(9.0)
$V(\varphi)$	46.7	(5.2)	45.9	(5.2)
$V(\chi)$	2.0	(1.7)	2.0	(1.7)
$V_{\text{Coulomb,NB}}$	−1516.1	(19.6)	−1388.8	(14.8)
$V_{\text{Coulomb,1-3}}$	2299.1	(19.1)	2290.5	(15.9)
V_{LJ}	−21.0	(2.0)	−19.9	(2.8)
total	874.1		992.3	
	Intermolecular Cu(Gly)$_2$·H$_2$O Contributions			
V_{Coulomb} (short-range)	−415.3	(43.6)	−587.8	(51.2)
V_{Coulomb} (long-range)	−35.4	(29.5)	−72.0	(32.6)
V_{LJ} (short-range)	−114.6	(15.9)	−114.0	(16.2)
V_{LJ} (long-range)	−6.3	(0.2)	−6.3	(0.2)
	Intermolecular H$_2$O·H$_2$O Contributions			
V_{Coulomb} (short-range)	−117708.0	(414.5)	−117588.0	(416.4)
V_{Coulomb} (long-range)	−3436.1	(230.0)	−3467.4	(228.5)
V_{LJ} (short-range)	19522.0	(256.3)	19518.3	(257.3)
V_{LJ} (long-range)	−284.4	(1.1)	−284.6	(1.1)

Table 4 presents average values and corresponding rms deviations of energy contributions for the *trans*- and *cis*-Cu(Gly)$_2$·2159H$_2$O systems calculated during 20 ns of MD simulations. Although the intramolecular potential energy contribution was lower for the trans isomer than for the cis isomer, the cis isomer had more favorable electrostatic interactions with the water molecules. Different interactions between the trans and cis conformations with the water molecules [also noticeable in the different organization structures of the water molecules around the isomers in Figure 3 from the radial distribution functions calculated for the distances between O_W and atoms of Cu(Gly)$_2$ that form hydrogen bonds with water molecules] influenced the water−water interactions as well. As a result, the two systems had similar total energies, with the average total energy of the solvated cis-isomer system being 3 kJ mol^{-1} lower than that of the solvated trans-isomer system (Table 4). The MD prediction is in line with the quantum chemical PCM energy estimations for the two isomers in approximate aqueous medium at 300 K, which predicted a small energy difference of 4.96 kJ mol^{-1} in favor of *trans*-Cu(Gly)$_2$. Moreover, recent EPR measurements of Cu(Gly)$_2$ in glycerol/water solution confirmed the simultaneous presence of both isomers at room temperature.[19]

Cis−Trans Isomerization of Cu(Gly)$_2$ in Aqueous Solution at 300 K. Table 5 lists the energy differences between the cis and trans minima and the TS structure of Cu(Gly)$_2$ calculated at the B3LYP/LanL2DZ{D95v+(d)} level of theory in the gas phase and in water medium at room temperature. The stationary points in the gas phase were calculated and compared with the geometries and energies obtained at the higher level of theory [denoted as G3(MP2)-B3$_{\text{LanL2DZ}}$ elsewhere][64] to verify the basis set and the method used. The B3LYP/LanL2DZ{D95v+(d)} gas-phase potential energy values differ by 1 kcal mol^{-1} from the higher theory's values,[64] yielding the same energy difference value of 56.74 kJ mol^{-1} between the cis and trans minima. The geometries of the stationary points derived at the lower and higher levels of theory are almost the same, as the maximum differences between their internal coordinates are up to 0.015 Å in bond lengths (Cu—N), 3.7° in valence angles (N—C$^{\alpha}$—H), and 2.9° in torsion angles (C′—C$^{\alpha}$—N—Cu).

Table 5. Potential Energy (ΔV) and Gibbs Free Energy (ΔG) Barriers for the Isomerization Reaction of Cu(Gly)$_2$ in the Gas Phase and in Aqueous Solution at 300 K Calculated at the B3LYP/LanL2DZ{D95v+(d)} Level of Theory and Reaction Rates Calculated from Eqs *2a* and *2b*

	gas phase	aqueous solution
	Energy Difference (kJ mol^{-1})	
$\Delta V_{\text{TS}\rightarrow\text{cis}}$	28.95	70.10
$\Delta V_{\text{TS}\rightarrow\text{trans}}$	85.69	75.06
$\Delta G_{\text{TS}\rightarrow\text{cis}}$	26.26	63.67
$\Delta G_{\text{TS}\rightarrow\text{trans}}$	81.92	69.47
	Reaction Rate (s^{-1})	
$k_{\text{cis}\rightarrow\text{trans}}$	1.7×10^{8}	51.3
$k_{\text{trans}\rightarrow\text{cis}}$	3.4×10^{-2}	5.0

The TS geometries obtained in the gas phase and in solution were additionally fully optimized in water medium. Interestingly, in both instances initial guess of the orbital energies directed the optimization to the cis final structure. Nevertheless, with a different choice of initial orbital energies (by using the Vshift option under the G03 SCF keyword), the optimization of the TS structure in aqueous medium ended in trans structure, as would be expected from the isomer energy estimations that the trans isomer had lower energy than the cis isomer (Table 5).

Compared to the gas-phase results, the PCM calculations in aqueous solution revealed a slightly lower energy differ-

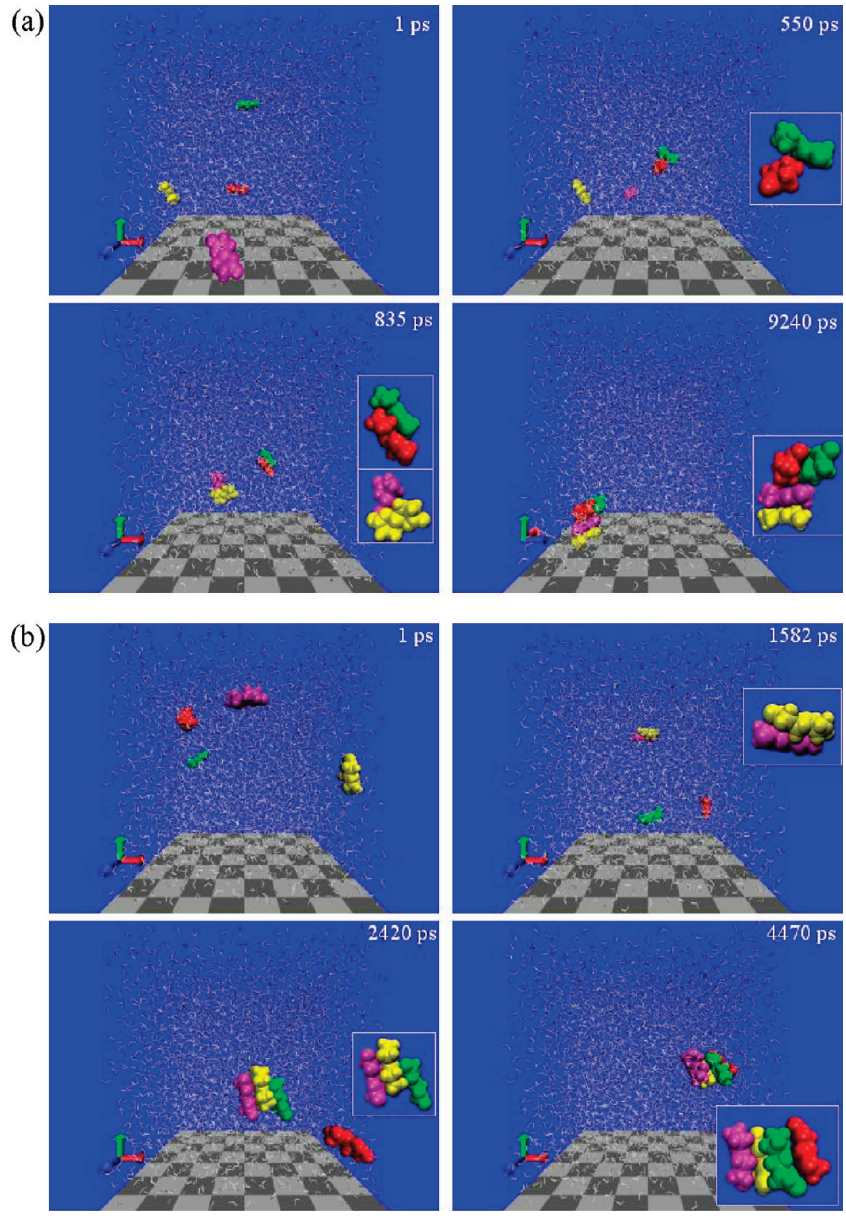


Figure 5. System of $4Cu(Gly)_2 \cdot 5457H_2O$ depicted at indicated times in MD simulations: (a) four trans isomers and (b) four cis isomers colored red, green, yellow, and magenta, with water molecules in white. The figure was prepared using the VMD program.[102]

Table 6. Aggregation Times (ps) of Four Solvated $Cu(Gly)_2$ Molecules during MD Simulations for Four Studied Systems Containing Four *trans*-$Cu(Gly)_2$ Molecules, Four *cis*-$Cu(Gly)_2$ Molecules, and Two *trans*-$Cu(Gly)_2$ and Two *cis*-$Cu(Gly)_2$ Molecules Having Different Initial Positions

$Cu(Gly)_2$	dimer	trimer	tetramer
four trans	550, 885		9240
four cis	1582	2420	4470
two trans, two cis	125 (trans−trans)	1465 (cis−trans−trans)	12045 (cis−trans−trans−cis)
two trans, two cis	580 (cis−trans)	1615 (cis−trans−cis)	4605 (cis−trans−cis−trans)

ence between the trans isomer and the TS structure, but a considerably larger energy difference between the cis isomer and the TS structure (Table 5), resulting in a dramatic lowering of the potential energy difference between the trans and cis conformations.

This result raised the following question: If the energy difference between the cis and trans isomers is considerably lower in aqueous solution than in the gas phase, as obtained by both MD and PCM calculations, why could a cis−trans interconversion of $Cu(Gly)_2$ not be obtained during the MD simulations? To find at least a qualitative answer to this question, we calculated the reaction rate constants for the isomerization reaction in aqueous solution and in the vacuum from eqs 2a and 2b (Table 5) and compared them with the

gas-phase rate constants determined by variational transition state theory including quantum effects such as tunneling and corner cutting.[64] The gas-phase calculations[64] showed that, at room temperature, the cis isomer spontaneously transforms into the trans isomer without breaking bonds, with reaction rate constants $k_{cis\rightarrow trans}$ and $k_{trans\rightarrow cis}$ equal to 2.1×10^6 and 2.7×10^{-4} s^{-1}, respectively. Hence, the gas-phase reaction rates calculated from the standard transition state theory (Table 5) are 2 orders of magnitude higher than the reaction rate constants obtained by more precise variational transition state theory. A contribution of 1 order of magnitude is caused by the difference between the high-level and low-level potential energy values of 1 kcal mol^{-1} (yielding 5.4 times larger rate constant values). Because the tunneling effect was calculated to be negligible at 300 K,[64] an additional contribution to the difference in the gas-phase rates should be due to a variational effect.[98] As the reaction rate constants were obtained using information only at the stationary points and assuming that the transmission factor was equal to unity, they can be considered as zerothe-order approximations of the actual rate constants. If we assume the same order-of-magnitude error for the calculation of the reaction rate values in aqueous solution (Table 5), then the interconversion between the cis and trans isomers in aqueous solution should be on the order from milliseconds to seconds, and this result is in accord with our not observing the isomerization during the MD simulations of 20 ns. Whether the assumption is correct remains to be examined by MD calculations of the free energy profile and activation free energy for the intramolecular cis−trans interconversion of $Cu(Gly)_2$ in aqueous solution in forthcoming studies. It would also be challenging to explore the possibility of cis−trans isomerization by an intermolecular process through a chelate ring-opening and bond-breaking and -forming mechanism. It would be appealing to probe the force field for creating a reactive force field (by adding new potential energy terms, and further reparameterization) to calculate a reactive potential energy surface and for proper configuration sampling in solution within, for example, an empirical valence bond (EVB) approach,[99−101] which uses classical MM force fields to model reactant and product configurations and can accurately describe reactive potential energy surfaces.

Twenty-Nanosecond MD Simulations of the 4Cu-$(Gly)_2$·5457H_2O System. The predictive properties of the FFWa-SPCE force field were examined on a system of four solvated $Cu(Gly)_2$ molecules in aqueous solution at room temperature. The MD trajectories were collected for systems containing four trans and four cis isomers (Figure 5). Interesting results were obtained as, after some time, the $Cu(Gly)_2$ complexes started to aggregate (Figure 5). The trans isomers formed two dimers that eventually aggregated to a tetramer. The cis isomers aggregated differently, from a dimer to a trimer to a tetramer. Two additional aggregation patterns were obtained for systems containing two trans and two cis isomers by taking different initial positions (Table 6). Although the aggregation patterns were different, these four MD trajectories offered some regularity. Interestingly, it was observed that, if a trans−trans dimer formed first, the aggregation to a tetramer required a longer time than if a cis isomer formed the initial dimer (Table 6). The average total MD energy of the solvated trans tetramer system was lower than those of the cis tetramer system and the mixed two trans, two cis tetramer by 24 and 12 kJ mol^{-1}, respectively.

The predicted results can be related to the experimental observations. Specifically, in the solid state, *trans*-$Cu(Gly)_2$ is thermodynamically more stable than the cis isomer.[103] However, aqua *cis*-$Cu(Gly)_2$ is the form that crystallizes upon slow evaporation of the aqueous solution at room temperature.[41,43] Thus, the MD-predicted aggregation results suggest that the crystallization process of $Cu(Gly)_2$ is a more kinetically than thermodynamically driven course of action.

Conclusions

The new force field parameterization, as a continuation of our previous work on molecular modeling of copper(II) complexes with aliphatic α-amino acids and their *N*-alkyl derivatives in the vacuum and simulated crystal surroundings,[37,38,59] enables the prediction of structural properties in aqueous solution as well. Specifically, the time-average bond distances and angles of $Cu(Gly)_2$ obtained during MD simulations in aqueous solution at room temperature with the new FFWa-SPCE force field are in good agreement with the experimental data obtained by X-ray absorption spectroscopy[20] and quantum chemical PCM calculations.

The quantum chemical PCM energy estimations of the trans and cis minima and TS structure of $Cu(Gly)_2$ in approximate water medium revealed a pronounced lowering of the energy difference between the two minima and an increase in the energy difference between the cis conformer and the TS structure with respect to the gas phase (Table 5). The MD-based estimations suggest that the decrease in energy difference was due to more favorable electrostatic interactions of the cis than the trans isomer with the water molecules. A small energy difference between the two solvated isomer systems is also predicted by the MD simulations and can be confirmed by experimental observations[19] that the trans and cis conformers of $Cu(Gly)_2$ are simultaneously present in aqueous solution at room temperature.

The use of the same set of relatively simple analytical functions with carefully selected empirical parameters for MM calculations in vacuo and in crystal, as well as MD simulations in aqueous solution, can provide structural and energetic information about bis(amino acidato)copper(II) compounds in these environments. Furthermore, the study can assist in understanding the self-association of the complexes in solution and identifying the formation of a nucleus of crystallization.

Acknowledgment. This work was supported by the Croatian Ministry of Science, Education and Sports (Project Grant 022-0222148-2822). We thank Dr Sanja Tomić (Ruđer Bošković Institute, Zagreb, Croatia) for encouragement and guidelines concerning the MD simulations and for critical reading and editing of the manuscript.

Supporting Information Available: Listing of means and standard deviations of the experimental and FFWa-SPCE

crystal bond lengths and angles of the copper(II) polyhedra in 14 anhydrous and 11 aqua copper(II) amino acidates (Table S1), reproduction of experimental in-crystal axial intermolecular $Cu\cdots O_{carbonyl}$ distances (Table S2), and $Cu\cdots O_W$ distances (Table S3) in 8 anhydrous and 11 aqua copper(II) amino acidates obtained using the FFW and FFWa-SPCE force fields, respectively. This material is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) *Molecular Biology and Toxicology of Metals*; Zalups, R. K.; Koropatnick, J., Eds.; Taylor & Francis: London, 2000.

(2) DiDonato, M.; Sarkar, B. *Biochem. Biophys. Acta* **1997**, *1360*, 3–16.

(3) Deschamps, P.; Kulkarni, P. P.; Gautam-Basak, M.; Sarkar, B. *Coord. Chem. Rev.* **2005**, *249*, 895–909.

(4) Kodama, H.; Fujisawa, C. *Metallomics* **2009**, *1*, 42–52.

(5) Gaetke, L. M.; Chow, C. K. *Toxicology* **2003**, *189*, 147–163.

(6) *Copper, Environmental Health Criteria 200*; International Programme on Chemical Safety (IPCS), World Health Organization: Geneva, Switzerland, 1998; available at http://www.inchem.org/documents/ehc/ehc/ehc200.htm (accessed Feb 28, 2008).

(7) Dokmanić, I.; Šikić, M.; Tomić, S. *Acta Crystallogr.* **2008**, *D64*, 257–263.

(8) Farkas, E.; Sóvagó, I. *Amino Acids, Pept. Proteins* **2007**, *36*, 287–345.

(9) Szabó-Plánka, T.; Rockenbauer, A.; Korecz, L.; Nagy, D. *Polyhedron* **2000**, *19*, 1123–1131.

(10) Szabó-Plánka, T.; Rockenbauer, A.; Korecz, L. *Polyhedron* **1999**, *18*, 1969–1974.

(11) Szilágyi, I.; Labádi, I.; Hernadi, K.; Pálinkó, I.; Nagy, N. V.; Korecz, L.; Rockenbauer, A.; Kele, Z.; Kiss, T. *J. Inorg. Biochem.* **2005**, *99*, 1619–1629.

(12) Altun, Y.; Köseoğlu, F. *J. Solution Chem.* **2005**, *34*, 213–231.

(13) Branica, G.; Paulić, N.; Grgas, B.; Omanović, D. *Chem. Speciation Bioavailability* **1999**, *11*, 125–134.

(14) Mirosavljević, K.; Sabolović, J.; Noethig-Laslo, V. *Eur. J. Inorg. Chem.* **2004**, 3930–3937.

(15) Sabolović, J.; Noethig-Laslo, V. *Cell. Mol. Biol. Lett.* **2002**, *7*, 151–153.

(16) Goodman, B. A.; McPhail, D. B. *J. Chem. Soc., Dalton Trans.* **1985**, 1717–1718.

(17) Goodman, B. A.; McPhail, D. B.; Powell, H. K. J. *J. Chem. Soc., Dalton Trans.* **1981**, 822–827.

(18) Noethig-Laslo, V.; Paulić, N. *J. Chem. Soc., Dalton Trans.* **1992**, 2045–2047.

(19) Pezzato, M.; Della Lunga, G.; Baratto, M. C.; Pogni, R.; Basosi, R. *Magn. Reson. Chem.* **2007**, *45*, 846–849.

(20) D'Angelo, P.; Bottari, E.; Festa, M. R.; Nolting, H.-F.; Pavel, N. V. *J. Phys. Chem. B* **1998**, *102*, 3114–3122.

(21) Kaitner, B.; Kamenar, B.; Paulić, N.; Raos, N.; Simeon, Vl. *J. Coord. Chem.* **1987**, *15*, 373–381.

(22) Kaitner, B.; Ferguson, G.; Paulić, N.; Raos, N. *J. Coord. Chem.* **1992**, *26*, 105–115.

(23) Kaitner, B.; Paulić, N.; Raos, N. *J. Coord. Chem.* **1991**, *22*, 269–279.

(24) Kaitner, B.; Meštrović, E.; Paulić, N.; Sabolović, J.; Raos, N. *J. Coord. Chem.* **1995**, *36*, 117–124.

(25) Hitchman, M. A.; Kwan, L.; Engelhardt, L. M.; White, A. H. *J. Chem. Soc., Dalton Trans.* **1987**, 457–465.

(26) Fawcett, T. G.; Ushay, M.; Rose, J. P.; Lalancette, R. A.; Potenza, J. A.; Schugar, H. J. *Inorg. Chem.* **1979**, *18*, 327–332.

(27) Levstein, P. R.; Calvo, R.; Castellano, E. E.; Piro, O. E.; Rivero, B. E. *Inorg. Chem.* **1990**, *29*, 3918–3922.

(28) Oliva, G.; Castellano, E. E.; Zukerman-Schpector, J.; Calvo, R. *Acta Crystallogr.* **1986**, *C42*, 19–21.

(29) Gillard, R. D.; Mason, R.; Payne, N. C.; Robertson, G. B. *J. Chem. Soc. A* **1969**, 1864–1871.

(30) Moussa, S.; Fenton, R. R.; Kennedy, B. J.; Plitz, R. O. *Inorg. Chim. Acta* **1999**, *288*, 29–34.

(31) Kaitner, B.; Paulić, N.; Pavlović., G.; Sabolović, J. *Polyhedron* **1999**, *18*, 2301–2311.

(32) Kaitner, B.; Pavlović, G.; Paulić, N.; Raos, N. *J. Coord. Chem.* **1995**, *36*, 327–338.

(33) Kaitner, B.; Paulić, N.; Raos, N. *J. Coord. Chem.* **1992**, *25*, 337–347.

(34) Kaitner, B.; Pavlović, G.; Paulić, N.; Raos, N. *J. Coord. Chem.* **1998**, *43*, 309–319.

(35) Kamenar, B.; Penavić, M.; Škorić, A.; Paulić, N.; Raos, N.; Simeon, Vl. *J. Coord. Chem.* **1988**, *17*, 85–94.

(36) Kaitner, B.; Ferguson, G.; Paulić, N.; Raos, N. *J. Coord. Chem.* **1992**, *26*, 95–104.

(37) Sabolović, J.; Kaitner, B. *Polyhedron* **2007**, *26*, 1087–1097.

(38) Sabolović, J.; Kaitner, B. *Inorg. Chim. Acta* **2008**, *361*, 2418–2430.

(39) Krishnakumar, R. V.; Natarajan, S.; Bahudur, S. A.; Cameron, T. S. *Z. Kristallogr.* **1994**, *209*, 443–444.

(40) Calvo, R.; Levstein, P. R.; Castellano, E. E.; Fabiane, S. M.; Piro, O. E.; Oseroff, S. B. *Inorg. Chem.* **1991**, *30*, 216–220.

(41) Freeman, H. C.; Snow, M. R.; Nitta, I.; Tomita, K. *Acta Crystallogr.* **1964**, *17*, 1463–1470.

(42) Weeks, C. M.; Cooper, A.; Norton, D. A. *Acta Crystallogr.* **1969**, *B25*, 443–450.

(43) Casari, B. M.; Mahmoudkhani, A. H.; Langer, V. *Acta Crystallogr.* **2004**, *E60*, 1949–1951.

(44) Banci, L. *Curr. Opin. Chem. Biol.* **2003**, *7*, 143–149.

(45) Hay, B. P. *Coord. Chem. Rev.* **1993**, *126*, 177–236.

(46) Zimmer, M. *Chem. Rev.* **1995**, *95*, 2629–2649.

(47) Rappé, A. K.; Casewit, C. J. *Molecular Mechanics Across Chemistry*; University Science Books: Sausalito, CA, 1997.

(48) Boeyens, J. C. A.; Comba, P. *Coord. Chem. Rev.* **2001**, *212*, 3–10.

(49) Norrby, P.-O.; Brandt, P. *Coord. Chem. Rev.* **2001**, *212*, 79–109.

(50) Ledecq, M.; Lebon, F.; Durant, F.; Giessner-Prettre, C.; Marquez, A.; Gresh, N. *J. Phys. Chem. B* **2003**, *38*, 10640–10652.

(51) Nielson, K. D.; van Duin, A. C. T.; Oxgaard, J.; Deng, W.-Q.; Goddard, W. A., III. *J. Phys. Chem. A* **2005**, *109*, 493–499.

(52) Comba, P.; Remenyi, R. *J. Comput. Chem.* **2002**, *23*, 697–705.

(53) Comba, P.; Rememyi, R. *Coord. Chem. Rev.* **2003**, *238*−*239*, 9–20.

(54) Deeth, R. J. *Chem. Commun.* **2006**, *2551*, 2551–2553.

(55) Deeth, R. J. *Inorg. Chem.* **2007**, *46*, 4492–4503.

(56) Deeth, R. J.; Anastasi, A.; Diedrich, C.; Randell, K. *Coord. Chem. Rev.* **2009**, *253*, 795–816.

(57) Sabolović, J.; Rasmussen, Kj. *Inorg. Chem.* **1995**, *34*, 1221–1232.

(58) Sabolović, J.; Liedl, K. R. *Inorg. Chem.* **1999**, *38*, 2764–2774.

(59) Sabolović, J.; Tautermann, C. S.; Loerting, T.; Liedl, K. R. *Inorg. Chem.* **2003**, *42*, 2268–2279.

(60) Murphy, B.; Hathaway, B. *Coord. Chem. Rev.* **2003**, *243*, 237–262.

(61) Siegbahn, P. E. M. *Q. Rev. Biophys.* **2003**, *36*, 91–145.

(62) Ryde, U. *Curr. Opin. Chem. Biol.* **2003**, *7*, 136–142.

(63) de Bruin, T. J. M.; Marcelis, A. T. M.; Zuilhof, H.; Sudhölter, E. J. R. *Phys. Chem. Chem. Phys.* **1999**, *1*, 4157–4163.

(64) Tautermann, C. S.; Sabolović, J.; Voegele, A. F.; Liedl, K. R. *J. Phys. Chem. B* **2004**, *108*, 2098–2102.

(65) Hattori, T.; Toraishi, T.; Tsuneda, T.; Nagasaki, S.; Tanaka, S. *J. Phys. Chem. A* **2005**, *109*, 10403–10409.

(66) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(67) Tomasi, J.; Mennucci, B.; Cancès, E. *J. Mol. Struct. (THEOCHEM)* **1999**, *464*, 211–226.

(68) Sabolović, J.; Mrak., Ž.; Koštrun, S.; Janeković, A. *Inorg. Chem.* **2004**, *43*, 8479–8489.

(69) Niketić, S. R.; Rasmussen, Kj. *The Consistent Force Field: A Documentation*; Lectures Notes in Chemistry; Springer-Verlag: Berlin, 1977; Vol. 3.

(70) Rasmussen, Kj. *Potential Energy Functions in Conformational Analysis*; Lectures Notes in Chemistry; Springer-Verlag: Berlin, 1985; Vol. 37.

(71) Rasmussen, Kj.; Engelsen, S. B.; Fabricius, J.; Rasmussen, B. The Consistent Force Field: Development of Potential Energy Functions for Conformational Analysis. In *Recent Experimental and Computational Advances in Molecular Spectroscopy*; Fausto, R., Ed.; NATO ASI Series C: Mathematical and Physical Sciences; Kluwer Academic Publisher: Dordrecht, The Netherlands, 1993; Vol. 406, pp 381−419.

(72) Williams, D. E. *Top. Curr. Phys.* **1981**, *26*, 3–40.

(73) Pietilä, L.-O.; Rasmussen, Kj. *J. Comput. Chem.* **1984**, *5*, 252–260.

(74) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735–746.

(75) Halgen, T. A.; Damm, W. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236–242.

(76) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Mod.* **2001**, *7*, 306–317.

(77) Berendsen, H. J. C.; van Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

(78) Input Gromacs files with the empirical parameter set are available from J.S. upon request.

(79) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(80) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952–962.

(81) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(82) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

(83) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys* **1980**, *58*, 1200–1211.

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

(85) Dunning, T. H., Jr.; Hay, P. J. Gaussian Basis Sets for Molecular Calculations. In *Methods of Electronic Structure Theory*; Schaefer, H. F., III., Ed.; Modern Theoretical Chemistry; Plenum Press: New York, 1977; Vol. 3, pp 1−27.

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

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

(88) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 270–283.

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

(90) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299–310.

(91) 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.

(92) Cossi, M.; Scalmani, G.; Rega, N.; Barone, V. *J. Chem. Phys.* **2002**, *117*, 43–54.

(93) Adamo, C.; Scuseria, G. E.; Barone, V. *J. Chem. Phys.* **1999**, *111*, 2889–2899.

(94) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(95) Adamo, C.; Cossi, M.; Scalmani, G.; Barone, V. *Chem. Phys. Lett.* **1999**, *307*, 265–271.

(96) Peng, C.; Schlegel, H. B. *Isr. J. Chem.* **1993**, *33*, 449–454.

(97) Peng, C.; Ayala, P. Y.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **1996**, *17*, 49–56.

(98) Albu, T. V.; Corchado, J. C.; Truhlar, D. G. *J. Phys. Chem. A* **2001**, *105*, 8465–8487.

(99) Warshel, A. *Computer Modeling of Chemical Reactions in Enzymes and Solutions*; John Wiley & Sons, Inc: New York, 1991.

(100) Åqvist, J.; Warshel, A. *Chem. Rev.* **1993**, *93*, 2523−2544.

(101) Olsson, M. H. M.; Mavri, J.; Warshel, A. *Philos. Trans. R. Soc. B* **2006**, *361*, 1417–1432.

(102) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.

(103) Delf, B. W.; Gillard, R. D.; O'Brien, P. *J. Chem. Soc., Dalton Trans.* **1979**, 1301–1305.

CT9000203