JCTC Journal of Chemical Theory and Computation

# Influence of Strong Electron Correlation on Magnetism in Transition-Metal Doped Si Nanocrystals

R. Leitsmann,[†] F. Küwen,[‡] C. Rödl, C. Panse, and F. Bechstedt*

*European Theoretical Spectroscopy Facility (ETSF) and Institut für Festkörpertheorie und -optik, Friedrich-Schiller-Universität Jena, Max-Wien-Platz 1, 07743 Jena, Germany*

Received July 31, 2009

**Abstract:** We studied the influence of strong electron correlation on magnetic properties of Si nanocrystals doped with the transition metal (TM) atoms Mn and Fe. Different approaches to describe exchange and correlation (XC) effects are compared within a density-functional framework. Beside a semilocal treatment, two different methods to include the influence of electron correlation on the localized TM 3d states are studied. They are based on XC functionals with the inclusion of on-site Coulomb repulsion or short-range screened exchange. We demonstrate a strong dependence of both electronic structure and magnetization on the used XC functional. The inclusion of strong correlation drastically changes position and occupation of the TM or TM−Si-bond-derived levels as well as the total magnetic moments.

## Introduction

Nanostructuring of materials can lead to novel properties that do not exist in bulk-phase materials. In particular, nanocrystals (NCs) have a high potential for multimodal biological applications by the addition of functionality to augment their optical efficiency.[1−3] Due to the quantum confinement, NCs exhibit intense photoluminescence at wavelengths that can be tuned throughout the visible spectrum by changing the particle size.[4,5] Their biocompatibility, the high photoluminescence quantum efficiency, and the stability against photobleaching make silicon NCs ideal candidates for many biological imaging techniques.[6,7] The incorporation of magnetic 3d transition metal (TM) impurities in Si NCs would allow a combination of optical detection with magnetic resonance imaging techniques or magnetic separation. It has been shown experimentally and theoretically that doping of Si nanostructures with nonmagnetic impurities already leads to significantly modified properties with respect to the bulk Si case.[8−12]

The modification of the magnetic and electronic properties of Si nanostructures by TM atom doping is an exciting field. A central question concerns the influence of electron confinement on magnetism on a nanoscale, for example, the combination of possible ferromagnetism with a half-metallic character of the NC. Similar questions have been studied by means of spin-polarized density functional theory (DFT) for $\delta$-doped layers of Mn in Si[13,14] and TM-doped Si nanowires.[15−17] Also, the spin polarization in Mn-doped Ge NCs and TM-doped Si NCs as well as its consequences have been investigated recently.[18−21] Self-organized $Ge_{1-x}Mn_x$ nanocolumns are found to tend to high-Curie-temperature ferromagnetism.[22]

However, there are serious limitations of such DFT[23] studies for localized electrons using local or semilocal approximations for exchange and correlation (XC) such as the local spin density approximation (LSDA) or the (spin-polarized) generalized gradient approximation (GGA). The electrons of the open 3d shell of TM atoms such as Mn and Fe are rather strongly localized. In transition metals and their oxides, for example, the 3d electrons experience strong Coulomb repulsion because of their spatial localization. Such strongly “interacting” or “correlated” electrons cannot be simply described as embedded in a mean field generated by the other electrons.[24] Frequently, this electron correlation is characterized by an empirical intra-atomic d−d Coulomb interaction $U$ within DFT-based descriptions, so-called LDA+$U$ or GGA+$U$ methods,[25] or dynamical mean-field theory (DMFT).[24,26] Furthermore, for antiferromagnetic TM oxides, it has been demonstrated that effects related to the strong localization of the TM 3d states can be described by a spatially nonlocal potential derived from a hybrid XC functional which contains a screened exchange contribution.[27−30] Also, ferromagnetic systems may be modeled using such an approach.[31] The idea is, however, not to include simply the nonlocal Hartree−Fock exchange in which no correlation part is present. Rather, the use of a spatially nonlocal XC potential in the Kohn−Sham equation is considered as a zeroth-order approximation for the XC self-energy of the quasiparticle equation.[29,32] That allows the description of electronic single-quasiparticle excitations. Thereby, the screened Fock exchange

* Correponding author e-mail: bechsted@ifto.physik.uni-jena.de.

[†] Current address: GWT-TUD GmbH, Material Calculations, Annabergerstr. 240, 09125 Chemnitz, Germany.

[‡] Current address: Energieforschungszentrum Niedersachsen, Technische Universität Clausthal, Am Stollen 19, 38640 Goslar, Germany.

10.1021/ct9003993
Published on Web 01/22/2010

with a modified Coulomb potential and a prefactor α, whose reciprocal value can be identified with a static background dielectric constant, is used. In any case, screened exchange as the most important effect for the widening of the energetic distances between occupied and empty single-electron states is taken into account.

## Theoretical Description

The influence of strong electron correlation effects of d electrons in nanoscale systems and its interplay with the spatial confinement of s and p electrons are barely investigated so far. For that reason, we study the electronic and magnetic properties of TM-doped Si NCs in the framework of three different approaches to XC in this work: semilocal GGA,[33] the inclusion of an additional Coulomb repulsion $U$ within GGA+$U$,[34] as well as a description of XC using the hybrid functional HSE03,[35,36] which accounts for nonlocal screened exchange. Within the GGA+$U$ scheme of Dudarev et al.,[34] which is applied here, only an effective parameter $U$ representing the difference between the on-site Coulomb repulsion and the exchange parameter is meaningful. For both Mn 3d and Fe 3d electron systems, we use an effective $U = 3$ eV that is somewhat smaller than that from earlier suggestions.[25] With the study of quasiparticle band structures of TM monoxides,[30] however, it has been found that larger values of $U$ give rise to wrong band orderings. Test calculations showed that an increase of $U$ up to 5 eV does not change the electronic properties of the NCs qualitatively.

The atomic positions of the atoms in the Si NCs are determined by a shell-by-shell construction procedure which starts from a central atom and successively adds shells of Si atoms.[37] This results in faceted Si NCs with six {100} and eight {111} facets whose surface dangling bonds are passivated by H atoms. Periodic arrangements of simple cubic supercells with sufficiently large edge lengths guarantee a distance larger than 1 nm between the surfaces of NCs in adjacent supercells. The atomic geometry of the clean and doped NCs is optimized using a DFT-GGA framework as implemented in the Vienna ab initio simulation package.[38] Pseudopotentials are generated within the projector-augmented wave method,[39] which allows for an accurate description of the (all-electron) wave functions in the core region. An energy cutoff of 200 eV is used for the plane-wave expansion.

## Results and Discussion

The energetic stability of the dopant arrangement has been studied in the framework of DFT-GGA and GGA+$U$ for both Mn and Fe atoms for different substitutional doping positions, as indicated in Figure 1. The results yield a tendency of the TM atoms to occupy either the NC center or subsurface positions.[40] In the light of these results, we focus our attention on substitutional sites in the center position of the NC to retain an atomic geometry with $T_d$ point-group symmetry. Interstitial sites and arbitrary sites outside the NC center would give rise to a lowering of the point-group symmetry and, hence, a splitting of the defect levels. Such splittings hamper a clear identification of the effects resulting from different treatments of XC.[21] Therefore, the $T_d$ symmetry of the NCs is enforced during the minimization of the total energy with respect to the structural degrees of freedom. In the case of the variation of the electronic degrees of freedom, we lift the symmetry constraint in order to allow for arbitrary level occupancies and accompanying level splittings. Nevertheless, the impurity levels, especially those derived from TM 3d states, will be still classified by $e$ and $t_2$ states. This procedure allows us a more precise discussion of the effects of the electron−electron interaction beyond the semilocal XC approach. Our GGA studies (not presented here) show a relatively weak dependence of the qualitative and absolute arrangement of the impurity levels with respect to energetic position and spin channel on the NC size. The main effect concerns the gap size, as for undoped Si NCs. On the other hand, numerical calculations within the hybrid XC functional framework are prohibitive for large NCs for computer-time reasons. Therefore, we restrict the studies to the model system of $Si_{16}H_{36}TM$ (TM = Mn, Fe) nanocrystals where the central Si atom is replaced by a TM atom.

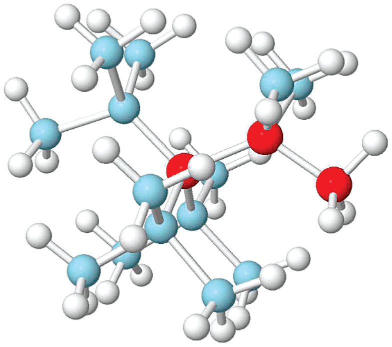


*Figure 1.* Stick-and-ball model of a $Si_{17}H_{36}$ nanocrystal with Si atoms (cyan/light gray) and H atoms (white). The three possible substitutional positions for TM atoms are indicated as red/dark gray balls.

The energy levels obtained for the NCs within the three different XC treatments GGA, GGA+$U$, and HSE03 are presented in Figures 2 (Mn doping) and 3 (Fe doping). The states with $e$ and $t_2$ symmetry and their occupation are indicated for both the majority and minority spin channels. Because of the interaction with the surrounding Si atoms, $t_2$ states with bonding and antibonding character appear. One observes a significant influence of the description of exchange and correlation using a hybrid functional (HSE03) or adding an on-site Coulomb repulsion (GGA+$U$) in comparison to the pure semilocal approximation (GGA).

Within the GGA approach, the TM 3d-derived impurity states appear in the vicinity of the fundamental gap of the undoped NC. These impurity levels, which are described and classified in terms of nonbonding states with $e$ character and $t_2$ bonding and antibonding states between TM 3d and Si $3sp^3$ orbitals, and their occupation can be explained within a defect molecule model.[41] Nine (ten) electrons are available to occupy these Mn(Fe)-derived defect levels in the two spin channels. For both spin channels, the $t_2$ levels with strong bonding (antibonding) character are fully occupied (remain empty). Consequently, the chemical bonding causes a violation of Hund's rule, which is valid for the free TM atoms. The almost half-metallic (Mn) or insulating (Fe) character of the TM-doped Si NCs is related to the occupation of the nonbonding $e$ levels. For doping with Fe, that is, a dopant with an even number of 3d electrons, the $e$

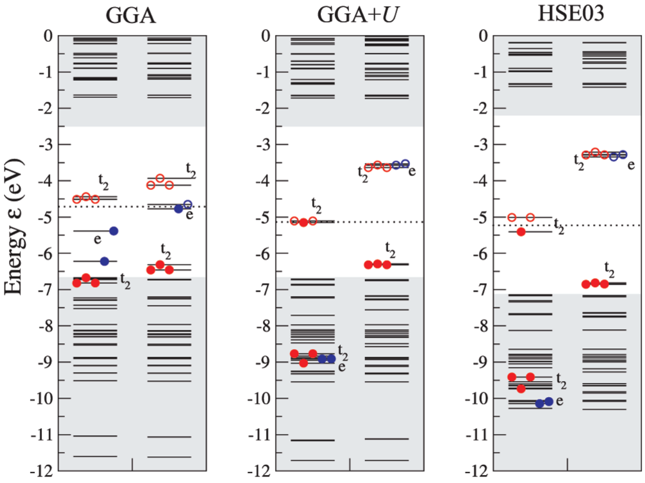


***Figure 2.*** Energy level schemes for a $Si_{16}H_{36}Mn$ nanocrystal with a central Mn atom obtained within (a) GGA, (b) GGA+$U$, and (c) HSE03 for the majority spin channel (left) and minority spin channel (right). The vacuum level is used as common energy zero. The Fermi level is given as a dotted horizontal line. The fundamental gap region of the undoped $Si_{17}H_{36}$ crystal is indicated by a white background. The occupation of levels with $e$ (blue) and $t_2$ (red) symmetry is denoted by filled red circles or empty circles, respectively.

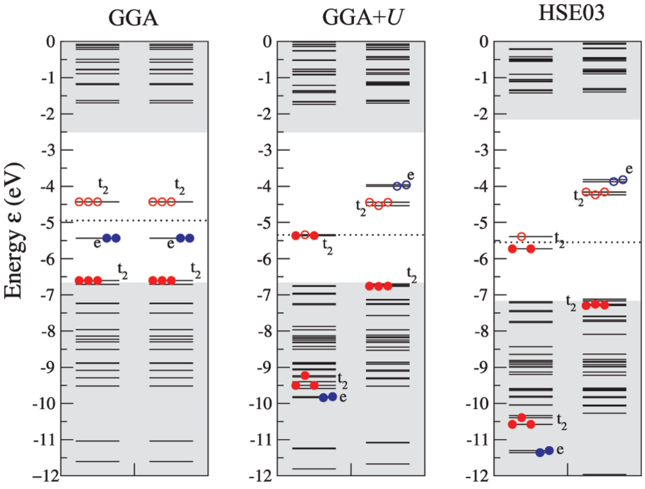


***Figure 3.*** As in Figure 2 but with Fe instead Mn.

levels in both spin channels are completely filled. The doped Si NC appears to be spin-unpolarized, that is, possesses a vanishing magnetic moment. In the Si NC doped with Mn featuring five 3d electrons, one $e$ level remains empty. Taking into account the very small gap between the two $e$ levels, this results in an almost half-metallic system with low spin polarization and a magnetic moment of about 1 $\mu_B$. In summary, low-spin configurations with $S = 1/2$ (Mn) or $S = 0$ (Fe) appear. The findings for the magnetic moments are almost in agreement with earlier GGA predictions.[20]

The inclusion of XC effects beyond GGA within the HSE03 functional or the GGA+$U$ approach yields a completely different electronic structure of the doped Si NCs in Figure 2 and Figure 3, since the positions and occupation of the impurity-derived levels are altered. Thereby, the differences due to the different XC treatment within the HSE03 and GGA+$U$ approaches are small. The main effect is already visible within the GGA+$U$ approach: Fully occupied (empty) levels with strong TM 3d character are shifted toward lower (higher) energies, while impurity states mainly localized at the four Si neighbors remain less influenced. In the majority and minority spin channels, splittings of the $t_2$-derived levels with their occupation appear for all studied treatments of XC. However, these splittings are much larger within the HSE03 approach, which is the result of the spin dependence of the Fock part in the HSE03 functional. Exchange only acts on parallel spins and mostly influences the occupied states by lowering them in energy. As a consequence, the occupied $t_2$ and $e$ levels with mainly TM 3d character shift toward lower energies and the empty levels in the opposite direction, similar to that within the GGA+$U$ scheme. Thus, the nonbonding $e$ states of the minority spin channel become unoccupied, and one additional electron occurs in the majority spin channel, resulting in spin-polarized NCs. A simple count of the difference of the electron numbers in the spin channels gives magnetic moments of 3 $\mu_B$ (Mn) or 4 $\mu_B$ (Fe). Consequently, the XC treatments beyond the semilocal GGA stabilize high-spin configurations with $S = 3/2$ (Mn) or $S = 2$ (Fe).

The Fermi level positions and the magnetic moments of the NCs depend strongly on their electronic structure and, hence, on the strong electron correlation effects. The account for on-site electron−electron interaction between the localized TM 3d electrons within the GGA+$U$ and HSE03 frameworks influences the absolute energetic positions of the occupied and empty mainly 3d-derived states with $e$ and $t_2$ symmetry. On the other hand, the $t_2$ levels in the minority spin channel with strong Si $sp^3$ character remain almost uninfluenced. Consequently, a complete change of the gap states occurs going beyond the (semi)local approximation for XC. A level mainly related to Si-derived $t_2$ defect states appears in a midgap position of the majority spin channel. The 3-fold degeneracy of this level is lifted due to an electronic Jahn−Teller effect. The lifted degeneracy of the Si-related $t_2$ defect level does not influence the high-spin state of the doped NC and the accompanying magnetic moment $\mu = 3$ $\mu_B$ (Mn) or 4 $\mu_B$ (Fe). However, in contrast to the almost half-metallic character within the GGA (Mn) and GGA+$U$ (Mn, Fe) approaches, the NC becomes insulating within the HSE03 treatment for both types of TM atoms, Mn and Fe.

Since the ground state is given by a single Slater determinant in DFT, problems concerning the description of spin multiplets arise.[43] It is also difficult to find the global minimum of the total energy with respect to the spin polarization. We have, therefore, carefully studied the total energy of the NC versus the local magnetic moment $\mu$ of the TM ions. We start the self-consistent procedure with large magnetic moments of the TM atoms (being typically by 2 $\mu_B$ larger than the “expected” value). During the electronic relaxation, the value of the magnetization is significantly reduced. Whereas within GGA only a global minimum for the low-spin configuration with $\mu = 1$ (0) $\mu_B$ was found for TM = Mn (Fe), several local minima occur for the GGA+U and HSE03 treatments, whereas the global one corresponds to the high-spin state. For TM = Mn, we compute an energy gain of 0.63 (0.65) eV for $\mu = 3$ $\mu_B$ compared to $\mu = 1$ $\mu_B$ within the GGA+$U$ (HSE03) framework. In the TM =

Fe case, the situation is more complex. The high-spin state $\mu = 4\ \mu_B$ is lower in energy by 0.14 (0.06) eV or 0.34 (−0.04) eV with respect to the intermediate-spin state $\mu = 2\ \mu_B$ or low-spin state $\mu = 0\ \mu_B$ using the GGA+$U$ (HSE03) method. That means that the total energy only weakly varies as a function of the local magnetization using the hybrid functional.

The question arises whether the obtained results follow a defect-molecule model with a central TM dopant and four nearest-neighbor Si atoms characterized by strong chemical bonding or whether such a picture is destroyed due to the strong correlation effects. To clarify this question, we neglect electronic confinement effects and compare with results for a substitutional TM doping in bulk Si, here, simulated by one TM atom in a simple cubic (sc) unit cell containing 216 atoms. For the Brillouin zone sampling, a 5 × 5 × 5 Monkhorst-Pack mesh is used. Thereby, we restrict ourselves to the GGA and GGA+$U$ treatments. The corresponding densities of states (DOS) are presented in Figure 4 for both spin channels. Because of computer-time limitations, we perform the hybrid-functional computations only for the Γ point. This allows us at least to compare the relative level position in the bulk case with those for the Si NCs. Also in the bulk limit, the inclusion of the on-site interaction changes the defect-induced levels in the fundamental gap region dramatically. The half-metallic character is conserved for Mn in Si, whereas for Fe in Si the on-site interaction $U$ destroys the insulating character and also gives rise to a half metal with partially occupied states in the majority spin channel. The results for the energy levels and their occupation are very similar to those observed in Figures 2 and 3 for the TM-doped $Si_{17}H_{36}$ nanocrystals. This holds for the relative level position and the level occupation; only the fundamental gap is much smaller. Qualitatively, the same holds for an HSE03 treatment. We conclude that the electron confinement effects influence the energy scale but not the qualitative impurity level arrangement. Such similarities of the TM impurity behavior in Si NCs and bulk Si are also observed for the magnetic moments with $\mu = 1.0$ (GGA, Mn), $\sim 0$ (GGA, Fe), 3.0 (GGA+$U$, Mn), 4.0 (GGA+$U$, Fe), 3.0 (GGA+HSE03, Mn), and 4.0 $\mu_B$ (GGA+HSE03, Fe). The reason for this congenerous behavior is closely related to the strong localization of the TM 3d states at the impurity sites: These states are hardly influenced by additional confinement effects due to the finite size of the NCs. Therefore, in Si NCs, the impurity levels exhibit a similar magnetic character as isolated TM impurities in bulk Si.

The comparison of these results with other theoretical investigations is somewhat puzzling, because of different numerical and methods and treatments of XC. For example, for bulk Si, the Green's function approach of Beeler et al.[42] predicts magnetic moments of 3 (Mn) and 0 $\mu_B$ (Fe), which agree completely neither with the GGA nor with the GGA+$U$ results. Only in the limit of higher TM concentrations, for example, described by one TM atom in a 32-atom supercell, GGA also yields $\mu = 3$ (Mn) and 0 $\mu_B$ (Fe). On the other hand, the GGA+$U$ results for bulk Si are fully consistent with the empirical rules of the Ludwig-Woodbury model, which predicts a total electron spin $S = N_e/2$, where $N_e$ is the number of electrons occupying the antibonding $t_2$ and nonbonding $e$ states (3 for Mn and 4 for Fe).[44] Strong electron correlation seems to be very important to fulfill the empirical rules for all situations. Due to the observed similarities of TM doping on the nanoscale and in bulk Si, the Ludwig-Woodbury model appears to be also applicable for nanostructures.

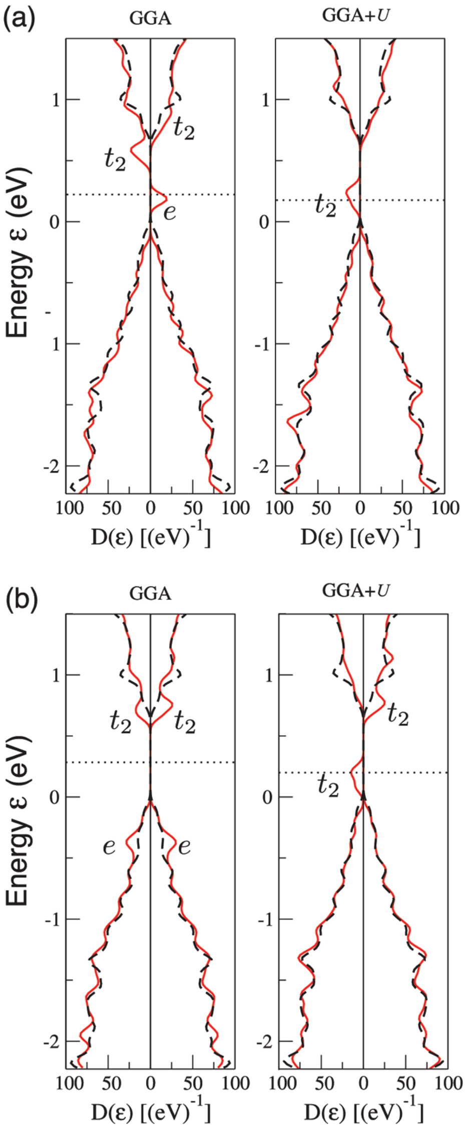


*Figure 4.* Densities of state $D(\varepsilon)$ of bulk Si doped with (a) Mn and (b) Fe at a substitutional position (red solid line) for the majority (left panel) and minority (right panel) spin channels. The bulk Si DOS (black dashed line) is shown for comparison. All DOSs are broadened using gaussians with the broadening parameter 0.1 eV. The horizontal dotted line indicates the Fermi level. The top of the valence bands of bulk Si is taken as energy zero. The main orbital character of the TM 3d-derived states close to the fundamental band gap is indicated.

## Summary

We have studied the influence of strong electron correlation effects on the electronic and magnetic properties of transition-metal doped Si NCs for highly symmetric atomic geometries. The results clearly indicate that the properties of the NCs, especially those related to the 3d shell of the dopants, depend strongly on the treatment of exchange and correlation. Thus, electron correlation due to the strong localization of the 3d electrons has to be taken into account beyond a (semi)local XC approximation. The GGA+$U$ approach gives rise to a completely changed level ordering in the fundamental gap region compared to a GGA treatment. A refined approach to exchange and correlation taking into account occupation-induced splittings within the HSE03 hybrid functional yields the same magnetic properties as GGA+$U$. Since the results are similar to the situation of TM doping of bulk silicon, the electron confinement seems not to enhance the effects of electron correlation and magnetic properties, and the magnetization still follows the Ludwig-Woodbury rules.

**Acknowledgment**. We gratefully acknowledge discussions with J. Furthmüller. Financial support was provided by the FWF Austria via SFB25 IR-ON, the Deutsche Forschungsgemeinschaft (Project No. Be1346/20-1), and the European Community within e-I3 project ETSF (GA No. 211956). We thank the Leibniz Rechenzentrum München and the HLRS in Stuttgart for grants of computer time.

## References

(1) Huh, Y.-M.; Jun, Y.-w.; Song, H.-T.; Kim, S; Choi, J.-s.; Lee, J.-H.; Yoon, S.; Kim, K.-S.; Shin, J.-S.; Suh, J.-S.; Cheon, J. In Vivo Magnetic Resonance Detection of Cancer by Using Multifunctional Magnetic Nanocrystals. *J. Am. Chem. Soc.* **2005**, *127*, 12387.

(2) Santra, S.; Yang, H.; Holloway, P. H.; Stanley, J. T.; Mericle, R. A. Synthesis of Water-Dispersible Fluorescent, Radio-Opaque, and Paramagnetic CdS:Mn/ZnS Quantum Dots: A Multifunctional Probe for Bioimaging. *J. Am. Chem. Soc.* **2005**, *127*, 1656.

(3) Michalet, X; Pinaud, F. F.; Bentolila, L. A.; Tsay, J. M.; Doose, S; Li, J. J.; Sundaresan, G; Wu, A. M.; Gambhir, S. S.; Weiss, S. Quantum Dots for Live Cells, in Vivo Imaging, and Diagnostics. *Science* **2005**, *307*, 538.

(4) Canham, L. T. Silicon quantum wire array fabrication by electrochemical and chemical dissolution of wafers. *Appl. Phys. Lett.* **1990**, *57*, 1046.

(5) Veinot, J. G. C. Synthesis, surface functionalization, and properties of freestanding silicon nanocrystals. *Chem. Commun.* **2006**, 4160.

(6) Li, Z. F.; Ruckenstein, E. Water-Soluble Poly(acrylic acid) Grafted Luminescent Silicon Nanoparticles and Their Use as Fluorescent Biological Staining Labels. *Nano Lett.* **2004**, *4*, 1463.

(7) Wang, L; Reipa, V; Blasic, J. Silicon Nanoparticles as a Luminescent Label to DNA. *Bioconjugate Chem.* **2004**, *15*, 409.

(8) Melnikov, D. V.; Chelikowsky, J. R. Quantum Confinement in Phosphorus-Doped Silicon Nanocrystals. *Phys. Rev. Lett.* **2004**, *92*, 046802.

(9) Ossicini, S.; Degoli, E.; Iori, F.; Luppi, E.; Magri, R.; Cantelle, G.; Trani, F.; Ninno, D. Simultaneously B- and P-doped silicon nanoclusters: Formation energies and electronic properties. *App. Phys. Lett.* **2005**, *87*, 173120.

(10) Fujii, M.; Yamaguchi, Y.; Takase, Y.; Ninomiya, K.; Hayashi, S. Photoluminescence from impurity codoped and compensated Si nanocrystals. *Appl. Phys. Lett.* **2005**, *87*, 211919.

(11) Pi, X. D.; Gresback, R.; Liptak, R. W.; Campell, S. A.; Kortshagen, U. Doping efficiency, dopant location, and oxidation of Si nanocrystals. *Appl. Phys. Lett.* **2008**, *92*, 123102.

(12) Xu, Q.; Luo, J.-W.; Li, S.-S.; Xia, J.-B.; Li, J.; Wei, S.-H. Chemical trends of defect formation in Si quantum dots: The case of group-III and group-V dopants. *Phys. Rev. B* **2007**, *75*, 235304.

(13) Qian, M. C.; Fong, C. Y.; Liu, K.; Pickett, W. E.; Pask, J. E.; Yang, L. H. Half-Metallic Digital Ferromagnetic Heterostructure Composed of a $\delta$-Doped Layer of Mn in Si. *Phys. Rev. Lett.* **2006**, *96*, 027211.

(14) Wu, H.; Kratzer, P.; Scheffler, M. Density-Functional Theory Study of Half-Metallic Heterostructures: Interstitial Mn in Si. *Phys. Rev. Lett.* **2007**, *98*, 117202.

(15) Durgun, E.; Cakir, D.; Akman, N.; Ciraci, S. Half-Metallic Silicon Nanowires: First-Principles Calculations. *Phys. Rev. Lett.* **2007**, *99*, 256806.

(16) Durgun, E.; Akman, N.; Ciraci, S. Functionalization of silicon nanowires with transition metal atoms. *Phys. Rev. B* **2008**, *78*, 195116.

(17) Xu, Q.; Li, J.; Li, S.-S.; Xia, J.-B. The formation and electronic structures of 3d transition-metal atoms doped in silicon nanowires. *J. Appl. Phys.* **2008**, *104*, 084307.

(18) Huang, X.; Makmal, A.; Chelikowsky, J. R.; Kronik, L. Size-Dependent Spintronic Properties of Dilute Magnetic Semiconductor Nanocrystals. *Phys. Rev. Lett.* **2005**, *94*, 236801.

(19) Arantes, J. T.; Dalpian, G. M.; Fazzio, A. Quantum confinement effects on Mn-doped InAs nanocrystals: A first-principles study. *Phys. Rev. B* **2008**, *78*, 045402.

(20) Ma, L. T.; Zhao, J.; Wang, J.; Wang, B.; Wang, G. Magnetic properties of transition-metal impurities in silicon quantum dots. *Phys. Rev. B* **2007**, *75*, 045312.

(21) Leitsmann, R.; Panse, C.; Küwen, F.; Bechstedt, F. Ab initio characterization of transition-metal-doped Si nanocrystals. *Phys. Rev. B* **2009**, *80*, 104412.

(22) Jamet, M.; Barski, A.; Devillers, T.; Poydenot, V.; Dujardin, R.; Bayle-Guillemaud, P.; Rothman, J.; Bellet-Amalric, E.; Marty, A.; Cibert, J.; Mattana, R.; Tatarenko, S. High-Curie-temperature ferromagnetism in self-organized Ge $1_x$ $Mn_x$ nanocolumns. *Nature Mat.* **2006**, *5*, 653.

(23) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133.

(24) Imada, M.; Fujimori, A.; Tokura, Y. Metal-insulator transitions. *Rev. Mod. Phys.* **1998**, *70*, 1039.

(25) Anisimov, V. I.; Zaanen, J.; Andersen, O. K. Band theory and Mott insulators: Hubbard U instead of Stoner I. *Phys. Rev. B* **1991**, *44*, 943.

(26) Biermann, S.; Aryasetiawan, F.; Georges, A. First-Principles Approach to the Electronic Structure of Strongly Correlated Systems: Combining the GW Approximation and Dynamical Mean-Field Theory. *Phys. Rev. Lett.* **2003**, *90*, 086402.

(27) Franchini, C.; Bayer, V.; Podloucky, R.; Paier, J.; Kresse, G. Density functional theory study of MnO by a hybrid functional approach. *Phys. Rev. B* **2005**, *72*, 045132.

(28) Marsman, M.; Paier, J.; Stroppa, A.; Kresse, G. Hybrid functionals applied to extended systems. *J. Phys.: Condens. Matter* **2008**, *20*, 064201.

(29) Rödl, C.; Fuchs, F.; Furthmüller, J.; Bechstedt, F. Ab initio theory of excitons and optical properties for spin-polarized systems: Application to antiferromagnetic MnO. *Phys. Rev. B* **2008**, *77*, 184408.

(30) Rödl, C.; Fuchs, F.; Furthmüller, J.; Bechstedt, F. Quasiparticle band structures of the antiferromagnetic transition-metal oxides MnO, FeO, CoO, and NiO. *Phys. Rev. B* **2009**, *79*, 235114.

(31) Wilson, N. C.; Russo, S. P. Hybrid density functional theory study of the high-pressure polymorphs of $\alpha$-$Fe_2O_3$ hematite. *Phys. Rev. B* **2009**, *79*, 094113.

(32) Fuchs, F.; Furthmüller, J.; Bechstedt, F.; Shishkin, M.; Kresse, G. Quasiparticle band structure based on a generalized Kohn-Sham scheme. *Phys. Rev. B* **2007**, *76*, 115109.

(33) Perdew, J. P.; Wang, Y. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B* **1992**, *45*, 13244.

(34) Dudarev, S. L.; Botton, G. A.; Savrasov, S. Y.; Humphreys, C. J.; Sutton, A. P. Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDA+U study. *Phys. Rev. B* **1998**, *57*, 1505.

(35) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **2003**, *118*, 8207.

(36) Krukau, A.; Vydrov, O.; Izmaylov, A.; Scuseria, G. Influence of the exchange screening parameter on the performance of screened hybrid functionals. *J. Chem. Phys.* **2006**, *125*, 224106.

(37) Ramos, L. E.; Furthmüller, J.; Bechstedt, F. Effect of backbond oxidation on silicon nanocrystallites. *Phys. Rev. B* **2005**, *70*, 033311.

(38) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **1996**, *6*, 15.

(39) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **1999**, *59*, 1758.

(40) Küwen, F.; Leitsmann, R.; Bechstedt, F. Mn and Fe doping of bulk Si: Concentration influence on electronic and magnetic properties. *Phys. Rev. B* **2009**, *80*, 45203.

(41) Enderlein, R.; Horing, N. In *Fundamentals of Semiconductor Physics and Devices*; World Scientific: London, 1997; p 285.

(42) Beeler, F.; Andersen, O.; Scheffler, M. Electronic and magnetic structure of 3d transition-metal point defects in silicon calculated from first principles. *Phys. Rev. B* **1990**, *41*, 1603.

(43) Zywietz, A.; Furthmüller, J.; Bechstedt, F. Spin state of vacancies: From magnetic Jahn-Teller distortions to multiplets. *Phys. Rev. B* **2000**, *62*, 6854.

(44) Ludwig, G.; Woodbury, H. In *Solid State Physics*; Academic: New York 1962; Vol. 13, p 331.

CT9003993

JCTC Journal of Chemical Theory and Computation

# Combined Quantum Mechanical and Molecular Mechanical Methods for Calculating Potential Energy Surfaces: Tuned and Balanced Redistributed-Charge Algorithm

Bo Wang and Donald G. Truhlar*

*Department of Chemistry and Supercomputing Institute, University of Minnesota, 207 Pleasant Street South East, Minneapolis, Minnesota 55455-0431*

Received July 15, 2009

**Abstract:** The combined quantum mechanical and molecular mechanical (QM/MM) method is one of the most powerful approaches for including correlation and polarization effects in simulations of large and complex systems, and the present article is concerned with the systematics of treating a QM/MM boundary that passes through a covalent bond, especially a polar covalent bond. In this study, we develop a new algorithm to treat such boundaries; the new method is called the balanced redistributed charge (balanced RC or BRC) scheme with a tuned fluorine link atom. The MM point charge on the MM boundary atom is modified to conserve the total charge of the entire system, and the modified charge is redistributed to the midpoints of the bonds between an MM boundary atom and its neighboring MM atoms. A pseudopotential is added to the fluorine link atom to reproduce the partial charge of the uncapped portion of the QM subsystem. We select proton affinities as the property used to validate the new method because the energy change associated with the addition of an entire charge (proton) to the QM system is very sensitive to the treatment of electrostatics at the boundary; we apply the new method to calculate proton affinities of 25 molecules with 13 different kinds of bonds being cut. The average proton affinity in the test set is 373 kcal/mol, and the test set provides a more challenging test than those usually used for testing QM/MM methods. For this challenging test set, common unbalanced schemes give a mean unsigned error (MUE) of 15−21 kcal/mol for H link atoms or 16−24 kcal/mol for F link atoms, much larger than the 5 kcal/mol obtained by simply omitting the MM region with either kind of link atom. Balancing the charges reduces the error to 5−7 kcal/mol for H link atoms and 4−6 kcal/mol for F link atoms. Balancing the charges and also tuning an F link atom lowers the MUE to 1.3−4 kcal/mol, with the best result for the balanced RC scheme. We conclude that properly tuning the link atom and correctly treating the point charges near the QM/MM boundary significantly improves the accuracy of the calculated proton affinities.

## 1. Introduction

The application of quantum chemistry to large and complex systems is one of the most challenging areas of current computational chemistry and also one that is seeing the most progress.[1] An important tool for such applications is the combined quantum mechanical/molecular mechanical (QM/MM) method for calculating potential energy surfaces and interatomic forces; the reader is directed to several reviews and overviews for background information.[2−23]

A stubborn issue in QM/MM calculations is the treatment of the boundary between the QM and MM regions when it

* Corresponding author phone: (612) 624-7555; fax: (612) 624-9390; e-mail: truhlar@umn.edu.

10.1021/ct900366m © 2010 American Chemical Society
Published on Web 01/07/2010

passes through a bond, which is practically unavoidable in the treatment of many solids, polymers, and complex systems. In general the QM region is capped to saturate dangling valences caused by the cut. Three different kinds of methods have been proposed to deal with capping the QM boundary atom. The first one is the link atom approach (LA).[24,25] The dangling bond of the QM region is capped with an additional atom (usually a hydrogen atom) and the QM calculations are performed on this capped system. The second method is localized orbitals.[26−28] The dangling bond is saturated by orbitals rather than by an atom. Examples of this approach are the local self-consistent field (LSCF) method[26] and the generalized hybrid orbitals (GHO) scheme.[27,28] The third kind of method involves a pseudobond or an effective core potential (ECP). In this approach, a parametrized atom, modified to mimic the behavior of the original MM boundary atoms or groups, is used to cap the QM system; examples of this approach are tuned capping atoms,[29−31] adjusted connection atoms,[32] a pseudobond,[33−35] an effective group potential,[36] a quantum capped potential,[37−39] and a variationally optimized effective atom-centered potential.[40] This third class of methods may be considered to be a second-generation link-atom method in which the link atom is optimized or tuned.

Though much progress has been made, there are still many problems in the treatment of QM−MM boundaries that pass through a bond. Most attention has been devoted to the cutting of C−C bonds, especially for modeling enzymatic binding and reactions, but some procedures are more general. The methods that have been developed exhibit a wide variety of differences in the precise way in which they have been implemented.

Pople has emphasized the importance of theoretical models, where a theoretical model is "an approximate but well-defined mathematical procedure for simulation. . . The approximate mathematical treatment must be precisely formulated. It should be general. . . . Particular procedures for particular molecules. . . should be avoided."[41] If tests of the model against a broad data set are successful, the model is said to be validated. The goal of this article is to develop and validate a new method, in the spirit of a theoretical model chemistry, for the treatment of a boundary between bonded atoms in QM/MM simulations. It is precisely defined in a general way applicable to all systems and all kinds of single bonds, and it is tested against a data set of 25 systems in which 13 different kinds of bonds are cut, in particular (where the atom listed first is in the QM subsystem, and the one listed second is in the MM subsystem): C−C, N−C, O−C, S−C, C−N, O−N, C−O, Al−O, Si−O, C−Si, O−Si, C−S, and S−S.

## 2. Methods

Our group has developed redistributed charge (RC) and redistributed charge and dipole (RCD) methods to treat the charges near a QM/MM boundary that passes through a bond.[42] These methods give good results even when large charges are present near the boundary. In the current work, we improve the RC and RCD methods by adding two new elements, a charge balancing step and a tuned link atom. In particular, the redistributed charges are used to conserve the charge of the entire system, and a tuned fluorine atom is used to saturate the free valence of the QM region and to reproduce the partial charge of the uncapped portion of the capped QM subsystem. The improved method is used to treat polar bonds between the QM and MM subsystems with large partial atomic charges near the boundary. In order to describe the algorithm, we label the atoms according to "tiers". The definition is the same as what is used in previous work;[4,42] in particular, the MM boundary atoms are denoted as M1 atoms, and the MM atoms directly bonded to M1 atoms are denoted as M2 atoms. M3 atoms are the third-tier MM atoms. The QM boundary atoms are denoted as Q1 atoms. The QM atoms directly bonded to Q1 atoms are labeled Q2 atoms. Q3 atoms are those bonded to Q2 atoms and so forth for Q4, Q5, etc. The QM region is also called the primary subsystem (PS) in this study. The sum of all QM atoms and MM atoms before the cutting and capping is called the original entire system. The sum of the capped QM subsystem and the whole MM subsystem after the charge redistribution is called the QM/MM entire system.

In the QM/MM calculations, we use an additive QM/MM scheme to define the total energy of the system:[23]

$$E = E^{\mathrm{QM}} + E^{\mathrm{QM/MM}} + E^{\mathrm{MM}} \tag{1}$$

$$E^{\mathrm{QM/MM}} = E_{\mathrm{el}}^{\mathrm{QM/MM}} + E_{\mathrm{vdW}}^{\mathrm{QM/MM}} + E_{\mathrm{val}}^{\mathrm{QM/MM}} \tag{2}$$

where $E^{\mathrm{QM}}$ is the quantum mechanical energy of the QM region, $E^{\mathrm{MM}}$ is the molecular mechanical energy of the MM region, and $E^{\mathrm{QM/MM}}$ accounts for the interaction energy between the QM and the MM regions. $E^{\mathrm{QM/MM}}$ is decomposed into three terms; $E_{\mathrm{el}}^{\mathrm{QM/MM}}$ represents electrostatic interactions, $E_{\mathrm{vdW}}^{\mathrm{QM/MM}}$ represents van der Waals interactions, and $E_{\mathrm{val}}^{\mathrm{QM/MM}}$ represents valence interactions. In this study, we will concentrate on the electrostatic coupling term $E_{\mathrm{el}}^{\mathrm{QM/MM}}$, which is the most technically involved term. The $E_{\mathrm{vdW}}^{\mathrm{QM/MM}}$ and $E_{\mathrm{val}}^{\mathrm{QM/MM}}$ terms will cancel out in the present work because we study fixed-geometry proton affinities to isolate the electrostatic terms, but these other QM/MM terms will be studied later when we consider QM/MM geometry optimization.

**2.A. Treatment of Boundary Charge.** It has been found that it is important to conserve the total charge of the QM/MM entire system in QM/MM calculations,[43] that is, the sum of the MM partial atomic charges of the MM region and the QM charge of the capped QM region should equal the total charge of the original entire system, as shown in eq 3:

$$q^{\mathrm{MM}} + q^{\mathrm{QM}} = q^{\mathrm{total}} \tag{3}$$

However, when the original entire system is divided into QM and MM regions, the sum of MM charges of the MM region does not necessarily equal zero or an integer. If MM charges are not modified, the total charge of the QM/MM entire system is not conserved. Several workers have recognized that this causes inaccuracies and have suggested various methods to remedy this.[33,43−45] Sherwood et al.[44] adjusted the charge on an M1 atom to conserve the total charge of the QM/MM entire system,

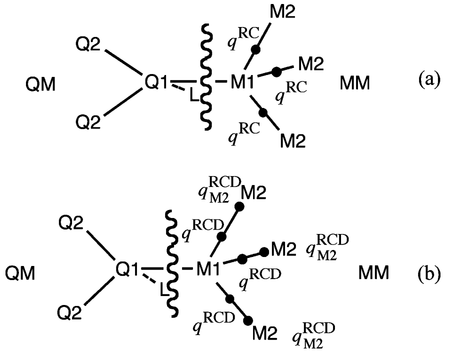


***Figure 1.*** QM/MM boundary treatments in (a) the balanced RC scheme and (b) the balanced RCD scheme.

and they redistributed the adjusted charge on the M1 atom evenly to M2 atoms; point dipoles were added at the M2 atoms to compensate the changes in the M1−M2 bond dipoles due to the movement of the charges. Zhang et al.[33] zeroed the charges on all MM atoms that are in the same group as the M1 atom. Das et al.[45] used a double link atom approach combined with delocalized Gaussian MM charges. Walker et al.[43] added the charge difference to the nearest M2 atom or evenly to all the MM atoms except the M1 atom.

In the previous RC scheme,[42] the charge on each M1 atom is redistributed to the midpoints of M1−M2 bonds. However, the total charge of the QM/MM entire system is not conserved when the sum of MM charges of the MM region is not zero or an integer. In the balanced RC scheme, introduced here, we first adjust the charge on the M1 atom so that

$$q_0 + \sum_i q_i + q^{\mathrm{QM}} = q^{\mathrm{total}} \tag{4}$$

where $q_0$ is the modified M1 charge, $\{q_i\}$ are the MM point charges of other MM atoms (except M1), $q^{\mathrm{QM}}$ is the charge of the QM region (that is, of the capped QM subsystem), and $q^{\mathrm{total}}$ is the charge of the original entire system. This charge balancing step conserves the charge of the QM/MM entire system.

Then the balanced RC scheme redistributes the charge $q_0$ evenly to the midpoints of all M1−M2 bonds, with each bond midpoint obtaining a charge $q^{\mathrm{RC}} = (q_0/n)$, where $n$ is the number of M1−M2 bonds. For the balanced redistributed charge and dipole (balanced RCD) method, we double the redistributed charges and adjust the charges $q_{\mathrm{M2}}^{\mathrm{RCD}}$ on M2 atoms to conserve the total charge of the QM/MM entire system, as shown in eqs 5 and 6:

$$q^{\mathrm{RCD}} = 2q^{\mathrm{RC}} = \frac{2q_0}{n} \tag{5}$$

$$q_{\mathrm{M2}}^{\mathrm{RCD}} = q_{\mathrm{M2}} - q^{\mathrm{RC}} \tag{6}$$

These two schemes are illustrated in Figure 1.

In this study, we compare balanced RC and balanced RCD to other methods that differ in how the redistributed charges are handled, e.g., to what location are they redistributed. These methods include: balanced straight electrostatic embedding (SEE), balanced RC2, Amber-1,[43] balanced RC3, Amber-2,[43] and balanced Shift.[44] Amber-1 and Amber-2 are the options called *adjust_q* = 1 and *adjust_q* = 2 in *AMBER 10*. The distinction between these methods is in the position of the redistributed charges and whether the dipoles of the M1−M2 bonds are corrected. In balanced SEE, the charge on the M1 atom is set to $q_0$, and it is not moved. In balanced RC2, we distribute $q_0$ evenly to all M2 atoms. In balanced RC3, we distribute $q_0$ evenly to all M2 and M3 atoms. In Amber-1, we move $q_0$ to the nearest M2 atom. In Amber-2, we distribute $q_0$ evenly to all MM atoms, except the M1 atom. (Note that Amber-2 is the default option in revision 10 of *AMBER*,[46] whereas Amber-1 can be selected in *AMBER 10* by specifying *adjust_q* = 1.) In balanced Shift, the redistributed charges are placed at M2 atoms, and dipoles are added around M2 atoms to compensate the movement of the charges. A summary of these charge schemes is shown in Table 1.

We call the methods in Table 1 balanced methods because they all conserve the total charge of the QM/MM entire system. Five unbalanced methods, in which the total charge of the QM/MM entire system is not necessarily conserved, are also tested, including SEE, Z1, Z2, Z3, and RC.[42] SEE is straight electronic embedding that makes no change of the charges of MM boundary atoms, Z1 denotes that the charge of the M1 atom is zeroed (this can be chosen by specifying *adjust_q* = 0 in *AMBER 10*, and it is the default method in CHARMM[47]), Z2 denotes that the charges of M1 and M2 atoms are zeroed (Z2 is the default scheme in both *Gaussion 03*[48] and *Gaussian 09*[49]), and Z3 denotes that all the charges of all M1, M2, and M3 atoms are zeroed. RC denotes that the charge on the M1 atom is redistributed to the midpoints of M1−M2 bonds without the balancing step. Balanced methods and unbalanced methods are compared to test the importance of conserving the charge of the QM/MM entire system. To make a comparison, we also carry out calculations on the capped primary system (CPS), in which the whole MM region is substituted by the link atom.

**2.B. Link Atom.** Another issue in the boundary treatment is the choice of the link atom. A hydrogen atom can be used as the link atom when a C−C bond is cut. However, a Q1−H bond may be a poor model for the cut bond when the M1 atom is electronegative, such as in a Si−O or C−O bond. Therefore, we use a tuned capping atom as the link atom to mimic a cut polar bond and to reproduce the electronic structure of the QM subsystem. Redondo et al.[29] used a tuned hydrogen atom to replace a silicon atom. Koga et al.[30] added a shift operator on the hydrogen atom to reproduce the effect of the substitution. Zhang et al.[33] and Nasluzov et al.[31] used tuned fluorine atoms and derived pseudopotentials for carbon and oxygen boundary atoms. Here, we provide a more general rule to tune a capping atom for boundary atoms. The capping atom is always a tuned F atom. We first replace the $1s^2$ core by a conventional pseudopotential $U$, and then a tuning pseudopotential $U_0'(r)$ is added to $U$. The con-

***Table 1.*** Charge Schemes

| | position of the redistributed charges | correction of bond dipole | ref |
|---|---|---|---|
| balanced SEE | M1 atom | no | present |
| balanced RC | midpoints of M1−M2 bonds | no | present |
| balanced RC2 | M2 atoms | no | present |
| Amber -1 | nearest M2 atom | no | Walker et al.[43] |
| balanced RC3 | M2, M3 atoms | no | present |
| Amber -2 | all MM atoms (except M1 atom) | no | Walker et al.[43] |
| balanced RCD | midpoints of M1−M2 atoms | yes | present |
| balanced Shift | M2 atoms | yes | Sherwood et al.[44] |

***Table 2.*** CRENBL Effective Core Potential[a]

| | $n_{lj}$ | $\alpha_{lj}$ | $C_{lj}$ |
|---|---|---|---|
| $U_0$ | 2 | 2.8835 | 12.685 306 |
| | 2 | 3.1077 | −19.302 589 |
| | 1 | 5.6122 | 1.002 179 |
| | 0 | 2.8146 | 2.245 349 |
| $U_1$ | 2 | 44.5166 | −6.723 024 |
| | 2 | 12.9487 | −0.929 649 |
| | 1 | 132.4967 | −1.526 734 |

[a] Reference 50.

ventional pseudopotential used here is the CRENBL effective core potential (ECP) for a fluorine atom developed by Pacios and Christiansen.[50]

The form of this potential is

$$U = U_0(r) + \sum_{m=-1}^{1} [U_1(r) - U_0(r)]|1m\rangle\langle m1| \qquad (7)$$

where

$$U_l(r) = r^{-2}\sum_{j} C_{lj} r^{n_{lj}} \mathrm{e}^{-\alpha_{lj} r^2} \qquad (8)$$

$r$ is the distance of an electron from the capping nucleus, and $|lm\rangle$ is a spherical harmonic. The parameters for this pseudopotential are listed in Table 2. The form of $U_0'(r)$ is

$$U_0'(r) = C\exp[-(r/r_0)^2] \qquad (9)$$

where $C$ and $r_0$ are parameters. The basis set used for the tuned F atom is the same as for a conventional F atom. For example, if the QM subsystem is treated by the 6-31G* basis set, then the tuned F atom has the 6-31G* basis set of a conventional F atom. To find an appropriate pseudopotential, we set $r_0$ equal to 1 bohr and tune the parameter $C$ of the pseudopotential.

The next key decision is how to choose $C$. In order to reproduce the electronic structure of the QM subsystem, we require that the total charge of the uncapped portion of the QM subsystem in a QM/MM calculation is equal to the total charge of the same subsystem in a QM calculation of the original entire system or, in practice, of a system that mimics the original entire system better than the capped QM subsystem does (see below for more details of this large system). Mulliken charges were used as the indicator. Because Mulliken charges become unphysical when large basis sets are used, we used small basis sets without diffuse functions for this tuning step, in particular, 6-31G* when M1 is from the second period (Li through F) and STO-3G otherwise. Since the STO-3G basis set is defined for the entire periodic table, the tuning step is well-defined for the entire periodic table.

We can perform the tuning process on either the reactant or the product. For the validation suite, the reactant is a neutral molecule, and the product is a deprotonated anion. In this study, we used the protonated neutral reactant to tune the F atom. All the parameters $C$ of the pseudopotentials are derived in the presence of MM background charges. Because the MM charges are redistributed differently in the various boundary charge schemes explained in section 2.A, the derived pseudopotentials are not the same for different charge schemes.

To enable the method to be applied to large systems for which it is difficult to perform a QM calculation on the original entire system, the tuning is performed on a model system created from the original entire system. This model system is called the tuning system or the entire system model (ESM). It consists of all QM atoms and all M1, M2, and M3 atoms, and all free valences on M3 atoms are capped by untuned H atoms. The scheme is illustrated in Figure 2.

The completely defined tuning process employed in the present study is as follows:

1. Choose a geometry and charge state for the tuning system and create the entire system model (ESM) by capping all M3 atoms with untuned hydrogens.

2. Do a full QM calculation on ESM and carry out Mulliken population analysis. For the basis set, use 6-31G*

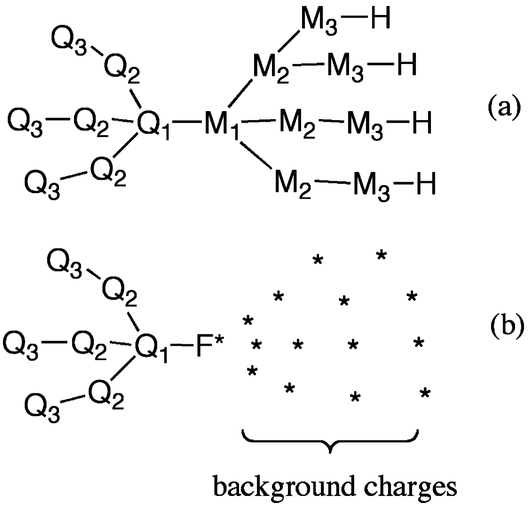


***Figure 2.*** Determining the pseudopotential for the tuned F atom in the entire system model (ESM): (a) ESM and (b) CPS**, which is the capped QM subsystem with background charges to replace the rest of the ESM.

if M1 is from the second period (e.g., C, N, O) and use STO-3G otherwise (that is, if M1 belongs to the third or higher period, for example, Si). This yields the total charge on the primary system (PS); call this $q_{\mathrm{PS}}^{\mathrm{ESM,MPA}}$, where MPA denotes Mulliken population analysis. It also yields $q_{\mathrm{SS}}^{\mathrm{ESM,MPA}}$, where SS is the secondary subsystem, equal to all the atoms in ESM except the PS atoms. By construction, $q_{\mathrm{PS}}^{\mathrm{ESM,MPA}} + q_{\mathrm{SS}}^{\mathrm{ESM,MPA}} = q_{\mathrm{ESM}}$ where $q_{\mathrm{ESM}}$ is the total charge of ESM.

3. Select an MM charge scheme. For the present calculations, the MM charge scheme is CM4M charges from a calculation on the ESM. The basis set used for the calculation of CM4M charges could in principle be the same as chosen for step 2, but in fact, we do not have CM4M charge schemes for STO-3G; therefore, the MM charges are always CM4M charges determined with M06-2X/6-31G* calculations on the ESM.

4. Define TSS as the truncated secondary system of the original entire system model, which includes all atoms in the secondary subsystem of the ESM except the M1 atom. For the chosen MM charge scheme of step 3, calculate $q_{\mathrm{TSS}}^{\mathrm{ESM,MM}}$, which equals the sum of the MM charges from step 3 (that is, the sum of CM4M charges) on all TSS atoms of ESM.

5. Cap the PS with an F* atom to create the capped primary system (CPS), where F* denotes a tuned F atom. Always set $r_0$ equal to 1 bohr in the pseudopotential. The other parameter ($C$) of the pseudopotential will be determined in step 7.

6. Select a charge modification scheme, for example, balanced RC or balanced RCD. For the balanced charge schemes, we set $q_0$ to make $q_{\mathrm{CPS}} + q_0 + q_{\mathrm{TSS}}^{\mathrm{ESM,MM}}$ equal to $q_{\mathrm{ESM}}$. In the usual case where $q_{\mathrm{CPS}} = q_{\mathrm{ESM}}$, then this yields $q_0 = -q_{\mathrm{TSS}}^{\mathrm{ESM,MM}}$.

7. Now, for a given MM charge scheme, and given the charge modification scheme, carry out a series of fixed-geometry CPS** calculations with various values of $C$. Note that CPS** here denotes the capped primary system in the modified charge environment of the secondary system of the ESM. Adjust $C$ until $q_{\mathrm{F*}}^{\mathrm{MPA}}$ equals $q_{\mathrm{SS}}^{\mathrm{ESM,MPA}}$, which was determined in step 2. Now the pseudopotential is known, so F* is properly tuned.

After the tuning step, the tuned F link atom can be used with the selected charge scheme on the QM/MM entire system to do QM/MM calculations on the proton affinities.

## 3. Details of Validation Calculations

We have implemented the proposed charge schemes and link atom treatments in the QMMM program,[51] which is based on the *Gaussian 03*[48] and TINKER[52] programs. Either density functional theory (DFT) or wave function theory (WFT) can be used for the QM calculations. In the study, the M06-2X density functional[53,54] was used for all the QM calculations. Proton affinities were used as the criterion to evaluate the methods, as they are sensitive to the boundary treatment.[47] The 6-31G** basis set was used for aluminosilicate clusters, and the 6-31G* basis set was used for organic molecules.

The geometry of all the molecules was fixed at the protonated geometry, so in eqs 1 and 2, the QM/MM valence term $E_{\mathrm{val}}^{\mathrm{QM/MM}}$, the QM/MM van der Waals term $E_{\mathrm{vdW}}^{\mathrm{QM/MM}}$ and the MM term $E^{\mathrm{MM}}$ are the same for the deprotonated and protonated forms, and they cancel out in the QM/MM calculations of proton affinities. The final expression for the proton affinity is

$$\begin{aligned} E(\text{proton affinity}) &= E(\text{deprotonated}) - E(\text{protonated}) \\ &= [E^{\mathrm{QM}}(\text{deprotonated}) + E_{\mathrm{el}}^{\mathrm{QM/MM}}(\text{deprotonated})] - \\ &\quad [E^{\mathrm{QM}}(\text{protonated}) + E_{\mathrm{el}}^{\mathrm{QM/MM}}(\text{protonated})] \end{aligned} \tag{10}$$

In this study, CM4M charges[55] were used for MM atoms. The protonated forms of the molecules in the test suite are illustrated in Figure 3. In each case, the QM region is on the left and the MM region is on the right. The selected molecules in the test suite contain different kinds of Q1−M1 bonds at the QM/MM boundary, in particular, C−C, N−C, O−C, S−C, C−N, O−N, C−O, Al−O, Si−O, C−Si, O−Si, C−S, and S−S. Both polar and nonpolar bonds are included in this test suite.

For the position of the link atom, the scaled bond distance method[56,57] was used, that is, the link atom is placed along the Q1−M1 bond, and the ratio of the Q1−link atom distance to the Q1−M1 distance is set to be the ratio of the standard bond length of the Q1−link atom bond to the standard bond length of the Q1−M1 bond. The standard bond lengths used in this study are listed in Table 3. For the tuned F link atom, it is placed at the same position as that of an ordinary F atom.

## 4. Results

**4.A. Example for Balancing the Charge and Tuning the Link Atom.** We use molecule **w** as an example to demonstrate the proposed algorithm. The O−Si bond is cut. In the tuning process, we create the entire system model (ESM) from the original QM system. The original entire system and the ESM are shown in Figure 4. A full QM calculation is performed on the ESM to get the total Mulliken charge on the QM region $q_{\mathrm{PS}}^{\mathrm{ESM,MPA}}$ and on the MM region $q_{\mathrm{SS}}^{\mathrm{ESM,MPA}}$. As the M1 atom is silicon, the STO-3G basis set was used. CM4M charges were calculated for all the MM atoms in ESM; the M1 atom has a charge of 0.515$e$, and the sum $q_{\mathrm{TSS}}^{\mathrm{ESM,MM}}$of the MM charges in the truncated secondary system (TSS) is −0.377$e$. Therefore, in the balanced schemes, the redistributed charge $q_0 = -q_{\mathrm{TSS}}^{\mathrm{ESM,MM}} = 0.377e$. Then the QM system is capped with a tuned F atom, and the capped QM subsystem is embedded in the redistributed MM charges using a boundary charge scheme. The parameter $C$ of the pseudopotential is adjusted to make the Mulliken charge $q_{\mathrm{F*}}^{\mathrm{MPA}}$ of the tuned F atom equal to $q_{\mathrm{SS}}^{\mathrm{ESM,MPA}}$. For example, in the balanced RC scheme, the parameter of the pseudopotential is 0.80, as shown in Table 6.

The tuned F* link atom is used to cap the QM system in the QM/MM entire system. The same charge scheme is used for the tuning and the calculations of proton affinities. The results with the tuned F atom are compared to those with untuned H and F link atoms.

**4.B. H and F Link Atoms.** Tables 4 and 5 show the proton affinities of the 25 molecules by full QM calculations

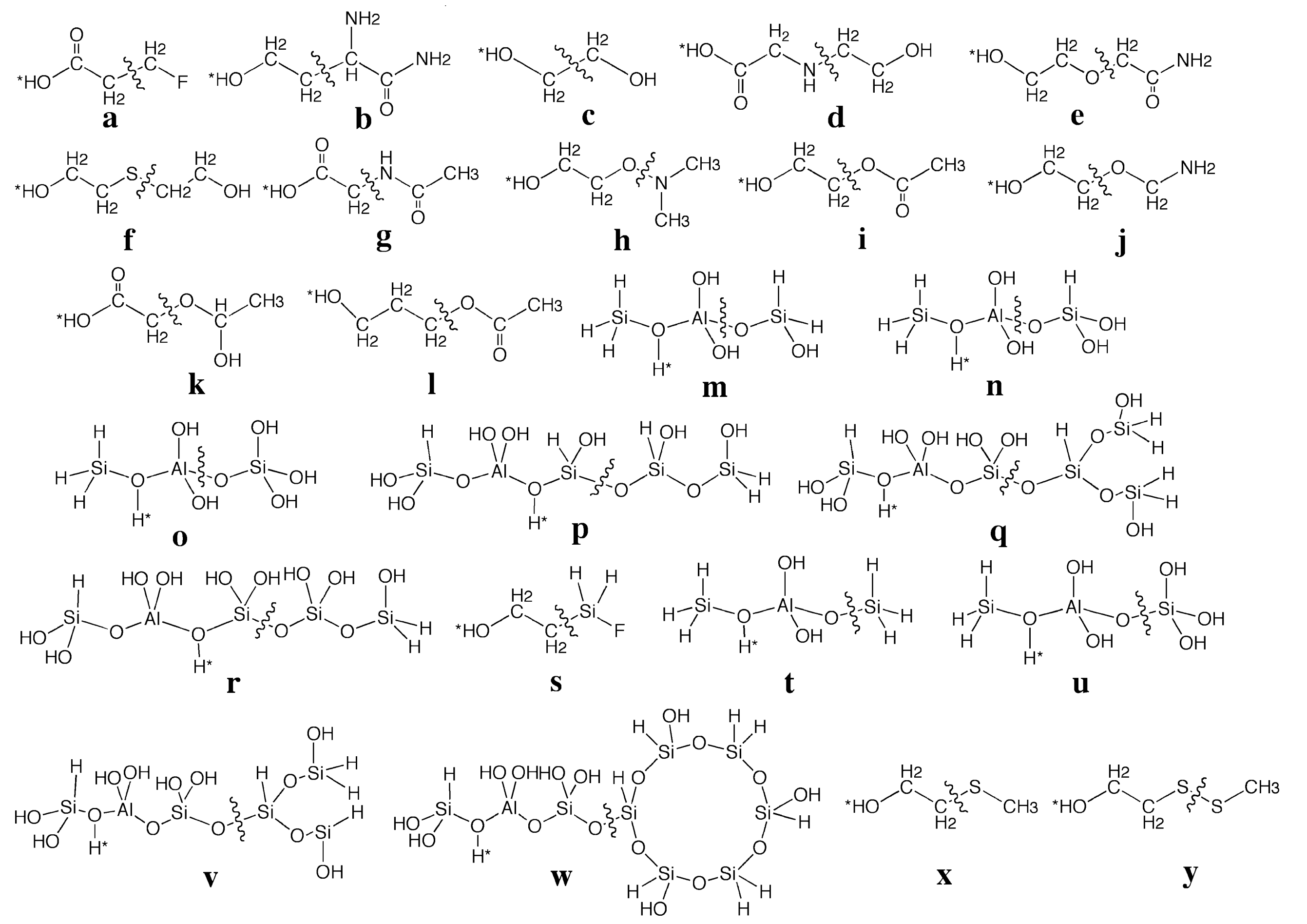


***Figure 3.*** The 25 molecules used in the test suite. The QM subsystem is on the left, and the MM subsystem is on the right. The * represents the proton involved in the protonation process.

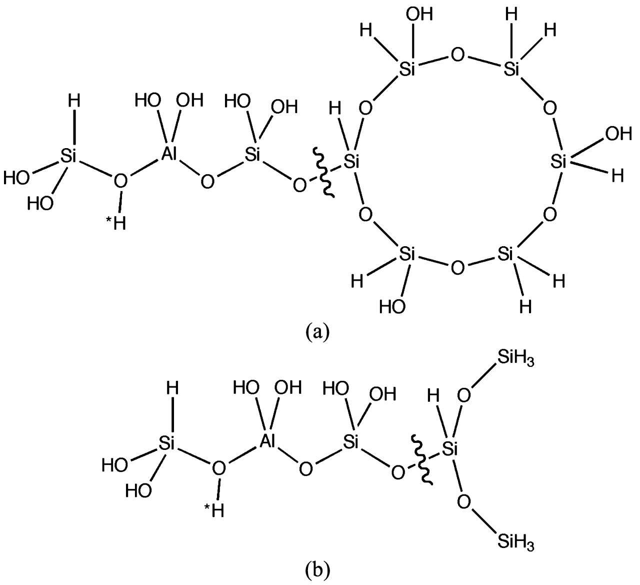
***Table 3.*** Standard Bond Lengths (Å)

| bond | distance | bond | distance | bond | distance | bond | distance |
|---|---|---|---|---|---|---|---|
| C−H | 1.09 | C−F | 1.33 | C−C | 1.53 | N−O | 1.47 |
| N−H | 1.01 | N−F | 1.41 | N−C | 1.45 | Al−O | 1.72 |
| O−H | 0.95 | O−F | 1.41 | O−C | 1.42 | Si−O | 1.61 |
| Al−H | 1.55 | Al−F | 1.67 | S−C | 1.81 | S−S | 2.04 |
| Si−H | 1.45 | Si−F | 1.56 | Si−C | 1.86 | | |
| S−H | 1.34 | S−F | 1.65 | | | | |

and the signed error by QM/MM calculations using untuned H and F atoms as link atoms.

*4.B.1. Balanced and Unbalanced Charge Schemes.* The balanced methods (balanced SEE, balanced RC, Amber-1, balanced RC2, balanced RC3, Amber-2, balanced RCD, balanced Shift) give much smaller errors in proton affinities than the unbalanced ones (SEE, Z1, Z2, Z3, RC). Both the H link atom and the F link atom schemes have the same trends. The mean unsigned errors (MUEs) given by all the balanced methods are 4−7 kcal/mol, while the MUEs given by all the unbalanced methods are 15−24 kcal/mol. This is because in the unbalanced methods, a net partial change is created near the QM region and the interactions between the QM and MM regions become unphysical.[4,33,43] The CPS method, in which any polarization of the QM region by the MM region is excluded, does not change the total charge of the QM/MM entire system and gives a smaller MUE than the unbalanced methods. Therefore, the conservation of the total charge of the QM/MM entire system is one of the key factors for the calculations of proton affinities.

*4.B.2. Different Link Atoms and Charge Schemes.* The comparison of the results using the H link atom (Table 4) and the F link atom (Table 5) shows that the proton affinities


(a)

(b)

***Figure 4.*** (a) The original entire system and (b) the entire system model (ESM) of the molecule **w**.

***Table 4.*** Full-QM Proton Affinities (PA, in kcal/mol), QM/MM Signed Errors (in kcal/mol), and Mean Unsigned Errors (MUE) (in kcal/mol) Averaged over 25 cases Using the H Link Atoms

| | | | | | balanced | balanced | balanced | | balanced | | balanced | balanced | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| case | bond | PA | site | CPS | SEE | RC | RC2 | Amber-1 | RC3 | Amber-2 | RCD | Shift | SEE | Z1 | Z2 | Z3 | RC |
| **a** | C–C | 362.3 | Q3 | 6.8 | 1.2 | 2.5 | 3.9 | 3.5 | 3.9 | 3.9 | 1.0 | 1.3 | –2.4 | 13.4 | 6.8 | 6.8 | –0.7 |
| **b** | C–C | 406.6 | Q3 | –0.6 | 3.9 | 5.1 | 6.3 | 5.5 | 6.9 | 7.3 | 3.9 | 4.2 | 4.8 | 13.1 | –3.6 | –40.7 | 5.8 |
| **c** | C–C | 402.6 | Q2 | 9.0 | 12.0 | 13.0 | 14.0 | 13.7 | 14.5 | 14.5 | 11.9 | 12.1 | 12.0 | 19.7 | –19.2 | 9.0 | 13.0 |
| **d** | N–C | 376.7 | Q4 | 3.4 | 6.6 | 4.8 | 2.9 | 3.8 | 1.7 | 1.2 | 6.6 | 6.2 | –14.7 | –9.7 | 7.2 | –15.1 | –14.1 |
| **e** | O–C | 398.6 | Q4 | 4.0 | 5.4 | 4.1 | 2.9 | 3.4 | 2.0 | 1.4 | 5.3 | 5.0 | –7.8 | –5.7 | 35.5 | –26.7 | –7.5 |
| **f** | S–C | 394.7 | Q4 | –1.1 | 2.2 | 0.6 | –1.2 | –0.5 | –2.6 | –3.2 | 2.3 | 1.9 | –6.8 | –15.0 | 2.0 | –16.4 | –7.6 |
| **g** | C–N | 355.2 | Q3 | 13.0 | 17.0 | 12.5 | 7.4 | 11.3 | 2.5 | –1.4 | 17.5 | 16.2 | 39.9 | –26.9 | 47.0 | –3.6 | 32.9 |
| **h** | O–N | 400.0 | Q4 | 3.0 | 13.8 | 9.6 | 5.4 | 5.2 | 3.8 | 3.8 | 13.5 | 12.4 | 6.8 | –19.1 | –23.0 | 3.0 | 3.4 |
| **i** | C–O | 394.9 | Q3 | 11.9 | 12.1 | 11.3 | 10.3 | 10.3 | 9.7 | 9.3 | 12.4 | 12.1 | 38.8 | 5.9 | 40.2 | –5.9 | 34.4 |
| **j** | C–O | 401.0 | Q3 | 5.5 | 13.9 | 11.1 | 7.7 | 7.7 | 6.1 | 5.1 | 14.4 | 13.5 | 37.5 | –6.8 | 12.0 | –30.5 | 31.3 |
| **k** | C–O | 366.7 | Q3 | 5.8 | 10.8 | 7.5 | 4.2 | 4.2 | 2.5 | 1.6 | 10.8 | 10.0 | 27.0 | –11.8 | 9.2 | –35.4 | 21.3 |
| **l** | C–O | 398.3 | Q4 | 7.1 | 7.6 | 7.1 | 6.5 | 6.5 | 6.2 | 5.9 | 7.7 | 7.6 | 28.1 | 3.5 | 30.6 | –7.8 | 25.1 |
| **m** | Al–O | 340.7 | Q2 | 5.1 | 6.0 | 3.8 | 1.1 | 1.1 | –0.2 | –0.3 | 6.5 | 5.7 | 33.6 | –10.8 | 19.9 | –18.2 | 28.5 |
| **n** | Al–O | 339.1 | Q2 | 6.8 | 6.0 | 3.8 | 1.0 | 1.0 | –0.4 | –0.3 | 6.6 | 5.7 | 33.6 | –11.1 | 26.1 | –42.0 | 28.4 |
| **o** | Al–O | 348.0 | Q2 | –2.2 | 5.5 | 3.7 | 1.5 | 1.5 | 0.4 | –0.2 | 5.9 | 5.1 | 35.1 | –8.0 | 35.1 | –61.5 | 29.9 |
| **p** | Si–O | 349.0 | Q2 | –4.9 | 4.0 | 0.7 | –2.9 | –2.9 | –4.5 | –6.3 | 4.3 | 2.8 | 27.2 | –15.9 | 27.9 | –35.5 | 20.1 |
| **q** | Si–O | 353.2 | Q4 | –4.0 | 0.7 | –0.4 | –1.8 | –1.8 | –2.5 | –3.8 | 1.0 | 0.7 | 18.3 | –11.0 | 22.8 | –22.2 | 15.6 |
| **r** | Si–O | 348.1 | Q2 | –4.1 | 4.0 | 0.6 | –3.2 | –3.2 | –5.0 | –6.8 | 4.4 | 2.8 | 23.9 | –17.3 | 30.7 | –60.6 | 17.4 |
| **s** | C–Si | 397.0 | Q3 | 8.6 | –2.9 | 2.3 | 7.2 | 6.2 | 7.2 | 7.2 | –3.0 | –1.3 | –20.4 | 33.4 | 8.6 | 8.6 | –12.6 |
| **t** | O–Si | 342.7 | Q3 | 5.8 | 3.9 | 4.9 | 5.8 | 6.0 | 5.8 | 5.8 | 4.0 | 4.3 | –8.1 | 13.2 | 5.8 | 5.8 | –5.7 |
| **u** | O–Si | 348.0 | Q3 | 0.5 | 1.7 | 6.2 | 9.8 | 10.8 | 11.7 | 11.7 | 2.4 | 3.5 | –9.1 | 38.7 | –59.5 | 0.5 | –3.1 |
| **v** | O–Si | 353.2 | Q5 | 1.6 | 2.2 | 4.6 | 6.6 | 7.9 | 8.2 | 10.4 | 2.4 | 2.9 | –8.1 | 29.2 | –16.8 | 24.5 | –4.7 |
| **w** | O–Si | 354.8 | Q5 | 0.0 | 2.7 | 5.3 | 7.5 | 9.0 | 9.3 | 14.2 | 3.0 | 3.4 | –5.0 | 32.5 | –13.0 | 38.6 | –1.6 |
| **x** | C–S | 395.8 | Q3 | 10.5 | 16.1 | 14.4 | 12.6 | 12.6 | 12.1 | 12.1 | 16.2 | 15.5 | 25.5 | 5.4 | –7.5 | 10.5 | 22.3 |
| **y** | S–S | 390.3 | Q4 | 2.8 | 6.1 | 5.1 | 4.0 | 4.0 | 3.7 | 3.7 | 6.1 | 5.8 | 5.3 | –2.8 | –14.8 | 2.8 | 4.4 |
| MUE | | | | 5.1 | 6.7 | 5.8 | 5.5 | 5.7 | 5.3 | 5.7 | 6.9 | 6.5 | 19.2 | 15.2 | 21.0 | 21.3 | 15.7 |

***Table 5.*** Full-QM Proton Affinities (PA, in kcal/mol), QM/MM Signed Errors (in kcal/mol), and Mean Unsigned Errors (MUE, in kcal/mol) Averaged over 25 cases Using the F Link Atoms

| | | | | | balanced | balanced | balanced | | balanced | | balanced | balanced | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| case | bond | PA | site | CPS | SEE | RC | RC2 | Amber-1 | RC3 | Amber-2 | RCD | Shift | SEE | Z1 | Z2 | Z3 | RC |
| **a** | C–C | 362.3 | Q3 | –2.6 | –7.8 | –7.0 | –5.6 | –6.0 | –5.6 | –5.6 | –8.4 | –8.1 | –11.1 | 3.8 | –2.6 | –2.6 | –10.1 |
| **b** | C–C | 406.6 | Q3 | –10.4 | –5.4 | –4.6 | –3.4 | –4.2 | –2.9 | –2.5 | –5.7 | –5.4 | –4.6 | 3.2 | –12.4 | –49.7 | –3.9 |
| **c** | C–C | 402.6 | Q2 | –8.5 | –5.3 | –4.8 | –3.7 | –4.0 | –3.3 | –3.3 | –5.8 | –5.6 | –5.3 | 1.8 | –36.3 | –8.5 | –4.8 |
| **d** | N–C | 376.7 | Q4 | –11.5 | –8.2 | –9.6 | –11.7 | –10.8 | –13.0 | –13.5 | –7.4 | –7.7 | –29.6 | –24.7 | –7.1 | –30.3 | –29.2 |
| **e** | O–C | 398.6 | Q4 | –10.4 | –9.7 | –10.7 | –12.3 | –11.7 | –13.3 | –14.0 | –9.2 | –9.4 | –23.7 | –21.5 | 22.6 | –42.3 | –23.5 |
| **f** | S–C | 394.7 | Q4 | –7.4 | –3.7 | –5.2 | –7.3 | –6.5 | –8.8 | –9.4 | –3.1 | –3.7 | –12.9 | –21.2 | –4.0 | –22.7 | –13.6 |
| **g** | C–N | 355.2 | Q3 | 4.0 | 5.4 | 3.4 | –1.5 | 2.3 | –6.2 | –9.9 | 8.4 | 7.1 | 26.8 | –34.7 | 37.4 | –12.5 | 23.6 |
| **h** | O–N | 400.0 | Q4 | –11.0 | 2.2 | –1.8 | –7.7 | –7.9 | –9.8 | –9.8 | 3.9 | 1.7 | –5.5 | –33.8 | –37.9 | –11.0 | –8.7 |
| **i** | C–O | 394.9 | Q3 | 3.1 | 3.3 | 2.9 | 1.9 | 1.9 | 1.3 | 1.0 | 3.8 | 3.6 | 27.6 | –2.3 | 30.5 | –14.3 | 25.1 |
| **j** | C–O | 401.0 | Q3 | –2.8 | 4.6 | 2.8 | –0.6 | –0.6 | –2.1 | –3.0 | 6.2 | 5.0 | 26.6 | –14.5 | 3.5 | –37.8 | 22.8 |
| **k** | C–O | 366.7 | Q3 | –0.8 | 3.5 | 1.4 | –2.3 | –2.3 | –4.2 | –5.1 | 5.2 | 4.0 | 19.2 | –18.4 | 2.7 | –41.9 | 15.7 |
| **l** | C–O | 398.3 | Q4 | 1.5 | 1.8 | 1.5 | 0.9 | 0.9 | 0.6 | 0.3 | 2.1 | 1.9 | 21.0 | –2.1 | 24.4 | –13.3 | 19.3 |
| **m** | Al–O | 340.7 | Q2 | 2.1 | 2.7 | 1.0 | –1.8 | –1.8 | –3.1 | –3.2 | 3.8 | 2.7 | 30.3 | –13.5 | 16.6 | –20.8 | 25.6 |
| **n** | Al–O | 339.1 | Q2 | 3.8 | 2.7 | 0.9 | –1.9 | –1.9 | –3.3 | –3.2 | 3.8 | 2.7 | 30.3 | –13.9 | 22.7 | –44.5 | 25.5 |
| **o** | Al–O | 348.0 | Q2 | –5.2 | 2.0 | 0.6 | –1.6 | –1.6 | –2.7 | –3.2 | 2.9 | 2.0 | 31.4 | –11.0 | 31.5 | –63.9 | 26.8 |
| **p** | Si–O | 349.0 | Q2 | –8.5 | –0.2 | –2.8 | –6.6 | –6.6 | –8.2 | –10.1 | 1.0 | –0.7 | 22.5 | –19.5 | 24.0 | –39.0 | 17.0 |
| **q** | Si–O | 353.2 | Q4 | –3.3 | 1.3 | 0.4 | –1.2 | –1.2 | –2.0 | –3.4 | 2.0 | 1.6 | 19.3 | –10.5 | 23.7 | –21.5 | 16.9 |
| **r** | Si–O | 348.1 | Q2 | –6.9 | 0.6 | –2.1 | –6.2 | –6.2 | –8.1 | –9.8 | 2.0 | 0.1 | 20.3 | –20.2 | 27.6 | –62.9 | 15.0 |
| **s** | C–Si | 397.0 | Q3 | –1.1 | –11.9 | –7.2 | –2.5 | –3.6 | –2.5 | –2.5 | –11.8 | –10.0 | –27.8 | 23.1 | –1.1 | –1.1 | –21.2 |
| **t** | O–Si | 342.7 | Q3 | –3.8 | –6.7 | –5.1 | –3.8 | –3.6 | –3.8 | –3.8 | –6.5 | –5.9 | –20.1 | 3.9 | –3.8 | –3.8 | –16.7 |
| **u** | O–Si | 348.0 | Q3 | –9.1 | –11.0 | –4.4 | 1.1 | 2.2 | 3.2 | 3.2 | –9.8 | –7.1 | –22.5 | 31.2 | –70.8 | –9.1 | –14.4 |
| **v** | O–Si | 353.2 | Q5 | –4.4 | –5.9 | –2.3 | 0.8 | 2.0 | 2.6 | 5.1 | –5.4 | –4.1 | –16.9 | 24.2 | –23.1 | 18.8 | –12.2 |
| **w** | O–Si | 354.8 | Q5 | –6.0 | –5.6 | –1.7 | 1.7 | 3.0 | 3.7 | 9.0 | –5.2 | –3.8 | –13.8 | 27.5 | –19.3 | 33.0 | –9.1 |
| **x** | C–S | 395.8 | Q3 | 1.4 | 6.8 | 5.0 | 3.3 | 3.3 | 2.8 | 2.8 | 6.8 | 5.9 | 16.3 | –3.7 | –16.1 | 1.4 | 12.8 |
| **y** | S–S | 390.3 | Q4 | –4.2 | 0.0 | –1.7 | –2.9 | –2.9 | –3.3 | –3.3 | –0.4 | –1.0 | –0.9 | –9.7 | –21.8 | –4.2 | –2.4 |
| MUE | | | | 5.4 | 4.7 | 3.6 | 3.8 | 4.0 | 4.8 | 5.6 | 5.2 | 4.4 | 19.4 | 15.8 | 20.9 | 24.4 | 16.6 |

are sensitive to the link atom. In most cases, the H link atom gives larger proton affinities than the full QM calculations, while the F link atom gives smaller proton affinities. As the link atom is directly connected to the QM region, it can greatly change the electronic structure of the QM region. The Q1−H and Q1−F bonds cannot adequately reproduce the properties of the original Q1−M1 bond. Electronegativity can be used as a qualitative criterion to decide which atom is better to be used as the link atom. For example, when the M1 atom is oxygen or nitrogen, the F link atom gives better results than the H link atom.

Moreover, compared with the charges on the other MM atoms, the charge on the link atom is closest to the QM region and greatly affects the electrostatic potential on the active site. König et al.[47] have compared different charge schemes to treat the boundary and found that all the balanced methods give, on average, comparable errors in proton affinities and deprotonation energies. Here, we also found

***Table 6.*** Parameters of Pseudopotentials for the Tuned F Link Atoms

| case | bond | CPS | balanced SEE | balanced RC | balanced RC2 | Amber-1 | balanced RC3 | Amber-2 | balanced RCD | balanced Shift |
|---|---|---|---|---|---|---|---|---|---|---|
| **a** | C–C | 0.80 | 0.95 | 0.95 | 0.85 | 0.85 | 0.85 | 0.85 | 1.00 | 1.00 |
| **b** | C–C | 0.65 | 0.65 | 0.65 | 0.60 | 0.60 | 0.55 | 0.55 | 0.70 | 0.65 |
| **c** | C–C | 0.65 | 0.65 | 0.65 | 0.65 | 0.65 | 0.65 | 0.65 | 0.70 | 0.70 |
| **d** | N–C | 1.40 | 1.20 | 1.25 | 1.40 | 1.35 | 1.40 | 1.40 | 1.15 | 1.20 |
| **e** | O–C | 1.35 | 1.35 | 1.45 | 1.50 | 1.50 | 1.55 | 1.55 | 1.35 | 1.35 |
| **f** | S–C | 1.05 | 0.75 | 0.85 | 1.05 | 1.00 | 1.05 | 1.10 | 0.70 | 0.75 |
| **g** | C–N | –0.20 | –0.25 | –0.15 | 0.05 | 0.00 | 0.10 | 0.15 | –0.35 | –0.35 |
| **h** | O–N | 1.00 | 0.35 | 0.60 | 0.85 | 0.85 | 0.95 | 0.95 | 0.30 | 0.45 |
| **i** | C–O | –0.40 | –0.30 | –0.30 | –0.25 | –0.25 | –0.25 | –0.25 | –0.30 | –0.30 |
| **j** | C–O | –0.25 | –0.50 | –0.45 | –0.30 | –0.30 | –0.25 | –0.25 | –0.55 | –0.45 |
| **k** | C–O | –0.25 | –0.50 | –0.40 | –0.25 | –0.25 | –0.20 | –0.20 | –0.55 | –0.45 |
| **l** | C–O | –0.45 | –0.30 | –0.30 | –0.25 | –0.25 | –0.25 | –0.25 | –0.35 | –0.30 |
| **m** | Al–O | –0.15 | –0.30 | –0.20 | –0.05 | –0.05 | –0.05 | –0.05 | –0.30 | –0.25 |
| **n** | Al–O | –0.10 | –0.40 | –0.20 | –0.10 | –0.10 | –0.10 | –0.10 | –0.25 | –0.15 |
| **o** | Al–O | –0.15 | –0.25 | –0.20 | –0.10 | –0.10 | –0.05 | –0.05 | –0.30 | –0.25 |
| **p** | Si–O | 0.00 | –0.25 | –0.15 | 0.05 | 0.05 | 0.10 | 0.10 | –0.30 | –0.25 |
| **q** | Si–O | –0.30 | –0.45 | –0.35 | –0.25 | –0.25 | –0.20 | –0.20 | –0.45 | –0.35 |
| **r** | Si–O | 0.00 | –0.25 | –0.15 | 0.05 | 0.05 | 0.05 | 0.10 | –0.30 | –0.15 |
| **s** | C–Si | 0.70 | 1.10 | 0.80 | 0.70 | 0.75 | 0.70 | 0.70 | 0.90 | 0.85 |
| **t** | O–Si | 0.60 | 0.80 | 0.60 | 0.60 | 0.60 | 0.60 | 0.60 | 0.70 | 0.60 |
| **u** | O–Si | 0.60 | 1.40 | 0.70 | 0.50 | 0.50 | 0.50 | 0.50 | 0.90 | 0.70 |
| **v** | O–Si | 0.65 | 1.40 | 0.80 | 0.65 | 0.65 | 0.60 | 0.55 | 1.00 | 0.80 |
| **w** | O–Si | 0.65 | 1.40 | 0.80 | 0.65 | 0.65 | 0.60 | 0.55 | 1.00 | 0.80 |
| **x** | C–S | 0.45 | 0.30 | 0.40 | 0.40 | 0.40 | 0.40 | 0.40 | 0.35 | 0.40 |
| **y** | S–S | 0.35 | 0.15 | 0.30 | 0.35 | 0.35 | 0.35 | 0.35 | 0.25 | 0.30 |

that in the proton affinity calculations, all balanced charge schemes give similar mean unsigned errors (MUEs) of 25 molecules in the test suite. As the link atom affects the charge distribution near the QM/MM boundary, it is possible that the errors brought by the different charge schemes and by the link atom are of the same order of magnitude. In the next section, we will first tune the link atom to make the total charge of the QM subsystem right and then compare different charge schemes.

The position of the active site for each molecule is also listed in the Tables 4 and 5. We found that even when the active site is far from the boundary, the error is still large if the boundary is not well treated. For example, in molecule **v**, the active site is at the Q5 atom but the errors range from −5.9 to 10.4 kcal/mol using different link atoms and different balanced charge schemes. Therefore, increasing the distance of the active site from the boundary cannot completely remove the error due to the link atom. When a polar bond is cut, a tuned link atom and balanced charge scheme should be used.

**4.C. Tuned F Link Atom.** We used the protocol presented in section 2.B to tune the F link atoms. The final parameters *C* of the pseudopotentials for the tuned F atoms are shown in Table 6. These parameters reflect the differences among various kinds of bonds. The same type of bond shows similar parameters even in different molecules. For example, in the balanced RC scheme, the parameter for the pseudopotential is 0.65−1.45 for a carbon boundary atom, −0.15−0.60 for a nitrogen boundary atom, −0.45 to ∼−0.15 for an oxygen boundary atom, 0.60−0.80 for a silicon boundary atom, and 0.30−0.40 for a sulfur boundary atom. These results are consistent with the electronegativities of the atoms. Also, we found that when the M1 atom is less electronegative than the Q1 atom, the pseudopotential for the tuned F atom is larger. For example, the pseudopotential needed for an O−N bond (0.60) is much larger than that needed for a C−N bond (−0.15).

The tuned F atom was used as the link atom to calculate the proton affinities, and the results are shown in Table 7. By examination of the mean unsigned error (MUE) of the 25 molecules, one finds that the tuned F link atom gives smaller errors than the H link atom (Table 5) or the F link atom (Table 6) in *all* the charge schemes. This indicates that the accurate treatment of the boundary is very important, especially after the total charge is conserved. The tuned F link atom makes the total charge of the QM region right, and it avoids the artifacts that can be introduced by the use of a link atom.

The balanced RC scheme gives the best results, and the MUE is only 1.3 kcal/mol. The good performance of the balanced RC scheme can be explained as follows. In order to correctly handle the electrostatic interactions between the QM region and the MM region in the QM/MM calculations, it is important to have an accurate charge distribution of the MM region. When we move the boundary charges to avoid overpolarization, the charge distribution of the MM region can be greatly changed if the redistributed charges are moved far from the boundary. In the balanced RC method, the redistributed charges are moved to the midpoints of the M1−M2 bonds, and compared to other charge schemes, they introduce smaller changes to the charge distributions in the MM regions. In the balanced RC2, RC3, and Amber-2 schemes, the redistributed charges are placed farther and farther from the boundary region and the MUE of the proton affinities increases. When we use balanced RC3, in which the charges are redistributed to M2 and M3 atoms, the error is approximately equal to that in the capped primary system (CPS). However, when the bond is nonpolar, such as a C−C bond, the redistributed charge is relatively small and different charge schemes give similar errors.

***Table 7.*** Full-QM Proton Affinities (PA, in kcal/mol), QM/MM Signed Errors (in kcal/mol), and Mean Unsigned Errors (MUE, in kcal/mol) Averaged over 25 cases Using the Tuned F Link Atoms

| case | bond | PA | CPS | balanced SEE | balanced RC | balanced RC2 | Amber-1 | balanced RC3 | Amber-2 | balanced RCD | balanced Shift |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **a** | C–C | 362.3 | 2.6 | –1.5 | –0.8 | 0.0 | –0.5 | 0.0 | 0.0 | –2.0 | –1.7 |
| **b** | C–C | 406.6 | –6.2 | –1.0 | –0.2 | 0.6 | –0.2 | 0.8 | 1.2 | –1.1 | –1.1 |
| **c** | C–C | 402.6 | –2.0 | 1.6 | 2.0 | 3.1 | 2.8 | 3.5 | 3.5 | 1.5 | 1.6 |
| **d** | N–C | 376.7 | 0.1 | 1.7 | 1.0 | –0.4 | 0.5 | –1.3 | –1.9 | 2.4 | 2.7 |
| **e** | O–C | 398.6 | –0.2 | 0.1 | –0.6 | –1.4 | –0.6 | –1.7 | –2.5 | 0.8 | 0.7 |
| **f** | S–C | 394.7 | –0.8 | 1.1 | 0.3 | –0.7 | –0.2 | –2.3 | –2.5 | 1.5 | 1.2 |
| **g** | C–N | 355.2 | 2.8 | 4.0 | 2.5 | –1.2 | 2.3 | –5.6 | –9.0 | 6.1 | 4.8 |
| **h** | O–N | 400.0 | –3.4 | 4.6 | 2.8 | –1.2 | –1.3 | –2.6 | –2.6 | 6.3 | 5.3 |
| **i** | C–O | 394.9 | 0.6 | 1.0 | 1.1 | 0.4 | 0.4 | –0.1 | –0.5 | 2.0 | 1.8 |
| **j** | C–O | 401.0 | –4.3 | 1.6 | 0.0 | –2.5 | –2.5 | –3.6 | –4.5 | 2.6 | 2.1 |
| **k** | C–O | 366.7 | –1.6 | 2.0 | 0.2 | –3.1 | –3.1 | –4.7 | –5.7 | 3.4 | 2.5 |
| **l** | C–O | 398.3 | –0.3 | 0.6 | 0.3 | –0.1 | –0.1 | –0.4 | –0.6 | 0.7 | 0.8 |
| **m** | Al–O | 340.7 | 1.7 | 1.8 | 0.3 | –2.0 | –2.0 | –3.3 | –3.3 | 2.8 | 1.8 |
| **n** | Al–O | 339.1 | 3.4 | 1.5 | 0.3 | –2.3 | –2.3 | –3.6 | –3.5 | 2.9 | 2.1 |
| **o** | Al–O | 348.0 | –5.6 | 1.2 | 0.0 | –1.9 | –1.9 | –2.8 | –3.4 | 1.9 | 1.2 |
| **p** | Si–O | 349.0 | –8.6 | –1.2 | –3.4 | –6.5 | –6.5 | –7.9 | –9.7 | –0.2 | –1.8 |
| **q** | Si–O | 353.2 | –3.2 | 1.6 | 0.5 | –1.1 | –1.1 | –1.9 | –3.3 | 2.1 | 1.7 |
| **r** | Si–O | 348.1 | –7.0 | –0.3 | –2.7 | –6.1 | –6.1 | –7.9 | –9.5 | 0.8 | –0.6 |
| **s** | C–Si | 397.0 | 3.5 | –5.0 | –2.2 | 2.1 | 1.3 | 2.1 | 2.1 | –6.5 | –4.8 |
| **t** | O–Si | 342.7 | –0.5 | –2.2 | –1.9 | –0.5 | –0.3 | –0.5 | –0.5 | –2.7 | –2.6 |
| **u** | O–Si | 348.0 | –5.8 | –2.7 | –0.6 | 3.8 | 4.8 | 5.9 | 5.9 | –5.1 | –3.4 |
| **v** | O–Si | 353.2 | –2.2 | –0.7 | 0.4 | 3.1 | 4.2 | 4.7 | 6.9 | –2.1 | –1.4 |
| **w** | O–Si | 354.8 | –3.8 | –0.4 | 1.0 | 3.9 | 5.2 | 5.7 | 10.9 | –1.8 | –1.1 |
| **x** | C–S | 395.8 | 4.3 | 8.8 | 7.7 | 5.8 | 5.8 | 5.3 | 5.3 | 9.2 | 8.6 |
| **y** | S–S | 390.3 | –1.9 | 1.0 | 0.4 | –0.6 | –0.6 | –1.0 | –1.0 | 1.3 | 1.0 |
|  | MUE |  | 3.1 | 2.0 | 1.3 | 2.2 | 2.3 | 3.2 | 4.0 | 2.8 | 2.3 |

In all three tables of mean unsigned errors (Tables 4, 5, and 7), the balanced RC scheme is better than the balanced RCD scheme, whereas previously[42] the RCD scheme performed slightly better. Since the testing is more thorough in the present paper, we now believe that the simpler RC scheme is to be preferred to the RCD one.

From the above results, we conclude that both a good charge scheme and an appropriate treatment of the link atom are needed to accurately treat the boundary. The balanced RC method with a tuned F atom gives the best results among all the methods.

However, there are still some problems. For example, we found that the error for the C−S bond is still quite large and there is no obvious reason for this error.

**4.D. Tuning the F Link Atom Based on Other Methods.** Having established that we can obtain better results with a tuned fluorine atom, we should note that in future work one could also consider other ways to do the tuning. For example, the tuned parameter could be tuned to make the proton affinity or a particular reaction energy come out right at representative geometries of the reactant or product. There would be three advantages of this approach: (i) the proton affinity, unlike the partial charges, is a physical observable that can be calculated straightforwardly; (ii) the proton affinity depends on both the initial and final states, so one does not have to make a decision whether to tune the pseudopotential in the reactant state or the product state; (iii) there is additional flexibility in that one could tune the calculated proton affinity or reaction energy either (a) to a calculation in which a portion of the MM system (e.g., the M1, M2, and M3 atoms or the functional groups that contain them) is treated quantum mechanically but at the same level as is to be used for the QM portion of the QM/MM calculations or (b) to experiments or high-level calculations.

## 5. Concluding Remarks

In QM/MM studies it is often inevitable to cut covalent bonds between the QM and MM parts. The present study addresses QM/MM boundary treatments in a systematic manner. All the ingredients considered are well-known in the literature from our own work and that of others (relevant references are cited above), in particular, the use of link atoms, the balancing and redistribution of charges close to the QM/MM boundary, and the tuning of the properties of the link atoms by suitable calibration (e.g., via a pseudopotential). In the present article, we present a new method of tuning and we combine it with the other ideas just mentioned in a new way to yield a method called the balanced redistributed charge (BRC) scheme with tuned fluorine link atom. If an acronym is needed to save space, this could be labeled the TBRC (tuned and balanced redistributed charge) method.

We apply the new method to calculate the proton affinities of 25 diverse compounds (at fixed geometries), and we compare its performance to that of other boundary charge and link atom schemes for treating the QM/MM boundary with regard to their ability to reproduce the quantum mechanical reference values. The balanced redistributed charge scheme with tuned fluorine link atoms outperforms the other treatments for the chosen validation suite of proton affinities, as shown in Tables 4, 5, and 7. Some of the methods listed in Table 4 are typical methods currently in use. For example, as mentioned in section 2.A, the Z1 scheme is the default in CHARMM,[47] the Z2 scheme is the default in both the *Gaussian 03*[48] and *Gaussian 09*[49] packages, and the Amber-2 method is the default in AMBER, version 10.[46]

One important finding is that the errors are generally larger for treatments without charge balancing, whereas the choice of the actual charge redistribution scheme is less crucial, but not insignificant. From Tables 4 and 5, we see that the mean unsigned error (MUE) in the unbalanced schemes ranges from 15 to 24 kcal/mol, much worse than even the CPS method, which has no MM region (just a capped quantum subsystem) and which gives an MUE of 5 kcal/mol. The importance of balancing, which was previously emphasized by others,[33,43−45] has a dramatic effect, reducing the MUE to 4−7 kcal/mol.

A key physical element of the new scheme is that we tune the link atom to try to make the charge on the primary subsystem (the atoms of the capped primary system but excluding the cap) be the same as it would be in a quantum mechanical calculation on the entire system. In the past there has been much more emphasis on redistributing the MM charge near the boundary than on the charge on the primary subsystem. If, however, the charge on the uncapped portion of the capped quantum mechanical system is inaccurate, then no treatment of the boundary can restore the correct physics. This motivation for tuning is substantiated by the finding that the use of a tuned fluorine link atom further reduces the mean unsigned error in all eight balanced charge schemes, leading to mean unsigned errors of 1−4 kcal/mol. In six of these eight schemes, the MUE is smaller than that for a CPS with a tuned F atom. All these eight schemes, give smaller MUEs than CPS capped with a hydrogen atom. The MUE of 1.3 kcal/mol for the combination of the balanced RC scheme for boundary charges and the tuned fluorine atom is particularly encouraging in light of the difficulty of the test set. In fact, the average proton affinity in all 25 cases of the test set is 373 kcal/mol, and the range of quantum mechanical reference values is 67.5 kcal/mol (from 339.1 to 406.6 kcal/mol). The mean unsigned error of 1.3 kcal/mol for the recommended new methods is only 1.9% of the range.

We conclude that two elements are very important in the energy calculations: balancing the charges of the MM region and tuning the link atom. A general rule (defined for all single bonds in the whole periodic table) is provided to tune the link atom when different types of single bonds are cut at the boundary. We can calculate accurate proton affinities after correctly handling the QM/MM boundary.

In the present work, motivated by interest in a variety of catalytic and redox systems, we intentionally chose a very difficult property, proton affinities, we considered QM/MM boundaries both close and far from the site of protonation, and we created a test set that is much more diverse than previous test sets used for QM/MM methods, in particular in that it includes boundaries that cut very polar bonds, including some in which neither atom is a carbon. Some of the methods found to be inadequate for this demanding test set will perform better for properties that are less sensitive than proton affinities to electrostatic potentials in the quantum subsystem or for cases where only a carbon−carbon bond is broken. However, simply enlarging the quantum system enough to move the QM/MM boundary far from the site of reaction, when affordable, is not sufficient to guarantee good accuracy. Furthermore, in simulations of complex systems, it is often desirable to use a method that has been validated for even the most challenging problems.

In future work, we will examine the problem of geometry optimization using the new QM/MM scheme. In the present article, the valence and van der Waals terms that involve interactions between the QM subsystem and MM subsystem cancel out, but we will need a protocol for defining them when we begin to optimize geometries or consider dynamics.

**Acknowledgment.** We thank Masahiro Higashi, Hai Lin, Manjeera Mantina, and Jingjing Zheng for helpful discussions. This work was supported in part by the National Science Foundation under Grant No. CHE07-04974 and by the Office of Naval Research under Award Number N00014-05-1-0538.

## References

(1) Truhlar, D. G. *J. Am. Chem. Soc.* **2008**, *130*, 16824.

(2) Gao, J. *Rev. Comp. Chem.* **1996**, *7*, 119.

(3) Reuter, N.; Dejaegere, A.; Maigret, B.; Karplus, M. *J. Phys. Chem. A* **2000**, *104*, 1720.

(4) Sherwood, P. In *Modern Methods and Algorithms of Quantum Chemistry*; Grotendorst, J., Ed.; Neuman Institute for Computing: Jülich, Germany, 2000; Vol. 3, p 285.

(5) Nicoll, R. M.; Hindle, S. A.; MacKenzie, G.; Hillier, I. H.; Burton, N. A. *Theor. Chem. Acc.* **2001**, *106*, 105.

(6) Colombo, M. C.; Guidoni, L.; Laio, A.; Magistrato, A.; Maurer, P.; Piana, S.; Röhrig, U.; Spiegel, K.; Sulpizi, M.; VandeVondele, J.; et al. *Chimia* **2002**, *56*, 13.

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

(8) Shurki, A.; Warshel, A. In *Protein Simulations*; Vol. 66; Academic Press Inc.: San Diego, CA, 2003; p 249.

(9) Sherwood, P.; de Vries, A. H.; Guest, M. F.; Schreckenbach, G.; Catlow, C. R. A.; French, S. A.; Sokol, A. A.; Bromley, S. T.; Thiel, W.; Turner, A. J.; Billeter, S.; Terstegen, F.; Thiel, S.; Kendrick, J.; Rogers, S. C.; Casci, J.; Watson, M.; King, F.; Karlsen, E.; Sjovoll, M.; Fahmi, A.; Schafer, A.; Lennartz, C. *J. Mol. Struct.: THEOCHEM* **2003**, *632*, 1.

(10) Amara, P.; Field, M. J. *Theor. Chem. Acc.* **2003**, *109*, 43.

(11) Hammes-Schiffer, S. *Curr. Opin. Struct. Biol.* **2004**, *14*, 192.

(12) Friesner, R. A.; Guallar, V. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389.

(13) Moret, M.-E.; Tapavicza, E.; Guidoni, L.; Röhrig, U.; Sulpizi, M.; Tavernelli, I.; Rothlisberger, U. *Chimia* **2005**, *59*, 493.

(14) Mulholland, A. J. *Drug Discovery Today* **2005**, *10*, 1393.

(15) Jorgensen, W. L.; Tirado-Rives, J. *J. Comput. Chem.* **2005**, *26*, 1689.

(16) Spoel, D. V. D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701.

(17) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; König, P.; Li, G.; Xu, D.; Guo, H.; et al. *J. Phys. Chem. B* **2006**, *110*, 6458.

(18) Zhang, Y. *Theor. Chem. Acc.* **2006**, *116*, 43.

(19) Senn, H. M.; Thiel, W. *Curr. Opin. Chem. Biol.* **2007**, *11*, 182.

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

(21) Hu, H.; Yang, W. *Annu. Rev. Phys. Chem.* **2008**, *59*, 573.

(22) Sousa, S. F.; Ramos, M. J. In *Computational Proteomics 2008*; Ramos, M. J., Ed.; Transworld Research Network: Kerala, India, 2008; p 101.

(23) Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198.

(24) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986**, *7*, 718.

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

(26) Théry, V.; Rinaldi, D.; Rivail, J. L.; Maigret, B.; Ferenczy, G. G. *J. Comput. Chem.* **1994**, *15*, 269.

(27) Gao, J.; Amara, P.; Alhambra, C.; Field, M. J. *J. Phys. Chem. A* **1998**, *102*, 4714.

(28) Pu, J.; Gao, J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 632.

(29) Redondo, A.; Goddard, W. A.; Swarts, C. A.; McGill, T. C. *J. Vac. Sci. Technol.* **1981**, *19*, 498.

(30) Koga, N.; Morokuma, K. *Chem. Phys. Lett.* **1990**, *172*, 243.

(31) Nasluzov, V. A.; Ivanova, E. A.; Shor, A. M.; Vayssilov, G. N.; Birkenheuer, U.; Rösch, N. *J. Phys. Chem. B* **2003**, *107*, 2228.

(32) Antes, I.; Thiel, W. *J. Phys. Chem. A* **1999**, *103*, 9290.

(33) Zhang, Y.; Lee, T.-S.; Yang, W. *J. Chem. Phys.* **1999**, *110*, 46.

(34) Zhang, Y. *J. Chem. Phys.* **2005**, *122*, 024114.

(35) Parks, J. M.; Hu, H.; Cohen, A. J.; Yang, W. *J. Chem. Phys.* **2008**, *129*, 154106.

(36) Alary, F.; Poteau, R.; Heully, J. L.; Barthelat, J. C.; Daudey, J. P. *Theor. Chem. Acc.* **2000**, *104*, 174.

(37) DiLabio, G. A.; Hurley, M. M.; Christiansen, P. A. *J. Chem. Phys.* **2002**, *116*, 9578.

(38) Moon, S.; Christiansen, P. A.; DiLabio, G. A. *J. Chem. Phys.* **2004**, *120*, 9080.

(39) DiLabio, G. A.; Wolkow, R. A.; Johnson, E. R. *J. Chem. Phys.* **2005**, *122*, 044708.

(40) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *J. Chem. Phys.* **2005**, *122*, 014113.

(41) Pople, J. A. *Rev. Mod. Phys.* **1999**, *71*, 1267.

(42) Lin, H.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 3991.

(43) Walker, R. C.; Crowley, M. F.; Case, D. A. *J. Comput. Chem.* **2008**, *29*, 1019.

(44) Sherwood, P.; de Vries, A. H.; Collins, S. J.; Greatbanks, S. P.; Burton, N. A.; Vincent, M. A.; Hillier, I. H. *Faraday Discuss.* **1997**, *106*, 79.

(45) Das, D.; Eurenius, K. P.; Billings, E. M.; Sherwood, P.; Chatfield, D. C.; Hodošček, M.; Brooks, B. R. *J. Chem. Phys.* **2002**, *117*, 10534.

(46) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; and Kollman, P. A.; *AMBER 10*; University of California: San Francisco, CA, 2008.

(47) König, P. H.; Hoffmann, M.; Frauenheim, T.; Cui, Q. *J. Phys. Chem. B* **2005**, *109*, 9082.

(48) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A.; *Gaussian 03*, revision D. 01; Gaussian, Inc.: Wallingford, CT, 2004.

(49) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J.; *Gaussian 09*, revision A.1; Gaussian, Inc.: Wallingford, CT, 2009.

(50) Pacios, L. F.; Christiansen, P. A. *J. Chem. Phys.* **1985**, *82*, 2664.

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

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

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

(54) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157.

(55) Olson, R. M.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2046.

(56) Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170.

(57) Dapprich, S.; Komáromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *J. Mol. Struct.: THEOCHEM* **1999**, *461*, 1.

CT900366M

JCTC Journal of Chemical Theory and Computation

# Electronic Transition Energies: A Study of the Performance of a Large Range of Single Reference Density Functional and Wave Function Methods on Valence and Rydberg States Compared to Experiment

Marco Caricato,*[,†,‡] Gary W. Trucks,[‡] Michael J. Frisch,[‡] and Kenneth B. Wiberg[†]

*Department of Chemistry, Yale University, 225 Prospect St., New Haven, Connecticut 06511 and Gaussian, Inc., 340 Quinnipiac. St Bldg. 40, Wallingford, Connecticut 06492*

Received September 28, 2009

**Abstract:** This work reports a comparison among wave function and DFT single reference methods for vertical electronic transition energy calculations toward singlet states, valence and Rydberg in nature. A series of 11 small organic molecules are used as test cases, where accurate experimental data in gas phase are available. We compared CIS, RPA, CIS(D), EOM-CCSD, and 28 multipurpose density functionals of the type LSDA, GGA, M-GGA, H-GGA, HM-GGA and with separated short and long-range exchange. The list of functionals is obviously not complete, but it spans more than 20 years of DFT development and includes functionals which are commonly used in the computation of a variety of molecular properties. Large differences in the results were found between the various functionals. The aim of this work is therefore to shed some light on the performance of the plethora of functionals available and compare them with some traditional wave function based methods on a molecular property of large interest as the transition energy.

## 1. Introduction

Vertical electronic transition energy calculations are presently routine in physical and theoretical chemistry research because they are a powerful tool for interpreting and often assisting predictions of experimental spectra. Theoretical approaches to treat excited states can be divided into single and multireference methods. The latter ones provide a balanced description of both ground and excited states and they are necessary to describe for instance two-electron excitations or conical intersections. However, multireference methods are computationally expensive and they can be ambiguous to apply, thus requiring considerable expertise in order to obtain meaningful results.

Single reference methods are, on the other hand, straightforward and the quality of the results is easier to evaluate. These methods are also completely defined given the level of theory and basis set, thus they are accessible even to nonspecialists. However, they can provide an unbalanced description of the ground state compared to the excited states, and they can usually be employed only to describe one-electron transitions. Fortunately, many excitation processes of chemical interest belong to this category, so that single reference methods can be successfully used to obtain quantitative results.

The simplest excited state method is the configuration interaction singles (CIS),[1] also known as Tamm−Dancoff approximation (TDA), where the excited state is described as a linear combination of singly excited determinants from the reference Hartree−Fock (HF) determinant. This method does not include any electronic correlation, since single excitations do not mix with the HF reference. A proposed improvement over CIS is to consider perturbative double excitations corrections to the transition energies, that gives rise to the CIS(D) method.[2] A different approach is to include selected doubly excited determinants to CIS, obtaining the

* Corresponding author e-mail: marco@gaussian.com.

† Department of Chemistry, Yale University.

‡ Gaussian, Inc.

10.1021/ct9005129 © 2010 American Chemical Society
Published on Web 01/22/2010

random phase approximation (RPA) method, also known as time-dependent HF (TDHF).[3]

The extension of density functional theory (DFT) to electronic excitation through the time-dependent formalism (TDDFT), within the adiabatic approximation,[4−6] represented a fundamental step considering the huge success of DFT in describing ground state phenomena. The advantage of TDDFT over CIS and RPA is the inclusion of correlation effects through the exchange-correlation potential for both the ground and the excited states, without adding significant computational effort. Since the exact density functional is not known, many approximated functionals have been proposed, and new ones continue to be developed. However, this point also represents the greatest weakness of DFT today, as it is not at all easy to decide which functional is better for a particular system in a particular process. And this is especially true for transition energy calculations.

Correlation can also be included in wave function methods, for instance through coupled cluster (CC) theory.[7] CC is exact when all of the possible excitations are included, but in practice the expansion is truncated at a given order. The most widely used truncation order implies singles and doubles excitations (CCSD). Because of the nonlinearity of the exponential wave operator, higher order excitations are included at this level of truncation, so that CCSD is a dramatic improvement over configuration interaction singles and doubles (CISD).[7] The success and limitation of CCSD is also related to its computational cost, as it scales iteratively as $O(N^6)$, where $N$ is the number of basis functions, so this method is only feasible for small and medium sized molecules. CC theory was extended to excited states calculation through the equation of motion (EOM) formalism,[8,9] or alternatively through the linear response (LR) formalism,[10,11] that are equivalent for transition energies. EOM-CC scales as the ground state at a certain level of truncation, thus the same systems which are accessible at CCSD level can be in general also treated at EOM-CCSD level for the excited states. Although the excitation expansion for the excited states is linear, contrary to the ground state, the inclusion of double excitations introduces enough flexibility in the wave function to allow for an accurate description of many one-electron transitions. The computational effort is much larger than in DFT, but since in CC theory the exact solution is known (it is obtained by including all the possible excitation operators) there is a systematic way to improve truncated CC wave functions, whereas this is not possible with the present formulation of DFT.

Recent studies[12−15] review and compare the performances of the plethora of DFT functionals on a variety of phenomena and molecular systems. This work fits into this group, presenting a comparison of computed and experimental transition energies for some small representative organic molecules. Only single reference methods are taken into account because their results are unambiguous and therefore they represent useful computational tools even for nontheoretically trained investigators. The range of functionals examined is wide, starting from local spin density approximation (LSDA), to include generalized gradient approximation (GGA), GGA with kinetic energy density or meta-GGA (M-GGA), hybrid GGA (H-GGA), and hybrid-meta-GGA (HM-GGA) as well as functionals that separate short and long-range exchange contribution (with and without the correct long-range limit). This allows one to draw some conclusions about the progress of DFT development, at least for this property. Even if the number we considered is large, this is nevertheless far from complete. In fact, it would be a nearly impossible task considering the number of existing funtionals and the rate of publication of new ones.[12] We point out that we choose *multipurpose* functionals, that is functionals which are not specifically designed for excited state calculations, but are generally used for the calculation of a variety of properties. This work, in connection with others, might thus help the interested investigator to choose a functional which behaves comparably well for ground and excited states and can be used, for instance, for the study of photochemical processes and reactions. Additionally, we also examined the most common single reference wave function methods, CIS, RPA, CIS(D), and EOM-CCSD, for which a hierarchy of accuracy can be more easily defined.

This work follows a previous one[16] where some of us studied the effect of the basis set on electronic transition energies, comparing only RPA, EOM-CCSD, and TDDFT with the B3P86 hybrid functional. The present study focuses on the comparison of different methods, using the 6-311(3+,3+)G** basis set, that was demonstrated[16] to be sufficiently accurate for this molecular property. The set of molecules has also been extended with respect to ref 16 to include trans-1,3-butadiene,[54] acetaldehyde,[55] and a series of azabenzenes.[56] The latter group of molecules includes excitation toward valence states ($n \rightarrow \pi^*$ or $\pi \rightarrow \pi^*$), whereas the rest mainly involves Rydberg states. In total, we considered 69 states. Among these, 30 are valence in nature and 39 are Rydberg.

We emphasize that this work aims to give a qualitative picture of the behavior of the examined methods. This is because the number and nature of the excitations considered is not enough for a statistically meaningful sampling. Alternatively, the rather accurate experimental data and the small number of systems allow a detailed analysis of the results for each molecule. Therefore, the format for the presentation of the results was chosen accordingly.

This work is organized as follows. Section 2 contains a description of the computational methods we considered and the details of the calculations performed. Section 3 collects the results of our calculations and presents a discussion on the performance of the various methods. Section 4 contains a summary of the results and concluding remarks.

## 2. Computational Details

The geometries of the systems under investigation were optimized for the ground states at the MP2/6-311+G** level of theory. These geometries were used for all the methods in the electronic transition calculations. They are reported in Tables 1−11 in the Supporting Information. This material is available free of charge via the Internet at pubs.acs.org. The basis set used for the latter calculations was 6-311(3+,3+)G**, previously demonstrated to be adequate for this kind of property.[16] We performed vertical excitation calculations

***Table 1.*** List of Functionals Used in This Work

| | year | type | % HF | | year | type | % HF |
|---|---|---|---|---|---|---|---|
| LSDA[17,21] | 1951 | LSDA | | B3VP86[18,20,21,23] | 1993 | H-GGA | 20 |
| BLYP[18,19] | 1988 | GGA | | PBE1PBE[27,28,43,44] | 1997 | H-GGA | 25 |
| OLYP[22,19] | 2001 | GGA | | B1B95[25] | 1996 | HM-GGA | 25 |
| BP86[18,23] | 1988 | GGA | | THCTHHYB[31] | 2002 | HM-GGA | 15 |
| BVP86[18,21,23] | 1988 | GGA | | TPSSh[33,46] | 2003 | HM-GGA | 10 |
| PBEPBE[27,28] | 1997 | GGA | | M05[34] | 2005 | HM-GGA | 28 |
| HCTH[24,29,30] | 2001 | GGA | | BH&H,[17,21,19a] | 1993 | H-GGA | 50 |
| THCTH[31] | 2002 | M-GGA | | BH&HLYP,[17,21,18,19a] | 1993 | H-GGA | 50 |
| BB95[18,25] | 1996 | M-GGA | | BMK[36] | 2004 | HM-GGA | 42 |
| VSXC[26] | 1998 | M-GGA | | M05−2X[35] | 2006 | HM-GGA | 56 |
| TPSSTPSS[33] | 2003 | M-GGA | | HSE1PBE[37] | 2003 | H-GGA | 25 − 0[a] |
| O3LYP[22,19,32] | 2001 | H-GGA | 11.61 | CAM-B3LYP[38] | 2004 | H-GGA | 19 − 65[b] |
| B3LYP[18−20,45] | 1994 | H-GGA | 20 | LC-BLYP[18,19,39,40] | 2001 | H-GGA | LC[c] |
| B3P86[18,20,23] | 1993 | H-GGA | 20 | LC-$\omega$PBE[39−42] | 2006 | H-GGA | LC[c] |

[a] Note that these are not the same as the half-and-half functionals proposed by Becke[47] [b] Short-range−long-range. [c] The percentage of HF exchange increases as described in refs 39−42.

from the ground state geometry toward singlet excited states. We limited the analysis to singlet states because the experimental data are more reliable.

As mentioned in the introduction, we examined four well-established wave function based methods, CIS, RPA, CIS(D), and EOM-CCSD. These methods are referred in the following as ab initio. One could argue that also pure functionals are ab initio, as they do not contain empirical parameters, but we prefer to put those functionals in the general DFT category, for the sake of simplicity.

The functionals[17−46] we considered are listed in Table 1. The table indicates the type of functional: LSDA, GGA, M-GGA, H-GGA, and HM-GGA. The year of publication is also reported and, for the hybrid functionals, the percentage of HF exchange. We note that the order of the functionals in Table 1 follows the one of the figures in the results section. We collected the pure functional together, in the order LSDA, GGA, and M-GGA. The hybrid functionals were also collected together, and they were divided into three subgroups: (i) the ones with small HF exchange amount (from O3LYP, 11.61%, to M05, 28%), (ii) the ones with large HF exchange amount (from BMK, 42%, to M05−2X, 56%), (iii) the ones with separated short and long-range exchange (from HSE1PBE to LC-$\omega$PBE). Groups (i) and (ii) were also ordered as H-GGA, HM-GGA. This choice is useful to rationalize the results for the various functionals. The functionals in group (iii) separate close and long-range exchange interaction in a very different way. CAM-B3LYP is hybrid in both short and long-range and it was designed to improve the B3LYP description for long-range phenomena, like polarizability of long chains or charge transfer excitations.[48] HSE1PBE is hybrid in the short-range and pure in the long-range, and it was designed to study solids, where the exchange interaction decays faster than that of HF. HSE1PBE is also supposed to behave like PBE1PBE at short-range, and thus also for the small molecules we considered. Both CAM-B3LYP and HSE1PBE maintain the same ratio of HF versus DFT exchange at all distances within the long and short ranges, respectively. However, for LC-BLYP and LC-$\omega$PBE this ratio varies so that at long distance they present 100% (exact) HF exchange.[39−42] The latter two functionals are thus effectively hybrid. We also note that PBE1PBE is often referred in literature as PBE0.

Most of the functionals we tested do not exhibit the correct asymptotic behavior and this may degrade the quality of the description for diffuse states, like Rydberg. The failure occurs due to the poor description of the virtual space. In fact, there is a significant deterioration above the ionization threshold, $-\varepsilon$(HOMO). Although methods exist to add in a correction,[49,50] it is observed that a sufficiently diffused basis set and the addition of exact exchange help to alleviate this problem.[16] Moreover, these long-range corrections introduce new issues.[50,51] However, we include two functionals with the correct asymptotic behavior, LC-BLYP and LC-$\omega$PBE, in order to test the effectiveness of such corrections.

All of the calculations were performed with a development version of the Gaussian suite of programs.[52]

## 3. Results

The test molecules are as follows: ethylene ($D_{2h}$), isobutene ($C_{2v}$), trans-1,3-butadiene ($C_{2h}$), formaldehyde ($C_{2v}$), acetaldehyde ($C_s$), acetone ($C_{2v}$), pyridine ($C_{2v}$), pyrazine ($D_{2h}$), pyrimidine ($C_{2v}$), pyridazine ($C_{2v}$), and symmetric tetrazine ($D_{2h}$). The first three molecules are alkenes, the second group of three molecules are carbonyl compounds and the rest are azabenzenes with a different number of nitrogen atoms.

These molecules were chosen because they have been intensively studied experimentally in gas phase. They also have molecular symmetry, which aids in matching the calculated and measured data. Only well established experimental data were taken into account, in order to ensure a correct matching with the calculated quantities. Experimental vertical transition energies were almost always compared with the calculated data, but in some cases only the adiabatic ones were measured and those are reported in the tables in the Supporting Information. For alkenes and carbonyls, only one or two states for each molecule are valence states ($n \rightarrow \pi^*$ or $\pi \rightarrow \pi^*$), whereas the others are Rydberg states. For the azabenzenes, all states are valence in nature. The assignment of the valence states is rather unambiguous, as it was checked by looking at the orbitals mainly involved in the transition (for low lying states, only few orbitals are involved, thus it is easier to consider the orbitals rather than the electron density differences, since there are many states and many methods). The Rydberg states are more difficult

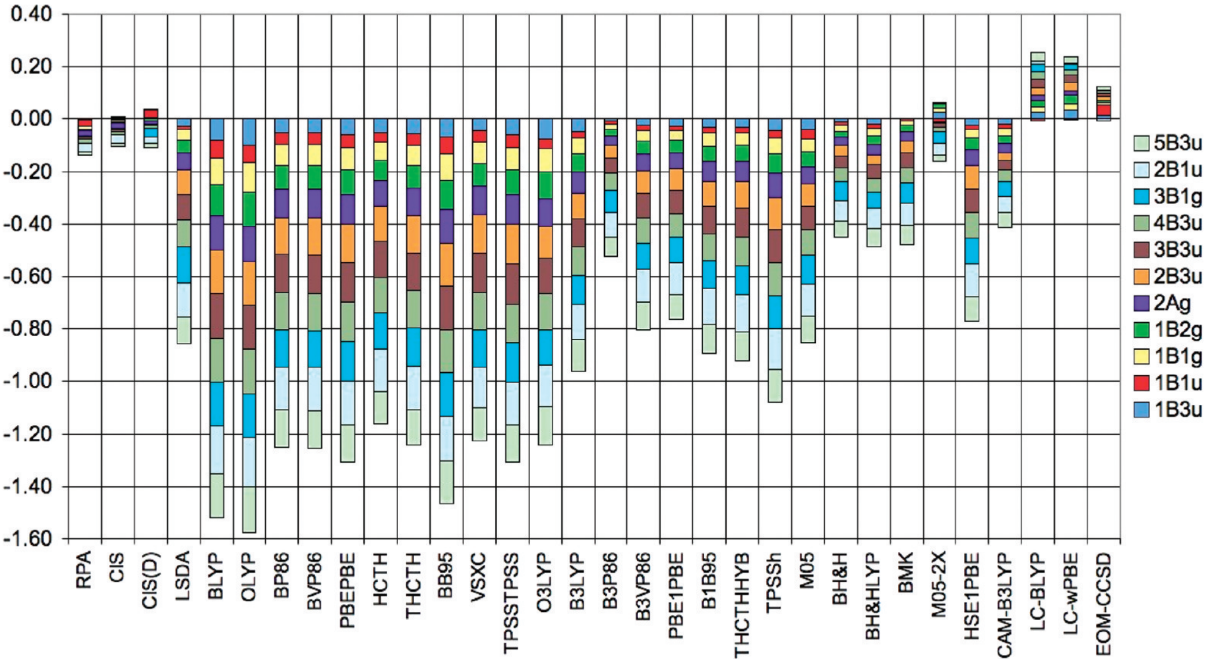


***Figure 1.*** Errors (eV) for ethylene. The $1B_{1u}$ transition is $\pi \rightarrow \pi^*$, the rest are Rydberg.

to assign. We put those states in energy order within each irreducible representation for each method and each molecule, and matched them against the experimental data, sorted in the same manner.[16,54,55] We made an attempt to use the oscillator strengths ($f$) to refine the matching, but the values of $f$ are very similar between different states for any particular method, thus the reordering of the states according to $f$ would be as arbitrary as sorting them in energy order.

A clarification about the azabenzenes set is also important. Vibronic effects might be not negligible for some transitions in this set, especially for the $n \rightarrow \pi^*$, which are in principle not allowed by symmetry. For these states, the experimental assignment of the transition energies might also have a larger uncertainty, as these bands have small intensities and may be broad. Our results do not include any vibrational correction, as this represents a large computational effort beyond the scope of this paper. The analysis of vibrational contributions for selected methods will be the subject of a following work.

The results are presented as error bars, where the errors for the various transitions stack one on top of the other in order to give an idea of the cumulative error for a certain method. The error for each transition is defined as follows:

$$\text{Error} = \frac{\Delta E^{\text{calc}} - \Delta E^{\text{exp}}}{n} \qquad (1)$$

where $n$ is the number of excitations considered. Thus, a positive error means that the experimental value is overestimated, and vice versa. Moreover by scaling the errors by the number of transitions all the molecules are more directly comparable. We chose this graphical representation of the errors because its visual impact makes the evaluation of the accuracy of a method straightforward. Nevertheless, the numerical values of all the experimental and calculated data and the statistical analysis can be found in Tables 12−25 in the Supporting Information. These tables also report the nature and the symmetry of the transitions. The performances of the methods are mainly discussed in terms of the cumulative errors, and not focusing on single transitions, because of the large amount of data we considered. Moreover, the performance of a method is judged in comparison to the other methods, in order to follow qualitative trends among the various test molecules. Therefore, a method can be considered to provide a good performance with respect to all the others, but still have a large absolute error for a particular transition.

The computational methods showed similar trends among the molecules within a particular set. In this respect, it is useful to group molecules with similar characteristics.

**3.1. Alkenes.** The ethylene molecule is one of the simplest systems to study, and many experimental and theoretical results are available, see ref 16 and references therein. We compared the measured and calculated data for eleven transitions. Among them, the $1B_{1u}$ transition is $\pi \rightarrow \pi^*$ in nature, the rest are Rydberg. The errors are reported in Figure 1. All of the ab initio methods, RPA, CIS, CIS(D), and EOM-CCSD, appear to perform quite accurately on this system. The total error is rather similar among those methods, but the first three often underestimate the experiments, whereas EOM-CCSD overestimates them. The same good performance is not shared by DFT functionals. The only ones that get close to the ab initio methods accuracy are M05−2X, LC-BLYP, and LC-$\omega$PBE. The pure functionals (GGA and M-GGA) perform poorly. Hybrid GGA functionals tend to improve the results, but they are still far from experiment. A good performance is also obtained with B3P86, BH&H, BH&HLYP, and CAM-B3LYP. Note that a previous report with PBE1PBE[53] suggested that this functional performed very well compared to experiments; although, the excited state symmetries were misassigned. HM-GGA functionals are divided. Most of them perform as poorly as H-GGA, from B1B95 to M05, whereas BMK performs as BH&H and

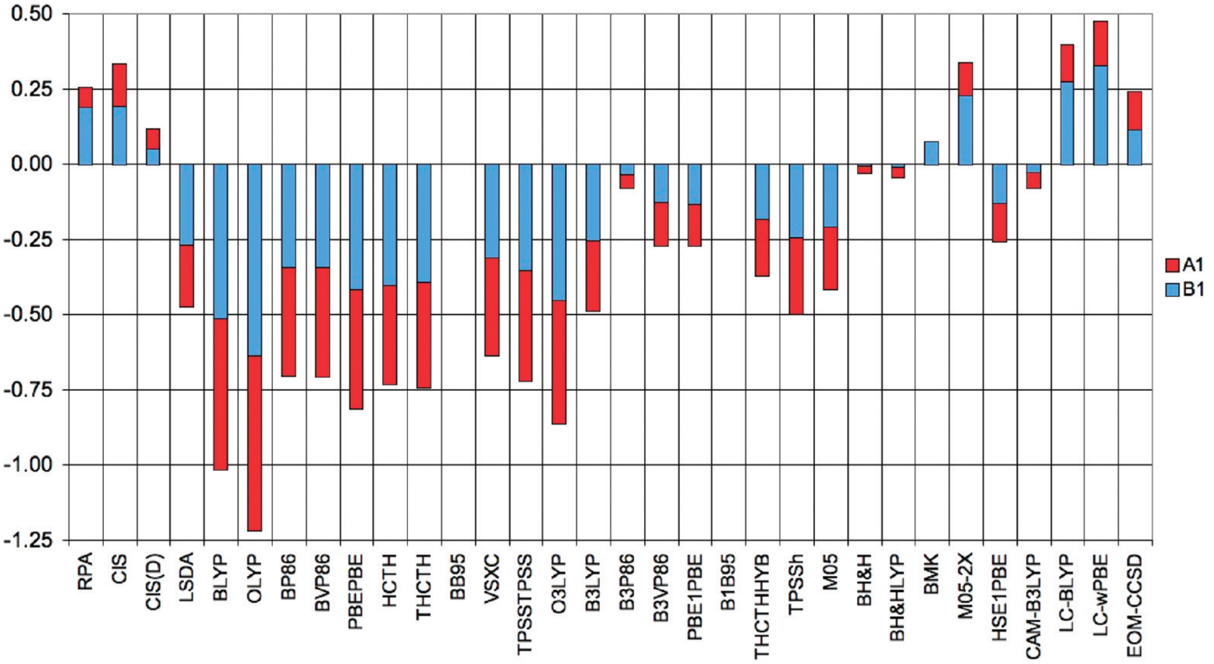


***Figure 2.*** Errors (eV) for isobutene. The SCF of BB95 and B1B95 did not converge. The transitions are Rydberg.

BH&HLYP. The overall performance of the DFT functionals is somehow disappointing for this molecule, considering its limited size and the availability of both experimental and high level computational data. With the only exception of B3P86, larger amount of HF exchange leads to better results.

Experimental data for the first transitions of $A_1$ and $B_1$ symmetry are available for isobutene. These states, both Rydberg, were compared with the computed data, see ref 16 and references therein. The error are plotted in Figure 2. For this molecule, we only have a limited set of experimental data, thus exhaustive conclusions cannot be drawn. However, the ab initio methods perform reasonably well with this molecule, with a small overestimation of the experimental results for both transitions. RPA and EOM-CCSD show a similar overall error but differently distributed among the two transitions. CIS performs slightly worse, whereas CIS(D) is even better than EOM-CCSD, probably due to a fortunate cancellation of errors. The functionals usually underestimate the measured transition energies, except for M05−2X, BMK, LC-BLYP, and LC-$\omega$PBE, that contain a large percentage of HF exchange and thus they tend to overestimate the transition energies as the HF methods do. However, some hybrid functionals perform quite well for this system. B3P86, BMK, BH&H, BH&HLYP, and CAM-B3LYP provide considerably smaller errors than the ab initio methods. Among them, B3P86 is the only one with a small contribution of HF exchange. BMK gives the exact excitation energy for the $A_1$ transition, which is most likely a fortunate combination of error cancellation. Remarkably, the two functionals with the correct long-range limit, LC-BLYP and LC-$\omega$PBE, show larger errors than most of the other hybrid functionals.

Seven experimental transition energies are used as reference to study the accuracy of computed data for the trans-1,3-butadiene.[54] We did not consider the $2^1A_g$ state as its assignment is controversial.[54] The $1B_u$ transition is $\pi \rightarrow \pi^*$ in nature, the rest are Rydberg. The errors are plotted in Figure 3. EOM-CCSD provides this time the best agreement with the experiments. Its largest error is on the first transition, of symmetry $B_u$, whereas all the others are very small. CIS, RPA, and CIS(D) perform considerably well, too. CIS and CIS(D) underestimate the experimental values except for the $1B_u$ transition. However, almost all DFT functionals show very large errors, generally underestimations of the experimental values. The only one that compares well with the ab initio methods is M05−2X. The general trend is that the quality of the results deteriorates for higher excited states. B3P86, BH&H, and BH&HLYP, together with the long-range corrected LC-BLYP and LC-$\omega$PBE are the H-GGA functionals that give a good performance. CAM-B3LYP significantly improves the B3LYP performance but its own total error is comparable to B3P86.

For this set, all of the ab initio methods perform well, with similar accuracy. Alternatively, pure functionals show large underestimation of the experimental results. LSDA is definitely better than GGA and M-GGA functionals. Hybrid functionals benefit from the HF good performance on this group of systems, so that larger percentage of HF exchange corresponds to smaller errors. M05−2X, the functional with the largest amount of HF exchange among the functionals we considered (56%), is the best DFT method for ethene and butadiene, whereas BH&H and BH&HLYP are the best ones for isobutene. LC-BLYP and LC-$\omega$PBE perform as well as M05−2X for ethene, as all the other functionals with large amount of HF exchange for butadiene and worse than most hybrid functionals for isobutene. We also note the good performance of B3P86, even with a modest 20% of HF exchange.

**3.2. Carbonyls.** Formaldehyde is another small molecule intensively studied experimentally and theoretically, see ref 16, and references therein. The comparison against the experiments is done for eleven electronic transitions

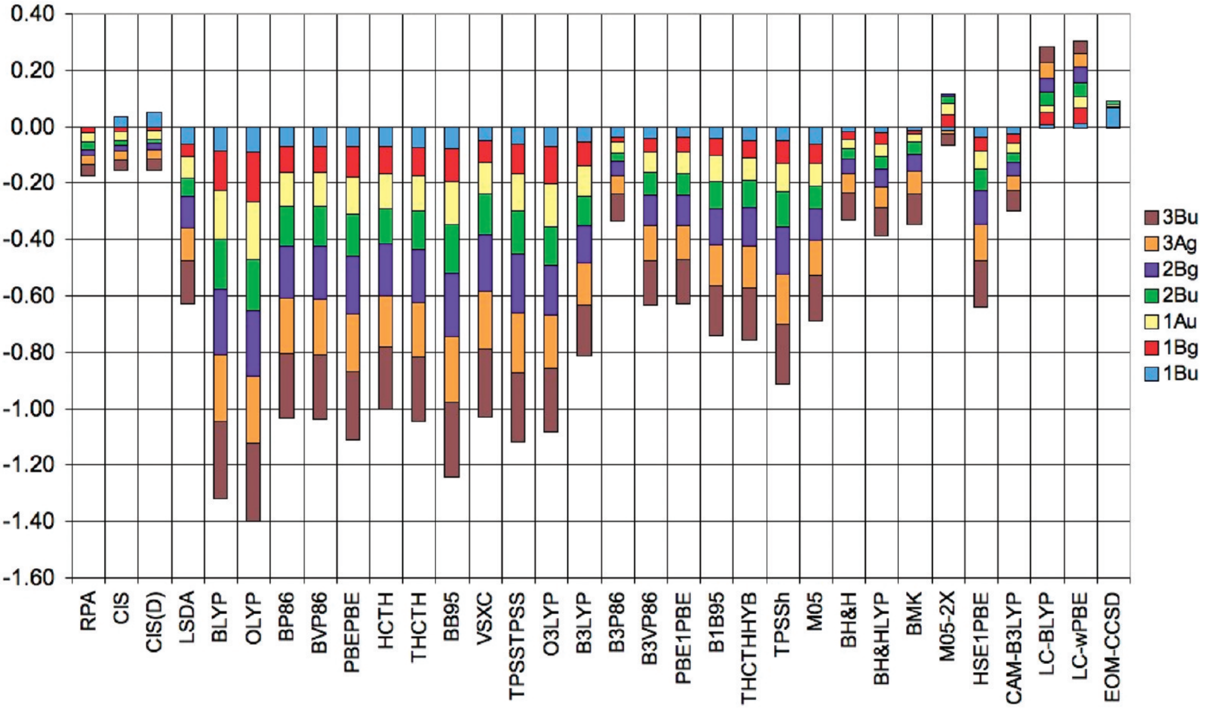


***Figure 3.*** Errors (eV) for trans-1,3-butadiene. The $1B_u$ transition is $\pi \rightarrow \pi^*$, the rest are Rydberg.

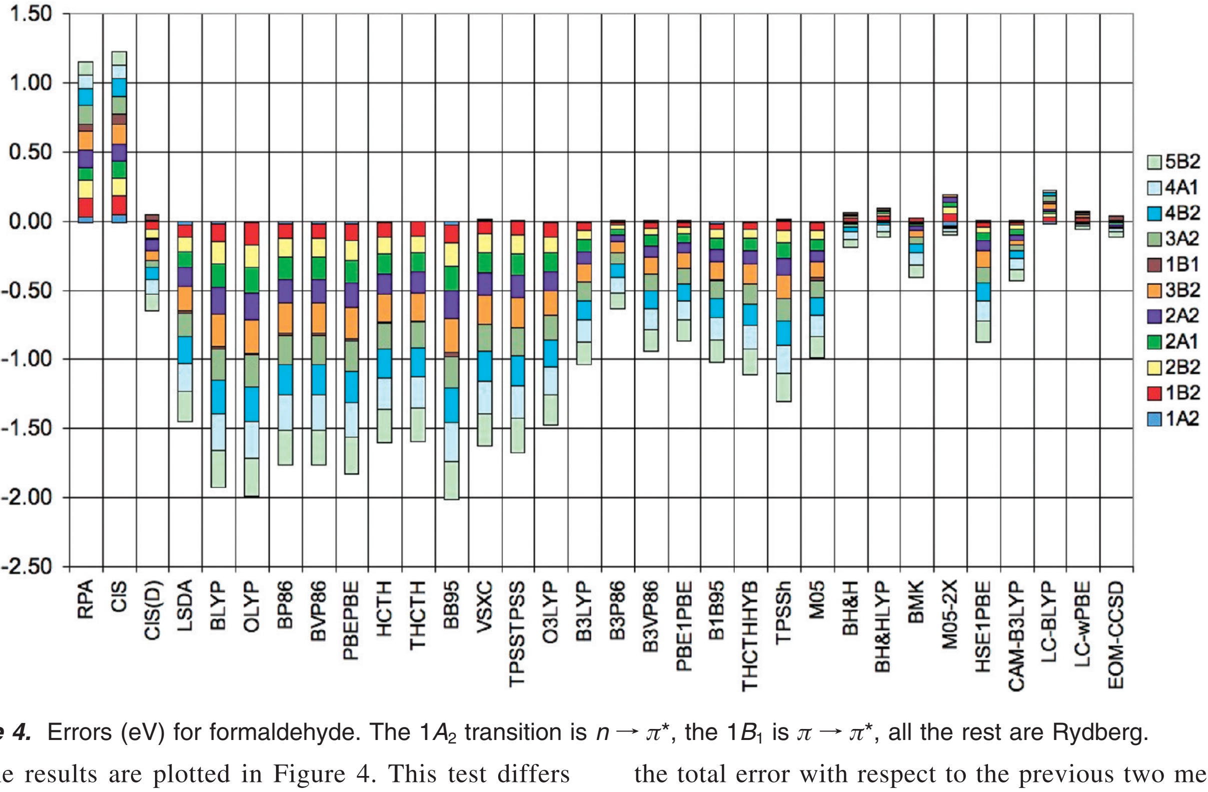


***Figure 4.*** Errors (eV) for formaldehyde. The $1A_2$ transition is $n \rightarrow \pi^*$, the $1B_1$ is $\pi \rightarrow \pi^*$, all the rest are Rydberg.

and the results are plotted in Figure 4. This test differs from the previous ones as this molecule has a carbonyl group, so that the experimental data set includes an $n \rightarrow \pi^*$ transition (corresponding to the lowest, $A_2$, excitation). Moreover, the double bond is more polarized than the alkene C=C bond. The $1B_1$ is also valence in nature ($\pi \rightarrow \pi^*$). Both of the valence excitations are well reproduced by most methods. Starting from the ab initio methods, Figure 4 shows very large overestimation of the measured data for RPA and CIS. CIS(D) on the other hand underestimates the experiment, at the same time halving the total error with respect to the previous two methods. However, only EOM-CCSD provides a very accurate description for all the transitions. As far as DFT is concerned, all of the GGA and M-GGA functionals largely underestimate the transition energies. These errors are even larger than the RPA and the CIS ones. The hybrid functionals perform better than the pure functionals, with the exception of O3LYP. LC-$\omega$PBE is even better than EOM-CCSD. BH&H and BH&HLYP are also quite close to EOM-CCSD. Also BMK, M05−2X, CAM-B3LYP and LC-BLYP perform well. The functionals with large HF

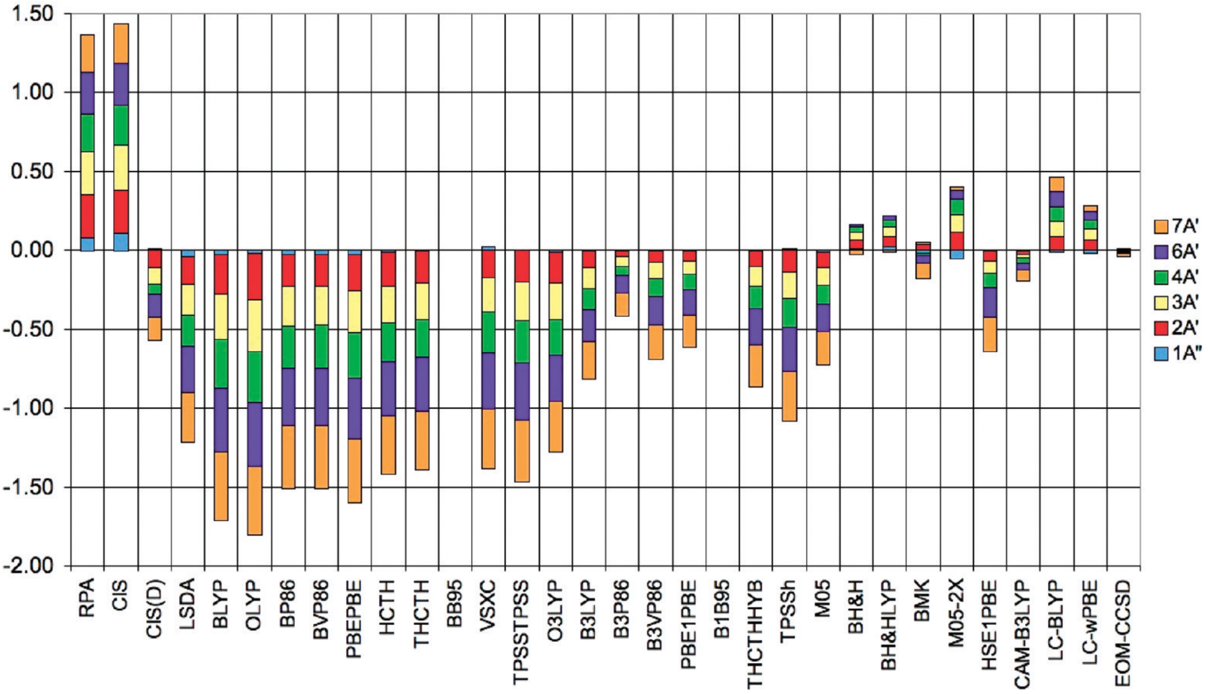


***Figure 5.*** Errors (eV) for acetaldehyde. The SCF of BB95 and B1B95 did not converge. The 1*A″* transition is $n \rightarrow \pi^*$, all the rest are Rydberg.

exchange contribution give a reasonable description of the transitions, close to the EOM-CCSD one.

For acetaldehyde six experimentally determined transition energies[55] are used to test the various computational methods. The errors are reported in Figure 5. The first transition, which is the only $n \rightarrow \pi^*$ (A″), is well reproduced by all the methods. The assignment of the nature of this transition was done by considering the differences of the excited and ground states CIS densities. CIS and RPA calculations give a poor description of the transitions. However, most of the GGA and M-GGA functionals show larger but opposite errors. CIS(D) improves the CIS performance, but the experiments are underestimated and the errors are still quite large. However, EOM-CCSD gives basically the correct description for all the transitions. Pure GGA and M-GGA classes yield poor results, often worse than RPA and CIS. All of the transition energies are underestimated. O3LYP is the only H-GGA functional with a total error comparable to the pure functionals. Alternatively, B3P86 performs again reasonably well, with total error smaller than CIS(D). M05−2X performance is better than the M05 one. However, M05−2X total error is larger than the H-GGA B3P86. BMK, BH&H, and BH&HLYP results are quite good, as well as the CAM-B3LYP ones. LC-BLYP and LC-$\omega$PBE also perform well, although with larger errors than the functionals with large amount of HF exchange, except M05−2X. Also, LC-BLYP performs worse than B3P86. Among the functionals the best performance is provided by BH&H.

Acetone is another widely studied small molecule, and we can compare the computational methods with experiments on eight electronic transitions, see ref 16 and references therein. The errors are reported in Figure 6. The first excited state is $n \rightarrow \pi^*$ ($A_2$) and it is well reproduced by all the methods. This case is quite similar to the previous one so it is not surprising that RPA and CIS largely overestimate all the transitions. CIS(D) does a better job reducing by more than half the total error, but most of the transitions energies are now underestimated. EOM-CCSD gives the best description of this system. The functionals performances are similar to the previous cases, too. Pure functionals errors are quite large, comparable to CIS and RPA. LSDA is the best functional among them. The excitation energies are underestimated. Hybrid functionals perform better, with the constant exception of O3LYP. Among the functionals with smaller percentage of HF exchange B3P86 (H-GGA) and B1B95 (HM-GGA) present the smallest total error, even better than the functionals with the correct asymptotic behavior. CAM-B3LYP provides the best performance among the functionals for this molecule, with BMK and B1B95 also very close.

For the carbonyl compounds, the first consideration is that all the methods, ab initio and DFT, are able to qualitatively reproduce the increase of the $n \rightarrow \pi^*$ transition energy with the number of methyl groups, that is observed experimentally. Approximate ab initio methods like CIS and RPA provide large errors in this case, overestimating the experimental data. CIS(D) improves the description, but there is huge shift with respect to CIS, so that most of the transitions are underestimated, and the total error is still large. However, EOM-CCSD shows very good agreement with experiments for all the transitions. Pure functionals errors are again very large, in many cases larger than the CIS and the RPA ones. Also for this set, LSDA is better than all the other pure functionals. Hybrid functionals are again in between CIS and the pure functionals. Large amount of HF exchange favors better error compensation, as the errors of CIS and pure functionals are comparable but opposite in sign. It is also interesting how functionals with large HF contribution tend

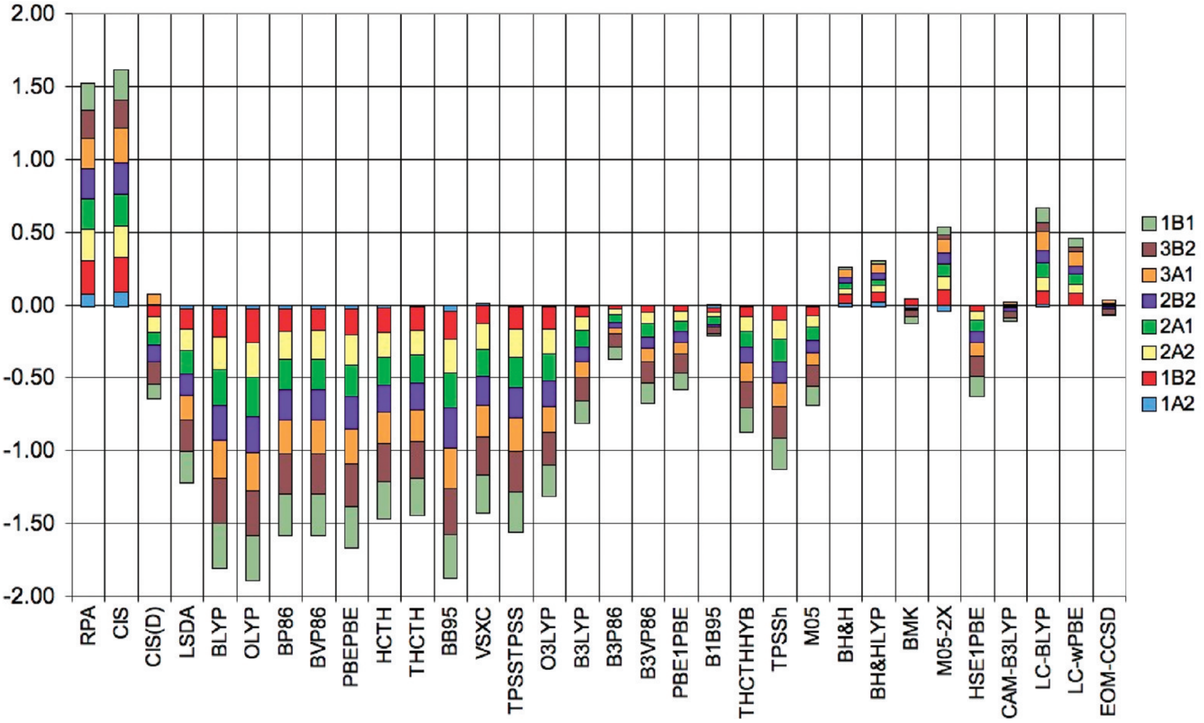


***Figure 6.*** Errors (eV) for acetone. The $1A_2$ transition is $n \rightarrow \pi^*$, all the rest are Rydberg.

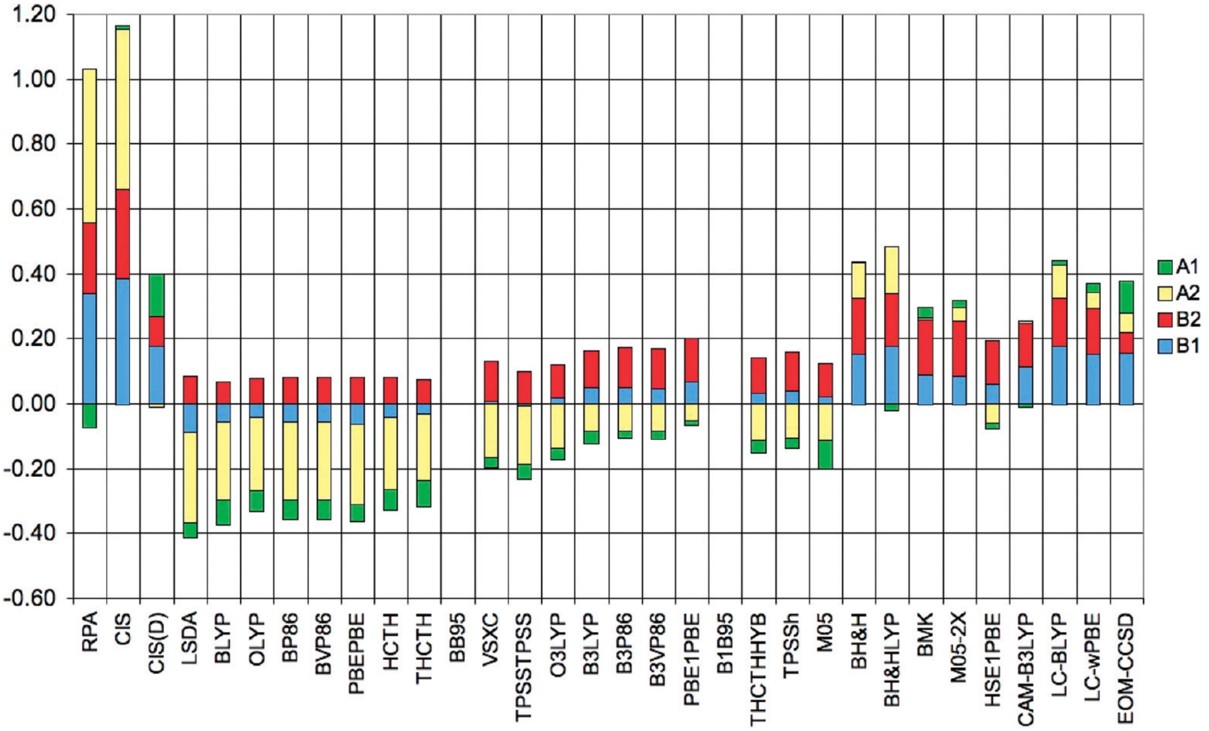


***Figure 7.*** Errors (eV) for pyridine. The SCF of BB95 and B1B95 did not converge.

to overestimate some of the transitions energies just as the ab initio methods do. We note again the good performance of B3P86 among the H-GGA functionals with small amount of HF exchange and of B1B95 for acetone. The functionals with the correct asymptotic behavior, LC-BLYP and LC-$\omega$PBE, do not seem to perform particularly better than noncorrected functionals with a large amount of HF exchange and, for acetaldehyde and acetone, the errors are comparable to or worse than B3P86.

**3.3. Azabenzenes.** Pyridine is the first azabenzene we considered. It has one nitrogen atom, and experimental data for four valence transitions are available, two of $n \rightarrow \pi^*$ type ($B_1$ and $A_2$) and two of $\pi \rightarrow \pi^*$ ($B_2$ and $A_1$).[56−59] The errors are reported in Figure 7. For this system, all of the ab initio methods present quite large errors. Almost all of the transitions are overestimated, except for the $A_1$ with RPA and $A_2$ for CIS(D). This molecule clearly shows how double excitation operators are not enough to obtain a good

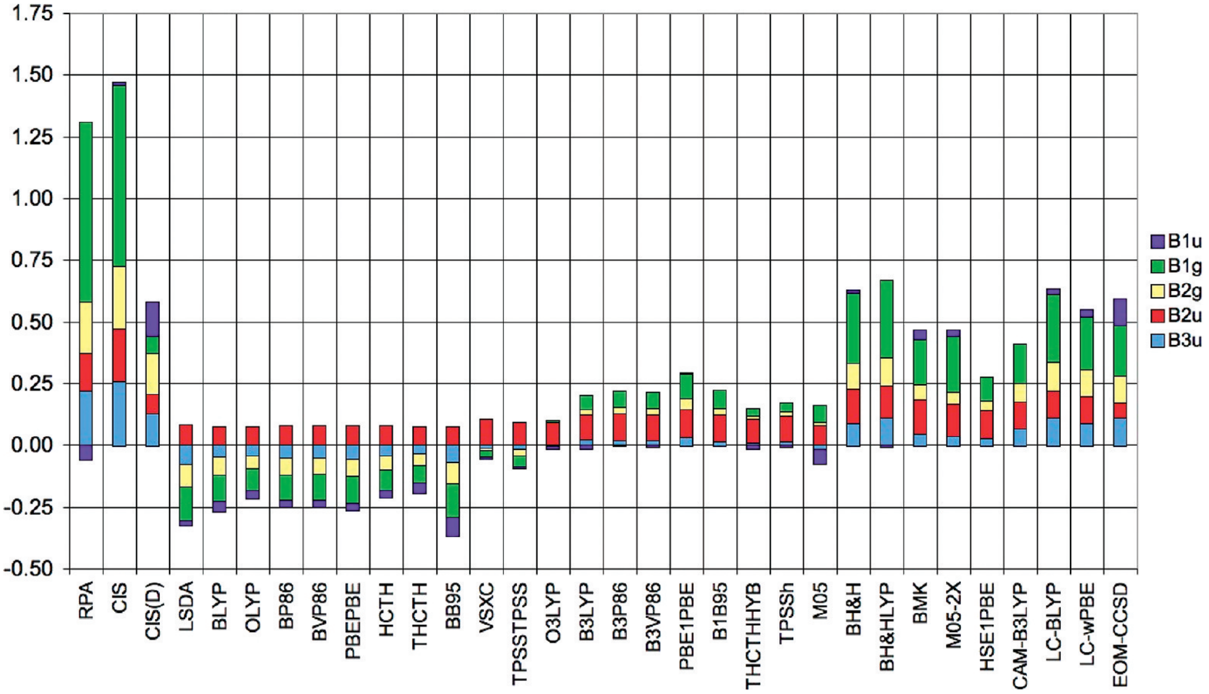


***Figure 8.*** Errors (eV) for pyrazine.

agreement with experiment. In fact, the improvement brought by CIS(D) over CIS is quite small. The case of EOM-CCSD is even more evident, whose performance is, this time, quite worse than the previous cases. Pure functionals underestimate all the excitations but the $B_2$ ($\pi \rightarrow \pi^*$). They also provide a significantly better description of this system than CIS, RPA, and CIS(D). M-GGA functionals perform slightly better than GGA. Hybrid functionals provide again smaller errors than pure functionals, but this time functionals with a small amount of HF exchange perform quite well, and often better than the ones with large amounts. From Figure 7, we can see for example the large error bars of BH&H and BH&HLYP, and of LC-BLYP and LC-$\omega$PBE, which performed among the best in the previous sections. For this molecule, the best agreement with experiments is achieved with CAM-B3LYP.

For pyrazine (1,4-diazine), five excitations are compared with experiments,[56,57,60] that include both $n \rightarrow \pi^*$ ($B_{3u}$, $B_{2g}$, and $B_{1g}$) and $\pi \rightarrow \pi^*$ types ($B_{2u}$ and $B_{1u}$). The errors are represented in Figure 8. For this system, most of the considerations reported for pyridine are valid and often enhanced. Ab initio calculated transition energies are quite far from the experimental data. Again, double excitations are not enough to get an accurate description of these transitions, as shown by the poor performances of CIS(D) and EOM-CCSD. However, pure functionals perform drastically better than the ab initio methods. They underestimate all the transition energies but the $B_{2u}$ ($\pi \rightarrow \pi^*$). The behavior of the hybrid functionals is similar to what we found for pyridine. In this case, the dependence on the HF exchange amount is even more evident since the difference between ab initio methods and pure functionals is larger. Indeed, the larger the amount of HF exchange, the larger the error. Thus the best results are obtained with O3LYP, THCTHHYB, and TPSSh. Slightly worse results are obtained with B3LYP, B3P86, B3VP86, B1B95, and M05. Larger errors are shown by M05−2X, BMK, BH&H, BH&HLYP, CAM-B3LYP, LC-BLYP, and LC-$\omega$PBE. For the latter functionals, the deviations from the experiment are larger than for many of the pure functionals. Interesting in this respect is also the comparison between BB95 and B1B95 and between M05 and M05−2X. The former two show a *decrease* of the total error by a factor of 2, passing from a pure functional to a hybrid one with 25% of HF exchange. The latter two show an *increase* of the error by a factor of 2, passing from 28% to 56% of HF exchange.

For pyrimidine (1,3-diazine), six experimental transition energies are available for comparison, again including $n \rightarrow \pi^*$ (two $B_1$ and two $A_2$) and $\pi \rightarrow \pi^*$ ($B_2$ and $A_1$) transitions.[56,57,60] The error bars are reported in Figure 9, and we can see trends that resemble the previous case (see Figure 8). CIS and RPA errors are larger than for pyrazine. CIS(D) and EOM-CCSD are relatively better. In this case, double excitations seem to play a more important role, but the lack of higher order excitations is still evident. Hybrid functionals perform better than the pure ones as long as the HF exchange amount is not larger than 28% (M05). The number of overestimated transition energies also raises with the percentage of HF exchange. B3LYP, B3P86, B3VP86, THCTHHYB, TPSSh, and M05 show the best performances for this molecule.

Pyridazine (1,2-diazine) is the third diazabenzene. Five transition energies are experimentally available, $n \rightarrow \pi^*$ (two $B_1$ and $A_2$) and $\pi \rightarrow \pi^*$ ($B_2$ and $A_1$) transitions, and they were used to compare the computed data.[56,61] The total error bars are plotted in Figure 10. The errors for the ab initio methods are still quite large, and most of the transition energies are overestimated. CIS and RPA are the least accurate. CIS(D) and EOM-CCSD improve the agreement with the experiments, but they are far from the expected accuracy. However, for the $B_2$ transition ($\pi \rightarrow \pi^*$), CIS(D) shows a larger error than CIS. The DFT functionals show

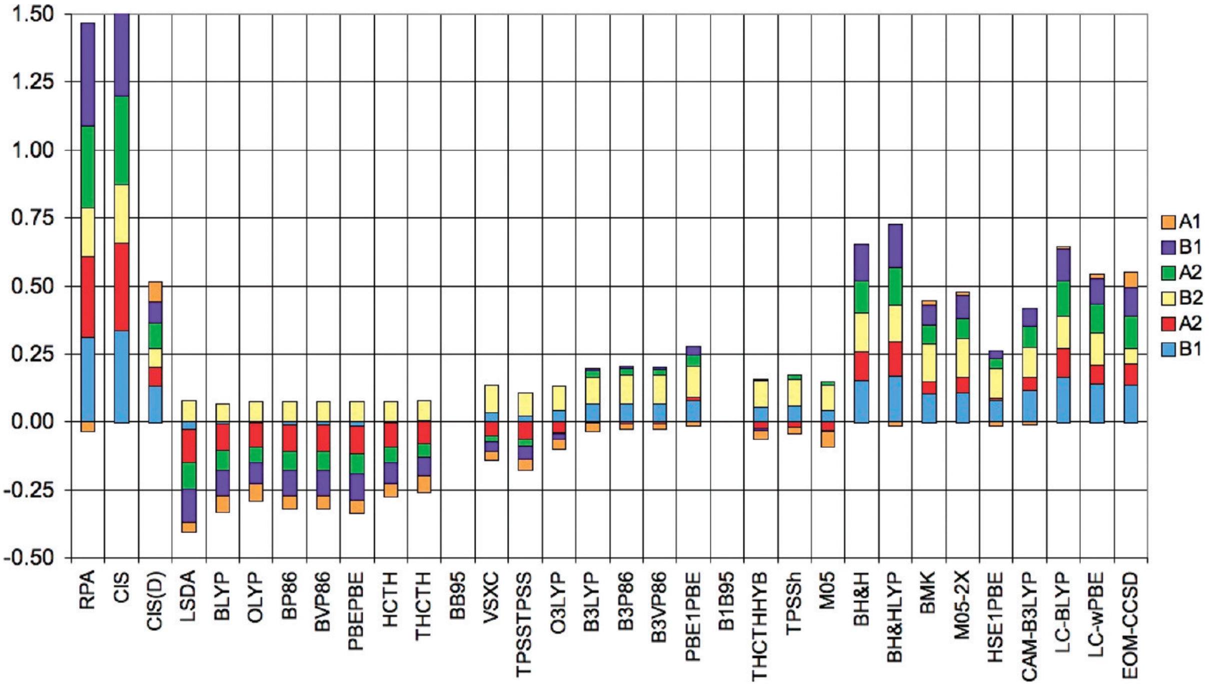


*Figure 9.* Errors (eV) for pyrimidine. The SCF of BB95 and B1B95 did not converge.

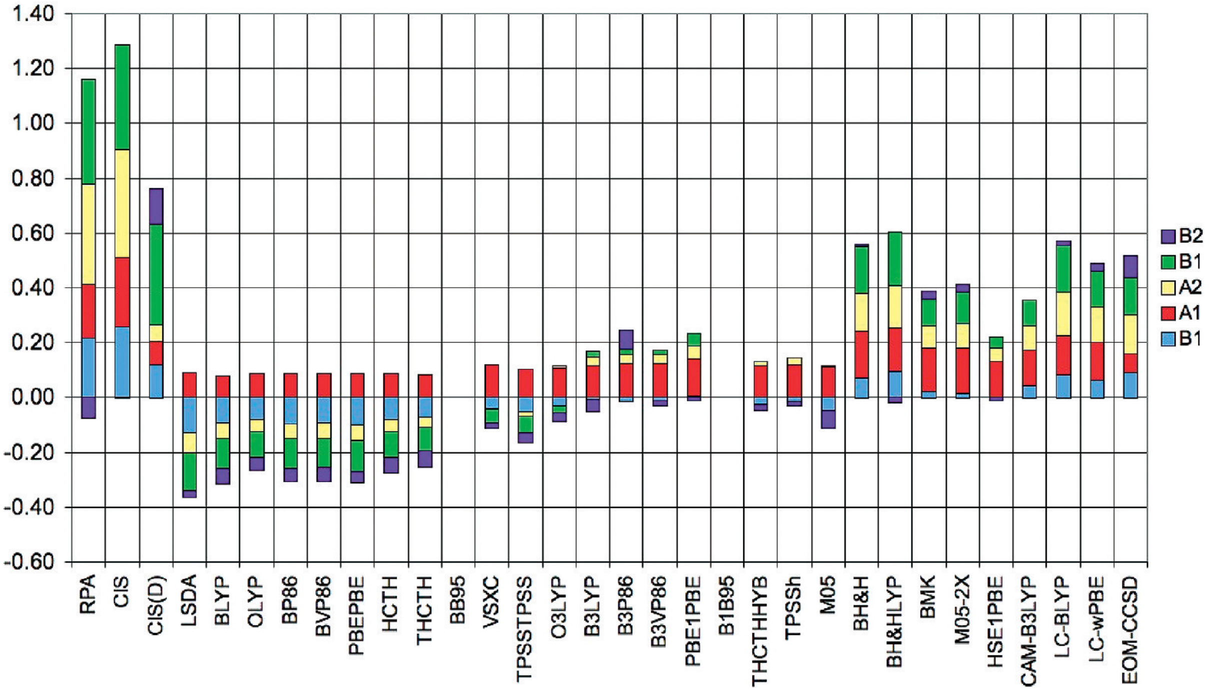


*Figure 10.* Errors (eV) for pyridazine. The SCF of BB95 and B1B95 did not converge.

trends similar to the previous three molecules. Pure functionals underestimate most of the transitions energies, except for the $A_1$ ($\pi \rightarrow \pi^*$). Hybrid functionals perform generally better than pure functionals. B3VP86, THCTHHYB, and TPSSh show the best agreement with the experiments, and B3LYP, B3P86, PBE1PBE, M05, and HSE1PBE errors are not much larger.

Symmetric tetrazine contains four nitrogen centers. Four transition energies, all $n \rightarrow \pi^*$,[56,60,62−66] are compared with computed data. The errors are reported in Figure 11. RPA and CIS errors are very large. CIS(D) drastically improves the description. This is the same range of error provided by EOM-CCSD. GGA and M-GGA functionals total errors are smaller than CIS(D) and EOM-CCSD. Almost all of the hybrid functionals provide superior accuracy than the pure functionals. The best among them, B3LYP, B3P86, B3VP86, PBE1PBE, HSE1PBE, THCTHHYB, and TPSSh have a small percentage of HF exchange. M05−2X, BMK, BH&H, BH&HLYP, CAM-B3LYP, LC-BLYP, and LC-$\omega$PBE show larger errors and they overestimate all of the transitions. Also, this case shows how the azabenzenes are a difficult test set for HF based methods (where among them we can include hybrid functionals with large part of HF exchange and exact long-

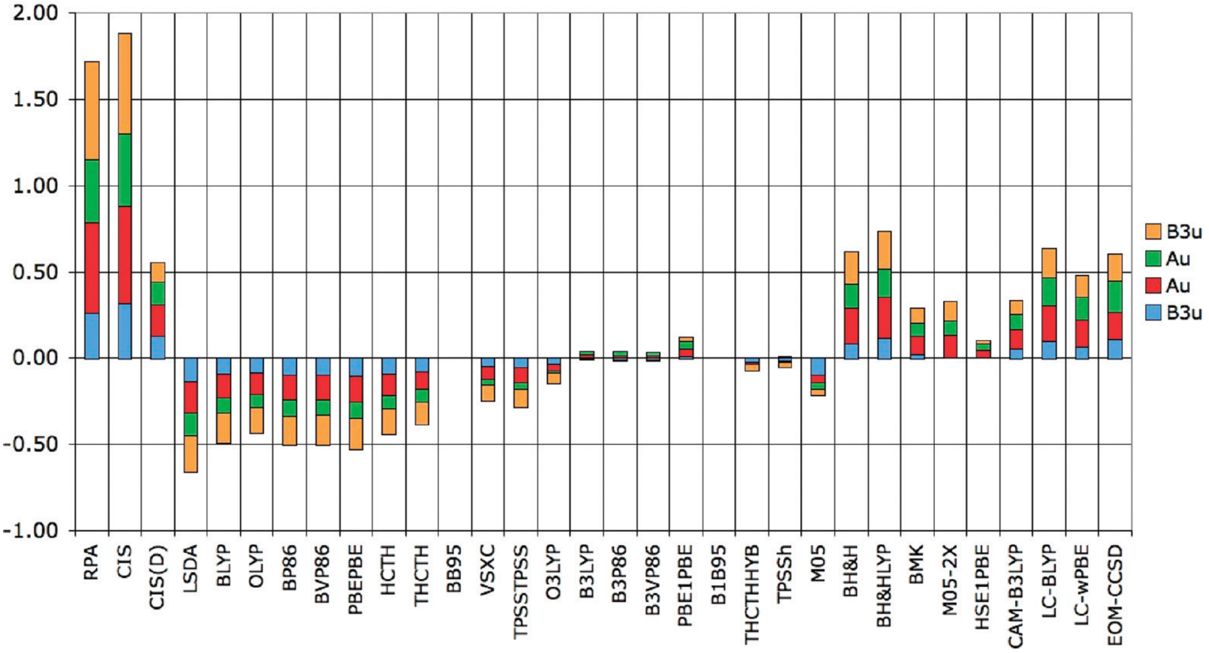


***Figure 11.*** Errors (eV) for S-tetrazine. The SCF of BB95 and B1B95 did not converge.

range functionals). Double excited determinants play a very important role, as evident by comparing CIS and RPA with CIS(D) and EOM-CCSD, but higher excitation are necessary to obtain a better agreement with experiments.

The azabenzenes add another piece of information to the puzzle. In this case, ab initio methods show quite large errors. CIS and RPA are particularly unsatisfactory, with large overestimation of almost all the excitation energies and total errors larger than the pure functionals. CIS(D) improves the results, but in many cases such improvement is small. Surprisingly, EOM-CCSD provides a considerably worse performance for this class of systems than for the previous two. The large basis set employed here allows us to exclude this as a possible cause of the disagreement with the experiments. Del Bene et al.[56] showed that higher order excitations are necessary to obtain an accurate description of the excitation energies in this type of systems. Nevertheless, the EOM-CCSD errors are not dramatic, if compared to the functional ones for alkenes and carbonyls. Pure functionals perform better than CIS and RPA, and sometimes even better than CIS(D) and EOM-CCSD, underestimating most of the measured data. For this set, there are a few cases where GGA and M-GGA functionals perform better than LSDA. Hybrid functionals in many cases show smaller errors than pure functionals and ab initio methods. However, this time the performance among the hybrid ones is inverted with respect to the previous two sets of molecules. In fact, the functionals with a small contribution of HF exchange often largely outperform the ones with a large contribution. The latter show very large errors when the total error of the ab initio methods is much larger than the one of the pure functionals. LC-BLYP and LC-$\omega$PBE behave like the functionals with fixed ratio of DFT and HF exchange and large percentage of the latter. We also note that in this case the functionals with small percentage of HF exchange are closer to the experiments than EOM-CCSD.

**3.4. Discussion.** Now that we have condensed the results we obtained into three sets of molecules, we can draw some general conclusions. We do not discuss any statistical analysis because the number and nature of the transitions considered is not a significant statistical sample, as mentioned in the Introduction. In fact, our test set includes only three alkenes and three carbonyl compounds, with mostly Rydberg states, and five azabenzenes with all valence states. The reduction of all of the collected results into few numbers for each method would be misleading, as these numbers would change dramatically by changing the size of the test set. For example, for a method like EOM-CCSD, which does not perform very well with the azabenzenes, but is extremely accurate for the other molecules, the statistical error would be large because the azabenzenes group would be overrepresented. However, for the interested reader, we report a comprehensive statistical analysis in Tables 23−25 in the Supporting Information. The tables report the analysis for all the molecules together, for the first and for all the states. Also, we report the analysis after separation of the azabenzenes group from the other two, again for the first and for all the states. In the following, however, only trends and qualitative behaviors are discussed.

CIS and RPA often provided large errors, except in the alkene set. They almost constantly overestimated the experimental data and there was not a significative difference between their results. Adding perturbative doubles corrections to CIS, CIS(D), led to an improvement of the results, but this was often not impressive, considering the computational cost added to the calculation and comparing with many functional performances. EOM-CCSD did a very good job for alkenes and especially for the aldehydes and acetone, but it was not able to accurately describe the azabenzenes. However, it was shown[56] that by increasing the excitation manifold, it is possible to obtain a better agreement with the experimental results. This is a very important feature of the coupled cluster theory, even for excited states, that it is always possible to systematically improve the description of the electronic structure of a molecule. However, this leads to a rapidly increasing computational effort that may become unbearable even for small systems.

Unfortunately, the opposite is true for many of the DFT functionals available nowadays, i.e., the improvements do not become overwhelming from the computational point of view, but they are also not systematic, at least for the excitation energies. Pure functionals showed performances very often worse than CIS, and it is in a way surprising how LSDA errors were in many cases smaller than gradient corrected functionals. Almost all of the experimental data were underestimated.

The behavior of hybrid functionals with a fixed amount of HF exchange mostly relied on the cancellation of errors between the HF and the pure DFT part. In fact, the examples in this work seem to show how the relative accuracy between the CIS and the pure functionals favored the hybrid functionals with more or less HF exchange contribution, depending on the cases. In particular, when CIS accuracy was good, as in the alkene set, hybrid functionals with large part of HF exchange performed better. Also, when both CIS and pure functionals errors were large but comparable and opposite in sign, the errors of functionals with large HF exchange were small, as for the case of the carbonyl compounds. On the other hand, when the CIS errors were much larger than the pure functionals ones, as for the azabenzenes, the trend reversed and the functionals with small percentage of HF exchange clearly outperformed the ones with a large contribution. The sign of the error is also significant for the hybrid functionals along the various sets of molecules. The errors went from negative to mixed to positive, going from the alkenes to the carbonyl systems to the azabenzenes, following the relative magnitude of the ab initio vs the pure functionals errors. This behavior is even more evident when we compare M05 and M05−2X, that mainly differ in the amount of HF exchange, 28% and 56%, respectively. Indeed, we found an almost complete change in the sign of the errors for isobutene, acetaldehyde, and acetone.

The functionals with the correct asymptotic bahavior, LC-BLYP and LC-$\omega$PBE, which have a variable amount of HF exchange depending on the distance from the nuclei, showed a consistent overestimation of the transition energies, with a few exceptions as in the formaldehyde, where the overlall error was already small. This is a good feature, as it already provides some information on the sign of the expected error. However, the general performance of such functionals is very similar to that of functionals with large (and fixed) amount of HF exchange, thus we can group all those functionals in the same category. We also point out that such correction for the asymptotic behavior cannot be directly applied to functionals which are already hybrid.

A functional that showed a consistently reasonable behavior throughout all of the test cases that we examined was B3P86, that has 20% HF exchange. This functional often performed as well as functionals with large amount of HF exchange, even for cases where the latter were favored, and it obviously performed much better than them for the azabenzenes. The popular B3LYP provided larger errors than B3P86, which is not surprising if we consider that pure BLYP errors were larger than the BP86 ones. The hybrid functional with the worst performance was O3LYP, which often provided errors of the same order of the pure functionals. BB95 and B1B95 also deserve a separated comment, as they showed numerical instabilities that prevented the convergence of the SCF in six cases out of eleven.

At this point, a comparison with the work in ref 15 may be useful, as it is on the same molecular property and our results may seem to lead to different conclusions. The largest difference is in the definition of the test set: Our set includes 30 valence states and 39 diffuse Rydberg states of small molecules in gas phase, whereas most of the test set in ref 15 is based on experimental data on the first excited state of large chromophores in solution, with a large oscillator strength. Additionally, we used a very large basis set because, as reported in ref 16, the lack of diffuse functions may lead to large errors, not only for EOM-CCSD but also for TDDFT, especially for higher and diffuse states. For low lying states, like the valence states, basis set issues are generally less dramatic. Therefore, the average error of 0.14−0.18 eV reported in the conclusions of ref 15 for PBE1PBE and LC-$\omega$PBE(20), although it still seems optimistic since such an accuracy is not even claimed for an ab initio method like EOM-CCSD,[56,67] may only apply to the lowest bright state of large chromophores. In fact, the statistical analysis on the first excited state for our test set, reported in the Supporting Information, is in agreement with this result, although our test set is extremely limited. However, note the better performance of B3P86 in this comparison.

Clearly many factors influence the relative accuracy of the various functionals, but this work shows that one of the most important seems to be the error compensation present in the hybrid functionals. For instance, as far as the electronic excitations are concerned, the kinetic energy density contribution does not significantly improve the results of both GGA and H-GGA types of functionals; whereas, it increases the computational effort. Also, the separation of long and short-range exchange seems to be less important than the percentage of HF exchange, as shown by CAM-B3LYP, LC-BLYP, and LC-$\omega$PBE. The case of HSE1PBE is slightly different, as for small molecules like the ones we considered this functional behaves like PBE1PBE.

## 4. Conclusions

Before summarizing the results reported in this work, we note that this work is not meant to be definitive, since new functionals appear almost monthly and many were left out from our representative set, and since other molecular systems with different characteristics may be studied. Nevertheless, it can bring some light to investigators who are struggling to decide which method is better for their electronic excitation studies, especially among the plethora of DFT functionals available in the computational packages. We show how CIS, RPA and pure functionals are not, in general, a good choice. Hybrid functionals with a fixed amount of HF exchange are often in between those methods, but their better performance seems to be mainly related to error compensation and, thus, it cannot be systematic. Asymptotically correct functionals seem to show a more consistent overestimation of the experimental data, thus providing a reference for the estimation of the error, although their overall performance is very similar to functionals with large and fixed amount of HF exchange. Remarkably, the best average performance is obtained with a hybrid functional with a small amounts of HF exchange, B3P86, that appeared in 1993 and was not specifically designed for excited state properties. CIS(D) also does not seem a good choice, as its performance is often worse than many hybrid functionals. EOM-CCSD results are very good for alkenes and carbonyls but less for the azabenzenes, where higher order excitations in the cluster expansion seem to be necessary. However, highly correlated wave function-based methods like EOM-CCSD are, at least so far, always more reliable than any DFT functional, because they represent a secure way to approach the experiment and they should be used when possible. However, even if successful approximations to such methods have been proposed and used in many circumstances, DFT still represents the best compromise between accuracy and computational effort. However, large differences in the results are found between the various functionals, thus the choice of the functional can largely affect the accuracy of a calculation. Therefore, we hope that this work can be helpful when it comes to making a decision about which method to use to compute electronic transition energies.

**Acknowledgment.** M.C. thanks Giovanni Scalmani for useful discussions about the differences among the DFT functionals. The authors also thank Fernando Clemente for reading the manuscript and providing interesting comments.

**Supporting Information Available:** Tables 1−25, reports the geometry of all the molecules, the calculated and experimental transition energies with their characterization and complete statistical analysis. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Foresman, J. B.; Head-Gordon, M.; Pople, J. A.; Frisch, M. J. *J. Phys. Chem.* **1992**, *96*, 135–149.

(2) Head-Gordon, M.; Rico, R. J.; Oumi, M.; Lee, T. J. *Chem. Phys. Lett.* **1994**, *219*, 21–29.

(3) McWeeny, R. *Methods of Molecular Quantum Mechanics*, 2nd Edition; Academic Press: London, 1992; pp 435−438.

(4) Casida, M. E. *Recent Advances in Density Functional Methods*; World Scientific: Singapore, 1995; Vol. 1.

(5) Casida, M. E.; *Recent Developments and Applications of Modern Density Functional Theory, Theoretical and Computational Chemistry*; Elsevier: Amsterdam, 1996; Vol. 4.

(6) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 8218–8224.

(7) Bartlett, R. J.; Musial, M. *Rev. Mod. Phys.* **2007**, *79*, 291–352.

(8) Sekino, H.; Bartlett, R. J. *Int. J. Quantum Chem.: Quantum Chem. Symp.* **1984**, *18*, 255–265.

(9) Stanton, J. F.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 7029–7039.

(10) Koch, H.; Jorgensen, P. *J. Chem. Phys.* **1990**, *93*, 3333–3344.

(11) Kallay, M.; Gauss, J. *J. Chem. Phys.* **2004**, *121*, 9257–9269.

(12) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Phys. Chem. A* **2007**, *111*, 10439–10452.

(13) Zhang, G.; Musgrave, C. B. *J. Phys. Chem. A* **2007**, *111*, 1554–1561.

(14) Silva-Junior, M. R.; Schreiber, M.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *129*, 104103.

(15) Jacquemin, D.; Wathelet, V.; Perpète, E. A.; Adamo, C. *J. Chem. Theory Comput.* **2009**, *5*, 2420–2435, and references therein.

(16) Wiberg, K. B.; de Oliveira, A. E.; Trucks, G. *J. Phys. Chem. A* **2002**, *106*, 4192–4199.

(17) Slater, J. C. *Phys. Rev.* **1951**, *81*, 385–390.

(18) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

(19) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

(20) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(21) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.

(22) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403–412.

(23) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.

(24) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264–6271.

(25) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040–1046.

(26) Van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400–410.

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

(28) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396–1396.

(29) Boese, A.; Doltsinis, N.; Handy, N.; Sprik, M. *J. Chem. Phys.* **2000**, *112*, 1670–1678.

(30) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2001**, *114*, 5497–5503.

(31) Boese, A. D.; Handy, N. C. *J. Chem. Phys.* **2002**, *116*, 9559–9569.

(32) Hoe, W. M.; Cohen, A. J.; Handy, N. C. *Chem. Phys. Lett.* **2001**, *341*, 319–328.

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

(34) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103.

(35) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364–382.

(36) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405–3416.

(37) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2003**, *118*, 8207–8215.

(38) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(39) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.

(40) Tawada, T.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425–8433.

(41) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 234109.

(42) Vydrov, O. A.; Heyd, J.; Krukau, V.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 074106.

(43) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(44) Ernzerhof, M.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 5029–5036.

(45) Stephens, P. J.; Devlin, F. J.; Ashvar, C. S.; Chabalowski, C. F.; Frisch, M. J. *Faraday Discuss.* **1994**, *99*, 103–119.

(46) Staroverov, V. N.; Scuseria, G. E.; Tao, J. M.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129–12137.

(47) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

(48) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *J. Chem. Phys.* **2008**, *128*, 044118.

(49) Casida, M. E.; Jamorski, C.; Casida, K. C.; Salahub, D. R. *J. Chem. Phys.* **1998**, *108*, 4439–4449.

(50) Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 10180–10189.

(51) Rohrdanz, M. A.; Herbert, J. M. *J. Chem. Phys.* **2008**, *129*, 034107.

(52) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Scalmani, G.; Mennucci, B.; Barone, V.; Petersson, G. A.; Caricato, M.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Li, X.; Hratchian, H. P.; Peralta, J. E.; Izmaylov, A. F.; Kudin, K. N.; Heyd, J. J.; Brothers, E.; Staroverov, V. N.; Zheng, G.; Kobayashi, R.; Normand, J.; Sonnenberg, J. L.; Ogliaro, F.; Bearpark, M.; Parandekar, P. V.; Ferguson, G. A. Mayhall, N. J.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Burant, J. C.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; 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.; Chen, W.; Wong M. W.; and Pople, J. A. *Gaussian Development Version, Revision G.01+*, Gaussian, Inc., Wallingford CT 2008.

(53) Adamo, C.; Scuseria, G. E.; Barone, V. *J. Chem. Phys.* **1999**, *111*, 2889–2899.

(54) Wiberg, K. B.; Hadad, C. M.; Ellison, G. B.; Foresman, J. B. *J. Phys. Chem.* **1993**, *97*, 13586–13597.

(55) Hadad, C. M.; Foresman, J. B.; Wiberg, K. B. *J. Phys. Chem.* **1993**, *97*, 4293–4312.

(56) Del Bene, J. E.; Watts, J. D.; Bartlett, R. J. *J. Chem. Phys.* **1997**, *106*, 6051–6060.

(57) Bolovinos, A.; Tsekeris, P.; Philis, J.; Pantos, E.; Andritsopoulos, G. *J. Mol. Spectrosc.* **1984**, *103*, 240–256.

(58) Walker, I. C.; Palmer, M. H.; Hopkirk, A. *Chem. Phys.* **1990**, *141*, 365–378.

(59) Goodman, L. *J. Mol. Spectrosc.* **1961**, *6*, 109–137.

(60) Innes, K. K.; Ross, I. G.; Moomaw, W. R. *J. Mol. Spectrosc.* **1988**, *132*, 492–544.

(61) Palmer, M. H.; Walker, I. C. *Chem. Phys.* **1991**, *157*, 187–200.

(62) Palmer, M. H.; McNab, H.; Reed, D.; Pollacchi, A.; Walker, I. C.; Guest, M. F.; Siggel, M. R. F. *Chem. Phys.* **1997**, *214*, 191–211.

(63) Spencer, G. H.; Cross, P. C.; Wiberg, K. B. *J. Chem. Phys.* **1961**, *35*, 1925–1938.

(64) Adamo, C.; Barone, V. *Chem. Phys. Lett.* **2000**, *330*, 152–160.

(65) Nooijen, M. *J. Phys. Chem. A* **2000**, *104*, 4553–4561.

(66) Devarajan, A.; Gaenko, A. V.; Khait, Y. G.; Hoffmann, M. R. *J. Phys. Chem. A* **2008**, *112*, 2677–2682.

(67) Stanton, J. F.; Gauss, J.; Ishikawa, N.; Head-Gordon, M. *J. Chem. Phys.* **1995**, *103*, 4160–4174.

CT9005129

JCTC Journal of Chemical Theory and Computation

# An Efficient Parallel All-Electron Four-Component Dirac−Kohn−Sham Program Using a Distributed Matrix Approach

Loriano Storchi,[†] Leonardo Belpassi,[†] Francesco Tarantelli,[*,†] Antonio Sgamellotti,[†] and Harry M. Quiney[‡]

*Dipartimento di Chimica and I.S.T.M.−C.N.R., Università di Perugia, 06123, Italy, and ARC Centre of Excellence for Coherent X-ray Science School of Physics, The University of Melbourne, Victoria, 3010, Australia*

Received October 12, 2009

**Abstract:** We show that all-electron relativistic four-component Dirac−Kohn−Sham (DKS) computations, using G-spinor basis sets and state-of-the-art density fitting algorithms, can be efficiently parallelized and applied to large molecular systems, including large clusters of heavy atoms. The performance of the parallel implementation of the DKS module of the program BERTHA is illustrated and analyzed by some test calculations on several gold clusters up to $Au_{32}$, showing that calculations with more than 25 000 basis functions (i.e., DKS matrices on the order of 10 GB) are now feasible. As a first application of this novel implementation, we investigate the interaction of the atom Hg with the $Au_{20}$ cluster.

## I. Introduction

Understanding the electronic structure and properties of molecules, clusters, and nanoscale materials containing heavy atoms represents a particularly challenging task for theory and computational science because the systems of interest have typically very many electrons, and both relativistic effects and electron correlation play a crucial role in their dynamics. The most rigorous way to introduce relativity in the modeling of molecular systems is to use the four-component formalism derived from the Dirac equation. The method of choice is density functional theory (DFT) if many electrons are involved, as is the case with large metal clusters. In DFT, which is normally cast in the form of the independent-particle Kohn−Sham model, all of the exchange-correlation effects are expressed implicitly as a functional of the electron density or, more generally, of the charge-current density.[1,2] The relativistic four-component generalization of the Kohn−Sham method, usually referred to as the Dirac−Kohn−Sham (DKS) model, was introduced several years ago.[3] Several modern implementations of this theory are available,[4–7] including the one contained in our own program, BERTHA.[8–16] The full four-component DKS formalism is particularly appealing because it affords great physical clarity and represents the most rigorous way of treating explicitly and ab initio all interactions involving spin, which are today of great technological importance.

The full four-component DKS calculations have an intrinsically greater computational cost than analogous nonrelativistic approaches or less rigorous quasi-relativistic approaches, mainly because of the four-component structure in the representation of the DKS equation, the complex matrix representation that usually arises as a consequence, the increased work involved in the evaluation of the electron density from the spinor amplitudes, and the intrinsically larger basis sets usually required. This greater computational cost, however, essentially involves only a larger prefactor in the scaling with respect to the number of particles or the basis set size, not a more unfavorable power law. Schemes have been devised in order to reduce the computational cost (see, e.g., refs 17–19 and references therein). A significant step forward in the effective implementation of the four-component DKS theory is based on the electron-density fitting approach that is already widely used in the nonrelativistic context. Numerical density fitting approaches based on an atomic multipolar expansion[7] and on a least-squares fit[20] have in fact been employed in the four-component

[*] Corresponding author e-mail: franc@thch.unipg.it.

[†] Università di Perugia.

[‡] The University of Melbourne.

10.1021/ct900539m
Published on Web 01/22/2010

relativistic domain. Recently, we have implemented the variational Coulomb fitting approach in our DKS method,[14] with further enhancements resulting from the use of the Poisson equation in the evaluation of the integrals,[15,21,22] and also from the extension of the density fitting approach to the computation of the exchange-correlation term.[16]

The above algorithmic advances have represented a leap forward of several orders of magnitude in the performance of the four-component DKS approach and have suddenly shifted the applicability bottleneck of the method toward the conventional matrix operations (DKS matrix diagonalization, basis transformations, etc.) and, especially, to the associated memory demands arising in large-system/large-basis calculations. One powerful approach to tackle these problems and push significantly further forward the applicability limit of all-electron four-component DKS is parallel computation with memory distribution. Analogous parallelization efforts of nonrelativistic DFT codes have already been described in the literature.[23,34] The purpose of the present paper is to describe the successful implementation of a comprehensive parallelization strategy for the DKS module of our code BERTHA.

In section II, we briefly review the DKS method as currently implemented in BERTHA and the computational steps making up a SCF iteration. In section III, we describe in detail the parallelization strategies adopted here. In section IV, we discuss the efficiency of the approach as results from some large all-electron test calculations performed on several gold clusters. Finally, we will present an actual chemical application of the method for the all-electron relativistic study of the electronic structure of Hg−Au$_{20}$ with a large basis set.

## II. Overview of the SCF Procedure

In this section, we will briefly review the DKS method as it is currently implemented in BERTHA. We will mainly underline its peculiarities, especially in relation to the density-fitting procedure, and summarize the steps making up a SCF iteration and their typical serial computational cost for a large case. Complete details of the formalism can be found in refs 8, 9, 11, 14–16.

In BERTHA, the large ($L$) and small ($S$) components of the spinor solutions of the DKS equation are expanded as linear combinations of Gaussian $G$-spinor basis functions. A peculiar and important feature of the BERTHA approach is that the density elements, $\varrho_{\mu\nu}^{TT}(\mathbf{r})$, which are the scalar products of pairs of $G$ spinors (labeled by $\mu$ and $\nu$, with $T = L, S$), are evaluated exactly as finite linear combinations of scalar auxiliary Hermite Gaussian-type functions (HGTF). This formulation[9,11] enables the highly efficient analytic computation of all of the required multicenter $G$-spinor interaction integrals.

In the current implementation of BERTHA, the computational burden of the construction of the Coulomb and exchange-correlation contributions to the DKS matrix has been greatly alleviated with the introduction[14,15] of some effective density fitting algorithms based on the Coulomb metric, which use an auxiliary set of HGTF fitting functions. The method results in a symmetric, positive-definite, linear system,

$$\mathbf{A}\mathbf{c} = \mathbf{v} \tag{1}$$

to be solved in order to obtain the vector of fitting coefficients, $\mathbf{c}$. The procedure involves only the calculation of two-center Coulomb repulsion integrals over the fitting basis set, $A_{ij} = \langle f_i \| f_j \rangle$, and three-center integrals between the fitting functions and $G$-spinor density overlaps, $I_{i,\mu\nu}^{TT} = \langle f_i \| \varrho_{\mu\nu}^{TT} \rangle$. The vector $\mathbf{v}$ in eq 1 is simply the projection of the electrostatic potential (due to the true density) on the fitting functions:

$$v_i = (f_i \| \varrho) = \sum_{T=L,S} \sum_{\mu\nu} I_{i,\mu\nu}^{TT} D_{\mu\nu}^{TT} \tag{2}$$

where $D_{\mu\nu}^{TT}$ are the density matrix elements.[14] In our implementation, we take further advantage of a relativistic generalization of the $\mathbf{J}$-matrix algorithm[11,25,26] and an additional simplification arising from the use of sets of primitive HGTFs of common exponents and spanning all angular momenta from zero to the target value (for details, see ref 14).

The auxiliary fitted density can be directly and efficiently used for the calculation of the exchange-correlation potential,[27–29] and we have implemented this procedure in our DKS module.[16] It is based on the solution a linear system similar to the Coulomb fitting one:

$$\mathbf{A}\mathbf{z} = \mathbf{w} \tag{3}$$

where the only additional quantity to be computed is the vector $\mathbf{w}$ representing the projection of the “fitted” exchange-correlation potential onto the auxiliary functions:

$$w_i = \int v_{\mathrm{xc}}[\tilde{\varrho}(\mathbf{r})]\, f_i(\mathbf{r})\, \mathrm{d}\mathbf{r} \tag{4}$$

The elements of the vector $\mathbf{w}$, involving integrals over the exchange-correlation potential $v_{\mathrm{xc}}$, are computed numerically by a standard cubature scheme.[8] The cost of this step tends to become negligible, scaling linearly with both the number of auxiliary functions and the number of integration grid points. In the integration procedure, we again take advantage of our particular choice of auxiliary functions: the use of primitive HGTFs that are grouped together in sets sharing the same exponent minimizes the number of exponential evaluations at each grid point. Further computational simplification arises from using the recurrence relations for Hermite polynomials in the evaluation of the angular part of the fitting functions and their derivatives.[28] Using the above combined approach, the Coulomb and exchange-correlation contributions to the DKS matrix can be formed in a single step:

$$\tilde{J}_{\mu\nu}^{TT} + \tilde{K}_{\mu\nu}^{TT} = \sum_i I_{i,\mu\nu}^{TT} (c_i + z_i) \tag{5}$$

We have found that the reduction of the computational cost afforded by the above density fitting scheme is dramatic. Besides reducing the scaling power of the method from O(N$^4$) to O(N$^3$), it reduces enormously the prefactor (by up to 2 orders of magnitude) without any appreciable accuracy loss.[16] The application of the method to very large systems is, however, still impeded by the substatial memory require-

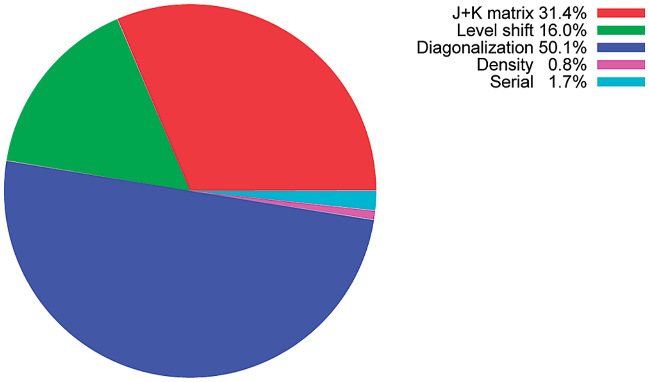


***Figure 1.*** CPU time percentages for the various phases of a serial DKS calculation of the gold cluster $Au_{16}$. All linear algebra operations are performed with the Intel Math Kernel Library.

ments imposed by huge basis sets and by the consequently large matrix dimensions. A truly practical parallelization scheme must inevitably tackle these aspects of the problem.

Before we proceed to describe the parallelization of our code, it is useful to take a brief look at the time analysis of a typical serial SCF iteration of our DKS program, shown in Figure 1 for a realistically sized case: the $Au_{16}$ cluster, with a DKS matrix dimension of 12 480. Here, we see that three $O(N^3)$ steps dominate the computation: diagonalization, $\mathbf{J} + \mathbf{K}$ matrix computation, and the level-shifting phase. Thanks to the significant progress reviewed above, the $\mathbf{J} + \mathbf{K}$ matrix computation time, formerly dominant, has been drastically reduced and, consequently, a larger time fraction (about half in the present example) is taken up by the diagonalization step. In our present serial implementation, the full DKS eigenvalue spectrum, comprising both the negative and positive energy halves, is computed. As we shall see later, we have adopted, in the parallel code, an effective reduction of the computed spectrum to the sole positive-energy occupied spinors, which affords considerable time savings. Projection methods such as those proposed by Peng et al.,[30] halving the size of the diagonalization problem, could also usefully be employed. We have not further investigated this point because, as hinted at above, our emphasis here is on the effective removal of the memory bottleneck for very large-scale applications, through data distribution. The $\mathbf{J} + \mathbf{K}$ matrix computation, which takes about a third of the time, mainly involves the calculation of the three-index integrals $I^{TT}_{i,\mu\nu}$ and the implementation of eq 5. The remaining sixth of the time is used almost entirely in the level-shifting phase, which involves the double matrix multiplication transforming the DKS matrix from basis function space to spinor space. Clearly, the parallelization effort must target these three time-consuming phases. It is remarkable that the entire density-fitting procedure, comprising the computation of the $\mathbf{A}$, $\mathbf{v}$, and $\mathbf{w}$ arrays and the solution of the associated linear systems of eqs 1 and 3, takes up an almost negligible fraction of the time. This phase, together with the HGTF expansion of the density, is bundled in the slice which we have labeled "Serial" in Figure 1, because we have left it unparallelized in the work described here. The remaining computation, labeled "Density", involves essentially the matrix multiplication necessary to obtain the density matrix from the occupied positive-energy spinors.

## III. Parallelization Strategy

The code has been developed on a local HP Linux Cluster with an Intel Pentium D, 3.00 GHz CPU with a central memory of 4 GB on each node. The parallel implementation has been ported with success on a parallel SGI Altix 4700 (1.6 GHz Intel Itanium2 Montecito Dual Core) equipped with the SGI NUMAlink[31] interconnect. All of the results in terms of scalability and speedup reported in the following have been obtained on the latter architecture.

In the parallelization of the DKS module of BERTHA, we used the SGI implementation of the Message Passing Interface (MPI)[32] and the ScaLAPACK library.[33] The overall parallelization scheme we used can be classified as master−slave. In this approach, only the master process carries out the "Serial" portion of the SCF described in the previous section. All of the concurrent processes share the burden of the other calculation phases in Figure 1. We decided to use this approach because it is the easiest to code in order to make memory management especially convenient and favorable. Only the master process needs to allocate all of the large arrays, that is, the overlap, density, DKS, and eigenvector matrices. Each slave process allocates only some temporary small arrays when needed. In using this approach to tackle large molecular systems, it is crucial to be able to exploit the fast memory distribution scheme offered by the hardware. In particular, the SGI Altix 4700 is classified as a cc-NUMA (cache-coherent Non-uniform Memory Access) system. The SGI NUMAflex architecture[34] creates a system that can "see" up to 128 TB of globally addressable memory, using the NUMAlink[31] interconnect. The master process is thus able to allocate as much memory as it needs, regardless of the actual amount of central memory installed on each node, achieving good performances in terms of both latency and bandwidth of memory access. Some aspects of performance degradation related to nonlocal memory allocation will be pointed out later on.

**A. $\mathbf{J} + \mathbf{K}$ Matrix Calculation.** To parallelize the $\mathbf{J} + \mathbf{K}$ matrix construction, the most elementary and efficient approach is based on the assignment of matrix blocks computation to the available processes. The optimal integral evaluation algorithm, exploiting HGTF recurrence relations on a single process, naturally induces a matrix block structure dictated by the grouping of *G*-spinor basis functions in sets characterized by common origin and angular momentum (see also ref 12). The master process broadcasts the $\mathbf{c} + \mathbf{z}$ vector to the slaves at the outset of the computation. After this, an on-demand scheme is initiated. Each slave begins the computation of a different matrix block, while the master sets itself listening for messages. When a slave has finished computing one block, it returns it to the master and receives the sequence number identifying the next block to be computed. The master progressively fills the global DKS array with the blocks it receives from the slaves. A slave only needs to temporarily allocate the small blocks it computes.

This approach, as will be evident in the next sections, has several advantages. The communication time is essentially independent of the number of processes involved. The matrix blocks, although all relatively small, have different sizes, which tends to minimize communication conflicts, hiding communications behind computations. The small size of the blocks ensures optimal load balance and also permits a much more efficient use of the cache with respect to the serial implementation.

**B. Matrix Operations.** The parallelization of the matrix operations which make up the bulk of the “level shift”, “diagonalization”, and “density” phases, has been performed using the ScaLAPACK library routines.[33] There are two main characteristics of ScaLAPACK that we need to briefly recall here. First, the $P$ processes of an abstract parallel computer are, in ScaLAPACK, mapped onto a $P_r \times P_c$ two-dimensional grid, or process grid, of $P_r$ process rows and $P_c$ process columns ($P_r \cdot P_c = P$). The shape of the grid for a given total number of processors affects ScaLAPACK performance, and we shall briefly return to this point shortly. The second fundamental characteristic of ScaLAPACK is related to the way in which all of the arrays are distributed among the processes. The input matrices of each ScaLAPACK routine must be distributed, among all of the processes, following the so-called two-dimensional block−cyclic distribution.[33] The same distribution is applied to the result arrays in output. For example, in the case of a matrix−matrix multiplication, the two input matrices must be distributed following the two-dimensional block cyclic distribution, and when the computation is done, the result matrix will be distributed among all of the processes following the same scheme.

To simplify and generalize the distribution of arrays, and the collection of the results, we first of all implemented two routines using MPI, named DISTRIBUTE(mat,rank) and COLLECT(mat,rank). The first distributes the matrix mat, originally located on process rank, among all processes, according to the block cyclic scheme; the second carries out the inverse operation, gathering on process rank the whole matrix mat block-distributed among the available processes. It is worthwhile to note that both operations, as implemented, exhibit a running time that, for a given matrix size, is very weakly dependent on the number of processes involved. Except for a larger latency overhead, this time is in fact roughly the same as that required to transfer the whole matrix between two processes. Sample timings for the distribution and collection of four double-precision complex matrices of varying sizes are shown in Table 1. The four matrices are in fact the matrices occurring in the DKS calculations of the gold clusters described later in this work.

***Table 1.*** Times in Seconds for the COLLECT (C) and DISTRIBUTE (D) Routines As a Function of Matrix Size and Number of Processors

| | matrix dimension | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1560 | | 3120 | | 6240 | | 12480 | |
| number of processors | C | D | C | D | C | D | C | D |
| 4 | 0.12 | 0.12 | 0.36 | 0.33 | 1.41 | 1.29 | 5.90 | 5.07 |
| 16 | 0.12 | 0.22 | 0.44 | 0.76 | 1.60 | 1.29 | 6.26 | 5.04 |
| 32 | 0.10 | 0.27 | 0.39 | 0.98 | 1.56 | 1.38 | 6.32 | 5.40 |
| 64 | 0.13 | 0.27 | 0.51 | 1.07 | 2.12 | 3.78 | 8.23 | 6.03 |
| 128 | 0.14 | 0.28 | 0.52 | 1.13 | 2.13 | 4.47 | 8.62 | 5.97 |

The ScaLAPACK routines we used for the DKS program are PZHEMM in the “level shift” phase, PZGEMM in the “density” phase, and finally PZHEGVX to carry out the complex DKS matrix diagonalization. Before and after the execution of these routines, we placed calls to our DISTRIBUTE and COLLECT routines to handle the relevant matrices as required. The workflow is extremely simple. Initially, the DKS matrix is distributed to all processors. After this, the “level shift”, “diagonalization”, and “density” steps are performed in this order, exploiting the intrinsic parallelism of the ScaLAPACK routines. At the end, we collect on the master both the density matrix and the eigenvectors. Thus, apart from the internal communication activity of the ScaLAPACK routines, there are just four explicit communication steps, namely, the initial distribution of the DKS and overlap matrices and the final gathering of the resulting eigenvectors and density matrices. Note that the only communication time in the entire calculation that depends appreciably on the number of processors involved is that of the largely insignificant initial broadcast of the $\mathbf{c} + \mathbf{z}$ vector, which is necessary to carry out the $\mathbf{J} + \mathbf{K}$ matrix construction.

Before we proceed, it is necessary to add a final note about the block cyclic decomposition. The scheme is driven, besides by the topology of the processes, by the size of the blocks into which the matrix is subdivided. This size is an important parameter for the overall performance of the ScaLAPACK routines. After some preliminary tests, we chose a block dimension equal to 32 (see also refs 24 and 35), and all of the results presented here are obtained using this block dimension.

**C. Notes about PZHEGVX Diagonalization.** As we have seen at the end of section II, the DKS diagonalization phase takes up the largest fraction of computing time, and therefore all factors that affect its performance are important. While, for practical reasons, we could not investigate exhaustively all of these factors, we would like to highlight two of them which are particularly relevant in the present case.

The PZHEGVX routine needs some extra work arrays to carry out the diagonalization. One can execute a special preliminary dummy call to PZHEGVX in order to obtain the routine’s estimate of the optimal size of this extra memory space for the case at hand. In our test applications, we noticed, however, that the estimated auxiliary memory was in fact insufficient to guarantee, especially for the larger systems, the full reorthogonalization of the eigenvectors (i.e., INFO > 0 and MOD(INFO/2,2) ≠ 0). This appeared to cause some inaccuracies and instabilities in the final results. In order to avoid these problems and also to establish common comparable conditions for all of the test cases studied, we decided to require the accurate orthogonalization of the eigenvectors in all cases. This could readily be achieved by suitably enlarging the size of the work arrays until INFO = 0. For example, in the case of the $Au_{16}$ cluster (with a 2.4 GB DKS matrix), when 16 processors were used, the

ScaLAPACK estimate of 180 MB per processor for the work arrays had to be raised to 400 MB to achieve complete reorthogonalization. Clearly, this strict requirement is quite costly in terms of both memory and computation time, and less tight constraints might be investigated and found to be acceptable in any given practical application.

Another important aspect of ScaLAPACK diagonalization is that the PZHEGVX routine permits the selection of a subset of eigenvalues and eigenvectors to be computed. In the DKS computation, it is only essential, in order to represent the density and carry out the SCF iterations, to compute the "occupied" positive-energy spinors. This clearly introduces great savings in eigenvector computation, both in time and memory, because the size of such subset is a small fraction (about 10% in our tests) of the total. In principle, of course, selecting a subset of eigensolutions to be computed should not affect their accuracy. However, in order to ensure this numerically, so as to reproduce the results obtained by the serial code to working precision, we imposed the requirement that the computed positive-energy eigenvectors be strictly orthogonal to the ones left out. Orthogonality to the negative-energy part of the spectrum is guaranteed in practice by the very nature of the spinors and the very large energy gap that separates them. To ensure orthogonality between the occupied and virtual spinors, we found it sufficient to include a small number of lowest-lying virtuals in the computed spectrum, for which orthogonalization is explicitly carried out (as explained in the previous paragraph). We conservatively set this number of extra spinors at 10% of the number of occupied ones. As stated above, in all cases, the choice of parameters and conditions for the parallel calculations described here ensured exact reproduction of the serial results to double-precision accuracy.

**D. Notes on the ScaLAPACK Grid Shape.** The shape of the processor grid arrangement presented to ScaLAPACK, for a given number of processors, may affect appreciably the performance of the routines. As suggested by the ScaLAPACK Users' Manual,[33] different routines are differently influenced in this regard. The performance dependence on the grid shape is, in turn, related to the characteristics of the physical interconnection network. While an exhaustive investigation of these aspects is outside the scope of the present work, we briefly explored the effect of the grid shape on the DKS steps of BERTHA which depend on ScaLAPACK. The results are summarized in Figure 2. This essentially reports, for some of the gold-cluster calculations discussed in depth in section IV, the relative performance of the diagonalization, level-shifting, and aggregate matrix-operation steps (including the small contribution of density-matrix evaluation), observed with three different arrangements of a 16-processor array and, in the inset, a 64-processor array. In the 16-processor case, we looked at the square 4 × 4 grid shape and at the two rectangular shapes, 2 × 8 and 8 × 2, for the clusters $Au_2$, $Au_4$, $Au_8$, and $Au_{16}$. In the 64-processor setup, we examined the 8 × 8, 2 × 32, and 32 × 2 arrangements in the case of $Au_{16}$. The data obtained, as the figure indicates, do not allow us to draw definitive conclusions, but it is clear that the processor grid shape does indeed affect the performance of the various ScaLAPACK steps in different ways, also in dependence on the total number of processors. Lacking a general a priori model to predict the optimal grid shape in any given case and architecture, performing preliminary test calculations may be an important part of the optimization process. Both the data for $P = 16$ and for $P = 64$ seem to indicate that rectangular grids with $P_r < P_c$ perform systematically better than the reverse arrangement for which $P_r > P_c$. The former is in fact also preferable over the square $P_r = P_c$ grid for the level-shifting phase. The data for $P = 16$ suggest, however, that in the diagonalization step the square arrangement is to be preferred for small systems, while both rectangular shapes tend to become more efficient as the size of the problem increases. As a result, a $P_r < P_c$ arrangement appears to be the globally optimal choice for large computational cases. The data for $P = 64$ and $Au_{16}$ in the inset, on the other hand, show that varying the number of processors may significantly alter the emerging pattern. In this case, for example, both rectangular grids appear again relatively unfavorable for the diagonalization step, so that the square arrangement is to be preferred overall. On the basis of the above results, we decided, for our further analysis, to use a square grid whenever possible and grids with $P_r < P_c$ otherwise.

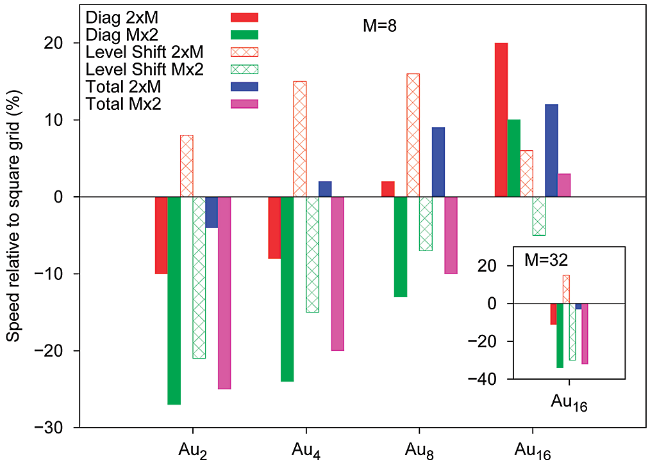


***Figure 2.*** Performance of DKS matrix operation steps with different ScaLAPACK processor grid shapes for a series of gold-cluster calculations discussed in the present work. Shown are the data for 16 processors ($M = 8$) and for 64 processors ($M = 32$, inset, for the sole $Au_{16}$ cluster). The histogram bars labeled "Total" refer to the sum of all three matrix operation steps, including "Density".

## IV. Discussion

We performed several computations for the gold clusters $Au_2$, $Au_4$, $Au_8$, $Au_{16}$, and $Au_{32}$. To achieve fully comparable results throughout, we used neither integral screening techniques nor molecular symmetry in the calculations. The large component of the *G*-spinor basis set on each gold atom (22s19p12d8f) is derived by decontracting the double-$\zeta$-quality Dyall basis set.[36] The corresponding small component basis was generated using the restricted kinetic balance relation.[11] The density functional used is the Becke 1988 exchange functional (B88)[37] plus the Lee−Yang−Parr (LYP) correlation functional[38] (BLYP). As auxiliary basis set, we use the HGTF basis called *B20*, optimized previously by us.[14]

***Table 2.*** CPU Times (s) for the Various Phases of DKS Calculations on Some Gold Clusters as a Function of the Number of Concurrent Processes Employed (Indicated by the ScaLAPACK Grid Shape)

| cluster | step | serial | 2 × 2 | 4 × 4 | 4 × 8 | 8 × 8 | 8 × 16 |
|---|---|---|---|---|---|---|---|
| $Au_2$ | **J** + **K** matrix | 24.75 | 8.63 | 2.43 | 1.69 | 0.90 | 0.90 |
| | diagonalization | 33.32 | 3.49 | 1.78 | 1.55 | 1.30 | 1.35 |
| | level shift | 10.20 | 3.59 | 1.14 | 0.71 | 0.58 | 0.51 |
| | density | 0.57 | 0.24 | 0.12 | 0.13 | 0.15 | 0.12 |
| | serial | 5.19 | 5.45 | 5.62 | 5.74 | 5.76 | 5.85 |
| | total iteration | 74.03 | 21.40 | 11.09 | 9.82 | 8.69 | 8.73 |
| $Au_4$ | **J** + **K** matrix | 183.50 | 62.15 | 13.82 | 7.55 | 5.04 | 3.27 |
| | diagonalization | 262.07 | 24.87 | 9.02 | 6.86 | 5.32 | 4.82 |
| | level shift | 83.87 | 27.50 | 8.46 | 4.13 | 2.66 | 1.52 |
| | density | 4.48 | 1.55 | 0.72 | 0.56 | 0.66 | 0.49 |
| | serial | 22.01 | 23.26 | 23.72 | 23.65 | 23.72 | 23.16 |
| | total iteration | 555.93 | 139.33 | 55.74 | 42.75 | 37.40 | 33.26 |
| $Au_8$ | **J** + **K** matrix | 1400.64 | 470.24 | 99.57 | 47.54 | 26.05 | 15.16 |
| | diagonalization | 1976.78 | 256.65 | 67.56 | 41.11 | 28.29 | 24.13 |
| | level shift | 679.52 | 214.64 | 63.44 | 32.16 | 18.88 | 9.56 |
| | density | 35.64 | 12.54 | 4.02 | 2.99 | 2.85 | 2.20 |
| | serial | 106.36 | 122.32 | 111.14 | 112.14 | 111.94 | 109.45 |
| | total iteration | 4198.94 | 1076.39 | 345.73 | 235.94 | 188.01 | 160.50 |
| $Au_{16}$ | **J** + **K** matrix | 10946.55 | 3655.87 | 752.26 | 364.35 | 182.52 | 95.11 |
| | diagonalization | 17463.86 | 2639.84 | 660.06 | 312.19 | 214.92 | 186.94 |
| | level shift | 5598.50 | 2254.54 | 477.62 | 238.30 | 133.56 | 71.84 |
| | density | 293.90 | 91.20 | 25.83 | 19.82 | 13.28 | 9.72 |
| | serial | 580.43 | 598.18 | 597.56 | 599.88 | 597.57 | 596.09 |
| | total iteration | 34883.24 | 9239.63 | 2513.33 | 1534.54 | 1141.85 | 959.70 |
| $Au_{32}$ | **J** + **K** matrix | | 28850.31 | 5848.33 | 2851.76 | 1413.57 | 708.57 |
| | diagonalization | | 28851.37 | 9950.42 | 5100.27 | 3197.00 | 1926.66 |
| | level shift | | 17406.51 | 5471.35 | 2677.06 | 1794.60 | 695.77 |
| | density | | 855.82 | 256.04 | 168.68 | 127.80 | 73.31 |
| | serial | | 3632.59 | 3684.21 | 3729.45 | 3692.64 | 3629.83 |
| | total iteration | | 79596.60 | 25210.35 | 14527.22 | 10225.61 | 7034.14 |

A numerical integration grid has been employed with 61 200 grid points for each gold atom. The five Au clusters chosen offer testing ground for a wide range of memory requirements and double-precision complex array handling conditions: the DKS matrix sizes were 1560 for $Au_2$ (37.1 MB), 3120 for $Au_4$ (148.5 MB), 6240 for $Au_8$ (594.1 MB), 12480 for $Au_{16}$ (2.3 GB), and 24 960 for $Au_{32}$ (9.3 GB).

Table 2 reports some elapsed times of the phases of the SCF iterations, for the various gold clusters and different numbers of processors employed (shown in their $P_r \times P_c$ ScaLAPACK arrangement). Note that the parallel diagonalization times are disproportionately smaller than, and not comparable with, the serial times because the latter involve the computation of the whole spectrum of eigensolutions, while in the parallel cases we could adopt the selection described in section III.C. It should further be noted that the $Au_{32}$ case was too demanding to be run on a single processor and that the times shown here are averages obtained from four SCF iterations.

Figure 3 shows the corresponding speedup for the various cases. Because of the remarks just made concerning the serial diagonalization and the $Au_{32}$ case, the speedup for the diagonalization step and for the entire $Au_{32}$ calculation could not be computed with reference to the serial calculation. In these cases, the figure shows the relative speedup with respect to the four-processor case. This appears to be a consistent and reliable procedure. Note, for example, that the total-iteration serial time for $Au_{32}$ estimated from the data of the 2 × 2 processor performance ($2.79 \times 10^5$ s) agrees within 0.1% with the estimate one obtains by fitting an $N^3$ power law to the serial times for the smaller clusters. In Figure 3, we see that, with some exceptions, the speedup generally tends to increase with the size of the system under study. In particular, the time for the construction of the DKS matrix scales extremely well for large systems, reaching 91% and 95% of the theoretical maximum with 128 processors for $Au_{16}$ and $Au_{32}$, respectively. It should be noted that the maximum speedup is one less than the number of processors in the array because of the master−slave approach used here. The other phases of the calculation scale less satisfactorily, reflecting the performance limitations of the underlying ScaLAPACK implementation. This is especially evident for the calculation of the density matrix, which is however a particularly undemanding task and a small fraction (<1%) of the whole workload. Using 32 and 64 processors, the performance of the diagonalization and level-shifting steps for the largest, $Au_{32}$, cluster is found to be particularly poor, appearing however to have promisingly ample room for improvement when more processors are involved. As already mentioned, the performance of the ScaLAPACK routines appears to be affected by the interplay of several contributing factors, which is not easy to unravel. Besides the array distribution block size and the shape of the processor grid, another crucial factor to consider is the global memory requirement per processor, which depends, besides application demands, on the ScaLAPACK implementation and requisites. When this exceeds the locally available memory, so that the system must resort to remote allocation through NUMAflex, or other devices in general, performance is likely to degrade, possibly quite significantly. By contrast, the diagonalization step for $Au_{16}$ and 32 processors shows a slightly superlinear performance, which is probably related

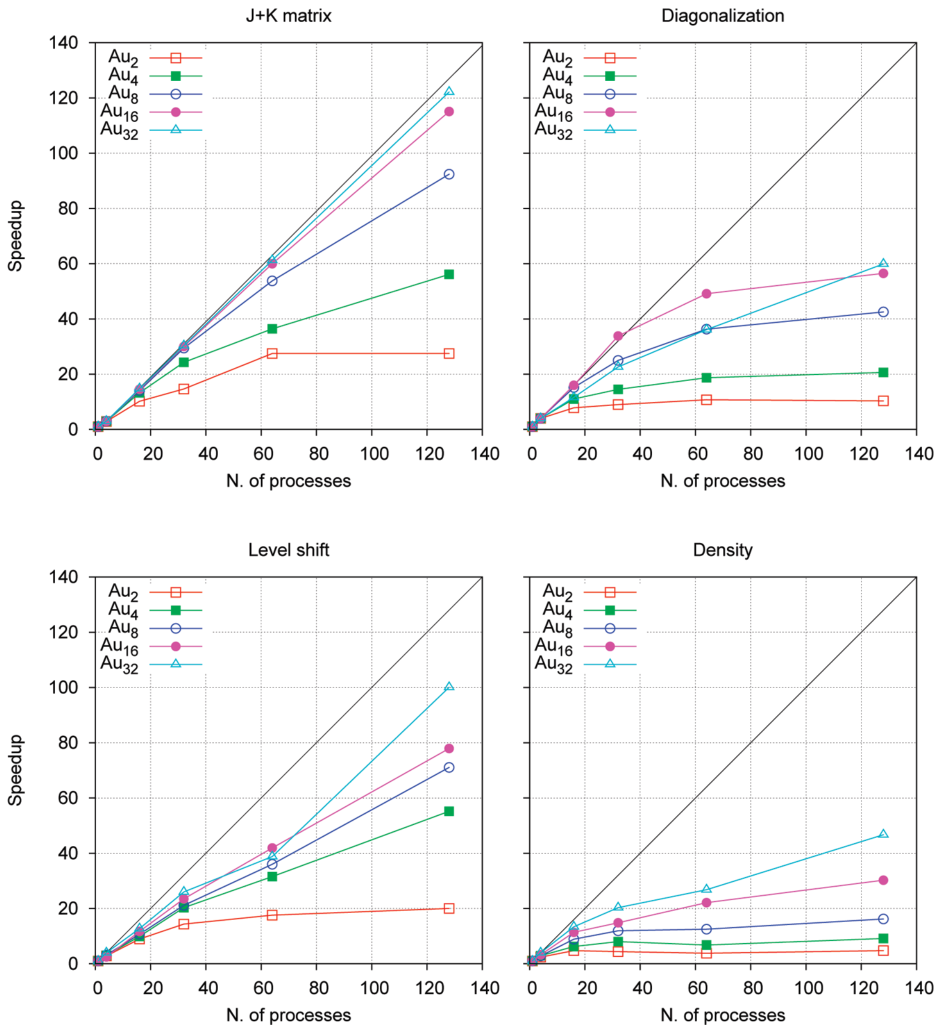


***Figure 3.*** Speedup of the DKS computation steps for the gold clusters studied in the present work.

to the fact that the 32-processor grid is the smallest one having a nonsquare shape (see Table 2 and section III.D).

Finally, Figure 4 displays a plot of the resulting global speedup for the SCF DKS iterations. The speedup for each gold cluster is shown as the band enclosed by the measured value on the lower limit and the theoretical maximum on the upper limit. The latter is computed by taking into account the unparallelized fraction ("Serial" phase) according to Amdahl's law.[39] As can be seen, the performance for the larger clusters appears to converge to more than 80% of the theoretical maximum on 128 processors, and it turns out to be about 60% of the limit value for an infinite number of processors, when the execution time reduces to that of the unparallelized portion. Considered together with the great step forward represented by the memory distribution scheme, we deem this to be a very satisfactory result, clearing the way to tackle previously unfeasible systems.

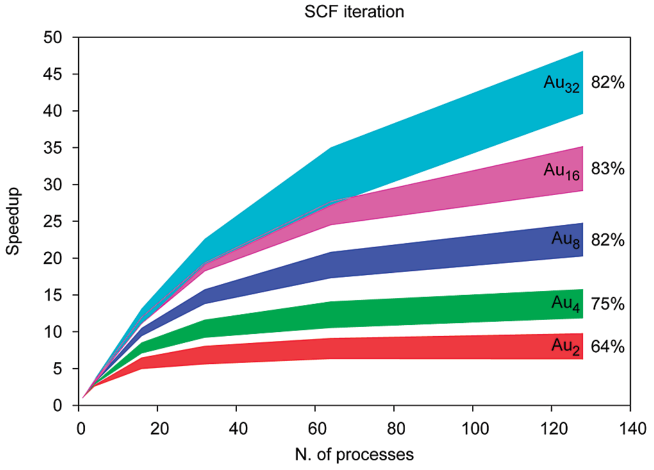


***Figure 4.*** Overall speedup for the DKS calculation of some gold clusters. The speedup for each cluster is shown as a band delimited, on the lower side, by the measured value and, on the upper side, by the theoretical maximum (Amdahl's law). The upper limit for $Au_{16}$ is evidenced as a line running over the $Au_{32}$ band in the region where the two bands overlap. The measured efficiency (percentage of the upper limit) with 128 processors is shown on the right side of each band.

## V. Application: Hg Atom Interacting with the $Au_{20}$ Cluster

Understanding the nature of the interactions involving mercury and, in particular, its interaction with gold clusters and surfaces is currently of great interest and is a formidable computational task (see for instance ref 40–46 and references therein). Here, we demonstrate the practicality and effectiveness of our new parallel implementation by applying it to the characterization of the interaction of the mercury atom with the $Au_{20}$ cluster. In addition to the determination of some spectroscopic parameters like equilibrium bond length ($R$) and dissociation energy ($D_e$), we also perform a detailed analysis of the modification of the full relativistic all-electron density of Hg and $Au_{20}$, which leads to the formation of the Hg−$Au_{20}$ bond.

**A. Description of the Calculations.** While the determination of the global equilibrium structure of gold clusters is itself a topic of great current interest (see for instance ref 47 and the references therein), the structure of the neutral $Au_{20}$ cluster appears to be well established, having been successfully identified in a gas phase vibrational spectroscopy experiment combined with quantum chemical calculations.[48] These confirmed that the neutral cluster retains the symmetric pyramidal geometry established for the anion.[49] We have performed a preliminary optimization of this $Au_{20}$ structure using the zero order relativistic approximation (ZORA) with small core and a QZ4P basis set, as implemented in the ADF package.[50−53] We then placed the Hg atom above the $Au_{20}$ pyramid vertex, which was found to be a preferred position for an interacting noble-gas atom,[48] and further fully optimized the whole adduct using the same method. Using our parallel DKS program, as described in the previous sections, we then further reoptimized the Hg−Au internuclear distance, keeping the Au-cluster structure fixed. The interaction energy was determined by the difference in total energy of Hg−$Au_{20}$ and the fragments Hg and $Au_{20}$.

The large component of the $G$-spinor basis set that we used was obtained by decontracting the Dyall basis set of triple-ζ quality (29s24p15d11f3g1h) on both gold and mercury atoms.[36] This is a larger basis set than that used in the $Au_2$−$Au_{32}$ test calculations. The corresponding small component basis was generated using the restricted kinetic balance relation.[11] This results in a DKS matrix of dimension 24 444 (about 8.9 GB double-precision complex numbers). As the basis set for density fitting we used the set B20 described elsewhere.[14,16] This comprises 307 Hermite Gaussian functions on each heavy atom. The density functional used is the Becke 1988 exchange functional (B88)[37] plus the Lee−Yang−Parr (LYP) correlation functional[38] (BLYP). All calculations were carried out with a total energy convergence threshold of $10^{-7}$ hartree.

The Au−Hg bond length was determined iteratively using a quadratic fit to the energy, requiring several DKS single points. The calculations have been performed on the SGI Altix 4700 described above, using 64 processors. The time required to complete a single SCF iteration was about 1.5 h, and each single-point DKS calculation required about 20 iterations to reach convergence.

**B. Results and Discussion.** The Hg−$Au_{20}$ bond length resulting from our calculations is 2.82 Å, and the corresponding interaction energy is 0.337 eV (32.5 kJ/mol). For comparison, the ZORA/QZ4P distance is 2.86 Å and the interaction energy is 27.5 kJ/mol. The complete optimized geometry of the $Au_{20}$ cluster and that of the $Au_{20}$−Hg adduct are available as Supporting Information (SI). These results clearly suggest the presence of a chemically relevant interaction. A first qualitative insight into the nature of this interaction is provided by the graphical representation of the electronic density difference between the complex and the noninteracting fragments placed at the same geometry. In Figure 5, we show a 3D contour plot of such a DKS/BLYP electron density difference. A surprising feature of this is its very rich and complex structure, reaching the most remote regions of the Au cluster far away from the interaction zone. The density accumulation is particularly pronounced in the Hg−Au internuclear region and in the zone between the first two gold layers (i.e., between the gold atom bound to Hg and the three neighbors below it). A significant density depletion zone is observed instead on the far side of the Hg atom, opposite the Hg−Au bond.

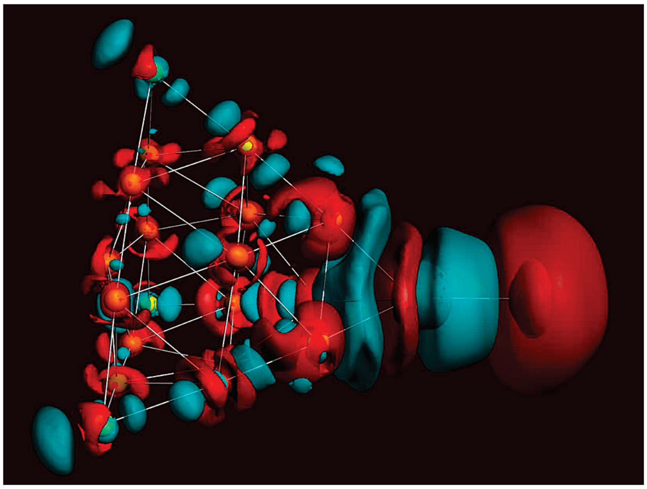


***Figure 5.*** DKS/BLYP contour plot of the electron density difference upon bond formation between the atom Hg and the $Au_{20}$ cluster. Red isodensity surfaces identify zones of density decrease, blue ones of density increase. The density value at the surfaces is $\pm 0.0001\, e/\mathrm{au}^3$.

It is common to investigate the changes in charge density that result from the binding of an adsorbate to a surface by computing the density difference along the binding direction $z$, $\Delta\varrho(z)$.[41] This is given by

$$\Delta\varrho(z) = \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y\, \Delta\varrho(x, y, z) \tag{6}$$

where $\Delta\varrho(x, y, z)$ is the electron density of the complex minus that of the two isolated fragments. $\Delta\rho(z)$ for the Hg−$Au_{20}$ system is shown in the top panel of Figure 6. Here, the $z$ axis is that passing through the positions of the Hg atom and the Au atom nearest to it. Positive values of the function denote accumulation of charge, while negative values indicate regions where the charge density is depleted. Inspection of this plot makes clearly more detailed the qualitative information obtained from Figure 5. Note in particular the marked density accumulation in the Hg−Au bond region and in the vicinity of the second gold layer, accompanied by density depletion around the Hg atom, especially on the far side, and to the left side of the nearest Au atom. This electron charge depletion at the Au site on the opposite side of a coordinating bond is an important feature of gold chemistry observed also in previous studies.[54,55] There are evident oscillations in the density difference around Hg and the nearest Au which may be put in relation with the structure of the electronic density of the atoms along $z$ already observed and discussed.

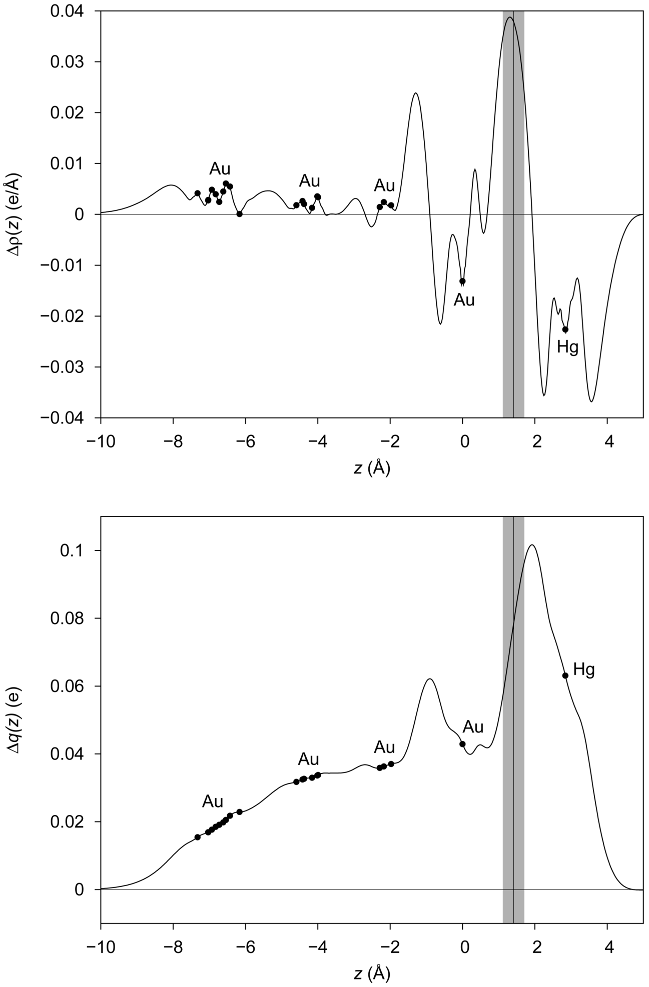


***Figure 6.*** Upper plot: DKS electron density change along the Hg−Au bond direction [eq 6] upon formation of the Hg−$Au_{20}$ system. The circles on the curve mark the projection of the position of the indicated atoms. The vertical gray strip marks a region of width equal to 20% of the Hg−Au distance centered about the $z$ position (vertical line) at which the densities of the noninteracting fragments cross (see the text for more details). Lower plot: Electron charge displacement function $\Delta q(z)$ [eq 7] for the same system.

A more immediately informative picture of the bonding may be obtained by a progressive integration of eq 6 along the internuclear axis, that is, by the function[54]

$$\Delta q(z) = \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y \int_{-\infty}^{z} \Delta\varrho(x, y, z')\, \mathrm{d}z' \qquad (7)$$

This measures the actual electron displacement taking place upon bond formation, that is, the amount of electron charge transferred into the integration region up to $z$ as one moves from left to right along the axis. In other words, $\Delta q(z)$ is the charge displaced from the right to the left side of the plane perpendicular to the axis in $z$. Thus, a negative value indicates a charge transfer (CT) of that magnitude from left to right, and similarly, the difference between two $\Delta q$ values gives the net electron influx into the region delimited by the corresponding planes.

The plot of $\Delta q(z)$ for Hg−$Au_{20}$ is shown in the lower panel of Figure 6. The most immediately eye-catching feature of the plot is that $\Delta q(z)$ is appreciably positive everywhere in the complex region. This means that there is a shift of charge from the Hg atom toward gold which does not stop at the nearest Au layers but, surprisingly, extends appreciably even beyond the fourth layer. The $\Delta q$ function shows two peaks, one corresponding to the already observed charge fluctuation between the first and second Au layers, and the second corresponding to the charge accumulation in the Hg−Au region followed by a decrease of about $0.1e$ around the Hg atom. If one defines an arbitrary boundary separating the Hg atom from the gold cluster, a corresponding effective CT value between the two fragments may be quantified. In Figure 6, we have shown one such plausible boundary, which has already been proposed in other cases,[54–56] corresponding to the point along $z$ where equal isodensity surfaces of the noninteracting fragments become mutually tangent. Remarkably, this point almost coincides with the peak of density accumulation (maximum of $\Delta\varrho(z)$ or maximum slope of $\Delta q$). The CT value at this point is $0.08e$, which is quite close to the value $0.09e$ recently proposed by Steckel[41] for a Hg atom interacting with a Au(001) surface. Interestingly, similar CT values were found in the comparably weak covalent bond between AuF and the heavier noble gases.[54] For those adducts, a detailed comparison was carried out between DKS/BLYP and four-component coupled-cluster results concerning a vast array of molecular properties. As noted above, in spite of the weakness of the Hg−Au interaction, this transferred charge is delocalized over a surprisingly large distance in the gold cluster. The $\Delta q$ curve is quite flat between the second and third gold layers, meaning that electrons do not accumulate here, so that nearly half of the total charge transferred is still found beyond the third gold layer. This insightful picture of the large spatial extent of the chemical interaction between Hg and a gold cluster represents a major novel result of the present investigation. The extent of charge delocalization found here may be overestimated by the BLYP functional self-interaction error (see for example refs 57 and 58 and references therein). It would indeed be of great interest to investigate exhaustively the behavior of different functionals with regard to the nature of similar chemisorption bonds. The computational advances documented in the present work will make such developments feasible and worthwhile in the rigorous four-component, all-electron framework. The all-electron character of the present approach is especially useful to bring to light changes in the electronic structure close to the nuclei which in general cannot be expected to be properly described by effective-core-potential or frozen-core methods. This is naturally of relevance for the calculation of properties which sensitively probe the electron density near heavy nuclei.

## VI. Conclusions and Outlook

In this work, we presented an effective, essentially complete, parallelization of a relativistic all-electron four-component Dirac−Kohn−Sham program called BERTHA. We used MPI for all data communication of the parallel routines, and the ScaLAPACK library to perform the most demanding matrix operations. The main aim of the parallelization scheme adopted was the effective reduction of memory requirements per processor through array distribution, enabling the handling of the DKS matrices arising in large-scale calculations on heavy-atom chemical systems using large basis sets. The scheme employs a master−slave model for the DKS matrix construction, relying on the automatic distributed memory allocation available on the SGI Altix architecture used. This procedure may be immediately transported to a generic conventional cluster provided a single node has enough memory to accommodate the arrays. It also lends itself straightforwardly, with minor changes, to explicit memory distribution by keeping track of array block location. In both cases, effective interprocess communication would even be reduced compared to the present implementation. The performance of the DKS matrix construction algorithm, including density fitting, was found to be excellent, reaching 95% of the linear limit for a large-basis $Au_{32}$ cluster on 128 processors.

The block-cyclic array decomposition required by ScaLAPACK was handled by explicitly written routines, both for the distribute and collect operations. The overall communication for the ScaLAPACK-dominated steps of the calculation amounts to one distribution of the relevant arrays and one final collection, per SCF iteration. Using our routines, communication time was conveniently found to be largely independent of the number of processors used, as ideally expected. The performance of the ScaLAPACK steps degrades, however, somewhat compared to the DKS construction step. It further depends appreciably on the complicated and not easily foreseeable interplay of several factors, including distribution blocksize, processor grid shape, number of processors, and memory requirement per processor. We undertook a preliminary investigation of the effect of some of these variables, which may be of interest also for other chemical applications. The global performance of the DKS calculation in large test cases was however quite satisfactory, reaching over 80% of the theoretical limit dictated by Amdahl's law.

As a first chemically significant application of the new all-electron DKS parallel code, we have studied the interaction of a Hg atom with a gold cluster of 20 atoms, using a triple-$\zeta$ basis set of 24 444 functions (a DKS matrix of about 8.9 GB). A detailed study of the electronic density modification caused by the interaction shows clearly that the bond exhibits a marked covalent character and that it is characterized by a significant charge transfer, of about 0.08 electron charges in magnitude, from the mercury atom to the gold cluster. We were able to demonstrate the rather unexpected and interesting feature that the charge transferred is significantly delocalized over the entire cluster, about half of it to be found as far away from the interaction site as the third and fourth gold atom layers.

**Acknowledgment.** This work has been carried out as part of an Extreme Computing Initiative project of the Distributed European Infrastructure for Supercomputing Applications (DEISA). We thank I. Girotto and G. Erbacci of CINECA and the staff of Leibniz Rechenzentrum (Garching, Germany) for continuous assistance, in particular, S. H. Leong, A. Block, and M. Allalen.

**Supporting Information Available:** Cartesian coordinates and contour plots of the elctron density difference at the Hg site. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Vignale, G.; Kohn, W. *Phys. Rev. Lett.* **1996**, *77*, 2037–2040.

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

(3) MacDonald, A. H.; Vosko, S. H. J. *Phys. C: Solid State* **1979**, *12*, 2977.

(4) Yanai, T.; Iikura, H.; Nakajima, T.; Ishikawa, Y.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 8267–8273.

(5) Saue, T.; Helgaker, T. *J. Comput. Chem.* **2002**, *23*, 814–823.

(6) Varga, S.; Engel, E.; Sepp, W.-D.; Fricke, B. *Phys. Rev. A* **1999**, *59*, 4288–4294.

(7) Liu, W.; Hong, G.; Dai, D.; Li, L.; Dolg, M. *Theor. Chem. Acc.* **1997**, *96*, 75–83.

(8) Quiney, H. M.; Belanzoni, P. *J. Chem. Phys.* **2002**, *117*, 5550–5563.

(9) Quiney, H. M.; Skaane, H.; Grant, I. P. J. *Phys. B: At. Mol. Opt.* **1997**, *30*, L829.

(10) Quiney, H. M.; Skaane, H.; Grant, I. P. Ab Initio Relativistic Quantum Chemistry: Four-Components Good, Two-Components Bad! In *Advanced Quantum Chemistry*; Lwdin, P.-O., Ed.; Academic Press: New York, 1998; Vol. 32, pp 1−49.

(11) Grant, I. P.; Quiney, H. M. *Phys. Rev. A* **2000**, *62*, 022508.

(12) Belpassi, L.; Storchi, L.; Tarantelli, F.; Sgamellotti, A.; Quiney, H. M. *Future Gener. Comp. Sy.* **2004**, *20*, 739–747.

(13) Belpassi, L.; Tarantelli, F.; Sgamellotti, A.; Quiney, H. M. *J. Chem. Phys.* **2005**, *122*, 184109.

(14) Belpassi, L.; Tarantelli, F.; Sgamellotti, A.; Quiney, H. M. *J. Chem. Phys.* **2006**, *124*, 124104.

(15) Belpassi, L.; Tarantelli, F.; Sgamellotti, A.; Quiney, H. M. *J. Chem. Phys.* **2008**, *128*, 124108.

(16) Belpassi, L.; Tarantelli, F.; Sgamellotti, A.; Quiney, H. M. *Phys. Rev. B* **2008**, *77*, 233403.

(17) Iliaš, M.; Saue, T. *J. Chem. Phys.* **2007**, *126*, 064102.

(18) Liu, W.; Peng, D. *J. Chem. Phys.* **2006**, *125*, 044102.

(19) Liu, W.; Peng, D. *J. Chem. Phys.* **2009**, *131*, 031104.

(20) Varga, S.; Fricke, B.; Nakamatsu, H.; Mukoyama, T.; Anton, J.; Geschke, D.; Heitmann, A.; Engel, E.; Bastug, T. *J. Chem. Phys.* **2000**, *112*, 3499–3506.

(21) Mintmire, J. W.; Dunlap, B. I. *Phys. Rev. A* **1982**, *25*, 88–95.

(22) Manby, F. R.; Knowles, P. J. *Phys. Rev. Lett.* **2001**, *87*, 163001.

(23) Geudtner, G.; Janetzko, F.; Köster, A. M.; Vela, A.; Calaminici, P. *J. Comput. Chem.* **2006**, *27*, 483–490.

(24) Inaba, T.; Sato, F. *J. Comput. Chem.* **2007**, *28*, 984–995.

(25) Challacombe, M.; Schwegler, E.; Almlöf, J. *J. Chem. Phys.* **1996**, *104*, 4685–4698.

(26) Ahmadi, G. R.; Almlöf, J. *Chem. Phys. Lett.* **1995**, *246*, 364–370.

(27) Laikov, D. N. *Chem. Phys. Lett.* **1997**, *281*, 151–156.

(28) Köster, A. M.; Reveles, J. U.; del Campo, J. M. *J. Chem. Phys.* **2004**, *121*, 3417–3424.

(29) Birkenheuer, U.; Gordienko, A. B.; Nasluzov, V. A.; Fuchs-Rohr, M. K.; Rösch, N. *Int. J. Quantum Chem.* **2005**, *102*, 743–761.

(30) Peng, D.; Liu, W.; Xiao, Y.; Cheng, L. *J. Chem. Phys.* **2007**, *127*, 104106.

(31) Silicon Graphics, SGI NUMAlink, White Paper 3771, 2005.

(32) MPI: A Message-Passing Interface Standard, version 2.2. *Message Passing Interface Forum*; University of Tennessee: Knoxville, TN, 2009. http://www.mpi-forum.org (accessed Jan 2010).

(33) Blackford, L. S.; Choi, J.; Cleary, A.; D'Azevedo, E.; Demmel, J.; Dhillon, I.; Dongarra, J.; Hammarling, S.; Henry, G.; Petitet, A.; Stanley, K.; Walker, D.; Whaley, R. C. *ScaLAPACK Users' Guide*; Society for Industrial and Applied Mathematics: Philadelphia, PA, 1997.

(34) Silicon Graphics, Powering the Real-time Enterprise, White Paper 3935, 2006.

(35) Choi, J.; Demmel, J.; Dhillon, I.; Dongarra, J.; Ostrouchov, S.; Petitet, A.; Stanley, K.; Walker, D.; Whaley, R. C. *Comput. Phys. Commun.* **1996**, *97*, 1–15.

(36) Dyall, K. G. *Theor. Chem. Acc.* **2004**, *112*, 403–409.

(37) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

(38) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

(39) Amdahl, G. Validity of the Single Processor Approach to Achieving Large-Scale Computing Capabilities. *AFIPS Conference Proceedings*; Thompson Books: Washington, DC, 1967; Vol. 30, pp 483−485.

(40) Zaleski-Ejgierd, P.; Pyykko, P. *J. Phys. Chem. A* **2009**, *113*, 12380–12385.

(41) Steckel, J. A. *Phys. Rev. B* **2008**, *77*, 115412.

(42) Rykova, E. A.; Zaitsevskii, A.; Mosyagin, N. S.; Isaev, T. A.; Titov, A. V. *J. Chem. Phys.* **2006**, *125*, 241102.

(43) Sarpe-Tudoran, C.; Fricke, B.; Anton, J.; Persina, V. *J. Chem. Phys.* **2007**, *126*, 174702.

(44) Munro, L. J.; Johnson, J. K.; Jordan, K. D. *J. Chem. Phys.* **2001**, *114*, 5545–5551.

(45) Gaston, N.; Schwerdtfeger, P.; Saue, T.; Greif, J. *J. Chem. Phys.* **2006**, *124*, 044304.

(46) Gaston, N.; Paulus, B.; Rosciszewski, K.; Schwerdtfeger, P.; Stoll, H. *Phys. Rev. B* **2006**, *74*, 094102.

(47) Assadollahzadeh, B.; Schwerdtfeger, P. *J. Chem. Phys.* **2009**, *131*, 064306.

(48) Gruene, P.; Rayner, D. M.; Redlich, B.; van der Meer, A. F. G.; Lyon, J. T.; Meijer, G.; Fielicke, A. *Science* **2008**, *321*, 674–676.

(49) Li, J.; Li, X.; Zhai, H.-J.; Wang, L.-S. *Science* **2003**, *299*, 864–867.

(50) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931–967.

(51) Fonseca Guerra, C.; Snijders, J. G.; te Velde, G.; Baerends, E. J. *Theor. Chem. Acc.* **1998**, *99*, 391–403.

(52) *ADF User's Guide, Release 2008.1*; SCM, Theoretical Chemistry, Vrije Universiteit: Amsterdam, The Netherlands, 2008. http://www.scm.com (accessed Jan 2010).

(53) van Leeuwen, R.; van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *J. Chem. Phys.* **1994**, *101*, 1272–1281.

(54) Belpassi, L.; Infante, I.; Tarantelli, F.; Visscher, L. *J. Am. Chem. Soc.* **2008**, *130*, 1048–1060.

(55) Salvi, N.; Belpassi, L.; Tarantelli, F. To be published.

(56) Belpassi, L.; Tarantelli, F.; Pirani, F.; Candori, P.; Cappelletti, D. *Phys. Chem. Chem. Phys.* **2009**, *11*, 9970–9975.

(57) Cohen, A. J.; Mori-Sanchez, P.; Yang, W. *Science* **2008**, *321*, 792–794.

(58) Jiménez-Hoyos, C. A.; Janesko, B. G.; Scuseria, G. E. *J. Phys. Chem. A* **2009**, *113*, 11742–11749.

CT900539M

JCTC Journal of Chemical Theory and Computation

# Integration Grid Errors for Meta-GGA-Predicted Reaction Energies: Origin of Grid Errors for the M06 Suite of Functionals

Steven E. Wheeler* and K. N. Houk

*Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095*

Received November 30, 2009

**Abstract:** We have assessed integration grid errors arising from the use of popular DFT quadrature schemes for a set of 34 organic reaction energies. The focus is primarily on M05-2X and the M06 suite of functionals (M06-L, M06, M06-2X, and M06-HF). M05-2X, M06, and M06-2X outperform popular older DFT functionals for the reaction energies studied and offer accuracies comparable to results from perturbative hybrid DFT functionals. However, these new functionals are more sensitive to the choice of quadrature grid than previous generations of DFT functionals. Errors in predicted reaction energies arising from the use of the popular SG-1 integration grid, which is the default in the Q-Chem package, are significant. In particular, M06-HF reaction energies computed with the SG-1 grid exhibit errors ranging from −6.7 to 3.2 kcal mol$^{-1}$, relative to results computed with a very fine integration grid. This grid sensitivity is not a general problem for meta-generalized gradient approximation functionals, but is instead due to the specific functional forms used in these functionals. The large grid errors are traced to the kinetic energy density enhancement factor utilized in the exchange component of the M05-2X and the M06 functionals. This term contains empirically adjusted parameters that are of large magnitude for all of these functionals and for M06-HF in particular. The product of these large constants and modest integration errors for the kinetic energy density results in very large errors in individual contributions to the exchange energy. This gives rise to the large errors in reaction energies exhibited by these functionals for certain integration grids.

## I. Introduction

Kohn−Sham density functional theory (DFT) has emerged as the preeminent choice for the computational study of organic reactions. The popularity of DFT over traditional ab initio methods in this context stems from a number of factors, including favorable scaling with system size combined with relatively high-accuracy, widespread availability of analytic first and second energy derivatives for efficient geometry optimizations and vibrational frequency computations, and efficient implementations in popular electronic structure theory packages. Because the necessary integrals over exchange−correlation functionals cannot be evaluated in closed form, Kohn−Sham DFT computations typically rely on numerical quadrature schemes. In most quantum chemistry programs, these integrals are approximated as a sum of contributions from atom-centered grids:

$$\int F(\mathbf{r})d\mathbf{r} \approx \sum_{A}^{\text{atoms}} \sum_{g}^{\text{grid}} w_g p_A(\mathbf{r}_g) F(\mathbf{r}_g) \qquad (1)$$

where $w_g$ is the quadrature weight at the corresponding grid point $\mathbf{r}_g$, and the atomic partitioning function, $p_A(\mathbf{r}_g)$, is defined such that $\sum_A^{\text{atoms}} p_A(\mathbf{r}_g) = 1$ at each point in space.

Various quadrature methods and atomic partitioning functions have been devised, several of which enjoy widespread use.[1−5] It has long been known (though not always appreciated!) that the choice of integration grid can significantly affect computed molecular properties.[6−9] Martin, Bauschlicher, and Ricca[6] studied the grid sensitivity of B3LYP-

* Corresponding author. E-mail: swheele2@chem.ucla.edu.

10.1021/ct900639j © 2010 American Chemical Society
Published on Web 01/15/2010

computed molecular properties and highlighted problems with popular grids for several third-row transition-metal systems and for simple hydrocarbon radicals. More recently, Papas and Schaefer[7] compared BLYP and B3LYP energies from the Gaussian, Molpro, NWChem, Q-Chem, and GAMESS packages, using default and finer integration grids, to assess the precision that can be expected from the integration schemes in these packages. Dressler and Thiel[8] analyzed the effect of integration grids on DFT-computed anharmonic force fields, and Termath and Sauer[9] examined their effect on DFT-based direct molecular dynamics simulations.

Some new meta-GGAs (meta-generalized gradient approximation), which explicitly depend on the kinetic energy density, offer significant advantages over previous generations of functionals.[10,11] A shortcoming of these functionals that is often neglected, however, is the increased sensitivity to the choice of integration grid. By default, popular electronic structure programs utilize quadrature grids that were developed and refined based on previous generations of DFT functionals. Unfortunately, integration grids that proved adequate for these older functionals can lead to significant errors when utilized with some new meta-GGAs.[12−17]

In 2004, Johnson et al.[12] demonstrated that potential energy curves for dispersion-bound complexes computed with VS98[18] and other meta-GGA functionals are prone to spurious oscillations unless very large integration grids are used. Gräfenstein and Cremer subsequently proposed[17] the use of locally augmented radial integration grids to combat these issues in a cost-effective way. In 2009, Johnson and co-workers[13] revisited these grid errors and showed that the grid sensitivity originates from singularities near the intermonomer midpoint in kinetic energy density-dependent functional forms present in many meta-GGAs. Similarly, Gräfenstein, Izotov, and Cremer[14] attributed irregularities in certain meta-GGA-predicted energies for stretched covalent bonds to singularities in the self-interaction correction term present in the correlation part of these functionals.

The oscillations in meta-GGA-computed interaction potentials for dispersion-bound complexes[12] were addressed in the design of the M06 suite of functionals through the elimination of problematic terms in the VS98 functional, on which the M06 functionals are in part based.[10] The resulting changes in the M06 functionals offer improved performance in this regard, although with some popular grids problems still arise.[13,15] For example, Merz and co-workers[15] recently analyzed the performance of a number of methods for the prediction of potential energy curves for model noncovalent interactions. The M06-2X and M06-L DFT functionals outperformed the other methods tested but yielded discontinuous energy curves when used with the popular SG-1 integration grid.[16]

Grid issues with these functionals are not limited to dispersion-bound complexes. For example, Csonka and co-workers[19] benchmarked a variety of DFT functionals for the prediction of geometries and conformational energies in a series of saccharides. Although M05-2X yielded the most reliable results of the tested methods when a dense integration grid was used, results computed using the default integration grid in the Jaguar program package lead to larger errors in energies and problems in geometry optimizations. Similarly, Scuseria and co-workers[20] recently reported geometry optimization convergence problems and spurious imaginary frequencies when pairing functionals from the M06 suite with various popular integration grids.

In the present work, we focus on the grid requirements for meta-GGA-predicted energies for a set of 34 organic isomerization reactions recently published by Grimme and co-workers.[21] This set, shown in Scheme 1, constitutes a diverse collection of reactions that are small enough for the application of accurate benchmark computations yet representative of the diverse changes in bonding that occur in organic reactions. Reference “experimental reaction energies” were derived from standard enthalpies of formation corrected for zero point vibrational and thermal effects by Grimme.[21] All of the reactions are endothermic as written. Our primary focus is on grid errors for M05-2X and the M06 family of functionals,[10] which have emerged as promising new functionals for diverse chemical applications.[22]

## II. Theoretical Methods

Single point energies were computed for the molecular species in Scheme 1 using five meta-GGA DFT functionals paired with various DFT integration grids. The TZV(2df,2pd) basis set[23,24] was used for all computations, which were executed at B3LYP/TZV(p,d) optimized geometries taken from ref 21. The TZV(2df,2pd) basis set comprises the Alrichs TZV triple-ζ quality basis set[23] plus polarization functions from the cc-pVTZ basis set.[24] Grid errors for other popular basis sets are expected to be similar. The meta-GGAs tested are VS98,[18] M05-2X,[25] M06-L,[26] M06-HF,[27] M06,[10] and M06-2X.[10] For comparison, grid errors for B3LYP,[28] PBE,[29,30] and TPSS[31] are also presented.

The atom-centered grids utilized in popular DFT codes are constructed as a direct product of sets of $N^r$ radial and $N^\Omega$ angular grid points:

$$\sum_{g}^{\text{grid}} w_g F(\mathbf{r}_g) = \sum_{i=1}^{N^r} w_i^r \sum_{i=1}^{N^\Omega} w_j^\Omega F(\mathbf{r}_i, \theta_j, \phi_j) \qquad (2)$$

Defining an integration grid requires a choice of atomic partitioning function and the number, weights, and distribution of radial and angular grid points.

Four popular integration grids were tested (see Table 1). Grids labeled Q-Chem, NWChem, and Gaussian are equivalent to the default grids in those packages.[32−35] The default Q-Chem grid is the popular SG-1 grid of Gill, Johnson, and Pople.[16] All of the tested grids rely on Lebedev’s angular quadrature.[1] The radial components of these grids are from either an Euler−Maclaurin quadrature with the coordinate transformation of Murray, Handy, and Laming (Euler)[2] or a modification of the Murray−Handy−Laming scheme published by Mura and Knowles (MK).[3] For the atomic partitioning functions, the tested grids use the scheme of Becke,[5] the modification introduced by Stratmann, Scuseria, and Frisch (SSF),[4] or an unpublished modification of the SSF scheme (Erf1) implemented in NWChem,[33,34] in which

***Scheme 1***

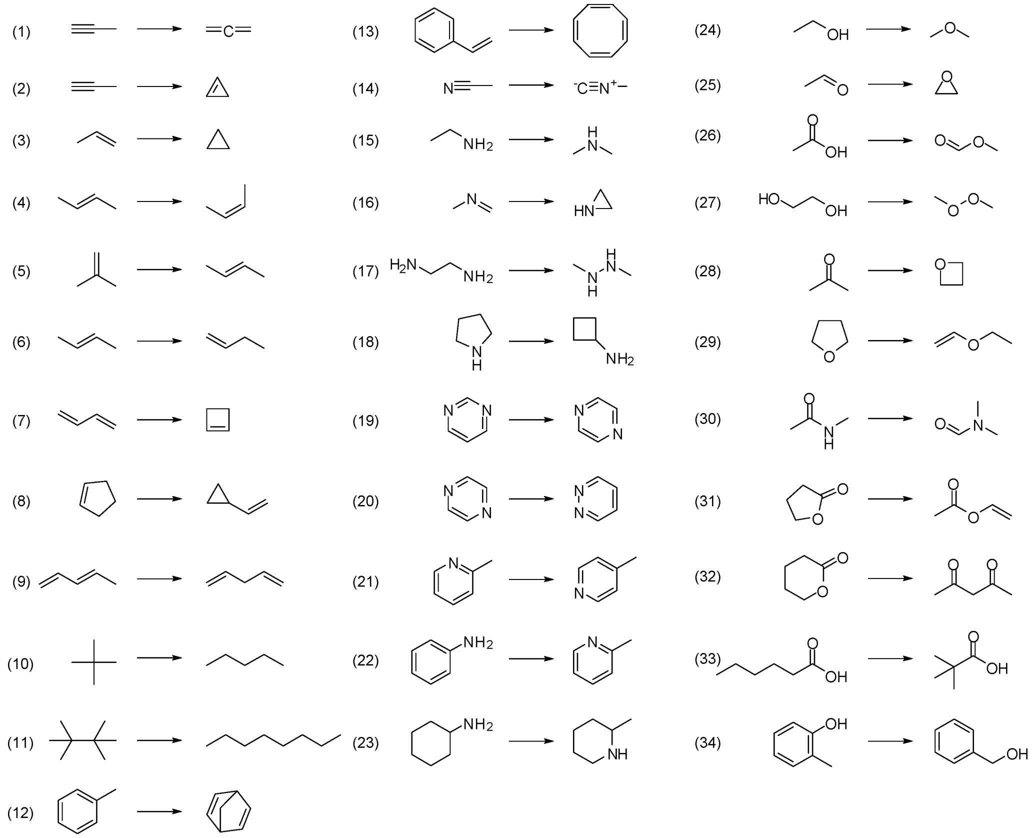


***Table 1.*** Partitioning Function, Radial Quadrature Method, and Number of Radial and Angular Points for Tested Grids[a]

| | partitioning | radial quadrature | radial points | angular points |
|---|---|---|---|---|
| Q-Chem | Becke | Euler−Maclaurin | 50 | 194 |
| NWChem | Erf1 | Mura−Knowles | 49 | 434 |
| Gaussian03 | SSF | Euler−Maclaurin | 75 | 302 |
| Fine | Erf1 | Mura−Knowles | 70 | 590 |
| Xfine | Erf1 | Mura−Knowles | 100 | 1 202 |

[a] All grids are pruned and utilize Lebedev's angular quadrature.[1]

the partition function weights are written in terms of products of error functions.

Most program packages utilize automatically pruned integration grids, in which the number of angular points is dependent on the radial coordinate.[16] This leads to significant reductions in computational cost with minimal loss in accuracy compared to the use of the unpruned grid. The effect of pruning was tested for selected reaction energies. Errors arising from pruning were less than 0.1 kcal mol$^{-1}$ across all grids considered and are considered inconsequential. Presented results are based on pruned grids.

Energies computed using the NWChem "Xfine" grid (see Table 1) were used as a benchmark, i.e., $\Delta E_{\text{grid}}^{\text{error}} = \Delta E_{\text{grid}} - \Delta E_{\text{Xfine}}$. To confirm that reaction energies computed using this grid are converged with respect to integration grid density, the M06-2X reaction energies, which are the most sensitive to the choice of grid, were also computed using a very large grid with 300 radial points and 1202 angular points. The mean absolute deviation in reaction energies between this very large integration grid and the Xfine grid is only 0.0003 kcal mol$^{-1}$; the maximum deviation is 0.003 kcal mol$^{-1}$. With these meta-GGA functionals, use of default convergence criteria sometimes lead to convergence problems and resulted in erratic and seemingly irreproducible grid errors. This was particularly true for the less dense grids and the M06-HF functional. Care was taken to ensure that all energies are converged to the correct Kohn−Sham solution to a precision of at least $10^{-10}$ au. Converging DFT energies to this precision is often hampered by the screening thresholds for integrals and the electron density employed by popular DFT programs. In this work, screening thresholds of at least $10^{-14}$ au were used.[36] Much larger errors than reported here could be encountered with these functionals if care is not taken to tightly converge results.

All computations were carried out using a locally modified version of NWChem 5.1.[33,34]

## III. Results and Discussion

The present work is primarily concerned with grid errors in computed reaction energies, not the error in the reaction energies themselves or the grid errors in absolute energies. However, since the accuracy of most of the tested functionals has not been previously assessed for the reactions in Scheme 1, a brief analysis of the reaction energy errors compared to experiment is presented first. This allows the subsequent analysis of grid errors to be put into the perspective of the inherent error in the DFT computations.

***Table 2.*** Experimental Reaction Energies (Δ*E*, from ref 21) and Errors in Predicted Energy for the Reactions in Scheme 1, Relative to Experiment[a]

| | Δ*E* | B3LYP | PBE | TPSS | VS98 | M05-2X | M06-L | M06 | M06-2X | BMK | mPW2-PLYP | B2-PLYP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.6 | −3.6 | −4.7 | −4.5 | −4.7 | −2.1 | −2.1 | −0.6 | −1.1 | −2.7 | −0.9 | −0.9 |
| 2 | 21.9 | 1.9 | −2.7 | −2.4 | 0.2 | −0.8 | −3.3 | −2.6 | −1.7 | 0.8 | 2.5 | 2.5 |
| 3 | 7.2 | 1.8 | −1.9 | −1.5 | 1.4 | −2.0 | −4.4 | −4.3 | −3.4 | −1.1 | 0.5 | 0.6 |
| 4 | 1.0 | 0.3 | 0.1 | 0.1 | −0.7 | 0.2 | −0.3 | 0.0 | 0.1 | 0.1 | 0.2 | 0.2 |
| 5 | 0.9 | −0.6 | −0.5 | −0.7 | 1.2 | 0.3 | −0.4 | 0.0 | 0.1 | −0.2 | −0.1 | −0.2 |
| 6 | 2.6 | 0.6 | 1.0 | 0.7 | 1.2 | 0.0 | 1.4 | 0.9 | 0.3 | 0.3 | 0.2 | 0.3 |
| 7 | 11.1 | 4.2 | −0.7 | 0.5 | 4.9 | 1.3 | −1.5 | −1.2 | −0.4 | −4.7 | 1.9 | 2.1 |
| 8 | 22.9 | −3.0 | −2.8 | −4.0 | −5.2 | −2.2 | −6.1 | −4.9 | −4.3 | 0.6 | −1.6 | −1.8 |
| 9 | 6.9 | 1.3 | 2.1 | 1.9 | 2.7 | −0.1 | 2.7 | 1.3 | 0.0 | 0.4 | 0.6 | 0.7 |
| 10 | 3.6 | −2.8 | −2.1 | −2.8 | 6.6 | 0.2 | −1.5 | 0.3 | −0.2 | −1.2 | −1.1 | −1.2 |
| 11 | 1.9 | −9.8 | −7.2 | −8.2 | 29.3 | −0.7 | −2.0 | 0.4 | −0.8 | −3.9 | −4.6 | −5.0 |
| 12 | 46.9 | 10.2 | 4.0 | 4.5 | 3.8 | 4.0 | 7.3 | 3.3 | 3.3 | −0.4 | 6.1 | 6.0 |
| 13 | 36.0 | 3.2 | 3.1 | 3.9 | 5.3 | 3.9 | 3.9 | 2.1 | 2.0 | 3.4 | 3.7 | 3.7 |
| 14 | 24.2 | −0.7 | 0.8 | −0.9 | 1.5 | −1.1 | 0.9 | −1.2 | −2.2 | −4.1 | 0.3 | 0.6 |
| 15 | 7.3 | 0.0 | 0.3 | −1.2 | 0.8 | 0.4 | −0.1 | −0.2 | −0.1 | 0.2 | 0.4 | 0.5 |
| 16 | 10.8 | 1.6 | −2.3 | −1.0 | 1.2 | −2.3 | −4.6 | −4.4 | −3.6 | −1.1 | 0.4 | 0.5 |
| 17 | 27.0 | −1.4 | −1.5 | −4.1 | 2.7 | 1.0 | −0.6 | −0.1 | −0.1 | 0.2 | 0.2 | 0.1 |
| 18 | 11.2 | 0.3 | −0.6 | 0.8 | −1.4 | 1.7 | −1.2 | 0.0 | 0.7 | −1.4 | 0.8 | 0.6 |
| 19 | 4.6 | −0.5 | −0.8 | −0.7 | −0.6 | 0.4 | −0.6 | −0.2 | 0.0 | −0.2 | −0.2 | −0.3 |
| 20 | 20.2 | −1.8 | −3.1 | −2.9 | −2.6 | −1.3 | −1.9 | −0.7 | −0.9 | 0.4 | −1.6 | −1.9 |
| 21 | 0.9 | 0.3 | 0.2 | 0.1 | 0.5 | 0.2 | 0.3 | 0.4 | 0.2 | 0.3 | 0.2 | 0.2 |
| 22 | 3.2 | 0.4 | 1.2 | −1.6 | 1.8 | 1.6 | −0.8 | 0.2 | 0.7 | −0.9 | 0.1 | 0.0 |
| 23 | 5.3 | −0.7 | −0.5 | −1.9 | 0.3 | −0.1 | −0.5 | −0.5 | −0.6 | −0.5 | −0.2 | −0.3 |
| 24 | 12.5 | −1.9 | −1.1 | −4.3 | −1.4 | −1.1 | −2.4 | −2.0 | −1.9 | −2.0 | −1.3 | −1.3 |
| 25 | 26.5 | 1.5 | −1.8 | −3.6 | 1.9 | −2.9 | −2.6 | −1.9 | −3.7 | 0.4 | 1.1 | 1.1 |
| 26 | 18.2 | −2.2 | −1.4 | −5.3 | −1.9 | −1.6 | −2.1 | −1.4 | −2.1 | −2.2 | −1.7 | −1.8 |
| 27 | 64.2 | −3.5 | −5.5 | −11.2 | −1.1 | 4.5 | −4.3 | 2.3 | 2.0 | 0.8 | 0.0 | −0.9 |
| 28 | 31.2 | 2.2 | −0.2 | −2.2 | 6.6 | −1.0 | 1.6 | 0.5 | −1.4 | −1.5 | 1.5 | 1.7 |
| 29 | 11.9 | −3.0 | −0.4 | −1.2 | −9.6 | 1.3 | −1.9 | −0.1 | 0.0 | 1.7 | −0.5 | −0.9 |
| 30 | 9.5 | 0.1 | −0.5 | −2.1 | 0.0 | 0.3 | −0.2 | −0.2 | −0.3 | 0.1 | 0.1 | 0.0 |
| 31 | 14.0 | −3.0 | 0.6 | −0.8 | −5.7 | 1.3 | −0.6 | 0.6 | 0.2 | 2.0 | −0.3 | −0.7 |
| 32 | 7.1 | −3.7 | −1.5 | −1.4 | −10.7 | 2.4 | −3.9 | −1.4 | 1.2 | −0.6 | −1.6 | −2.3 |
| 33 | 5.6 | 4.6 | 5.0 | 1.7 | −4.2 | 2.3 | 2.7 | 2.7 | 2.3 | 2.7 | 3.5 | 3.6 |
| 34 | 7.3 | −0.4 | 1.4 | 0.2 | 3.0 | −0.5 | 2.4 | 1.2 | 0.1 | 0.5 | −0.4 | −0.4 |
| MAD | | 2.3 | 1.9 | 2.5 | 3.7 | 1.4 | 2.2 | 1.3 | 1.2 | 1.3 | 1.2 | 1.3 |
| MSD | | −0.2 | −0.7 | −1.6 | 0.8 | 0.2 | −0.8 | −0.3 | −0.5 | −0.4 | 0.2 | 0.2 |
| min | | −9.8 | −7.2 | −11.2 | −10.7 | −2.9 | −6.1 | −4.9 | −4.3 | −4.7 | −4.6 | −5.0 |
| max | | 10.2 | 5.0 | 4.5 | 29.3 | 4.5 | 7.3 | 3.3 | 3.3 | 3.4 | 6.1 | 6.0 |

[a] Mean absolute deviations (MAD), mean signed deviations (MSD), and minimum and maximum deviations (min and max, respectively) are also provided for each functional. The Xfine grid and TZV(2df,2pd) basis set were used. BMK, mPW2-PLYP, and B2-PLYP results are from ref 21. All values are in kcal mol$^{-1}$.

**A. Performance of Meta-GGA Functionals for Reaction Energies.** Errors in DFT-computed reaction energies, relative to experimental results,[21] are shown in Table 2. These values were computed using the Xfine integration grid. Mean signed deviations (MSD), mean absolute deviations (MAD), and the range of deviations from experiment are plotted in Figure 1. Of the functionals tested here, the M05-2X, M06, and M06-2X functionals offer the best performance, although even these functionals exhibit errors approaching ±5 kcal mol$^{-1}$ for selected reactions. The mean deviations for the M05-2X, M06, and M06-2X functionals are comparable to those from mPW2-PLYP,[37] B2-PLYP,[38] and BMK[39] for this test set.[21] The B3LYP, PBE, TPSS, M06-L, and M06-HF functionals perform poorly and are all comparable. VS98 performs the worst for these reactions, delivering the largest mean error and a very large error for reaction 11.

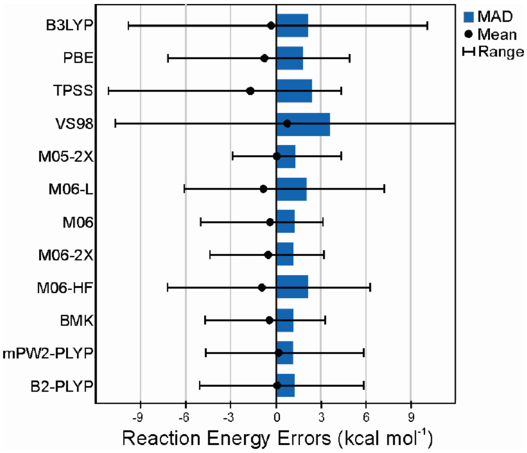


***Figure 1.*** MAD, MSD, and range of deviations from experiment for computed energies of the reactions in Scheme 1 (kcal mol$^{-1}$).

**B. Errors with Popular Integration Grids.** Integration grid errors have been assessed for the 34 organic isomerizations in Scheme 1. The grid errors for four popular quadrature grids are summarized in Table 3. Details for each reaction are provided in the Supporting Information. Normalized error distributions for the B3LYP, PBE, and TPSS functionals are shown in Figure 2. Analogous plots for VS98, M05-2X, and the four M06 functionals are shown in

***Table 3.*** Analysis of Grid Errors for the Energies of the Reactions in Scheme 1 for Four Popular DFT Integration Grids[a]

| | Q-Chem | NWChem | Gaussian | fine |
|---|---|---|---|---|
| B3LYP | | | | |
| MAD | 0.08 | 0.00 | 0.01 | 0.00 |
| MSD | 0.04 | 0.00 | 0.00 | 0.00 |
| min | −0.17 | −0.02 | −0.02 | −0.01 |
| max | 0.69 | 0.04 | 0.08 | 0.01 |
| PBE | | | | |
| MAD | 0.10 | 0.00 | 0.01 | 0.00 |
| MSD | 0.05 | 0.00 | 0.00 | 0.00 |
| min | −0.21 | −0.02 | −0.03 | −0.01 |
| max | 0.85 | 0.05 | 0.10 | 0.01 |
| TPSS | | | | |
| MAD | 0.09 | 0.01 | 0.01 | 0.00 |
| MSD | 0.05 | 0.00 | 0.00 | 0.00 |
| min | −0.19 | −0.03 | −0.03 | −0.01 |
| max | 0.85 | 0.05 | 0.09 | 0.01 |
| VS98 | | | | |
| MAD | 0.25 | 0.03 | 0.04 | 0.01 |
| MSD | 0.13 | 0.00 | 0.00 | 0.00 |
| min | −0.38 | −0.13 | −0.22 | −0.07 |
| max | 2.80 | 0.21 | 0.37 | 0.05 |
| M05-2X | | | | |
| MAD | 0.16 | 0.05 | 0.01 | 0.01 |
| MSD | 0.02 | −0.02 | 0.01 | 0.00 |
| min | −0.39 | −0.45 | −0.02 | −0.03 |
| max | 0.52 | 0.10 | 0.10 | 0.02 |
| M06-L | | | | |
| MAD | 0.20 | 0.05 | 0.02 | 0.01 |
| MSD | 0.09 | 0.03 | 0.00 | 0.00 |
| min | −0.84 | −0.07 | −0.10 | −0.02 |
| max | 0.76 | 0.65 | 0.11 | 0.03 |
| M06 | | | | |
| MAD | 0.25 | 0.04 | 0.02 | 0.00 |
| MSD | 0.06 | 0.02 | 0.01 | 0.00 |
| min | −1.04 | −0.10 | −0.08 | −0.01 |
| max | 0.75 | 0.31 | 0.12 | 0.02 |
| M06-2X | | | | |
| MAD | 0.29 | 0.08 | 0.02 | 0.01 |
| MSD | 0.03 | −0.02 | 0.00 | 0.00 |
| min | −2.02 | −0.39 | −0.09 | −0.02 |
| max | 0.81 | 0.23 | 0.11 | 0.02 |
| M06-HF | | | | |
| MAD | 1.20 | 0.20 | 0.03 | 0.02 |
| MSD | 0.05 | −0.07 | 0.01 | 0.00 |
| min | −6.70 | −1.51 | −0.12 | −0.05 |
| max | 3.21 | 0.42 | 0.13 | 0.06 |

[a] Grid errors are relative to Xfine results. All values are in kcal mol$^{-1}$.

Figure 3. For B3LYP, PBE and TPSS, grid errors arising from the use of the NWChem, Gaussian, and fine grids are negligible, never exceeding ±0.1 kcal mol$^{-1}$. The errors are slightly larger for the Q-Chem grid, and there is one outlier (reaction 11) at 0.7, 0.9, and 0.9 kcal mol$^{-1}$ for B3LYP, PBE, and TPSS, respectively. Overall, the grid requirements for these three functionals are modest, and the errors resulting from any of these grids are far less than the errors in the computed reaction energies.

For VS98, M05-2X, and the four M06 functionals (Figure 3), the grid errors are significantly larger than for B3LYP, PBE, or TPSS. Even so, the errors arising from the use of the Gaussian and fine integration grids are tightly grouped around 0 and never exceed 0.15 kcal mol$^{-1}$ for M05-2X or the M06 suite of functionals. For the NWChem grid, the errors are only slightly larger than the Gaussian and fine grid results, although the magnitude exceeds 0.5 kcal mol$^{-1}$ in several cases. The most troubling results arise from use of the Q-Chem (SG-1) grid,[16] which leads to significant errors in computed reaction energies. The problems are most severe for the M06-HF functional, for which the grid errors range from −6.7 to 3.2 kcal mol$^{-1}$. For M06-HF computations employing this popular grid, these grid errors will be the dominant source of error, and the predicted reaction energies will generally be qualitatively different than those computed with finer integration grids.

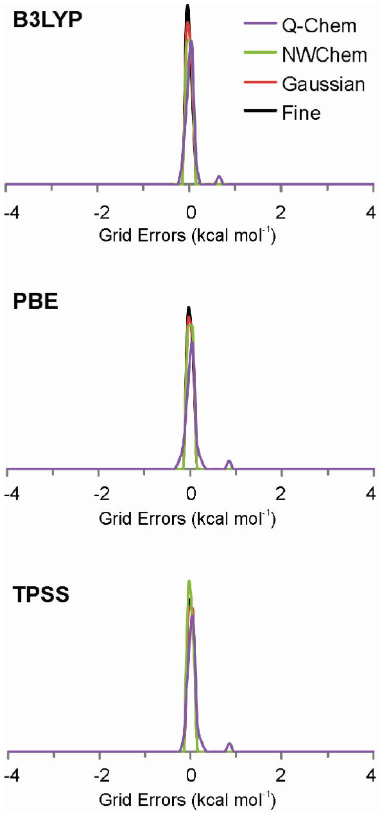


***Figure 2.*** Normalized distributions of grid errors (kcal mol$^{-1}$) for the energies of the reactions in Scheme 1, computed using four popular quadrature grids paired with the B3LYP, PBE, and TPSS functionals. Grids labeled Q-Chem, NWChem, and Gaussian are equivalent to the default grids in those packages. All computations were carried out using the NWChem program.[33,34]

Across all of these functionals, the largest grid errors occur for reaction 10, followed by 11. However, not all functionals behave uniformly. For example, with the Q-Chem grid, the M06-HF predicted energy for reaction 11 is within 1.7 kcal mol$^{-1}$ of the Xfine reference, while the energy for reaction 10 deviates by 5.6 kcal mol$^{-1}$. For the M05-2X functional, the energy for reaction 11 falls 0.4 kcal mol$^{-1}$ from the reference value, even though the value for reaction 10 is in error by only −0.1 kcal mol$^{-1}$. This nonsystematic behavior is indicative of an underlying cancelation of more sizable errors, which is discussed in detail below. Reaction 10 has been highlighted previously[16] as a case prone to grid errors. Gill and co-workers[16] attributed this to the large difference in shape of the two isomers of pentane. Essentially, for more

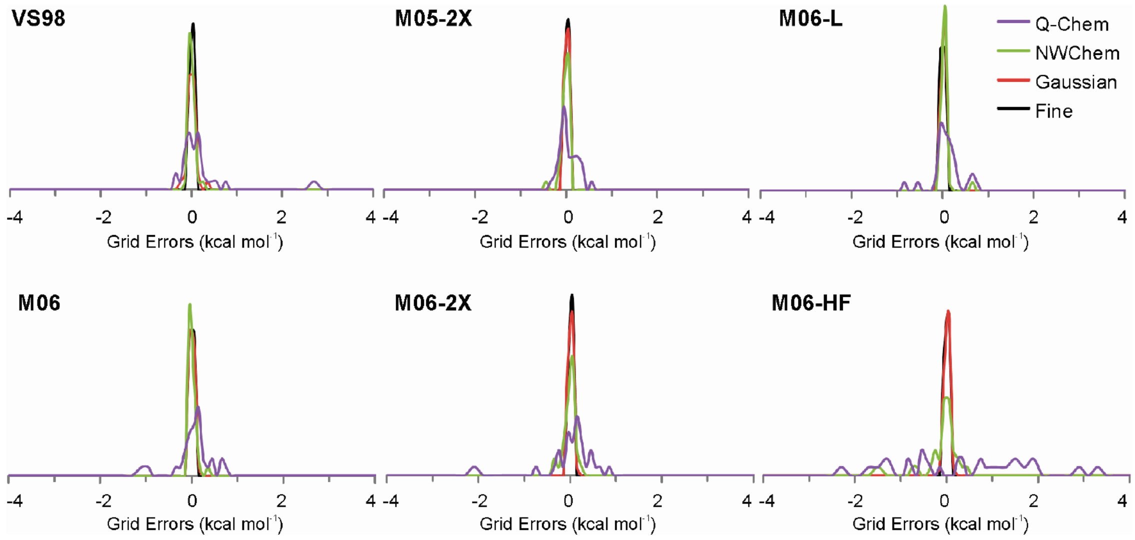


***Figure 3.*** Normalized distributions of grid errors (kcal mol$^{-1}$) for the energies of the reactions in Scheme 1 using four popular quadrature grids.

"compact" molecules (e.g., neopentane), the same number of grid points covers a smaller amount of space, so the grid error in the absolute energy should be reduced compared to that of the error for less compact species (e.g., *n*-pentane).

**C. Effects of Atomic Partitioning Function and Radial Quadrature Methods.** To quantify the effect of different atomic partitioning functions and radial quadrature schemes, the errors associated with grids comprising 50 radial points and 194 angular points are examined more closely. Errors for six combinations of atomic partitioning function and radial quadrature scheme are summarized in Table 4. The left-most column (Becke–Euler) is the Q-Chem (SG-1) grid.[16] Normalized error distributions for the six meta-GGAs paired with these six grids are shown in Figure 4. The Euler and MK radial quadrature schemes[2,3] perform similarly, regardless of the choice of partitioning function. The MK scheme provides a minimal reduction in mean grid errors compared to that of the Euler method. On the other hand, the choice of partitioning function has a significant impact on grid errors for all functionals considered. The SSF or Erf1 partitioning functions[4] result in significantly smaller mean errors compared to that of the Becke results.[5] These results are in accord with previous findings of Martin et al.,[6] and the assertion by Scuseria and co-workers[4] that the SSF scheme[4] is numerically more stable than that of Becke.[5] However, for the M06-HF functional, none of the grids tested delivers reaction energies with errors consistently below 0.5 kcal mol$^{-1}$.

**D. Origin of Grid Errors.** In order to unravel the origins of the large grid errors arising from the use of the Q-Chem (SG-1) grid,[16] the contributions to these errors are examined in more detail. Normalized error distributions for the exchange ($E_x$) and correlation ($E_c$) components of the reaction energies are shown in Figure 5.[40] Mean grid errors in $E_x$ and $E_c$ are given in the Supporting Information (Table S2). For all but the VS98 functional, the errors in $E_x$ swamp those arising from $E_c$, for which the MADs are less than 0.1 kcal mol$^{-1}$. For the VS98 functional, the situation is reversed; in this case, $E_c$ exhibits a larger MAD than $E_x$. Regardless, the unsettling grid errors exhibited by M05-2X and M06 functionals arise from the exchange energies.

***Table 4.*** Analysis of Grid Errors for the Energies of the Reactions in Scheme 1 for Combinations of Three Partitioning Functions and Two Radial Quadrature Schemes[a]

| | Euler–Maclaurin | | | Mura–Knowles | | |
|---|---|---|---|---|---|---|
| | Becke | SSF | Erf1 | Becke | SSF | Erf1 |
| VS98 | | | | | | |
| MAD | 0.26 | 0.18 | 0.19 | 0.27 | 0.15 | 0.16 |
| MSD | 0.14 | 0.10 | 0.12 | 0.13 | 0.05 | 0.05 |
| min | −0.38 | −0.24 | −0.27 | −0.35 | −0.28 | −0.33 |
| max | 2.80 | 1.19 | 1.37 | 3.12 | 0.93 | 0.94 |
| M05-2X | | | | | | |
| MAD | 0.16 | 0.06 | 0.06 | 0.16 | 0.06 | 0.07 |
| MSD | 0.03 | −0.01 | −0.01 | 0.04 | −0.01 | −0.02 |
| min | −0.39 | −0.13 | −0.14 | −0.32 | −0.18 | −0.21 |
| max | 0.52 | 0.19 | 0.25 | 0.37 | 0.11 | 0.15 |
| M06-L | | | | | | |
| MAD | 0.21 | 0.09 | 0.08 | 0.20 | 0.07 | 0.07 |
| MSD | 0.09 | 0.05 | 0.04 | 0.08 | 0.02 | 0.02 |
| min | −0.84 | −0.13 | −0.21 | −0.67 | −0.16 | −0.17 |
| max | 0.76 | 0.54 | 0.34 | 0.71 | 0.30 | 0.27 |
| M06 | | | | | | |
| MAD | 0.25 | 0.08 | 0.08 | 0.23 | 0.09 | 0.09 |
| MSD | 0.06 | 0.02 | 0.01 | 0.07 | 0.01 | 0.01 |
| min | −1.04 | −0.17 | −0.19 | −0.82 | −0.18 | −0.16 |
| max | 0.75 | 0.35 | 0.32 | 0.69 | 0.48 | 0.43 |
| M06-2X | | | | | | |
| MAD | 0.29 | 0.07 | 0.08 | 0.26 | 0.07 | 0.07 |
| MSD | 0.05 | 0.00 | 0.00 | 0.05 | −0.02 | −0.02 |
| min | −2.02 | −0.36 | −0.49 | −1.85 | −0.20 | −0.28 |
| max | 0.81 | 0.30 | 0.33 | 0.66 | 0.14 | 0.17 |
| M06-HF | | | | | | |
| MAD | 1.24 | 0.24 | 0.30 | 1.13 | 0.20 | 0.25 |
| MSD | 0.05 | −0.04 | −0.03 | 0.05 | −0.05 | −0.04 |
| min | −6.70 | −1.35 | −1.71 | −6.57 | −1.24 | −1.61 |
| max | 3.21 | 0.69 | 0.89 | 2.92 | 0.51 | 0.65 |

[a] Results for B3LYP, PBE, and TPSS are available in the Supporting Information. All grids have 50 radial points and 194 angular points. Errors are relative to Xfine results. All values are in kcal mol$^{-1}$.

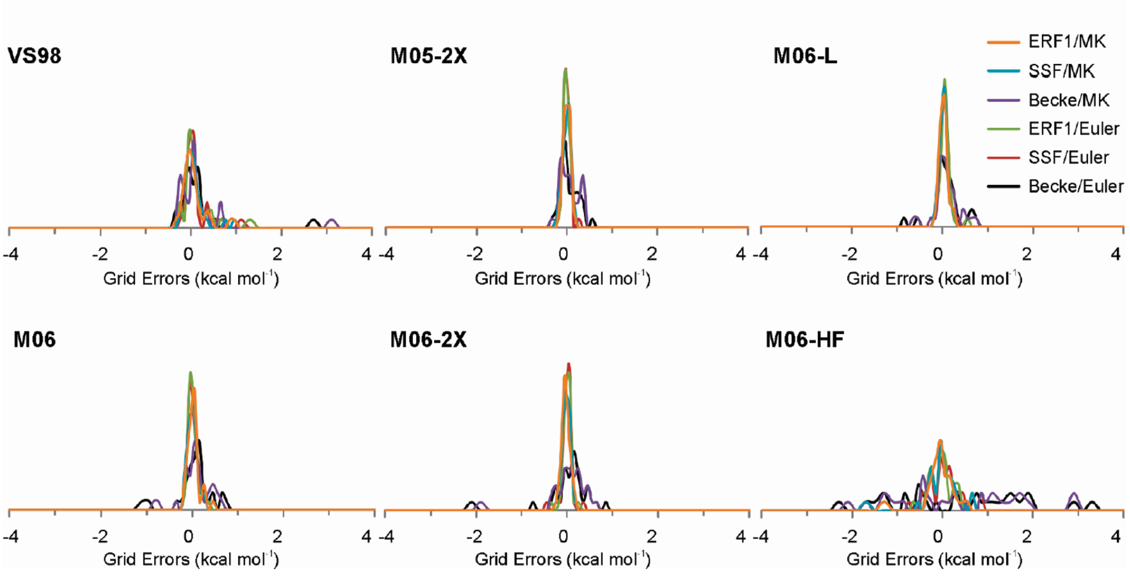


***Figure 4.*** Normalized distributions of grid errors (kcal mol$^{-1}$) for the energies of the reactions in Scheme 1 using six combinations of partitioning functions (Becke,[5] SSF,[4] or Erf1) and radial quadrature schemes [Euler−Maclaurin (Euler)[2] or Mura−Knowles (MK)[3]]. All grids are pruned and have 50 radial points and 194 angular points. The Becke/Euler combination corresponds to the default Q-Chem (SG-1) grid.[16]

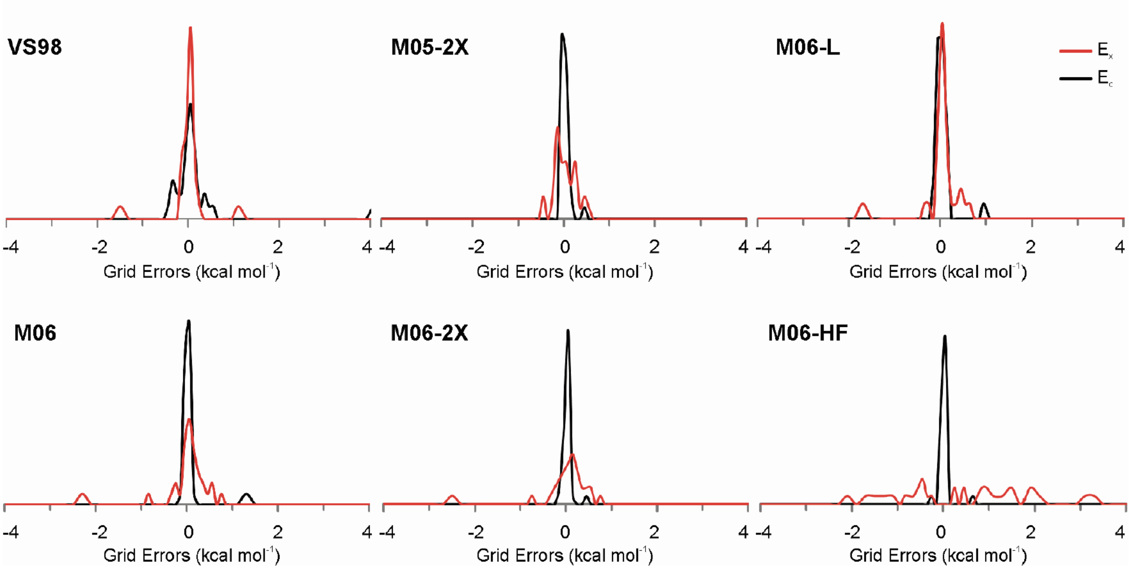


***Figure 5.*** Normalized distributions of grid errors (kcal mol$^{-1}$) for the contribution of $E_x$ and $E_c$ to the total energies for the reactions in Scheme 1, computed with the Q-Chem grid.

The exchange functional utilized for the M06 suite[10] is a linear combination of the functional forms of the M05-2X[25] and VS98[18] exchange functionals:

$$E_x^{\mathrm{M06}} = E_x^{\mathrm{M05-2X}} + E_x^{\mathrm{VS98}} \quad (3)$$

$E_x^{\mathrm{M05\text{-}2X}}$ is the PBE exchange functional, $F_{X\sigma}^{\mathrm{PBE}}$, multiplied by a kinetic energy density-dependent term, $f(w_\sigma)$. This function is referred to by Truhlar and co-workers[10] as the "kinetic energy density enhancement factor" and is written as a power series in $w_\sigma$. The $w_\sigma$, in turn, is a function of the spin kinetic energy density.[41] The form for the M06 exchange functional is given in eqs 4−11, where $\rho_\sigma$, $\nabla\rho_\sigma$, and $\tau_\sigma$ are the spin density, gradient, and kinetic energy density, respectively, and $d_i$ and $a_i$ are empirically determined parameters. Setting $a_0$ to 1.0 and the other $a_i$ and $d_i$ constants to zero gives the standard PBE GGA exchange functional, while setting the $a_i$ constants to zero (and the $d_i$ to the appropriate values) yields VS98 exchange. The M05-2X exchange functional is obtained by setting the $d_i$ constants to zero, and the $a_i$ to the appropriate values.

$$E_x^{\mathrm{M06}} = \sum_{\sigma}^{\alpha,\beta} \int d\mathbf{r}\Bigg[F_{X\sigma}^{\mathrm{PBE}}(\rho_\sigma, \nabla\rho_\sigma)f(w_\sigma) + \varepsilon_{X\sigma}^{\mathrm{LSDA}}\left(\frac{d_0}{\gamma(x_\sigma, z_\sigma)} + \frac{d_1x_\sigma^2 + d_2z_\sigma}{\gamma^2(x_\sigma, z_\sigma)} + \frac{d_3x_\sigma^4 + d_4x_\sigma^2z_\sigma + d_5z_\sigma^2}{\gamma^3(x_\sigma, z_\sigma)}\right)\Bigg] \quad (4)$$

$$f(w_\sigma) = \sum_{i=0}^{11} a_i w_\sigma^i \qquad (5)$$

$$\varepsilon_{X\sigma}^{\mathrm{LSDA}} = \frac{3}{2}\left(\frac{3}{4\pi}\right)^{1/3} \rho_\sigma^{4/3} \qquad (6)$$

$$x_\sigma = \frac{|\nabla \rho_\sigma|}{\rho_\sigma^{4/3}} \qquad (7)$$

$$z_\sigma = \frac{2\tau_\sigma}{\rho^{5/3}} - \frac{3}{5}(6\pi^2)^{2/3} \qquad (8)$$

$$\gamma(x_\sigma, z_\sigma) = 1 + \alpha(x_\sigma^2 + z_\sigma) \qquad (9)$$

$$w_\sigma = (t_\sigma - 1)/(t_\sigma + 1) \qquad (10)$$

$$t_\sigma = \frac{3(6\pi^2)^{2/3}\rho_\sigma^{5/3}}{10\tau_\sigma} \qquad (11)$$

Errors arising from the use of the Q-Chem grid for each term in eq 4 are summarized in the Supporting Information (Tables S2 and S3). These were computed using converged densities from Xfine computations with a locally modified version of NWChem 5.1.[33,34] Distributions of grid errors arising from the two main components of the M06-HF exchange functional, $E_{\mathrm{x}}^{\mathrm{M05\text{-}2X}}$ and $E_{\mathrm{x}}^{\mathrm{VS98}}$, are shown in Figure 6a. The grid errors arise primarily from the M05-2X component.

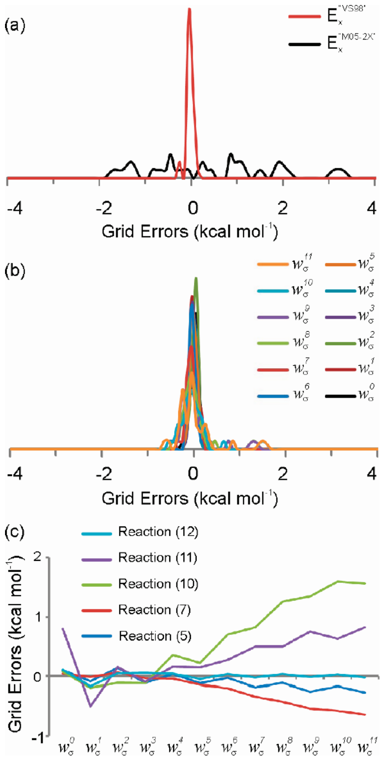


***Figure 6.*** (a) Normalized Q-Chem grid error distributions for the $E_{\mathrm{x}}^{\mathrm{M05\text{-}2X}}$ and $E_{\mathrm{x}}^{\mathrm{VS98}}$ components of M06-HF. (b) Q-Chem grid error distribution for each power of $w_\sigma$ in eq (12). (c) Q-Chem grid errors for each power of $w_\sigma$ in eq (12) for representative reactions from Scheme 1.

***Table 5.*** Coefficients in the Kinetic Energy Density Enhancement Factor (eq 5) in M05-2X (see ref 25) and the M06 Suite of Functionals (see ref 10)

| | M05-2X | M06-L | M06 | M06-2X | M06-HF |
|---|---|---|---|---|---|
| $a_0$ | 1.0 | 0.3987756 | 0.5877943 | 0.46 | 0.1179732 |
| $a_1$ | −0.56833 | 0.2548219 | −0.1371776 | −0.2206052 | −1.066708 |
| $a_2$ | −1.30057 | 0.3923994 | 0.2682367 | −0.09431788 | −0.1462405 |
| $a_3$ | 5.5007 | −2.103655 | −2.515898 | 2.164494 | 7.481848 |
| $a_4$ | 9.06402 | −6.302147 | −2.978892 | −2.556466 | 3.776679 |
| $a_5$ | −32.21075 | 10.97615 | 8.710679 | −14.22133 | −44.36118 |
| $a_6$ | −23.73298 | 30.97273 | 16.88195 | 15.55044 | −18.30962 |
| $a_7$ | 70.22996 | −23.18489 | −4.489724 | 35.98078 | 100.3903 |
| $a_8$ | 29.88614 | −56.7348 | −32.99983 | −27.22754 | 38.6436 |
| $a_9$ | −60.25778 | 21.60364 | −14.4905 | −39.24093 | −98.06018 |
| $a_{10}$ | −13.22205 | 34.21814 | 20.43747 | 15.22808 | −25.57716 |
| $a_{11}$ | 15.23694 | −9.049762 | 12.56504 | 15.22227 | 35.90404 |

Normalized grid error distributions for each power of $w_\sigma$ in $f(w_\sigma)$ are plotted in Figure 6b with all of the $a_i$ constants set to unity, i.e.:

$$\sum_{\sigma}^{\alpha,\beta} \int d\mathbf{r} \mathbf{F}_{X\sigma}^{\mathrm{PBE}}(\rho_\sigma, \nabla\rho_\sigma) w_\sigma^i \qquad (12)$$

The grid errors for each term are modest and exceed 0.5 kcal mol$^{-1}$ for only a few reactions. The tendency is for the magnitude of the integration error to gradually increase with the power of $w_\sigma$, although for some reactions the integration errors are negligible for all powers of $w_\sigma$ (errors for representative reactions are shown in Figure 6c). The empirical constants $a_i$, on the other hand, vary considerably in magnitude and oscillate between negative and positive values (see Table 5). The grid errors for each term in $E_{\mathrm{x}}^{\mathrm{M05\text{-}2X}}$ will be the product of the $a_i$ constant and the integration error for the corresponding power of $w_\sigma$. Because the magnitudes of the $a_i$ constants are large, grid errors for the individual contributions to $f(w_\sigma)$ are enormous. For example, for reaction 10 the errors in the $w_\sigma^7$ and $w_\sigma^9$ terms are +82 and −132 kcal mol$^{-1}$, respectively. These sizable errors mostly cancel due to the almost monotonic change in grid errors with each power of $w_\sigma$ (see Figure 6c) combined with the oscillations in the $a_i$ constants. This cancelation is incomplete, of course, and gives rise to the troubling spread of grid errors discussed above. The substantial grid errors exhibited by the M06-HF functional, in particular, are a result of the very large $a_i$ constants defining that functional (e.g.: see the $a_7$ and $a_9$ constants for M06-HF in Table 5).

The attribution of these errors to the particular functional form used in the M06 suite of functionals is further supported by the qualitatively different grid-dependence exhibited by the meta-GGA functional TPSS. This functional, which does not contain a polynomial expansion in terms of the kinetic energy density, exhibits very modest dependence on the choice of integration grid; TPSS grid errors are on par with those from popular GGAs.

These grid errors are qualitatively different from those underlying the discontinuities in meta-GGA-computed potential energy curves for dispersion-bound complexes,[12] which arise from singularities in the $\tau$-dependent functional forms.[13] The grid errors in reaction energies arise from modest errors in the integration of $w_\sigma$ amplified by the large empirical $a_i$ constants. One consequence is that these errors vary smoothly across potential energy surfaces. This is demonstrated for neopentane in the Supporting Information (Figures S1 and S2), in which grid errors in the M06-HF energy computed with the Q-Chem grid are plotted as a function of the C−C and C−H bond lengths.

## IV. Summary and Conclusions

DFT is invaluable in the computational study of organic reactions and can provide accurate energies for large molecular systems that are beyond the reach of traditional ab initio methods. New DFT functionals offer increased accuracy and broader applicability compared to those of previous generations of functionals and have enabled the application of DFT to myriad new problems in organic chemistry and molecular biology. However, these new functionals are not without drawbacks, one of which is the increased sensitivity of energies and other properties to the choice of integration grid. Previously documented[12−16] grid sensitivities exhibited by these functionals include the prediction of potential energy curves for dispersion-bound complexes with spurious oscillations as well problems with predicted energies, geometries, and vibrational frequencies.[19,20]

We have quantified integration grid errors for six meta-GGA functionals (VS98, M05-2X, M06-L, M06, M06-2X, and M06-HF) paired with popular integration grids for 34 organic isomerization energies. The popular SG-1 grid,[16] which is the default in the Q-Chem package, leads to large errors for all of these functionals and very large errors for M06-HF. This grid should not be used with any of these functionals. By contrast, the grid errors in B3LYP, PBE, and TPSS computed energies are small for all of the grids tested. Use of the SSF[4] or Erf1 atomic partitioning functions reduces grid errors compared to that of the standard SG-1 grid,[16] which utilizes the partitioning function of Becke.[5] However, the M06-HF functional still exhibits grid errors exceeding 0.5 kcal mol$^{-1}$ for several of the reactions.

The grid errors exhibited by the M06 suite of functionals arise from integration errors in the exchange component of the energy. In particular, the significant errors arising from the use of the Q-Chem (SG-1) grid are due to the large empirical constants in the kinetic energy density enhancement factor. Some of these constants are of considerable size and amplify modest errors in the integration of the kinetic energy density. This is not a general weakness of meta-GGAs but a problem arising from the particular functional form used in the M06 suite of functionals.

Zhao and Truhlar recently published[11] the M08-HX and M08-SO functionals, which incorporate more flexible functional forms than members of the M06 suite and contain an altered self-interaction correction term that avoids the numerical instabilities discussed by Gräfenstein, Izotov, and Cremer.[14] Although the grid errors associated with these new functionals were not tested here, the kinetic energy density enhancement factor in M08-HX and M08-SO is the same as in the M06 functionals. Moreover, the empirical coefficients in this factor are larger than in any of the M06 suite of functionals, so M08-HX and M08-SO grid errors are expected to be even more severe than observed for M06-HF.

The popularity of DFT and ease with which many computational chemistry program packages can be used has led to a continued increase in the application of DFT to chemical problems by nonspecialists. While this is certainly a welcome development and a testament to the maturity of the field of Kohn−Sham DFT, the present results offer a poignant reminder of the dangers of employing “default” options in any program package. These defaults are not suitable for all applications, and in the case of DFT integration grids, the defaults in some cases are woefully inadequate for some meta-GGA functionals. In particular, use of the SG-1 integration grid with M05-2X or the M06 suite of functionals can result in significant errors in predicted reaction energies.

**Acknowledgment.** This work was supported by the National Science Foundation (CHE-0548209). S.E.W. was also supported by an National Institute of Health National Research Service Award Postdoctoral Fellowship (NIH-5F32GM082114-2) and would like to thank K.M. Williams for assistance preparing figures. Computer time was provided by the University of California, Los Angeles Institute for Digital Research and Education (IDRE).

**Supporting Information Available:** Tables of grid errors for individual reactions and absolute energies. Plots of grid errors for neopentane as a function of the C−C and C−H bond lengths. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Lebedev, V. I.; Laikov, D. N. *Dokl. Math.* **1999**, *59*, 477.

(2) Murray, C. W.; Handy, N. C.; Laming, G. L. *Mol. Phys.* **1993**, *78*, 997.

(3) Mura, M. E.; Knowles, P. J. *J. Chem. Phys.* **1996**, *104*, 9848.

(4) Stratmann, R. E.; Scuseria, G. E.; Frisch, M. J. *Chem. Phys. Lett.* **1996**, *257*, 213.

(5) Becke, A. *J. Chem. Phys.* **1988**, *88*, 1053.

(6) Martin, J. M. L.; Bauschlicher, C. W., Jr.; Ricca, A. *Comput. Phys. Commun.* **2001**, *133*, 189.

(7) Papas, B. N.; Schaefer, H. F. *J. Mol. Struct. THEOCHEM* **2006**, *768*, 175.

(8) Dressler, S.; Thiel, W. *Chem. Phys. Lett.* **1997**, *273*, 71.

(9) Termath, V.; Sauer, J. *Chem. Phys. Lett.* **1996**, *255*, 187.

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

(11) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1849.

(12) Johnson, E. R.; Wolkow, R. A.; DiLabio, G. A. *Chem. Phys. Lett.* **2004**, *394*, 334.

(13) Johnson, E. R.; Becke, A.; Sherrill, C. D.; DiLabio, G. A. *J. Chem. Phys.* **2009**, *131*, 034111.

(14) Gräfenstein, J.; Izotov, D.; Cremer, D. *J. Chem. Phys.* **2007**, *127*, 214103.

(15) Fusti-Molnar, L.; He, X.; Wang, B.; Merz, K. M., Jr. *J. Chem. Phys.* **2009**, *131*, 065102.

(16) Gill, P. M. W.; Johnson, B. G.; Pople, J. A. *Chem. Phys. Lett.* **1993**, *209*, 506.

(17) Gräfenstein, J.; Cremer, D. *J. Chem. Phys.* **2007**, *127*, 164113.

(18) Van Voorhis, T.; Scuseria, G. *J. Chem. Phys.* **1998**, *109*, 400.

(19) Csonka, G. I.; French, A. D.; Johnson, G. P.; Stortz, C. A. *J. Chem. Theory Comput.* **2009**, *5*, 679.

(20) Jiménez-Hoyos, C. A.; Janesko, B. G.; Scuseria, G. E. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6621.

(21) Grimme, S.; Steinmetz, M.; Korth, M. *J. Org. Chem.* **2007**, *72*, 2118.

(22) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157.

(23) Schafer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.

(24) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.

(25) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.

(26) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.

(27) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126.

(28) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

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

(30) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.

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

(32) Shao, Y.; Fusti-Molnar, L.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; DiStasio, R. A., Jr.; Lochan, R. C.; Wang, T.; Beran, G. J. O.; Besley, N. A.; J. M., H.; Lin, C. Y.; Voorhis, T. V.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C.-P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L., III; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; Schaefer, H. F.; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172.

(33) Kendall, R. A.; Apra, E.; Bernholdt, D. E.; Bylaska, E. J.; Dupuis, M.; Fann, G. I.; Harrison, R. J.; Ju, J.; Nichols, J. A.; Nieplocha, J.; Straatsma, T. P.; Windus, T. L.; Wong, A. T. *Comput. Phys. Commun.* **2000**, *128*, 260.

(34) Bylaska, E. J.; de Jong, W. A.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Valiev, M; Wang, D.; Aprà, E.; Windus, T. L.; Hammond, J.; Nichols, P; Hirata, S.; Hackler, M. T.; Zhao, Y.; Fan, P.-D.; Harrison, R. J.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Wu, T; Van Voorhis, T; Auer, A. A.; Nooijen, M.; Brown, E; Cisneros, G.; Fann, G. I.; Fruchtl, H; Garza, J; Hirao, K; Kendall, R.; Nichols, J. A.; Tsemekhman, K; Wolinski, K.; Anchell, J.; Bernholdt, D; Borowski, P; Clark, T; Clerc, D.; Dachsel, H; Deegan, H.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R; Long, X.; Meng, B.; Nakajima, T; Niu, S; Pollack, L.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem, A Computational Chemistry Package for Parallel Computers, Version 5.1*; Pacific Northwest National Laboratory: Richland, Washington, 2007.

(35) 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.; and Pople, J. A. *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(36) Note that as implemented in NWChem 5.1.1, the density screening threshold used by other functionals is overridden by the VS98, M05-2X, and M06 suite of functionals. Instead, in NWChem 5.1.1, density screenings for the exchange component of these functionals are fixed at $10^{-7}$ or $10^{-8}$ au.

(37) Grimme, S.; Schwabe, T. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4398.

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

(39) Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405.

(40) The decomposition of grid errors for DFT energies into exchange and correlation components is complicated by the secondary effects of the grid errors on the SCF procedure. Even though the other components of the electronic energy (Coulomb and one-electron terms and HF exchange) are evaluated analytically and should exhibit no grid errors, when the Kohn−Sham orbitals are optimized, grid the errors in $E_x$ and $E_c$ lead to differences in all of the components of the energy for different grids. As such, to evaluate the errors arising from $E_c$ and $E_x$, energies for the Q-Chem grid were evaluated using converged Kohn−Sham orbitals from Xfine grid computations. This has only a minor effect on the grid errors in the total reaction energies.

(41) Becke, A. *J. Chem. Phys.* **2000**, *112*, 4020.

CT900639J

JCTC Journal of Chemical Theory and Computation

# Divide and Conquer Hartree−Fock Calculations on Proteins

Xiao He and Kenneth M. Merz, Jr.*

*Department of Chemistry and the Quantum Theory Project, 2328 New Physics Building, P.O. Box 118435, University of Florida, Gainesville, Florida 32611-8435*

Received December 9, 2009

**Abstract:** The ability to perform ab initio electronic structure calculations that scale linearly with the system size is one of the central aims in theoretical chemistry. In this study, the implementation of the divide and conquer (DC) algorithm, an algorithm with the potential to aid the achievement of true linear scaling within Hartree−Fock (HF) theory, is revisited. Standard HF calculations solve the Roothaan−Hall equations for the whole system; in the DC-HF approach, the diagonalization of the Fock matrix is carried out on smaller subsystems. The DC algorithm for HF calculations was validated on polyglycines, polyalanines, and 11 real three-dimensional proteins of up to 608 atoms in this work. We also found that a fragment-based initial guess using the molecular fractionation with conjugated caps (MFCC) method significantly reduces the number of SCF cycles and even is capable of achieving convergence for some globular proteins where the simple superposition of atomic densities (SAD) initial guess fails.

## Introduction

Ab initio quantum mechanical methods have been developed over the past several decades and successfully applied to the study of the chemical properties for small- to moderate-sized molecules. The routine application of these full quantum mechanical calculations on macromolecules (molecules containing greater than 500 atoms) continues to pose a great challenge for theoretical chemists. The major limitation of ab initio methods is the scaling problem, since the computational cost of ab initio methods increases considerably as the size of the molecule increases. For instance, Hartree−Fock (HF)[1] and density functional theory (DFT)[2] scale as $O(N^4)$, post-Hartree−Fock MP2[3] scales as $O(N^5)$, and the coupled cluster(CC)[4−9] method that includes single and double excitations (CCSD) scales as $O(N^6)$. In modern HF calculations, the computational cost for the 2-electron integrals can be reduced from $O(N^4)$ to $O(N^2)$ using a simple screening technique.[10] Hence, the dominant step for large molecules becomes the matrix diagonalization, which scales as $O(N^3)$. In this study, our goal was to reduce the computational cost of the diagonalization step in HF calculations to linear with system size.

The state-of-the-art linear-scaling algorithms, which make the computational cost scale linearly $O(N)$ with the system size, have attracted the focus of many theorists during the past decade.[11−21] Much effort has been devoted to the development of linear-scaling methods in order to compute the total energy of large molecular systems at the Hartree−Fock (HF) or density functional theory (DFT) level.[12,15,18,22−26] One of the challenges is to speed up the calculation of long-range Coulomb interactions when assembling the Fock matrix elements. Fast multipole-based approaches have successfully reduced the scaling in system size to linear[14,16−18,25] and made HF and DFT calculations affordable for larger systems when small- to moderate-sized basis sets are utilized. The more recently developed Fourier transform Coulomb method of Fusti and Pulay[27,28] reduced the steep $O(N^4)$ scaling in basis set size to quadratic and makes the calculations much more affordable with larger basis sets.[29] There is also a class of fragment-based methods for quantum calculation of protein systems including the divide and conquer (DC) method of Yang,[22] Yang and Lee,[23] Dixon and Merz,[30] Gogonea et al.,[31] Shaw and St-Amant,[32] and Nakai and co-workers,[33−36] the adjustable density matrix assembler (ADMA) approach method of Exner and Mezey,[26,37−39] the fragment molecular orbital (FMO) method of Kitaura and co-workers,[13,40,41] and the molecular frac-

* Correspondence author phone: 352-392-6973; fax: 352-392-8722, e-mail: merz@qtp.ufl.edu.

10.1021/ct9006635
Published on Web 01/07/2010

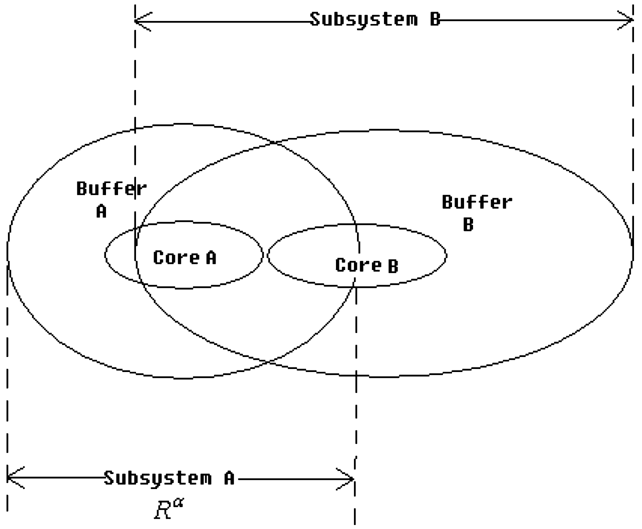


***Figure 1.*** Graphical representation of the subsetting scheme used in divide and conquer calculations.

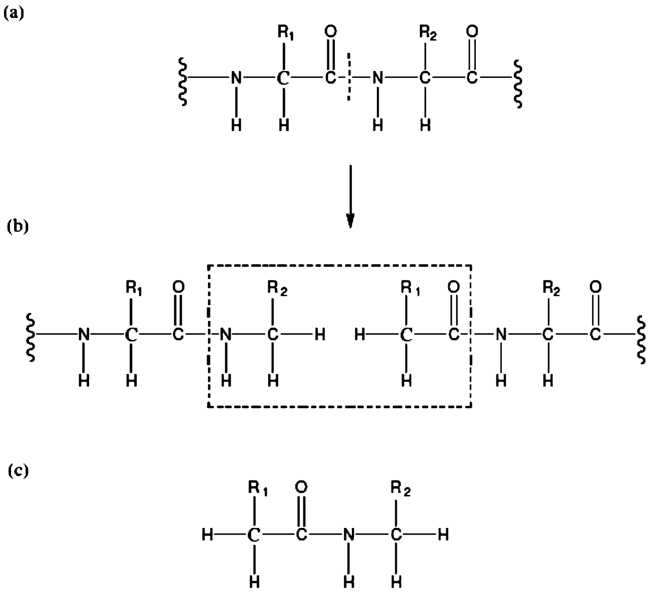


***Figure 2.*** MFCC scheme in which the peptide bond is cut (a) and the fragments are capped with $C_{cap}$ and its conjugate $C_{cap}$* (b). (c) Atomic structure of the concap. The concap is defined as the fused molecular species of $C_{cap}*-C_{cap}$.

tionation with conjugate caps (MFCC) approach developed by Zhang and co-workers.[42,43] Most applications of these methods to protein systems have been largely limited to semiempirical, HF, and DFT calculations. Among these approaches, FMO has been applied to higher level ab initio calculations such as second-order Møller−Plesset perturbation theory (MP2)[44] and coupled cluster theory (CC).[45] Nakai and co-workers recently proposed DC-MP2[33,36,46] and DC-CCSD[47] approaches; however, only systems of linear chains or near-linear chains have been tested so far for the divide and conquer algorithm for ab initio calculations.

In the DC algorithm, the total system is divided into small fragments. Atoms within adjustable buffer regions surrounding each fragment are included in the calculations to preserve the chemical environment of the divided subsystem. A set of local Roothaan−Hall equations is then solved for each subsystem, and an approximate full density matrix of the entire molecular system is built up from subsystem contributions. By solving the HF self-consistent field (SCF) equation iteratively, the final converged full density matrix is used to obtain the total energy of the entire system. In this manner, linear scaling of the Fock matrix diagonalization step is achieved as a result of the fact that a set of smaller subsystem Fock matrices is diagonalized in the DC-HF approach rather than the global Fock matrix diagonalization for traditional HF calculations. Furthermore, divide and conquer calculations may be efficiently parallelized because the individual subsystem calculations are solved separately. In the DC-HF approach, the memory usage will increase linearly as the size of the system increases, which is also an important feature of this approach.

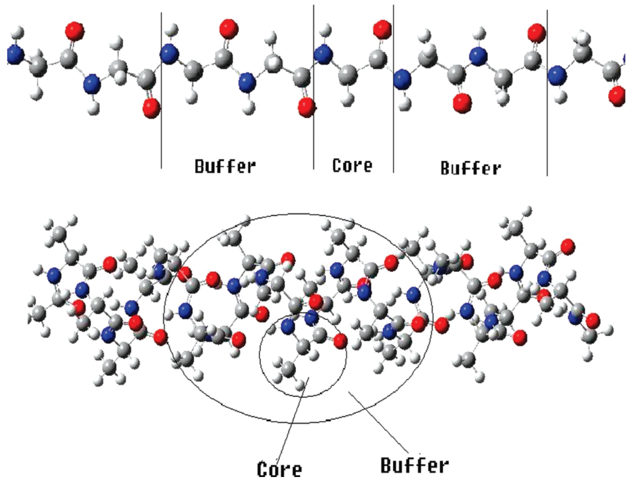


***Figure 3.*** Subsetting schemes for divide and conquer calculations on the extended polyglycine $(Gly)_n$(upper) and polyalanine in an α-helical structure (α-$(Ala)_n$, bottom).

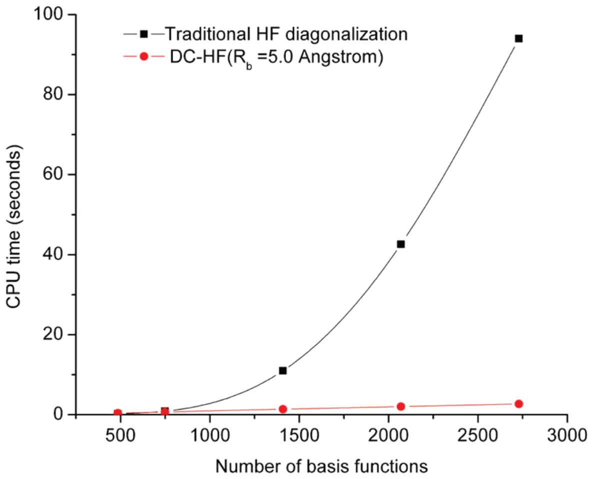


***Figure 4.*** Average computational time to diagonalize the Fock matrix in each SCF cycle using traditional HF and DC-HF for a series of extended polyglycines at the HF/6-31G* level.

The aim of our current research is to further develop and validate the divide and conquer (DC)[22,23,30,32,46−48] methodology to aid in the application of ab initio methods to biomacromolecules. In this study, our goal is to validate the divide and conquer algorithm for Hartree−Fock calculations on globular proteins. Moreover, we propose a fragment-based

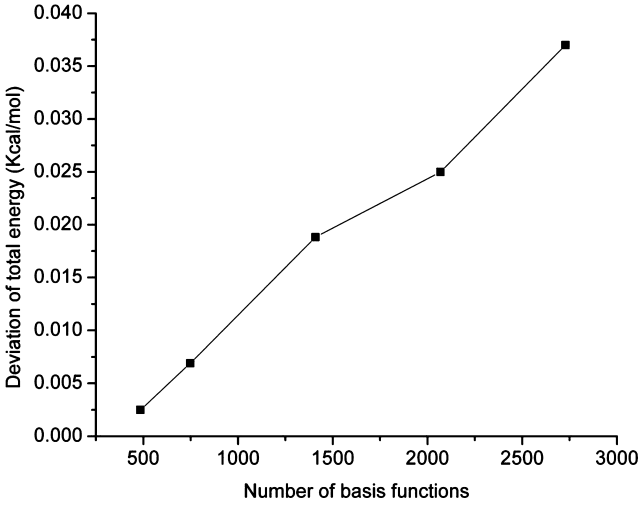


***Figure 5.*** Accuracy of the total energy calculated by the DC-HF approach on extended polyglycine systems compared to full system calculations.

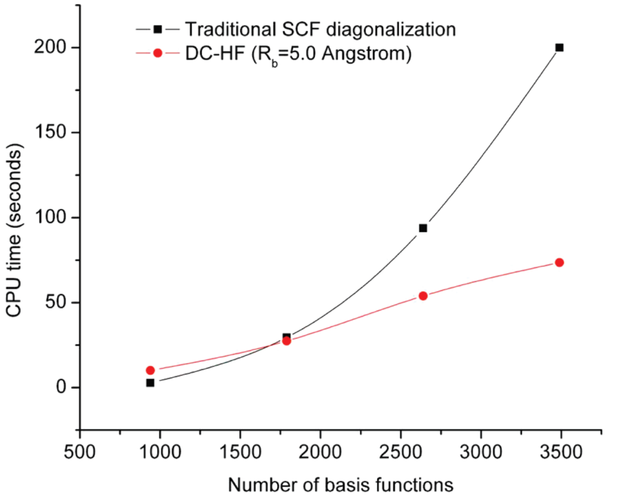


***Figure 6.*** Similar to Figure 4 but for the polyalanine systems in an α-helical structure α-(Ala)$_n$.

initial guess using molecular fractionation with conjugated caps (MFCC) method to reduce the number of SCF cycles, and different division schemes are compared.

## Computational Approaches

**Divide and Conquer Approach on the Hartree−Fock Calculations.** In protein systems, the divide and conquer approach is based on the chemical locality; this assumes that local regions of a protein are only weakly influenced by the atoms that are far away from the region of interest. The whole system is divided into fragments called core regions (Core$^\alpha$). A buffer region (Buffer$^\alpha$) is assigned for each core region to account for the environmental effects. The combination of every core region and its buffer region constitutes each individual subsystem ($R^\alpha$) as illustrated in Figure 1. Local MOs of each subsystem are solved by the Roothaan−Hall equation

$$F^\alpha C^\alpha = S^\alpha C^\alpha E^\alpha \tag{1}$$

where $F^\alpha$ and $S^\alpha$ are local Fock matrix and local overlap matrix, respectively.

$$F^\alpha_{\mu\nu} = \begin{cases} F_{\mu\nu} & \text{if } \chi_\mu \in R^\alpha \text{ and } \chi_\nu \in R^\alpha \\ 0 & \text{elsewhere} \end{cases} \tag{2}$$

After the local MO coefficient matrices $C^\alpha$ are obtained, the total density matrix of the whole system is given by

$$P_{\mu\nu} = \sum_{\alpha=1}^{N_{\text{sub}}} P^\alpha_{\mu\nu} = \sum_{\alpha=1}^{N_{\text{sub}}} D^\alpha_{\mu\nu} p^\alpha_{\mu\nu} \tag{3}$$

where $D^\alpha_{\mu\nu}$ is the partition matrix

$$D^\alpha_{\mu\nu} = \begin{cases} 1 & \phi_\mu \in \text{Core}^\alpha \text{ and } \phi_\nu \in \text{Core}^\alpha \\ {}^1/_2 & \phi_\mu \in \text{Core}^\alpha \text{ and } \phi_\nu \in \text{Buffer}^\alpha \text{ or} \\ & \phi_\mu \in \text{Buffer}^\alpha \text{ and } \phi_\nu \in \text{Core}^\alpha \\ 0 & \phi_\mu \notin \text{Core}^\alpha \text{ and } \phi_\nu \notin \text{Core}^\alpha \end{cases} \tag{4}$$

and $p^\alpha_{\mu\nu}$ is the local density matrix defined by

$$p^\alpha_{\mu\nu} = \sum_i^{\text{LMOs}} n^\alpha_i C^\alpha_{\mu i} C^{\alpha *}_{\nu i} \tag{5}$$

where $n^\alpha_i$ is a smooth approximation to the Heaviside step function

$$n^\alpha_i = \frac{2}{1 + \exp[(\varepsilon^\alpha_i - \varepsilon_F)/kT]} \tag{6}$$

$\varepsilon_F$ is determined through the normalization of the total number of electrons of the whole system

$$N_{\text{elec}} = \sum_\alpha \sum_\mu (P^\alpha S^\alpha)_{\mu\mu} \tag{7}$$

After the density matrix is converged, the total HF energy is given as

$$E^{\text{DC}}_{\text{HF}} = \frac{1}{2} \sum_\alpha \sum_{\mu\nu} P^\alpha_{\mu\nu}(H^\alpha_{\mu\nu} + F^\alpha_{\mu\nu}) \tag{8}$$

where $H^\alpha_{\mu\nu}$ is the local one-electron core Hamiltonian matrix similar to the definition of local Fock matrix in eq 2.

For HF calculations on large systems, the construction of the Coulomb matrix and exchange matrix along with the diagonalization of the Fock matrix constitute the three most time-consuming steps. The Hamiltonian matrix diagonalization intrinsically scales as O($N^3$). In the divide and conquer scheme the diagonalization calculation is performed on each submatrix, which will naturally make the SCF diagonalization step scale linearly with the number of subsystems. However, it is important to realize that the divide and conquer algorithm does not help to reduce the scale of computation of the Coulomb matrix and exchange matrix. The continuous fast multipole method (CFMM)[14,16−18,25,49−51] and the linear exchange K approach (LinK)[52,53] provide ways in which the Coulomb matrix and exchange matrix can be made linear scaling, respectively.

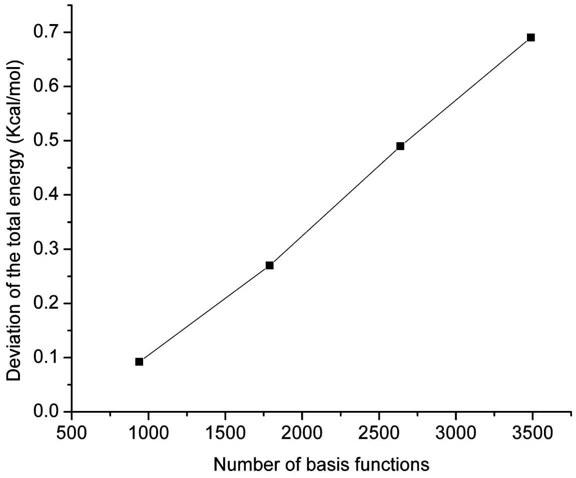


***Figure 7.*** Similar to Figure 5 but for the polyalanine systems in an α-helix structure α-(Ala)$_n$.

***Table 1.*** Number of SCF Cycles Needed To Reach Convergence for the SAD and MFCC Initial Guess at the HF/6-31G* Level

| | DC | | non-DC[a] | |
|---|---|---|---|---|
| system | SAD initial guess | MFCC initial guess | SAD initial guess | MFCC initial guess |
| Gly$_6$ | 18 | 10 | 12 | 7 |
| Gly$_{10}$ | 18 | 11 | 12 | 7 |
| Gly$_{20}$ | 18 | 10 | 12 | 6 |
| Gly$_{30}$ | 18 | 10 | 12 | 6 |
| Gly$_{40}$ | 18 | 8 | 12 | 7 |
| α-(Ala)$_{10}$ | 18 | 15 | 12 | 9 |
| α-(Ala)$_{20}$ | 16 | 12 | 12 | 9 |
| α-(Ala)$_{30}$ | 16 | 12 | 12 | 8 |
| α-(Ala)$_{40}$ | 15 | 12 | 12 | 8 |

[a] In the SCF procedure of the non-DC case every 10 previous Fock matrices were stored in the DIIS calculations, while for the DC case every 2 previous Fock matrices were stored in the DIIS calculations until the root-mean-squared change of the density matrix elements reaches $10^{-4}$ au, after which the DIIS technique was turned off.

**MFCC Initial Guess.** Here we introduce a fragment-based initial guess for ab initio calculations using the molecular fractionation with conjugate caps (MFCC) algorithm as described elsewhere.[42,54,55] In the spirit of the MFCC approach, the full density matrix of the protein can be assembled by a linear combination of fragment density matrices

$$P_{\mu\nu} = \sum_{i=1}^{N_f} P^f_{\mu\nu}(i) - \sum_{j=1}^{N_c} P^{cc}_{\mu\nu}(j) \qquad (9)$$

where $P^f_{\mu\nu}(i)$ is the density matrix element of the $i$th protein fragment and $P^{cc}_{\mu\nu}(j)$ is the density matrix element of the $j$th conjugate cap. $N_f$ and $N_c$ are the total number of the protein fragments and conjugate caps, respectively. The MFCC partition scheme to cut a protein into amino acid fragments and conjugate caps is shown in Figure 2. First, a series of single-point HF calculations on all the fragments and conjugate caps are performed; then the full density matrix of the protein obtained using the converged fragment density matrices based on eq 9 is taken as the initial guess for the subsequent divide and conquer HF calculations. All ab initio calculations were implemented in an in-house-developed quantum chemistry package QUICK.[56]

***Table 2.*** Converged Total Energies (au) (at the HF/6-31G* level) Using Two Different Subsetting Schemes: Residue Based with a Buffer of 5 Å and Atom Based with a Buffer of 7 Å[a]

| system | residue-centric core region | atom-centric core region | deviation (kcal mol$^{-1}$) |
|---|---|---|---|
| Gly$_{10}$ | −2314.783296 | −2314.783272 | −0.015 |
| Gly$_{20}$ | −4382.595749 | −4382.595726 | −0.014 |
| Gly$_{30}$ | −6450.407962 | −6450.407938 | −0.015 |
| Gly$_{40}$ | −8518.221662 | −8518.221679 | 0.011 |
| α-(Ala)$_{20}$ | −5164.086850 | −5164.086911 | 0.038 |
| α-(Ala)$_{30}$ | −7622.660188 | −7622.660373 | 0.116 |
| α-(Ala)$_{40}$ | −10081.238571 | −10081.238839 | 0.168 |
| MUD | | | 0.054 |

[a] MUD: mean unsigned deviation.

## Results and Discussion

**Accuracy and Timing Comparisons.** In this section we assess the DC-HF approach performance by performing calculations on two types of simple systems: extended polyglycine (gly)$_n$ and an α-helix of polyalanine (α -(ala)$_n$, see Figure 3). All calculations discussed here use the 6-31G* basis set. A buffer radius of $R_b$ = 5.0 Å was adopted for all DC-HF calculations. Within this definition we include all the residues that contain any atoms within 5 Å from the core region as part of the buffer region. A comparison of the CPU time required to solve the SCF equations on the extended polyglycine (gly)$_n$ ($n$ = 6−40) using the standard HF and DC-HF approaches is shown in Figure 4. As expected, the computational scale for the DC-HF diagonalization calculation is O(N), in contrast to O($N^{2.9}$) for the traditional HF SCF diagonalization on the full Fock matrix of the entire system. Moreover, since the polyglycine is extended, the basis set cross-over point is between 485 and 749. Figure 5 shows the deviation of DC-HF energy compared to the full system calculation on extended polyglycine systems. The error becomes larger as the size of the system increases; however, all of the deviations are within 0.04 kcal mol$^{-1}$. This good accuracy suggests that we can employ the divide and conquer scheme to study large, 3-dimensional systems. The computational cost and accuracy of DC-HF for α-(ala)$_n$ ($n$ = 10−40) systems are illustrated in Figures 6 and 7, respectively. Because of the helix structure, each subsystem contains a larger number of residues than in the extended system using the same buffer size. As illustrated in Figure 6, the cross-over point is around 1789, which is over 2 times larger than for the polyglycine example. Each DC-HF diagonalization SCF cycle in this example scales as $O(N^{1.1})$, in contrast to $O(N^{2.7})$ for the traditional HF diagonalization cost. Furthermore, the total energy errors for the α-helical polyalanines are slightly larger than those for the extended polyglycine systems (see Figure 7), but they are still in a good agreement with the full system calculations since the largest error is less than 0.7 kcal mol$^{-1}$ for α-(ala)$_{40}$.

In the current DC-HF approach, the scale for the computation of the Coulomb matrix is still $O(N^2)$ after prescreening

***Table 3.*** Total Energies (au) of Three-Dimensional Globular Proteins Obtained Using Standard Full System HF/6-31G* Calculations and Divide and Conquer HF/6-31G* Calculations Using the MFCC Initial Guess[a]

| system | number of atoms | number of basis functions | standard full system calculation | DC using MFCC initial guess | deviation (kcal mol$^{-1}$) |
|---|---|---|---|---|---|
| Trp-cage | 304 | 2610 | −7439.721780 | −7439.722124 | −0.22 |
| 1VTP | 396 | 3418 | −10014.756053 | −10014.755741[b] | 0.20 |
| 1BBA | 582 | 5033 | −15103.299186 | −15103.302595 | −2.14 |
| 1AML | 598 | 5178 | −15140.895905 | −15140.897305[b] | −0.88 |
| 1BHI | 591 | 5124 | −15989.697592 | −15989.696544 | 0.66 |
| 1BZG | 573 | 4851 | −13680.602670 | −13680.602916[b] | −0.15 |
| 2JPK | 589 | 5000 | −13854.809422 | −13854.810188[b] | −0.48 |
| 2KCF | 576 | 4991 | −14599.178617 | −14599.180118 | −0.94 |
| 2PPZ | 608 | 5111 | −14957.602116 | −14957.605696 | −2.25 |
| 2RLK | 588 | 5089 | −14589.701015 | −14589.702771[b] | −1.10 |
| 2YSC | 578 | 5108 | −14634.254517 | −14634.257181 | −1.67 |
| MUD | | | | | 0.97 |

[a] MUD: mean unsigned deviation. [b] Did not converge using the SAD initial guess.

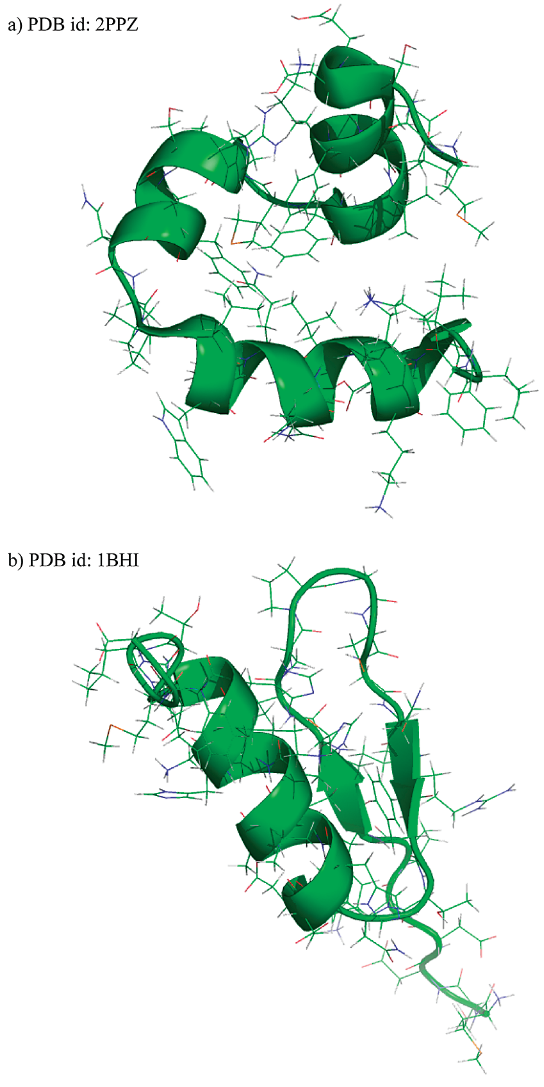


***Figure 8.*** Two representative three-dimensional protein structures studied in this work.

the two-electron integrals.[10] When we apply eq 2 to construct the subsystem Fock matrix, the long-range Coulomb interactions between the local subsystem and distant atoms cannot be circumvented; thus, it should be emphasized that the DC algorithm itself does not reduce the scale of Coulomb and exchange matrix evaluations, and other approaches are necessary to achieve this result (e.g., CFMM).[14,16,17,49]

**MFCC Initial Guess for DC-HF Calculations.** Next, we compare the number of SCF cycles necessary to reach convergence when the SAD (superposition of atomic densities) and MFCC initial guesses are used in the divide and conquer scheme using the 6-31G* basis set (see Table 1). The convergence criterion in all examples was set to $10^{-6}$ au on the root-mean-squared change of the density matrix elements and $10^{-4}$ au for the maximum change of the density matrix elements. Nakai and co-workers[35] and Shaw and St-Amant[32] noted that DIIS causes SCF calculations to oscillate at the final stage of the SCF convergence process due to the slight errors introduced by the DC approximation for assembling the density matrix (see eq 3). In our HF DC calculations, the DIIS technique was turned off when the root-mean-squared change of the density matrix elements reaches $10^{-4}$ au. We also found that although DIIS works in the early stages of the SCF procedure, we get the best performance when only two previous Fock matrices were stored in the DIIS calculations. One can see from Table 1 that the HF DC calculations usually require more SCF cycles than the non-DC runs; however, for the polyglycine and polyalanine systems, the MFCC initial guess helps to reduce the number of SCF cycles in both DC and non-DC cases.

**Residue-Centric Core Region versus Atom-Centric Core Region.** Previously, all calculations used a residue-based definition for the core region. We also examined an atom-based subsetting strategy for the core region in polyglycines and polyalanines. One can see from Table 2 that the converged total energies using the atom-centric core region were almost identical to those using a residue-based cutoff. Indeed, the overall mean unsigned deviation is as low as 0.054 kcal mol$^{-1}$. This is an attractive aspect of the divide and conquer approach since it allows for parallelization at the atom level rather than at the much larger reside-based cutoff scheme.

**Validation on Three-Dimensional Protein Systems.** No previous studies have utilized the divide and conquer HF approach on three-dimensional globular proteins. In order to address this point, we validated the accuracy of divide and conquer HF/6-31G* calculations on 11 real proteins. The systems ranged from 304 to 608 atoms and are listed in Table 3. The proteins consisted of α-helical structures (see Figure 8a) or are mixed α−β structures (see Figure 8b). As shown in Table 3, the largest deviation is 2.25 kcal mol$^{-1}$ and the overall mean unsigned deviation is only 0.97 kcal mol$^{-1}$ compared to standard full system calculations. Importantly, the observed error is larger than what was observed for the one-dimensional examples but is still within acceptable limits. This study sets the stage for the wide application of divide and conquer calculations on real protein systems. Furthermore, we found that for five proteins the divide and conquer HF calculations are not able to reach convergence using the simple SAD initial guess while all cases converged using the MFCC initial guess. Therefore, we conclude that the quality of the initial guess plays an important role in ensuring the convergence of divide and conquer calculations. Fragment-based electron density provides a much better quality initial guess with linear-scaling computational cost and, ultimately, much less computational time when compared to full system calculations.

## Conclusions

In this study, divide and conquer HF theory was revisited in order to examine its potential to study three-dimensional constructs and a new and effective initial guess scheme was introduced. We first validated the accuracy of the divide and conquer HF/6-31G* calculations on 11 three-dimensional globular proteins. The overall mean unsigned error was within 1 kcal mol$^{-1}$ when compared to standard full system calculations. Furthermore, we found that the fragment-based initial guess using the MFCC approach reduces the number of SCF cycles for polyglycine and polyalanine systems. Moreover, the MFCC initial guess facilitated SCF convergence for several of the globular proteins, where the SAD initial guess was unable to yield a converged wave function.

**Acknowledgment.** We thank the NIH (GM GM044974) for financial support of this research. Computing support from the University of Florida High Performance Computing Center is gratefully acknowledged.

## References

(1) Szabo, A.; Ostlund, N. S. *Modern quantum chemistry: introduction to advanced electronic structure theory*, 1st ed.; McGraw-Hill: New York, 1989.

(2) Parr, R. G.; Yang, W. T. *Annu. Rev. Phys. Chem.* **1995**, *46*, 701.

(3) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 0618.

(4) Bartlett, R. J.; Musial, M. *Rev. Mod. Phys.* **2007**, *79*, 291.

(5) Čížek, J. *J. Chem. Phys.* **1966**, *45*, 4256.

(6) Crawford, T. D.; Schaefer, H. F. *Rev. Comput. Chem.* **2000**, *14*, 33.

(7) Kállay, M.; Gauss, J. *J. Chem. Phys.* **2005**, *123*, 214105.

(8) Kállay, M.; Surján, P. R. *J. Chem. Phys.* **2001**, *115*, 2945.

(9) Bomble, Y. J.; Stanton, J. F.; Kállay, M.; Gauss, J. *J. Chem. Phys.* **2005**, *123*, 054101.

(10) Strout, D. L.; Scuseria, G. E. *J. Chem. Phys.* **1995**, *102*, 8448.

(11) Schwegler, E.; Challacombe, M. *J. Chem. Phys.* **1996**, *105*, 2726.

(12) Goedecker, S. *Rev. Mod. Phys.* **1999**, *71*, 1085.

(13) Fedorov, D. G.; Kitaura, K. *J. Phys. Chem. A* **2007**, *111*, 6904.

(14) Challacombe, M.; Schwegler, E. *J. Chem. Phys.* **1997**, *106*, 5526.

(15) Friesner, R. A.; Murphy, R. B.; Beachy, M. D.; Ringnalda, M. N.; Pollard, W. T.; Dunietz, B. D.; Cao, Y. X. *J. Phys. Chem. A* **1999**, *103*, 1913.

(16) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chem. Phys. Lett.* **1994**, *230*, 8.

(17) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chem. Phys. Lett.* **1996**, *253*, 268.

(18) Scuseria, G. E. *J. Phys. Chem. A* **1999**, *103*, 4782.

(19) Korchowiec, J.; Lewandowski, J.; Makowski, M.; Gu, F. L.; Aoki, Y. *J. Comput. Chem.* **2009**, *30*, 2515.

(20) Jiang, N.; Ma, J.; Jiang, Y. S. *J. Chem. Phys.* **2006**, *124*, 114112.

(21) Daniels, A. D.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 1321.

(22) Yang, W. T. *Phys. Rev. Lett.* **1991**, *66*, 1438.

(23) Yang, W. T.; Lee, T. S. *J. Chem. Phys.* **1995**, *103*, 5674.

(24) Kohn, W. *Phys. Rev. Lett.* **1996**, *76*, 3168.

(25) Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Science* **1996**, *271*, 51.

(26) Exner, T. E.; Mezey, P. G. *J. Phys. Chem. A* **2002**, *106*, 11791.

(27) Fusti-Molnar, L. *J. Chem. Phys.* **2003**, *119*, 11080.

(28) Fusti-Molnar, L.; Pulay, P. *J. Chem. Phys.* **2002**, *117*, 7827.

(29) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; DiStasio, R. A.; 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. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C. P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L.; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; Schaefer, H. F.; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172.

(30) Dixon, S. L.; Merz, K. M. *J. Chem. Phys.* **1996**, *104*, 6643.

(31) Gogonea, V.; Westerhoff, L. M.; Merz, K. M. *J. Chem. Phys.* **2000**, *113*, 5604.

(32) Shaw, D. M.; St-Amant, A. *J. Theor. Comput. Chem.* **2004**, *3*, 419.

(33) Kobayashi, M.; Nakai, H. *Int. J. Quantum Chem.* **2009**, *109*, 2227.

(34) Akama, T.; Fujii, A.; Kobayashi, M.; Nakai, H. *Mol. Phys.* **2007**, *105*, 2799.

(35) Akama, T.; Kobayashi, M.; Nakai, H. *J. Comput. Chem.* **2007**, *28*, 2003.

(36) Kobayashi, M.; Akama, T.; Nakai, H. *J. Chem. Phys.* **2006**, *125*, 204106.

(37) Exner, T. E.; Mezey, P. G. *J. Comput. Chem.* **2003**, *24*, 1980.

(38) Exner, T. E.; Mezey, P. G. *J. Phys. Chem. A* **2004**, *108*, 4301.

(39) Exner, T. E.; Mezey, P. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 4061.

(40) Nakano, T.; Kaminuma, T.; Sato, T.; Fukuzawa, K.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. *Chem. Phys. Lett.* **2002**, *351*, 475.

(41) Fedorov, D. G.; Kitaura, K. *Chem. Phys. Lett.* **2006**, *433*, 182.

(42) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599.

(43) He, X.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 031103.

(44) Fedorov, D. G.; Ishimura, K.; Ishida, T.; Kitaura, K.; Pulay, P.; Nagase, S. *J. Comput. Chem.* **2007**, *28*, 1476.

(45) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2005**, *123*, 134103.

(46) Kobayashi, M.; Imamura, Y.; Nakai, H. *J. Chem. Phys.* **2007**, *127*, 074103.

(47) Kobayashi, M.; Nakai, H. *J. Chem. Phys.* **2008**, *129*, 044103.

(48) Dixon, S. L.; Merz, K. M. *J. Chem. Phys.* **1997**, *107*, 879.

(49) Schwegler, E.; Challacombe, M. *J. Chem. Phys.* **1999**, *111*, 6223.

(50) Burant, J. C.; Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Chem. Phys. Lett.* **1996**, *248*, 43.

(51) Shao, Y. H.; White, C. A.; Head-Gordon, M. *J. Chem. Phys.* **2001**, *114*, 6572.

(52) Ochsenfeld, C. *Chem. Phys. Lett.* **2000**, *327*, 216.

(53) Ochsenfeld, C.; White, C. A.; Head-Gordon, M. *J. Chem. Phys.* **1998**, *109*, 1663.

(54) Chen, X. H.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**, *125*, 044903.

(55) Chen, X. H.; Zhang, Y. K.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 184105.

(56) He, X.; Ayers, K.; Brothers, E.; Merz, K. M. *QUICK*; University of Florida: Gainesville, FL, 2008.

CT9006635

JCTC Journal of Chemical Theory and Computation

# $HO_2 + O_3$ Reaction: Ab Initio Study and Implications in Atmospheric Chemistry

Luís P. Viegas and António J. C. Varandas*

*Departamento de Química, Universidade de Coimbra, 3004-535 Coimbra, Portugal*

Received July 17, 2009

**Abstract:** We report a theoretical investigation on the reaction between ozone and the hydroperoxyl radical, which is part of the ozone depletion cycle. This reaction represents a great challenge to the state of the art ab initio methods, while its mechanism remains unclear to both experimentalists and theoreticians. In this work we calculated the relative energies of the stationary points along the reaction coordinate of the oxygen- and hydrogen-abstraction mechanisms using different levels of theory and extrapolating some of the results to the complete one-electron basis set limit. Oxygen abstraction is shown to be preceded by formation of hydrogen-bonded complexes, while hydrogen abstraction shows a lower energy barrier than oxygen abstraction. Both mechanisms lead to formation of $HO_3 + O_2$ in a very troublesome region of the potential-energy surface that is not correctly described by single-reference methods. The implications of the results on reaction dynamics are discussed.

## 1. Introduction

The reaction of hydroperoxyl radical with ozone is the rate-limiting step in the natural cycle for ozone depletion:[1]

$$HO_2 + O_3 \xrightarrow{k_1} OH + 2O_2 \tag{1}$$

$$HO + O_3 \xrightarrow{k_2} HO_2 + O_2 \tag{2}$$

$$\text{net: } 2O_3 \rightarrow 3O_2 \tag{3}$$

This process is mostly active in the lower stratosphere over much of the globe, and it is believed to be responsible for approximately one-half of the global ozone loss in this atmospheric layer.[2] Knowledge of $k_1$ is therefore of crucial importance in order to calculate ozone profiles in the stratosphere and to make reliable modeling of atmospheric ozone phenomena.[3-6]

Not surprisingly, the title reaction became much studied experimentally over the years so as to determine $k_1$ and unravel the mechanistic details of eq 1.[7-15] An intriguing aspect of the title reaction is the positive curvature at low temperatures reported for the Arrhenius plot ln $k_1(T)$ vs $T^{-1}$. Such a curvature is often due to competition of different reaction channels, which led Sinha et al.[11] and Nelson and Zahniser[13] to carry out a mechanistic study of reaction 1 using isotopic labeling of the oxygen atoms. Their sophisticated experimental work suggests that the title reaction proceeds via two distinct mechanisms:

$$H^{18}O_2 + {}^{16}O_3 \rightarrow {}^{18}OH + {}^{18}O^{16}O + {}^{16}O_2 \tag{4}$$

$$\rightarrow {}^{16}OH + {}^{18}O^{18}O + {}^{16}O_2 \tag{5}$$

Thus, in mechanism 4 $^{18}$OH production occurs via oxygen abstraction from ozone, whereas $^{16}$OH formation in eq 5 would imply ozone hydrogen abstraction. In order to decide about the two mechanisms, Nelson and Zahniser[13] measured the $^{16}$OH/$^{18}$OH product branching ratio over the temperature range $226 \le T/\mathrm{K} \le 355$. The results have shown that both $^{16}$OH and $^{18}$OH are formed, with their analysis suggesting that hydrogen abstraction by ozone accounts for 88% of the reactive encounters, with this fraction increasing with decreasing temperature to ~95% at 226 K. Such an observation supports an earlier estimate of (75 ± 10)% by Sinha et al.[11] Nelson and Zahniser[13] further concluded that "the barrier to oxygen abstraction by ozone exceeds that for hydrogen abstraction by 1 ± 0.4 kcal mol$^{-1}$".

* Corresponding author phone/fax: +351-239-835867; e-mail: varandas@qtvs1.qui.uc.pt.

10.1021/ct900370q © 2010 American Chemical Society
Published on Web 01/22/2010

In spite of the relative abundance of experimental work, theoretical studies of reaction 1 are rather scarce in the literature. In fact, only three papers have been published so far concerning the study of $HO_2 + O_3$ on theoretical grounds.[16−18] The first one, by Varandas and Zhang,[16] presented a dynamics study of the title reaction using a test potential-energy surface (PES) for ground-state $HO_5(^2A)$ based on the double many-body expansion (DMBE) method.[19−21] In this work, they showed that the saddle-point structure representing ozone being attacked by $HO_2$ from the O end obtained from the $HO_5$ DMBE-4B (DMBE PES ignoring $n \geq 5$ terms) was similar to the structure obtained by performing exploratory complete active space self-consistent field (CASSCF) calculations considering 11 electrons and 11 active molecular orbitals and employing a 6-311G(2df) one-electron basis set, briefly CASSCF(11,11)/6-311G(2df). A six-body term was then added to the DMBE-4B PES so that the dynamics calculations would reproduce within error bars the recommended value of the rate constant for reaction 1 at room temperature. From these calculations, the authors predicted reaction 1 to occur almost exclusively via oxygen abstraction. This contrasts with the conclusions extracted from isotopically labeled experiments,[11,13] and they tentatively attributed this discrepancy to the presence of fast isotopic scrambling reactions (which are supposed to be absent from the mentioned experimental work), namely

$$^{18}OH(v) + {}^{16}O_3 \rightarrow H^{18}O^{16}O + {}^{16}O_2 \qquad (6)$$

where $^{18}OH$ is vibrationally excited. More recently, two theoretical studies on the title reaction have been reported. One of them, by Mansergas and Anglada,[17] focuses on the gas-phase hydrogen-bonded species (these and other intermediates will be hereinafter referred to as complexes) formed between $HO_2$ and $O_3$. The authors account for six stationary points, all energetically below the separated reactants: three minima interconnected by three saddle points. These stationary points were calculated by performing geometry optimizations at the CASSCF($m$, $n$)/6-311+G(2df,2p) level of theory (CASSCF(19,15) and CASSCF(15,13) for the minimum structures and CASSCF(11,10) for the saddle points). The reported minima and one of the saddle points were further optimized with the QCISD approach[22] using the same basis set, while the stability of such stationary points was calculated by performing single-point energy calculations at the geometries predicted at the CCSD(T)/aug-cc-pVTZ level of theory. This was considered by the authors as their best level of theoretical treatment and can be indicated by CCSD(T)/aug-cc-pVTZ//QCISD/6-311+G(2df,2p). These energies were then subtracted from the energies of the reactants calculated with the same level of theory and with basis set superposition error (BSSE) corrections according to the counterpoise method of Boys and Bernardi.[23] The hydrogen-bonded complexes were found do be stable by no more than ∼3.7 kcal $mol^{-1}$ and are suspected to be formed in the early stages of the mechanism of reaction 1 as a prereactive complex. For a review on the importance of the formation of radical−molecule complexes, see ref 24 and references therein. The latest study on the $HO_2 + O_3$ reaction is from Xu and Lin,[18] where the authors investigate the mechanism and kinetics of the title reaction. The geometries of the stationary points were optimized at the spin-unrestricted BH&HLYP/6-311++G(2df,2p) level of theory, while the relative energies were obtained with the G2M(CC2) method.[25] Two saddle points energetically above the reactants were reported: the first corresponds to the barrier to oxygen abstraction and has a similar structure to the one previously reported by Varandas and Zhang;[16] the second represents the barrier to hydrogen abstraction, which is reported by the authors to be 1 kcal $mol^{-1}$ below the former saddle point. This energy difference was found to be consistent with the one suggested in previous experimental work.[11,13]

In this work we performed a theoretical study of the title reaction, analyzing both its oxygen- and hydrogen-abstraction mechanisms. Such a task was performed by carefully mapping, connecting, and calculating the relative energies of the relevant stationary points of the $HO_5(^2A)$ PES. This theoretical work is part of an ongoing study that intends to clarify the mechanistic details of the $HO_2 + O_3$ reaction and to improve the DMBE $HO_5(^2A)$ PES[16] for future dynamics studies. The next two sections present the theory and computational methods used in this work and the results and discussion, respectively. They are both divided in the same way for an easy understanding and correspondence between the problematics addressed in each subsection. The last section presents the conclusions and a brief discussion about future work in the title system.

## 2. Theory and Computational Methods

The theoretical study of the title reaction consisted of two main steps: (1) finding the first-order saddle points of the oxygen- and hydrogen-abstraction mechanisms and performing intrinsic reaction coordinate (IRC) calculations in order to identify which minima they connect; (2) calculating the relative energies between the stationary points obtained in the previous step. Each of these steps carries its own difficulties and will be addressed separately. All calculations have been performed in the ground state of $HO_5(^2A)$ with the GAMESS[26] and molpro[27] packages. The MacMolPlt[28] graphical user interface was used for visualization of the geometric and electronic features of the different stationary points.

**2.1. Geometry Optimizations.** The choice of an electronic structure method applied to the mapping of a PES, especially in reaction pathways where one is looking for first-order saddle points, must be made with great care. At such geometries, often associated with bond breaking, it is natural to encounter wave functions with an increasing degree of multireference character. At such regions of the PES, the use of single-reference methods is often problematic. On the other hand, the use of a multireference approach may require expertise and be computationally expensive. Thus, one needs some pragmatism in choosing a method capable of being flexible to correctly describe the PES regions of interest while being computationally affordable.

In this study, all geometry optimizations and IRC calculations have been done with the CASSCF method, which we believe to be the method that best fits the requirements stated above. Following previous work,[16] the orbital space consists of 15 core orbitals, 11 active orbitals, and 11 active electrons. This active space has been chosen with the automatic procedure of Pulay and Hamilton,[29] who suggested that the unrestricted Hartree−Fock (UHF) natural orbitals can be used as a good starting point for the CASSCF calculations. The active space should then contain the fractionally occupied UHF natural orbitals, which in this work have been defined as the ones with occupation numbers between 1.998 and 0.002. Additionally, with the goal of getting more accurate geometries, we improved the basis set used before:[16] p-type polarization functions have been added to the hydrogen atom while diffuse functions were added to both the hydrogen and oxygen atoms. The geometry optimizations have then been carried out at the CASSCF(11,11)/6-311++G(2df,2p) level of theory without imposing constraints. The nature of each stationary point has been examined via analysis of the harmonic vibrational frequencies.

**2.2. Relative Energies.** The calculation of the relative energies of the stationary points of $HO_5(^2A)$ is a delicate subject. Generally speaking, at some regions of the PES, the wave function may have a high degree of single-reference nature while at regions of bond forming or breaking the wave function can be expected to have some degree of multireference character. In this case, one requires to account for both the dynamical and the nondynamical electron correlation effects. For small systems, the MRCI or even FCI approaches can be used to accurately map out the entire PES, but for larger systems, such as the one presented in this work, those methods can be prohibitive. One possibility is to use multireference perturbation theory (MRPT), although this method may not be free from problems.[30−32] The difficulties imposed by the title system and the lack of sufficient related theoretical work in the literature lead us to test and analyze several ab initio methods in the study of reaction 1. The different methods have all been utilized as implemented in the molpro[27] package for electronic structure calculations: CASSCF,[33−37] CASPT2,[38−41] CCSD,[42−44] and CCSD(T),[45,46] with the single-reference methods using the restricted open-shell Hartree−Fock (ROHF) method. We further employed density functional theory (DFT) in the study of such a reaction. As for choosing the BH&HLYP functional, two reasons may be advanced: first, the same functional has been utilized in previous work,[18] and hence, the published results can offer data for comparison; second, some preliminary tests have shown that the BH&HLYP functional yields accurate results. The BH&HLYP calculations have been carried out with the GAMESS[26] package.

The augmented correlated-consistent polarized valence *X*-tuple zeta (aug-cc-pV*X*Z or simply AV*X*Z) and the augmented correlated-consistent polarized core−valence *X*-tuple zeta (aug-cc-pCV*X*Z or ACV*X*Z) basis sets,[47−49] with $X = D, T, Q$ have been employed in all single-point energy calculations. Built in a manner that is intended to relate the correlation energy to the cardinal number *X* in a systematic way, such basis sets have prompted the search for laws to extrapolate the correlation energy to the complete one-electron basis set (CBS) limit[50−54] at $X = \infty$.

In this work, we applied the CBS extrapolation procedure to the coupled-cluster energies. For this, the electronic energy is first split as

$$E_X = E_X^{\mathrm{HF}} + E_X^{\mathrm{cor}} \tag{7}$$

To treat the uncorrelated Hartree−Fock energies, the two-point extrapolation formula recommended by Karton and Martin[55] has been utilized. For the AV(*T*, *Q*)Z pair, it assumes the form

$$E_X^{\mathrm{HF}} = E_\infty^{\mathrm{HF}} + B/X^{5.34} \tag{8}$$

In turn, the correlation energy employed the newly developed uniform singlet-pair and triplet-pair extrapolation (USTE) method,[56] which can be cast into the form

$$E_X^{\mathrm{cor}} = E_\infty^{\mathrm{cor}} + A_3 Y \tag{9}$$

with *Y* being defined by

$$Y = (X + \alpha)^{-3}\left[1 + \frac{A_5}{A_3}(X + \alpha)^{-2}\right] \tag{10}$$

Equation 9 is also a two-point extrapolation formula with parameters $E_\infty^{\mathrm{cor}}$ and $A_3$. The numerical values of the parameters in eq 10 are dependent on the ab initio method and can be obtained by consulting Table 1 of ref 56.

The energetics of the stationary points have been referred to the reactants of the oxygen-abstraction reaction ($\mathbf{R_O}$), $HO_2 + O_3$, which has been assumed as a supermolecule where the fragments are separated by 150 Å.

## 3. Results and Discussion

**3.1. Geometry Optimizations.** The calculated stationary points for the oxygen-abstraction mechanism can be seen in Figure 1.

We started our calculations by optimizing $\mathbf{SP_1}$. The geometry of the saddle point is in almost perfect agreement with the one found before,[16] the only relevant difference being the three dihedral angles which are now slightly bigger in absolute value (maximum difference is ∼14°). The O−O forming bond is predicted to be 1.935 Å, 0.231 Å larger than the one reported by Xu and Lin.[18] The associated imaginary frequency is 557i cm$^{-1}$, and IRC calculations in the direction of the reactants show that this saddle point is linked to a minimum structure which we refer as $\mathbf{MIN_1}$, corresponding to the structure **C1** reported by Mansergas and Anglada.[17] The distance from the hydrogen atom to the nearest oxygen from ozone in $\mathbf{MIN_1}$ is 2.404 Å, in excellent agreement with the value of 2.393 Å reported for **C1**, obtained at the CASSCF(19,15)/6-311+G(2df,2p) level of theory. A similar structure is obtained by Xu and Lin[18] (**LM1**), with the O−H distance being somewhat different, namely, 2.195 Å. However, these authors report **LM1** as a minimum that is connected to the hydrogen-abstraction saddle point. According to Mansergas and Anglada,[17] the two isomers of **C1** are linked by a saddle point (**TS1**), and indeed, we have confirmed this by obtaining the $\mathbf{SP_2}$ saddle point and running IRC calculations. The calculated imaginary

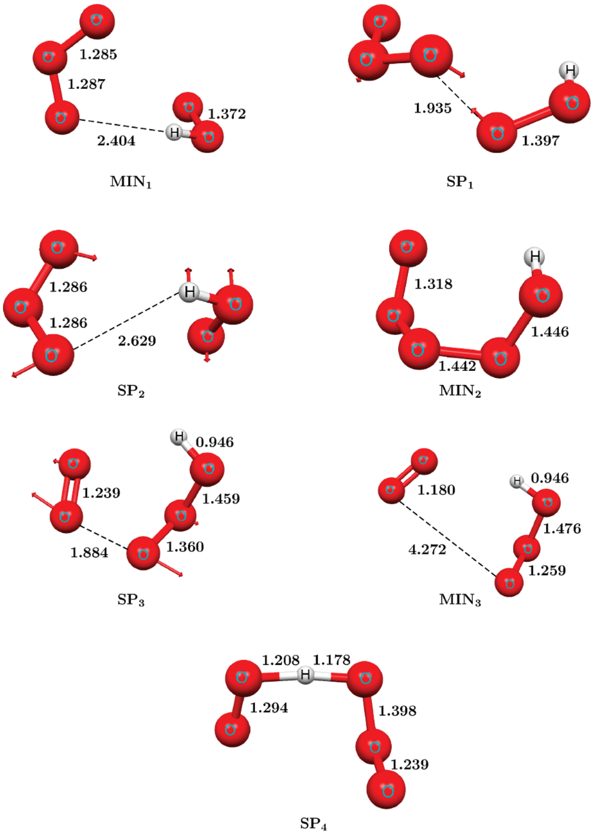


***Figure 1.*** Geometries of the stationary points of the oxygen- and hydrogen-abstraction mechanisms for the title reaction. The calculations were performed on the $HO_5$ doublet state PES and optimized at the CASSCF(11,11)/6-311++G(2df,2p) level of theory. The arrows show the vector displacements associated to the corresponding imaginary frequencies, except for $\mathbf{SP_4}$, which has the associated vector displacements hidden by the representation of the bonds between H and O. Distances are in Angstroms.

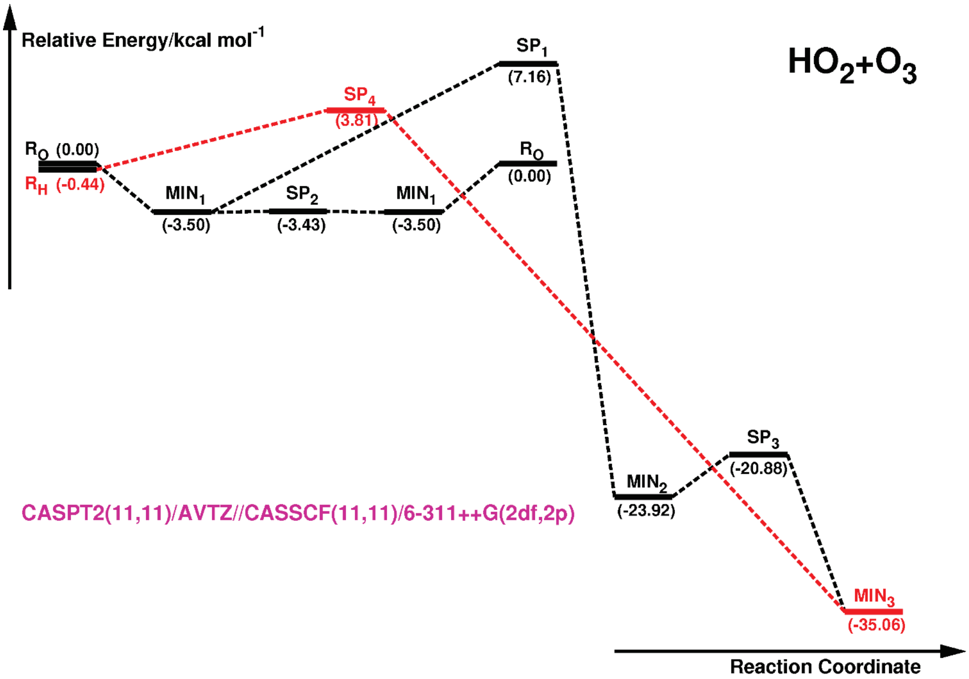
frequency is 42i cm$^{-1}$, and its geometry is very similar to the **TS1** structure, with a distance of 2.629 Å between the hydrogen atom and both terminal oxygen atoms of ozone. However, despite the fact that the calculations are well converged (calculations at DFT/BH&HLYP level carried out in the present work also predict a saddle point but with a frequency of 75 cm$^{-1}$), the low frequency of $\mathbf{SP_2}$ indicates that the PES is rather flat; thus, the possibility that it is artificial cannot be entirely ruled out. Such a $\mathbf{SP_2}$ structure resembles the one encountered for ground-state $HO_3$[57,58] and $HO_4$[59,60] before the complexes break to form products. The minimum $\mathbf{MIN_1}$ is therefore linked to both $\mathbf{SP_1}$ and $\mathbf{SP_2}$ saddle points (in the eventuality of $\mathbf{SP_2}$ being in fact a saddle point), which is a novel though somewhat questionable result as far as this mechanism is concerned. The details can be seen in Scheme 1. By following the vector displacements associated with the imaginary frequency of $\mathbf{SP_1}$ through IRC calculations, we conclude that this saddle point is linked to another minimum of the PES, $\mathbf{MIN_2}$, where a bond of 1.442 Å is now formed between one of the terminal atoms of ozone and the hydroperoxyl radical. This is a significant difference from the results of Xu and Lin,[18] since they do not find a minimum structure resulting from their oxygen-abstraction process. Instead, they obtain $HO_3 + O_2$ which is known[57,61−63] to easily fragment to $HO + 2O_2$. Our calculations show that $\mathbf{MIN_2}$ is connected to the $\mathbf{SP_3}$ saddle point, which represents the energy barrier to overcome so that reaction can proceed to $HO_3 + O_2$. This saddle point has an imaginary frequency of 333i cm$^{-1}$, while the breaking O−O bond distance is 1.884 Å. Only then the reaction path proceeds to a minimum structure ($\mathbf{MIN_3}$) where $O_2$ and $HO_3$ are well separated, with the O−O bond distance being 4.272 Å.

The calculated stationary points for the hydrogen-abstraction mechanism are represented at the bottom of Figure 1. As in the previous case, we started the calculations by optimizing the first-order saddle point, $\mathbf{SP_4}$, which now represents the attack of the hydroperoxyl radical to the ozone molecule from the H end. The optimized geometry shows similarities to the one found by Xu and Lin,[18] for example, the hydrogen atom is separated from the two closest oxygen atoms by 1.208 and 1.178 Å, while Xu and Lin[18] predict 1.063 and 1.356 Å, respectively. The imaginary frequency of this saddle point is 4227i cm$^{-1}$, as indicated in Table 1, together with the remaining imaginary frequencies calculated in this work and in the previous theoretical studies of $HO_5(^2A)$. The different imaginary frequencies of $\mathbf{SP_4}$ deserve a comment. Although this work and ref 18 refer to the same type of motion of the hydrogen atom (switching between the $O_2$ and the $O_3$ fragments), the directional characteristics of the atoms are different as the norms of the direction vectors of almost all oxygen atoms in ref 18 are bigger than the ones here reported. In $\mathbf{SP_4}$, practically only the two oxygen atoms closest to the hydrogen atom are displaced in the vibrational motion, making the imaginary frequency obtained in the present work resemble more the one of "pure" OH stretching in the diatomic, thus with a larger (imaginary) value.

Note that all frequencies have been calculated without scaling, which is known to generally lower the ab initio frequencies, bringing them to a closer agreement with the experimental counterparts. In the particular case of the CASSCF calculations in $HO_5$, an increase in the active space would also lower the frequencies, especially the one associated with the OH stretching.[17]

The IRC path in the products direction calculated in the present study is in close agreement with previous work.[18] The calculations show that $\mathbf{SP_4}$ is also connected to $\mathbf{MIN_3}$, which is similar to **LM2** obtained by Xu and Lin[18] (see Figure 1 for more details). However, the IRC results from the present work disagree in the backward direction as they show that, when distorting away from the saddle point, the ozone and hydroperoxyl radical fragments break away without forming any kind of complex, while in ref 18 it yields the **LM1** complex mentioned previously. In summary, both the oxygen- and hydrogen-abstraction mechanisms share the fact that at some point of the reaction coordinate the $O_2$ molecule separates from the $HO_3$ fragment. Because of this and the fact that the minimum energy path of the oxygen-abstraction mechanism of the $HO_5(^2A)$ DMBE PES[16] shows a similar behavior, the optimizations have been terminated at this stage of the reaction coordinate. It should be stressed that the emphasis has been on the $\mathbf{SP_1}$ and $\mathbf{SP_4}$ saddle points

***Scheme 1.*** Schematic Diagram Showing the Energetics of the Title Reaction[a]



[a] The energies (in kcal mol$^{-1}$) are relative to the reactants of the oxygen-abstraction mechanism calculated at the CASPT2(11,11)/AV*TZ*//CASSCF(11,11)/6-311++G(2df,2p) level. The red lines connect the structures belonging to the hydrogen-abstraction mechanism, while the black lines do the same for the oxygen-abstraction mechanism.

***Table 1.*** Comparison between the Imaginary Frequencies, in cm$^{-1}$, of All Saddle Points Obtained in This Work and the Ones Obtained in Previous Theoretical Studies

| | ref 16 | ref 17 | ref 18 | this work |
|---|---|---|---|---|
| **SP$_1$** | 573[a] | | 630[b]/741[c] | 557 |
| **SP$_2$** | | 117[d] | | 42 |
| **SP$_3$** | | | | 333 |
| **SP$_4$** | | | 2089[b]/2927[c] | 4227 |

[a] CASSCF(11,11)/6-311G(2df). [b] BH&HLYP/6-311++G(2df,2p). [c] MP2/6-311++G(2df,2p). [d] CASSCF(11,10)/6-311+G(2df,2p).

***Table 2.*** Electronic Energies, in kcal mol$^{-1}$, of the Different Stationary Points Relative to the Reactants (R$_O$) at Different Levels of Theory

| method | MIN$_1$ | SP$_1$ | SP$_2$ | MIN$_2$ | SP$_3$ | MIN$_3$ |
|---|---|---|---|---|---|---|
| CASSCF(11,11)/AV*TZ* | −1.48 | 11.28 | −1.46 | 8.24 | 16.77 | −6.81 |
| CASPT2(11,11)/AV*TZ* | −3.50 | 7.16 | −3.43 | −23.92 | −20.88 | −35.06 |
| CCSD/AV*TZ* | −3.40 | 15.30 | −3.37 | −11.22 | 12.24 | −10.38 |
| CCSD/ACV*TZ* | −3.32 | 15.66 | −3.30 | −11.13 | −3.94 | −10.86 |
| CCSD(T)/AV*TZ* | −3.74 | 11.97 | −3.73 | −5.53 | 17.60 | −1.58 |
| CCSD(T)/ACV*TZ* | −3.66 | 12.22 | −3.66 | −5.41 | −15.36 | −1.93 |
| BH&HLYP/AV*TZ* | −2.51 | 15.26 | −2.41 | −16.90 | −10.90 | −15.81 |

with a view to improve the HO$_5$($^2$A) DMBE PES,[16] since they play a key role in dynamics calculations.

**3.2. Relative Energies.** We begin our analysis by addressing the first part of the oxygen-abstraction mechanism, in which the formation of the hydrogen-bonded complexes takes part. By inspecting Table 2 we can see that all calculations place **SP$_2$** above **MIN$_1$**, except the CCSD(T)/ACV*TZ* one, which places the two stationary points at the same height. The stability of **MIN$_1$** is in good agreement with the best level of theoretical treatment reported by Mansergas and Anglada[17] for **C1**, −3.68 kcal mol$^{-1}$, and in reasonable agreement with the results of Xu and Lin[18] for **LM1**, −2.5 kcal mol$^{-1}$. The CASSCF results deviate the most from the ones obtained with the other methods, but this is to be expected, since a CASSCF calculation lacks dynamical correlation. Inspection of the CI coefficients of the CASSCF wave function of the reactants, **MIN$_1$** and **SP$_2$** also indicates that they mainly have single-reference character, and therefore, the use of an ab initio method that only promotes excitations from the Hartree−Fock determinant is, in principle, a good approximation. This is also the case with **MIN$_2$**, where the single-reference character of the CASSCF wave function is even higher than the previous cases. The USTE extrapolation method was then used along with the CCSD and CCSD(T) ab initio methods to obtain more accurate energies for these stationary points and also with the purpose of benchmarking this method. Note that the extrapolation to the CBS limit should also minimize the BSSE.[64] The results can be seen in Tables 3 and 4. The isomerization barrier, Δ$E^{\rm iso}$, increases as the CBS limit is reached and never surpasses 0.05 kcal mol$^{-1}$. Note that all coupled-cluster calculations employing double-ζ basis sets

***Table 3.*** Electronic Energies, in kcal mol$^{-1}$, of Selected Stationary Points Relative to the Reactants (R$_O$) with the CCSD and CCSD(T) Methods Using the AV*XZ* Basis Sets[a]

| method | MIN$_1$ | SP$_2$ | MIN$_2$ | Δ$E^{\rm iso}$ |
|---|---|---|---|---|
| CCSD/AV*DZ* | −3.70 | −3.71 | −9.47 | −0.01 |
| CCSD/AV*TZ* | −3.40 | −3.37 | −11.22 | 0.03 |
| CCSD/CBS | −3.23 | −3.18 | −11.11 | 0.05 |
| CCSD(T)/AV*DZ* | −4.06 | −4.09 | −4.27 | −0.03 |
| CCSD(T)/AV*TZ* | −3.74 | −3.73 | −5.53 | 0.01 |
| CCSD(T)/CBS | −3.56 | −3.53 | −5.09 | 0.03 |

[a] In the CBS calculation, the Hartree–Fock value was extrapolated from the (*T*, *Q*) pair and the correlation energy was extrapolated from the (*D*, *T*) pair, i.e., $E_\infty = E_\infty^{\rm HF}(T, Q)$ + USTE-(*D*, *T*). Δ$E^{\rm iso}$ is the barrier of isomerization, *E* (SP$_2$) – *E*(MIN$_1$).

***Table 4.*** Electronic Energies, in kcal mol$^{-1}$, of Selected Stationary Points Relative to the Reactants ($R_O$) with the CCSD and CCSD(T) Methods Using the ACV*X*Z Basis Sets[a]

| method | $MIN_1$ | $SP_2$ | $MIN_2$ | $\Delta E^{iso}$ |
|---|---|---|---|---|
| CCSD/ACV*D*Z | –3.79 | –3.80 | –9.87 | –0.01 |
| CCSD/ACV*T*Z | –3.32 | –3.30 | –11.13 | 0.02 |
| CCSD/CBS | –3.07 | –3.03 | –10.79 | 0.04 |
| CCSD(T)/ACV*D*Z | –4.15 | –4.18 | –4.68 | –0.03 |
| CCSD(T)/ACV*T*Z | –3.66 | –3.66 | –5.41 | 0.00 |
| CCSD(T)/CBS | –3.39 | –3.38 | –4.70 | 0.01 |

[a] In the CBS calculation, the Hartree–Fock value was extrapolated from the (*T*, *Q*) pair and the correlation energy was extrapolated from the (*D*, *T*) pair, i.e., $E_\infty = E_\infty^{HF}(T, Q)$ + USTE-(*D*, *T*). $\Delta E^{iso}$ is the barrier of isomerization, $E$ ($SP_2$) – $E$($MIN_1$).

***Table 5.*** $T_1$ and $D_1$ Diagnostics Obtained After CCSD Calculations of Selected Geometries with the AV*T*Z and ACV*T*Z Basis Sets

| | AV*T*Z | | | ACV*T*Z | | |
|---|---|---|---|---|---|---|
| geometry | $T_1$ | $D_1$ | $T_1/D_1$ | $T_1$ | $D_1$ | $T_1/D_1$ |
| **$R_O$** | 0.0325 | 0.1257 | 0.2584 | 0.0280 | 0.1237 | 0.2261 |
| **$MIN_1$** | 0.0330 | 0.1283 | 0.2573 | 0.0284 | 0.1262 | 0.2250 |
| **$SP_2$** | 0.0330 | 0.1283 | 0.2571 | 0.0284 | 0.1262 | 0.2248 |
| **$MIN_2$** | 0.0325 | 0.1370 | 0.2370 | 0.0278 | 0.1347 | 0.2065 |

place **$SP_2$** slightly below **$MIN_1$**, which is reflected on the negative sign of $\Delta E^{iso}$, indicating their possible inadequacy in this rather flat region of the PES. The CBS results for $\Delta E^{iso}$ are in close agreement with the ones obtained with CASSCF and CASPT2 using the AV*T*Z basis set: 0.02 and 0.07 kcal mol$^{-1}$, respectively. The BH&HLYP results place **$MIN_1$** and **$SP_2$** approximately 1 kcal mol$^{-1}$ above the other calculations, coinciding with **LM1**,[18] having $\Delta E^{iso} = 0.1$ kcal mol$^{-1}$. It should be emphasized that the small value of $\Delta E^{iso}$ in conjunction with the low imaginary frequency of **$SP_2$** casts doubt on the true existence of this stationary point. All of our $\Delta E^{iso}$ results are below the ones obtained by Mansergas and Anglada,[17] 1.26 or 1.33 kcal mol$^{-1}$, depending on the active space used for the geometry optimization of the hydrogen-bonded minimum. However, besides using a different basis set and active space for the geometry optimizations, they also use a different level of theory for the single-point energy calculations of this barrier: CASPT2(19,15)/6-311+G(2df,2p).

The second part of this mechanism comprises the attack of the hydroperoxyl radical to the ozone molecule from the O end, which is represented by **$SP_1$**. The weight of the Hartree−Fock determinant on the wave function decreases to 65% at this geometry as indicated by the square of the corresponding coefficient in the determinantal expansion of the CI wave function. The coupled-cluster and DFT energies are at least about 4 kcal mol$^{-1}$ above the CASPT2 energies, the latter being considered the most accurate energies for this point as they include the effect of the several determinants with considerable weight in the electronic wave function. The multireference character of **$SP_1$** can also be seen in the significant increase of the number and value of the $t_2$ cluster amplitudes for this stationary point. The CCSD and CCSD(T) results for **$MIN_2$** seem surprising at first, especially because at the Hartree−Fock level this point is 22.61 kcal mol$^{-1}$ below the reactants. This result is in very good agreement with the CASPT2 calculation, leading us to think that there is some kind of problem in the balance between the coupled-cluster energies of the reactants and **$MIN_2$**. To better understand these coupled-cluster energies, we investigated the $T_1$ and $D_1$ diagnostics[65−67] of the reactants, **$MIN_1$**, **$SP_2$**, and **$MIN_2$**, results which are presented in Table 5. For both basis sets, the $T_1$ diagnostics are identical between the four geometries, the same happening for the $D_1$ diagnostics, with the exception of **$MIN_2$**, which has a higher $D_1$ value. This has consequences in the $T_1/D_1$ ratio, which becomes smaller for **$MIN_2$**. According to Lee,[67] a value of $T_1/D_1$ which is much smaller than $1/\sqrt{2}$ "indicates that there is a large variation in orbital rotation parameters in the coupled-cluster wave function, or in other words, there are problem areas in the molecule and other areas where the coupled-cluster approach is performing better." Clearly, one finds two sets of ratios for both basis sets: the first set involves the reactants, **$MIN_1$** and **$SP_2$**, and the second set is composed only of **$MIN_2$**. This explains the excellent results obtained for the relative energies of **$MIN_1$** and **$SP_2$** with CCSD, since they are the result of performing differences between geometries belonging to the first set, where a favorable cancellation of errors occurs. The same does not hold for the relative energy of **$MIN_2$**, because it involves a difference between absolute energies of the two sets, where the $T_1/D_1$ ratio of **$MIN_2$** indicates more problem areas in this geometry. This unbalance in the coupled-cluster wave functions of the two sets leads to an increase of the relative energy of this point. This behavior is most likely a consequence of calculating the energy in a geometry optimized with a completely different electronic structure method. In fact, we observe that calculations carried out on a minimum geometry optimized with B3LYP rather than CASSCF yield energies of −17.80 and −24.32 kcal mol$^{-1}$, respectively, for CCSD and CCSD(T); cf. Table 2. In this case, the $T_1/D_1$ ratio increases, which is an indication that there are less problem areas in this new geometry. A similar problem occurs in the calculation of the perturbative corrections of connected triple excitations as it is known that the presence of large singles amplitudes can cause instability in this method. Again, it just so happens that the reactants, **$MIN_1$** and **$SP_2$**, have the same number of $t_1$ amplitudes above 0.05 with almost equal absolute values, contrasting with the **$MIN_2$** calculation which shows a large number of larger magnitude. The result is that the relative energy of **$MIN_2$** at the CCSD(T) level gets even higher because of this unbalanced treatment, and as already mentioned, the calculation of the relative energies of **$MIN_1$** and **$SP_2$** benefits from a cancellation of errors. The computation of coupled-cluster energies including quadruply excited clusters would help to clarify the extent of these considerations. Note that, as expected, Tables 3 and 4 show that the coupled-cluster calculations correlating all the orbitals with the ACV*X*Z basis set do not alter the results significantly.

The last two geometries obtained in this reaction coordinate are the **$SP_3$** and **$MIN_3$** stationary points, having less than 5% of single-reference character. The coupled-cluster results are thus completely untrustworthy. The BH&HLYP

***Table 6.*** Comparison between the Energies of the Common Stationary Points Calculated in This Work and in Ref 18[a]

| | $R_H$ | $MIN_1$ | $SP_1$ | $SP_4$ | $MIN_3$ |
|---|---|---|---|---|---|
| CASPT2(11,11)/AV*T*Z | −0.44 | −3.50 | 7.16 | 3.81 | −35.06 |
| BH&HLYP/AV*T*Z | −4.00 | −2.51 | 15.26 | 14.25 | −15.81 |
| ref 18 [G2M(CC2) method] | 0.00 | −2.50 | 5.00 | 4.00 | −36.40 |

[a] Energy units are in kcal $mol^{-1}$.

calculations are the only ones that can compete with the CASPT2 level of theory, but even in this case the results are still far away from the quality obtained with MRPT. The same reasoning holds for $\mathbf{SP_4}$, which is part of the hydrogen-abstraction mechanism. Table 6 shows a comparison between the energies of the common stationary points calculated in this work and in ref 18.

Note that the CASPT2 dissociation energy of the reactants in the hydrogen-abstraction mechanism ($\mathbf{R_H}$) is 0.44 kcal $mol^{-1}$ below its counterpart in the oxygen-abstraction mechanism ($\mathbf{R_O}$). Such a difference may be explained by the dissimilarity of the geometries of the fragments in each dissociation channel as a result of the different mechanisms involved and the fact that the products are not infinitely separated in practice. This difference rises to 4 kcal $mol^{-1}$ at the BH&HLYP level. Another important result is the confirmation of the lower energy of $\mathbf{SP_4}$ with respect to $\mathbf{SP_1}$ both in absolute as in relative terms. This means that there are effectively two competitive reaction channels, and their influence in the dynamics of reaction 1 should be investigated in the future. The G2M(CC2) energies used in ref 18 show an interesting agreement with the CASPT2 method, although it seems rather surprising that such an agreement occurs due to the strong multireference character of several stationary points.

It should be mentioned that new generations of the single-reference coupled-cluster methods are available[68−70] that can handle significant degrees of multireference character quite well while offering a high-level description of dynamical correlations. In fact, we tested one of the so-called completely renormalized coupled-cluster methods, CR-CC(2,3), for some geometries ($\mathbf{R_O}$, $\mathbf{MIN_1}$, $\mathbf{MIN_2}$, $\mathbf{SP_1}$, $\mathbf{SP_2}$) with the AV*D*Z basis set (the calculations employing V*T*Z or AV*T*Z basis sets are computationally too demanding), but no significantly different results have been found. For other geometries, which show a high degree of multireference character, we often had trouble in converging the lambda vector iterations.

## 4. Conclusions and Future Work

In this study, the oxygen- and hydrogen-abstraction mechanisms of the $HO_2 + O_3$ reaction have been computationally investigated with the CASSCF, CASPT2, CCSD, CCSD(T), and BH&HLYP theoretical methods. The optimizations were performed at the CASSCF(11,11)/6-311++G(2df,2p) level of theory, while the energetics of the title reaction was investigated with the previously mentioned methods.

Of the several saddle points located, two represent the attack of the hydroperoxyl radical to ozone from the O and H end. They are $\mathbf{SP_1}$ and $\mathbf{SP_4}$, respectively. Some stationary points of both mechanisms ($\mathbf{SP_3}$, $\mathbf{SP_4}$, $\mathbf{MIN_3}$, and, to a smaller extent, $\mathbf{SP_1}$) have a high degree of multireference character, leading to the expected failure of the single-reference methods used in this work, CCSD and CCSD(T). The best level of theoretical treatment is therefore assumed to be CASPT2. Conversely, the first part of the oxygen-abstraction mechanism, where the formation of the hydrogen-bonded complexes takes place, is mainly single-reference in nature. The agreement between all the ab initio methods in this region of the PES is therefore expected, with the exception of $\mathbf{MIN_2}$, for the reasons explained before. The connection of this part of the mechanism to the formation of $\mathbf{SP_1}$ is a novel result with possible implications in the reactions dynamics, since the energy barrier to the formation of $\mathbf{SP_1}$ is increased by a considerable amount and the barrier is likely to get narrower. The presence of the hydrogen-abstraction barrier, $\mathbf{SP_4}$, is also expected to have an impact on the dynamics, namely, lowering the value of $k_1$. This does not necessarily mean that a stronger agreement between theoretical and experimental interpretations will be observed, because of two main reasons. First, the open-chain structure of $\mathbf{MIN_2}$ suggests the possibility that OH will be formed vibrationally excited, therefore increasing $k_2$ from a factor of ~25 times[71,72] the value used by experimentalists to analyze their data. Second, reaction 6 will allow the isomerization reaction $H^{18}O^{16}O \leftrightarrow H^{16}O^{18}O$ to occur and, through reaction with ozone, formation of $H^{16}O$, which is expected by experimentalists to be formed only via hydrogen abstraction. These two reasons question not only the mechanistic interpretation given by experimentalists but also the barrier height for hydrogen abstraction that is considered to be 1 ± 0.4 kcal $mol^{-1}$ higher than the barrier for hydrogen abstraction.

Extrapolation of the coupled-cluster energies to the CBS limit was achieved through the USTE method. The improvements on the results are seen on the slight increase of $\Delta E^{iso}$, which is negative when the double-$\zeta$ basis sets are used, and becomes positive with the increase of the cardinal number *X*.

Overall, the CASPT2 method was revealed to be the only one with an accurate description of both dynamical and nondynamical correlation effects in the structures calculated in this work. As for the DFT results, the BH&HLYP functional performed a little worse than all methods in the single-reference regions of the PES (excluding the CASSCF results) and better than the coupled-cluster methods in the multireference regions of the PES. However, this functional did not achieve the quality of the CASPT2 results in the latter regions of $HO_5(^2A)$.

We hope that the present theoretical work contributes significantly to the understanding of the title reaction for both experimentalists and theoreticians. It would be interesting to extend it with a study of the effect of quadruply excited clusters in the calculations of coupled-cluster energies of the geometries with a high degree of single-reference character and also of $\mathbf{SP_1}$, which has still a considerable amount of single-reference character (65%). Furthermore, it would be interesting to thoroughly test different density functionals to compare the results with the CASPT2 energies obtained in this paper, since the use of DFT theory is simpler and computationally more affordable than MRPT, thus allowing a more extensive mapping of the PES. Such developments will be part of a planned improvement of the DMBE $HO_5(^2A)$ PES on which further dynamics studies will be carried out.

**Acknowledgment.** The authors acknowledge funding from Fundação para a Ciência e a Tecnologia, Portugal (contracts REEQ/128/QUI/2005 and SFRH/BPD/40807/2007).

**Supporting Information Available:** Coordinates, frequencies, and absolute energies for all computed structures. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Monks, P. S. *Chem. Soc. Rev.* **2005**, *34*, 376–395.

(2) Wennberg, P. O.; Cohen, R. C.; Stimpfle, R. M.; Koplow, J. P.; Anderson, J. G.; Salawitch, R. J.; Fahey, D. W.; Woodbridge, E. L.; Keim, E. R.; Gao, R. S.; Webster, C. R.; May, R. D.; Toohey, D. W.; Avallone, L. M.; Proffitt, M. H.; Loewenstein, M.; Podolske, J. R.; Chan, K. R.; Wofsy, S. C. *Science* **1994**, *266*, 398–404.

(3) Duewer, W. H.; Wuebbles, D. J.; Ellsaesser, H. W.; Chang, J. S. *J. Geophys. Res.* **1977**, *82*, 935–942.

(4) Crutzen, P. J.; Howard, C. J. *Pure Appl. Geophys.* **1978**, *116*, 497–510.

(5) Whitten, R. C.; Borucki, W. J.; Capone, L. A.; Turco, R. P. *Nature* **1978**, *275*, 523–524.

(6) Turco, R. P.; Whitten, R. C.; Poppoff, I. G.; Capone, L. A. *Nature* **1978**, *276*, 805–807.

(7) Simonaitis, R.; Heicklen, J. *J. Phys. Chem.* **1973**, *77*, 1932–1935.

(8) DeMore, W. B.; Tschuikow-Roux, E. *J. Phys. Chem.* **1974**, *78*, 1447–1451.

(9) Zahniser, M. S.; Howard, C. J. *J. Chem. Phys.* **1980**, *73*, 1620–1626.

(10) Manzanares, E. R.; Soto, M.; Lee, L. C. *J. Chem. Phys.* **1986**, *85*, 5027–5034.

(11) Sinha, A.; Lovejoy, E. R.; Howard, C. J. *J. Chem. Phys.* **1987**, *87*, 2122–2128.

(12) Wang, X.; Soto, M.; Lee, L. C. *J. Chem. Phys.* **1988**, *88*, 896–899.

(13) Nelson, D. D.; Zahniser, M. S. *J. Phys. Chem.* **1994**, *98*, 2101–2104.

(14) Nizkorodov, S. A.; Harper, W. W.; Blackmon, B. W.; Nesbitt, D. J. *J. Phys. Chem. A* **2000**, *104*, 3964–3973.

(15) Herndon, S. C.; Villalta, P. W.; Nelson, D. D.; Wayne, J. T.; Zahniser, M. S. *J. Phys. Chem. A* **2001**, *105*, 1583–1591.

(16) Varandas, A. J. C.; Zhang, L. *Chem. Phys. Lett.* **2004**, *385*, 409–416.

(17) Mansergas, A.; Anglada, J. M. *J. Phys. Chem. A* **2007**, *111*, 976–981.

(18) Xu, Z. F.; Lin, M. C. *Chem. Phys. Lett.* **2007**, *440*, 12–18.

(19) Varandas, A. J. C. *Adv. Chem. Phys.* **1988**, *74*, 255–338.

(20) Varandas, A. J. C. In *Lecture Notes in Chemistry*; Laganá, A., Riganelli, A., Eds.; Springer: Berlin, 2000; Vol. 75, pp 33−56.

(21) Varandas, A. J. C. In *Conical Intersections: Electronic Structure, Dynamics & Spectroscopy*; Domcke, W., Yarkony, D. R., Köppel, H., Eds.; Advanced Series in Physical Chemistry; World Scientific Publishing: Singapore, 2004; Vol. 15, pp 205−270.

(22) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968–5975.

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

(24) Hansen, J. C.; Francisco, J. S. *ChemPhysChem* **2002**, *3*, 833–840.

(25) Mebel, A. M.; Morokuma, K.; Lin, M. C. *J. Chem. Phys.* **1995**, *103*, 7414–7421.

(26) Schmidt, M. W.; Baldridge, K. K.; Boats, J. A.; Elbert, S. T.; Gorgon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J., Jr. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

(27) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M. Molpro, a package of ab initio programs, version 2008.1; see http://www.molpro.net.

(28) Bode, B. M.; Gordon, M. S. *J. Mol. Graphics Modell.* **1998**, *16*, 133–138.

(29) Pulay, P.; Hamilton, T. P. *J. Chem. Phys.* **1988**, *88*, 4926–4933.

(30) Rode, M. F.; Werner, H.-J. *Theor. Chem. Acc.* **2005**, *114*, 309–317.

(31) Cramer, C. J.; Włoch, M.; Piecuch, P.; Puzzarini, C.; Gagliardi, L. *J. Phys. Chem. A* **2006**, *110*, 1991–2004.

(32) Cramer, C. J.; Kinal, A.; Włoch, M.; Piecuch, P.; Gagliardi, L. *J. Phys. Chem. A* **2006**, *110*, 11557–11568.

(33) Roos, B. O.; Taylor, P. R.; Siegbahn, P. E. M. *Chem. Phys.* **1980**, *48*, 157–173.

(34) Ruedenberg, K.; Schmidt, M. W.; Gilbert, M. M.; Elbert, S. T. *Chem. Phys. Lett.* **1982**, *71*, 41–49.

(35) Roos, B. O.; Linse, P.; Siegbahn, P. E. M.; Blomberg, M. R. A. *Chem. Phys.* **1982**, *66*, 197–207.

(36) Werner, H.-J.; Knowles, P. J. *J. Chem. Phys.* **1985**, *82*, 5053–5063.

(37) Knowles, P. J.; Werner, H.-J. *Chem. Phys. Lett.* **1985**, *115*, 259–267.

(38) Wolinski, K.; Sellers, H. L.; Pulay, P. *Chem. Phys. Lett.* **1987**, *140*, 225–231.

(39) Wolinski, K.; Pulay, P. *J. Chem. Phys.* **1989**, *90*, 3647–3659.

(40) Werner, H.-J. *Mol. Phys.* **1996**, *89*, 645–661.

(41) Celani, P.; Werner, H.-J. *J. Chem. Phys.* **2000**, *112*, 5546–5557.

(42) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910–1918.

(43) Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1993**, *99*, 5219–5227.

(44) Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **2000**, *112*, 3106–3107.

(45) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.

(46) Watts, J. D.; Gauss, J.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 8718–8733.

(47) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(48) Kendall, R. A.; Dunning, T., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6769–6806.

(49) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1995**, *103*, 4572–4585.

(50) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.

(51) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olson, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243–252.

(52) Truhlar, D. G. *Chem. Phys. Lett.* **1998**, *294*, 45–48.

(53) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Olsen, J. *Chem. Phys. Lett.* **1999**, *302*, 437–446.

(54) Varandas, A. J. C. *J. Chem. Phys.* **2000**, *113*, 8880–8887.

(55) Karton, A.; Martin, J. M. L. *Theor. Chim. Acta* **2006**, *115*, 330–333.

(56) Varandas, A. J. C. *J. Chem. Phys.* **2007**, *126*, 244105.

(57) Yu, H. G.; Varandas, A. J. C. *Chem. Phys. Lett.* **2001**, *334*, 173–178.

(58) Varandas, A. J. C. *ChemPhysChem* **2005**, *6*, 453–465.

(59) Varandas, A. J. C.; Zhang, L. *Chem. Phys. Lett.* **2000**, *331*, 474–482.

(60) Mansergas, A.; Anglada, J. M. *ChemPhysChem* **2007**, *8*, 1534–1539.

(61) Mathisen, K. B.; Siegbahn, P. E. M. *Chem. Phys.* **1984**, *90*, 225–230.

(62) Chen, M. M. L.; Wetmore, R. W.; Schaefer, H. F., III *J. Chem. Phys.* **1981**, *74*, 2938–2944.

(63) Dupuis, M.; Fitzgerald, G.; Hammond, B.; Lester, W. A., Jr.; Schaefer, H. F., III *J. Chem. Phys.* **1986**, *84*, 2691–2697.

(64) Varandas, A. J. C. *Theor. Chem. Acc.* **2008**, *119*, 511–521.

(65) Lee, T. J.; Taylor, P. R. *Int. J. Quant. Chem. Symp.* **1989**, *23*, 199–207.

(66) Leininger, M. L.; Nielsen, I. M. B.; Crawford, T. D.; Janssen, C. L. *Chem. Phys. Lett.* **2000**, *328*, 431–436.

(67) Lee, T. J. *Chem. Phys. Lett.* **2003**, *372*, 362–367.

(68) Piecuch, P.; Kucharski, S. A.; Kowalski, K.; Musial, M. *Comput. Phys. Commun.* **2002**, *149*, 71–96.

(69) Piecuch, P.; Włoch, M. *J. Chem. Phys.* **2005**, *123*, 224105.

(70) Piecuch, P.; Włoch, M.; Varandas, A. J. C. In *Topics in the Theory of Chemical and Physical Systems*; Lahmar, S., Maruani, J., Wilson, S., Delgado-Barrio, G., Eds.; Progress in Theoretical Chemistry and Physics; Springer: Dordrecht, The Nederlands, 2007; Vol. 16, pp 63−121.

(71) Varandas, A. J. C. *ChemPhysChem* **2002**, *3*, 433–441.

(72) Zhang, L.; Varandas, A. J. C. *Phys. Chem. Chem. Phys.* **2001**, *3*, 1439–1445.

CT900370Q

JCTC Journal of Chemical Theory and Computation

# QM/MM Studies on the $\beta$-Galactosidase Catalytic Mechanism: Hydrolysis and Transglycosylation Reactions

Natércia F. Brás, Pedro A. Fernandes, and Maria J. Ramos*

*REQUIMTE, Departamento de Química, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal*

Received October 6, 2009

**Abstract:** Carbohydrates perform a wide range of crucial functions in biological systems and are of great interest for the food and pharmaceutical industries. $\beta$-Galactosidase from *Escherichia coli* catalyzes both the hydrolytic breaking of the very stable glycosidic bond of lactose and a series of transglycosylation reactions. These reactions are crucial for the development of new carbohydrate molecules, as well as the optimization of their syntheses. In this work we have used computational methods to study the catalytic mechanism of hydrolysis and a set of distinct transglycosylation reactions of a retaining galactosidase, with atomic detail, with lactose as the natural substrate. The ONIOM method (BB1K:AMBER//B3LYP:AMBER calculations) was employed to address such a large enzymatic system. Such a methodology can efficiently account for the stereochemistry of the reactive residues, as well as the long-range enzyme−substrate interactions. The possible importance of the magnesium ion in the catalytic reaction was investigated, and it was found that, indeed, the magnesium ion catalyzes the transformation, lowering the activation barrier by 14.9 kcal/mol. The calculations indicate that the formation of $\beta$(1−3) glycosidic linkages is thermodynamically very unfavorable. In contrast, the formation of $\beta$(1−6) glycosidic bonds is the most favored, in complete agreement with the enantioselectivity observed experimentally. The data also clearly show the importance of the enzyme scaffold beyond the first-shell amino acids in the stabilization of the transition states. It is fundamental to include the enzyme explicitly in computational studies.

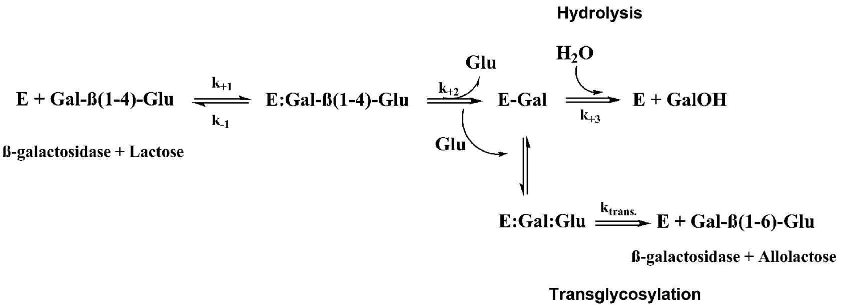
## 1. Introduction

Oligosaccharides play a large number of crucial functions in biological systems and have attracted the attention of the pharmaceutical industry due to their potential application as therapeutics.[1,2] *Escherichia coli* (*lacZ*) $\beta$-galactosidase (EC 3.2.1.23) catalyzes the hydrolysis and transglycosylation of $\beta$-D-galactosides.[3,4] Both the amino acid and nucleotide sequences have been determined. The enzyme is a homotetramer, each monomer weighing 116353 Da and having 1023 amino-acid residues displayed in five sequential domains, with an extended segment at the amino terminus. The monomers work independently. The three-dimensional structure of $\beta$-galactosidase shows that the active site is located in a deep pocket within a distorted "TIM" barrel. Divalent and monovalent cations are required for full catalytic efficiency. The $Mg^{2+}$ or $Mn^{2+}$ cations provide 5−100-fold activation depending on the substrate used, whereas the $Na^{+}$ or $K^{+}$ cations provide only 0.3−6-fold activation depending on the ion and substrate involved. The active site has two subsites: the first binding site is highly specific for the galactose moiety, whereas the second binding site lacks specificity.[4–6] This ambiguous specificity allows the binding of a wide variety of $\beta$-D-galactosides beyond the natural substrate lactose (Figure SI-1 in the Supporting Information). An example is X-Gal (5-bromo-4-chloro-3-indoyl-$\beta$-D-galactopyranoside), which is a substrate that incorporates a chromophore. With this substrate, the activity of the enzyme is easily recognized by a distinct change in color. The

* Corresponding author e-mail: mjramos@fc.up.pt.

10.1021/ct900530f © 2010 American Chemical Society
Published on Web 01/13/2010

***Scheme 1.*** General Scheme for the Catalytic Mechanism of $\beta$-Galactosidase and a Lactose Molecule



capacity of the $\beta$-galactosidase to cleave chromogenic substrates has clearly contributed to its usefulness as a tool of research in molecular biology.[3,6]

Glycosidic linkages are enzymatically hydrolyzed by one of the two major groups of glycosidases: retaining and inverting glycoside hydrolases.[7] $\beta$-Galactosidase from *E. coli* is a retaining glycosidase. Thus, it maintains the initial conformation on the anomeric carbon. The proposed catalytic mechanism of this enzyme is believed to occur through a double-displacement reaction involving galactosylation and degalactosylation steps, with the reaction probably proceeding through a covalent galactosyl−enzyme intermediate.[8,9] The active site contains two carboxylic acid residues, which have been identified as Glu461 (proton-donor residue) and Glu537 (catalytic nucleophile group).[6] These two amino acids are approximately 5.5 Å apart. It is well-known that more basic groups, such as glucose, require acid/base catalytic assistance for departure of the leaving group.[7] A magnesium ion is also found close to the active site, although the exact role of this ion in catalysis has always remained unclear.[6,10,11] Kinetic data support the possibility of Lewis catalysis promoted by the $Mg^{2+}$ ion, the possibility of Brönsted catalysis through protonation by the Glu461 residue, or eventually both. Crystallographic studies suggest that the latter hypothesis is the most probable to occur.[6] Simultaneously, Glu537 acts as a nucleophile, forming a covalent galactosyl−enzyme intermediate.

Experimental and computational works suggest that a hydrogen bond between the $C_2$ hydroxyl group and this nucleophilic residue, already present in the reactants, is shortened at the transition state and further catalyzes the reaction.[7,12] It has also been suggested on the basis of experimental observations and computational work that the enzyme has another strategy for stabilizing the transition state, which corresponds to substrate distortion from a chair to a half-chair conformation, promoted by substrate binding.[7,12] This first step is characterized by the departure of the glucose group. The second mechanistic step of the reaction is supposed to involve the attack of the covalent carbohydrate−enzyme intermediate by a water molecule (hydrolysis) or a sugar molecule (transglycosylation), concomitantly with or followed by the transfer of a proton from a water/sugar to the proton donor, in a reverse mode of the first step (Scheme 1). Some researchers consider that both transition states (TSs) have a mainly dissociative character.[7] The experimental measurements of α secondary deuterium kinetic isotope effects (αDKIEs) suggest that both transition states have substantial oxocarbenium ion character.[13,14] Furthermore, these studies suggest that the reaction rate of the degalactosylation step depends on the acceptor concentration, which indicates that this transition state somehow involves the acceptor molecule and, consequently, is not a pure $S_N1$ TS, as has been proposed for many other glycosidases.[6]

A large number of oligosaccharides can be synthesized by transglycosylation reactions, even though the yields obtained are typically low (5−20%), as the products are also substrates for the enzyme and undergo hydrolysis. These reactions are supposed to be kinetically controlled, with the yields of the synthesized oligosaccharides depending on the relative rates of the transglycosylation and hydrolysis reactions.[15] The stereoselectivity of the glycosidic bond is a difficulty associated with oligosaccharides synthesis. Experimental data suggest that the transglycosylation reaction catalyzed by $\beta$-galactosidase produces preferably allolactose [$\beta$-galactopyranosyl-(1−6)-α-glucopyranose] with a yield of ∼97%.[3,4,6] This is physiologically important to *E. coli*, as this molecule is the natural inducer for the *lac* operon, which is responsible for the $\beta$-galactosidase expression. In that case, it might be expected that, in the transglycosylation enrichment on $\beta$-(1−6), the new linkage would be created by a rotation of the glucose molecule immediately after the cleavage followed by the chemical reaction, without it ever leaving the binding pocket.

Other disaccharides can be formed at low levels, particularly when glucose is added to the reaction. Lactose and allolactose are hydrolyzed with equal efficiency ($k_{cat} = 60$ $s^{-1}$ and $K_m = 1$ mM for both molecules), whereas only allolactose is produced from the transglycosylation reaction. Furthermore, the values of $k_{cat}$ for allolactose production and hydrolysis are similar, but this balance can be altered by the pH and by the absence of magnesium ions.[6,16]

Mutagenesis studies performed with $\beta$-galactosidase show that Glu461, Glu537, Tyr503, Asn460, His357, His391, and His540 residues interact with specific hydroxyl groups of the substrate.[17−20] Juers et al. cocrystallized series of ligands bound to the active site of $\beta$-galactosidase, and in the X-ray structures, it was possible to confirm that all of the residues previously identified as catalytically important were indeed positioned in or near the active site.[6] These crystallographic structures show that there are two distinct manners for the binding of ligands, namely, "shallow" and "deep" modes. Phe601 and Trp999 are also considered to be important for the catalytic mechanism. The substrate initially binds in a shallow mode, characterized by stacking on Trp999, establishing interactions between the side chain of the tryptophan and the ring of the glucose group. Subsequently, the substrate moves into a deep mode, further into the active site. To allow this transition, the side chain of Phe601 rotates, and the 794−804 loop also rearranges.[6] In this conformation, after the first step, the access to the galactosyl−enzyme intermediate by other molecules becomes more restricted.

Our goal with this article is to present a set of calculations that completely elucidate and clarify with atomic-level detail the chemical events involved in the catalytic mechanism of *E. coli* $\beta$-galactosidase using lactose as the natural substrate. For that purpose, we have used a large enzyme model and employed the ONIOM method to deal with such a large system, namely, the BB1K:AMBER//B3LYP:AMBER methodology. This enzymatic model can efficiently account for the restrained mobility of the reactive residues, as well as the long-range enzyme−substrate interactions.

## 2. Theoretical Calculations

X-ray crystallographic structures of the enzyme $\beta$-galactosidase complexed with substrate analogues show that the substrate has to bind in a shallow mode. Subsequently, the substrate must move to the deep mode, that is, further into the active site, for catalysis to occur. The 1DP0 protein databank structure of $\beta$-galactosidase (at 1.7-Å resolution) was used as the starting point for all computational studies.[6] All water molecules were deleted, with the exception of some conserved waters near the active site that coordinate the magnesium and sodium ions. Hydrogen atoms were added using Insight II software,[21] with all residues in their physiological protonation state. The only exception was the proton-donor residue (Glu461), which was protonated. The geometry optimization of the protein was done in three stages to release the bad contacts in the crystallographic structure. In the first stage, only the hydrogen atoms were minimized; in the second stage, the backbone was minimized; and in the third and final stage, the entire system was minimized. About 1500 steps were used for each stage, with the first 500 steps performed using the steepest-descent algorithm and the remaining steps carried out using conjugate gradient.

A lactose molecule was initially docked into the structure of the optimized unligated $\beta$-galactosidase, mimicking the deep mode of binding. To do this, we used GOLD docking software[22] and ChemScore as the scoring function.[23] The program is based on a genetic algorithm that is used to place different ligand conformations in the protein binding site, recognized by a fitting-points strategy.

It is well-known that the hydrogen bridge established between the nucleophile and the 2′-OH group of the substrate is essential for the stabilization of the transition states structures. Therefore, this interaction is conserved among retaining glycoside hydrolases.[12,24] To constrain the docking solutions toward this particular pose, this bridge was maintained through a distance constraint (2.00−2.50 Å) between the two atoms involved. Furthermore, in many glycosidases, the presence of several aromatic residues (Trp, Phe, or Tyr) close to the active site provides the hydrophobic platform common to carbohydrate−protein interactions. Trp568 plays a similar role and displays a position and orientation that promote the packing of the galactose moiety in the deep binding mode. To obtain the correct position for binding, a distance constraint (3.40−3.70 Å) was also included between this aromatic residue and the galactose moiety of the docked lactose. Furthermore, in the docking method, we have taken into consideration whether the acid/base was close and directly positioned for the attack on the glycosidic oxygen. After analysis of all of the solutions obtained, the best docking solution was chosen as the starting structure for the subsequent molecular dynamics study to release the bad contacts in the structure. In the end, we used the final structure to design a large enzymatic model. All molecular dynamics simulations were performed with the parametrization adopted in Amber 8,[25] using the Amber 1999 force field (parm99) for proteins and the Glycam 2004 force field (Glycam-04 parameters) for carbohydrates.[26−28] In this simulation, an explicit solvent model with pre-equilibrated TIP3P water molecules was used, filling a truncated octahedral box with a minimum 12-Å distance between the box faces and any atom of the protein.[29] The complex structure was minimized in two stages. In the first stage, the protein was kept fixed, and only the positions of the water molecules and counterions were minimized. In the second stage, the full system was minimized. Subsequently, using the Langevin temperature equilibration scheme, a 20-ps molecular dynamics (MD) simulation at constant volume and with periodic boundary conditions was run starting from the optimized structures.[30,31] After this, 1 ns of MD simulation was performed. Langevin dynamics was used (collision frequency of 1.0 ps$^{-1}$) to control the temperature.[30,31] The simulation was carried out using the sander module, implemented in the Amber8 simulations package, with the Cornell force field.[32] Bond lengths involving hydrogens were constrained using the SHAKE algorithm,[33] and the equations of motion were integrated with a 2-fs time step using the Verlet leapfrog algorithm.

The QM/MM calculations performed to determine the enzyme−substrate potential energy surface (PES) were executed using Gaussian 03 software.[34] To perform the study, we used the final structure of the MD simulation, from which we cut a model including a 15-Å radius around the lactose molecule. The system was composed of a total of 2707 atoms. To explore the PES of the catalytic reaction, our system was divided into two layers, within the ONIOM formalism[35,36] as implemented in Gaussian 03. In geometry optimizations, the higher-level layer included almost the entire substrate and the side chains of Glu461 and Glu537 for a total of 49 atoms, and it was treated with density functional theory (DFT) at the unrestricted B3LYP/6-31G(d) level.[37−39] The rest of the system was treated at the molecular mechanics level with the parm99 and Glycam04 force fields. We further froze the positions of the atoms in the outer 5-Å shell of the entire system (Figure 1). For each reaction step, we performed a linear transit scan along the reaction coordinate with a step value of 0.05 Å to locate the geometry of the transition state. In the literature, this procedure is common for large models with QM/MM calculations.[40−42] All linear transit schemes obtained here were smooth and are included in the Supporting Information (Schemes SI-

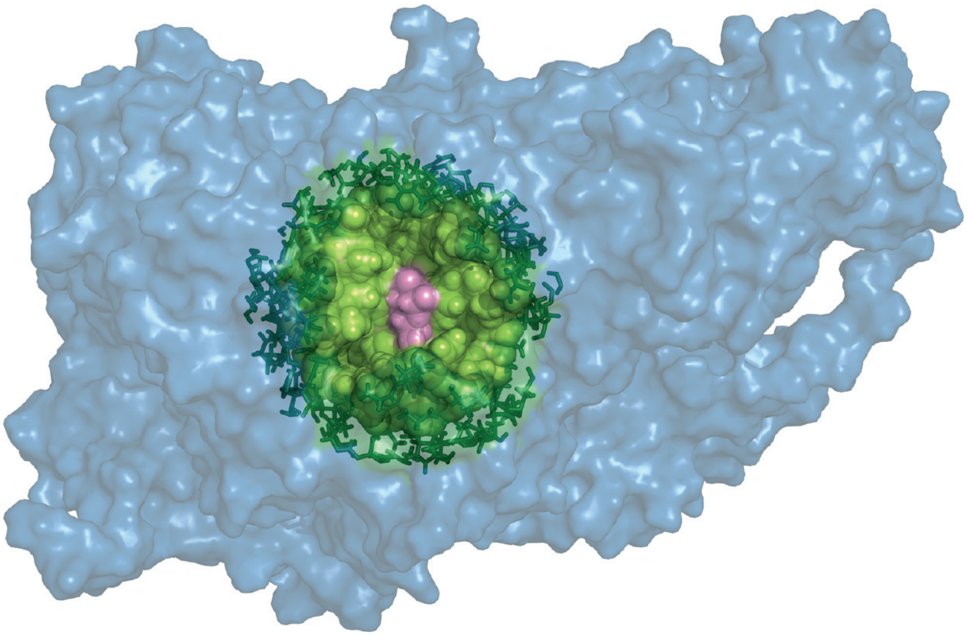


***Figure 1.*** Representation of the enzymatic model studied, which includes a 15-Å radius of the amino acids around lactose. The model system is shown in green. The frozen region (the outer 5-Å shell of the entire system) is represented as dark green sticks. The substrate is colored pink.

1−SI-5). As the interaction between the layers was described at the MM level, we had to recalculate the atomic point charges for all of the atoms involved in bond breaking/bond formation during the scan. The atomic point charges were taken from a Mulliken population analysis of the electronic density of the higher-level region.[43]

Single-point energy calculations were then performed on the optimized geometries, increasing the higher-level region to 168 atoms and treating this layer at the density functional theory level, with the BB1K functional[37,39,44] and the larger 6-311+G(2d,2p) basis set. We used this hybrid-meta density functional because it was shown to lead to good agreement in activation and reaction energies for this reaction, when compared to higher-level post-Hartree−Fock methods.[12,45] The level of theory used to obtain the geometry was lower than that used to calculate the energy, and even though this procedure might introduce some inaccuracies in the calculation, it is well-known that the energy is not very sensitive to the quality of the geometry, provided that the geometry has a good accuracy. This approach is implicit when someone calculates the energy with a larger basis set that the one used to optimize the geometry. This procedure has also been used in other enzymatic studies.[40–42]

## 3. Results and Discussion

It is well-known that $\beta$-galactosidase from *E. coli* belongs to the retaining glycosidases class, which catalyzes the hydrolysis of $\beta$-D-galactosides with retention of the same stereochemistry as the reactants. In this work, we try to understand the atomistic detail of the mechanism by which this enzyme catalyzes the hydrolysis of the extremely stable glycosidic bond, and to do so, we performed docking simulations, MD simulations, and QM/MM calculations. The mutated enzyme (E537Q), with the crystallographic structure available in the Protein Data Bank (PDB code 1JYN), binds lactose in the shallow mode.[6] To study the catalytic mechanism, we needed to have the substrate closer to the active site, that is, in the deep binding mode. For that purpose, we docked a lactose molecule farther inside the active site of the optimized unligated $\beta$-galactosidase (PDB code: 1DP0).[4] As mentioned, the aromatic Trp568 residue close to the active site provides a hydrophobic platform and displays a position and orientation that promote the packing of the galactosyl ring of the lactose in a favorable position for binding. Consequently, a distance constraint (3.40−3.70 Å) was included between this residue and the docked galactose moiety.[24] Furthermore, the crucial hydrogen bridge between the nucleophile and the $C_2$−OH group of the galactosyl ring was also considered in the docking protocol, with a distance constraint (2.00−2.50 Å) between the two atoms involved.[12] Subsequently, we performed an MD simulation to release the bad contacts in the structure, and in the end, we used the final structure to design a large model that included a radius with 15 Å of protein around the lactose substrate for the subsequent QM/MM calculations, as seen in Figure 1.

**3.1. Reactants.** Analyzing the reactants' structures, one can see that the galactose moiety stacks with Trp568 and the galactosyl hydroxyl groups make specific contacts with the enzyme and with the water molecules coordinated to the $Mg^{2+}$ ion. The glucose moiety is stacked against the side chain of Trp999. Figure 2 shows that the galactosyl 2′-OH group establishes hydrogen bridges with the Glu537 nucleophilic carboxylate (1.71 Å) and with the Asn460 amine (1.94 Å). One can see also that the 3′-OH group is hydrogen-bonded to one $Mg^{2+}$-bound water molecule and to the His391 side chain (1.85 Å), the 4′-OH group is H-bound to the Asp201 side-chain carboxylate (1.66 Å), and the 6′-OH group interacts with one water molecule and, intramolecularly, with the 3′-HO group of the glucosyl molecule (2.11 Å). This last glucosyl hydroxyl group is also hydrogen-bonded to the

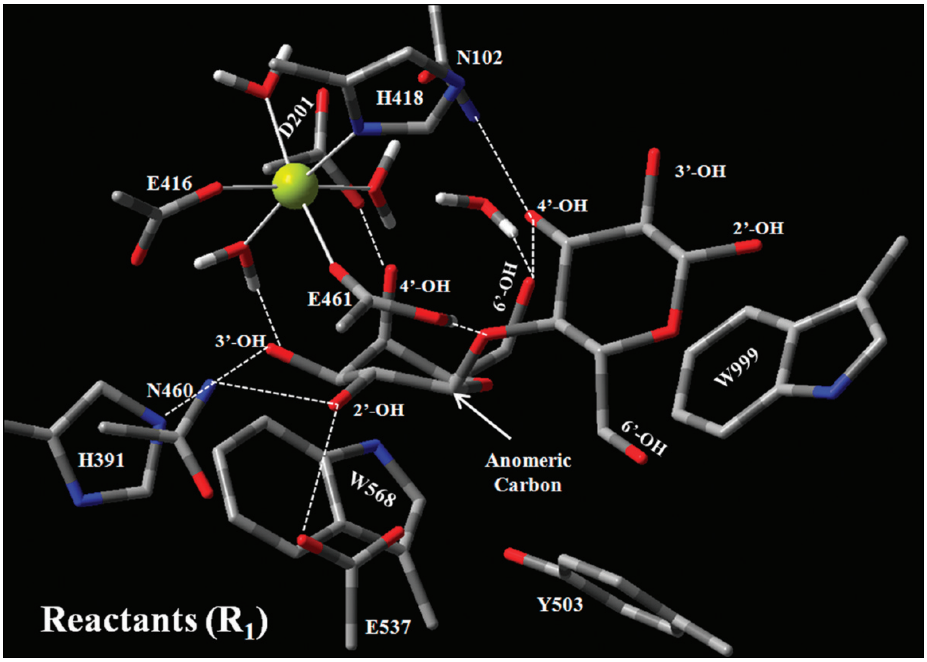


***Figure 2.*** Representation of the optimized structure of the reactants (R$_1$).

Asn102 amine group (2.15 Å). The glycosidic oxygen establishes a hydrogen bridge (1.67 Å) with the carboxylate group of the catalytic acid/base Glu461, which is protonated. In summary, a complex network of hydrogen interactions with the Asn102, Asp201, His391, Asn460, Glu461, and Glu537 residues determines the binding pose of lactose. These are complemented by hydrophobic interactions between the sugars rings and the side chains of Trp568 and Trp999. This situation is in sharp contrast with the glucosyl moiety, which contains five hydroxyl groups but establishes only one hydrogen interaction with the enzyme residues.

Additionally, the subsite for the glucosyl moiety is significantly larger than the substrate, suggesting that the latter has significant freedom to move. Therefore, one can understand why the second binding subsite is less specific for the bound substrate, which is in agreement with other studies.[6] Experimental data suggest that the glucose subsite is somewhat ill-defined, being specified mostly by a stacking interaction with the aromatic Trp999.[46] These facts are consistent with the behavior of the enzyme, which can bind several aglycon groups.

One can also observe how close the docking/molecular simulations placed the substrate to the magnesium ion. The bivalent ion is hexacoordinated, coordinating to Glu416, His418, Glu461, and three water molecules, with an octahedral geometry. Contrarily, the sodium ion is distant from the substrate, at a distance from the glycosidic oxygen of ca. 8 Å. This ion presents a tetrahedral geometry, coordinated to Asp201, Asn604, and two water molecules. It is worth noticing that all of these water molecules that already existed in the X-ray crystallographic structure were conserved throughout the entire MD simulation.

The antiperiplanar lone-pair hypothesis (ALPH) is a stereoelectronic concept that requires the glycosidic bond to be antiperiplanar in relation to a lone pair of electrons of the oxygen atom of the ring in order to achieve the transition state (TS).[47,48] Furthermore, it is well-known that a conformational change of the glycosidic bond into an equatorial orientation leads to a more planar ring structure, facilitating direct in-line nucleophilic attack.[48] Interestingly, in this $\beta$-galactosidase, one can see that a substrate distortion and rotation is necessary to maximize interactions and to fit the active-site pocket. According to the Cremer–Pople polar coordinates[49] that describe sugar conformations, the bound galactosyl ring is almost in a $^4H_3$ conformation. This pretransition state (pre-TS) helps reduce the number of steric clashes between the substrate and the enzyme. Juers et al. proposed that, when the lactose moves from the shallow mode to the deeper mode of binding, an enzyme conformational modification is required, namely, the rotation of the lateral side chain of Phe601.[6] As we started with the lactose molecule bound inside the active site of the enzyme (in the deep mode), no substantial alterations were observed in this lateral side chain, probably because this rotation is only necessary for the movement of the substrate from the shallow to the deeper mode of binding. Subsequently, the side chain should return to the original rotamer.

**3.2. First Mechanistic Step.** The first step of this catalytic mechanism involves a cleavage of the glycosidic bond of the lactose molecule, as well as the formation of the covalent galactosyl–enzyme intermediate. Given that we used a glucose molecule as the leaving group, acid catalysis is required for this mechanistic step to occur. Various experimental studies on this mechanism have been performed, but the results remain controversial.[6] A possibility, supported by kinetic data, proposes that the reaction is promoted by Lewis catalysis by $Mg^{2+}$, in which the breakage of the glycosidic bond is facilitated by a direct $Mg^{2+}$ electrophilic attack on the glycosidic oxygen, leading to an Mg–OR complex.[6] On the other hand, the kinetic data are also consistent with Brönsted catalysis, in which the glycosidic bond cleavage is promoted by proton donation to the

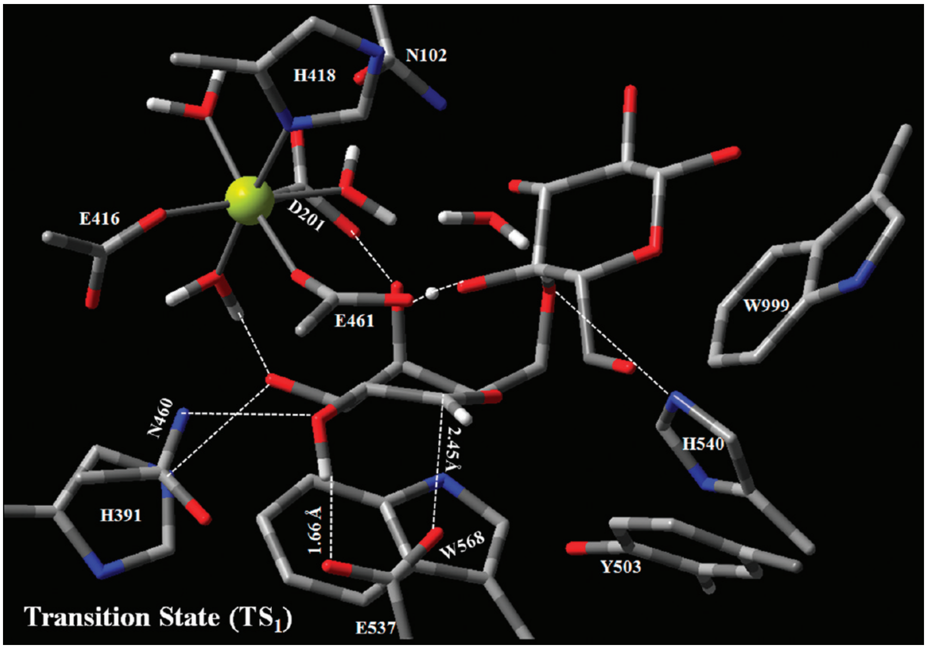


***Figure 3.*** Representation of the optimized structure of the transition state ($TS_1$).

glycosidic oxygen by Glu461.[10,11] Taking into account the positions of the magnesium ion, Glu461, and the glycosidic oxygen in the crystallographic structures of various ligands bound to the active site of $\beta$-galactosidase, catalysis promoted by the proton donor is clearly the most plausible. However, the importance of the $Mg^{2+}$ ion for the enzymatic activity of this galactosidase is well-known. Therefore, its precise function remains unclear.[50,51] In our calculations, we tried to clarify these doubts, as well as understand the role of the magnesium ion in the catalytic mechanism. The reaction coordinate adopted was the glycosidic bond length, that is, the distance between the anomeric carbon and the glycosidic oxygen. Analyzing the transition state for the first mechanistic step ($TS_1$), which emerges from such a PES, one can see that the breaking of the glycosidic bond occurs with the release of the glucose group. For the reactants, the length of the glycosidic bond is 1.47 Å, increasing to 2.25 Å in $TS_1$; the bond is almost broken at this stage. Although the proton of the acidic residue comes closer to the glycosidic oxygen, it has not been transferred yet, having a hydrogen-bond distance of 1.46 Å. These results are not identical to the calculations performed with a simple general model for this reaction using the DFT level of calculation, because in that small model, the proton was transferred from the acidic residue to the glycosidic oxygen in the transition state.[12] The difference arises from the greater stabilization of the glycosyl oxygen anion by the enzyme scaffold, which plays a clear catalytic role. Only in the products does the Glu461 residue donate its proton to the glycosidic oxygen, with a distance of 1.03 Å between these two atoms. In the same step, the nucleophilic group comes closer to the anomeric carbon, with one of its oxygens attacking the anomeric carbon. The distance between these two atoms is 3.01 Å in the reactants ($R_1$), evolving to 2.45 Å in the transition state ($TS_1$) and to 1.53 Å in the products of this step ($P_1$), in which a covalent bond is established. Therefore, at the end of this step, a covalent galactosyl−enzyme intermediate is formed. These results confirm the predicted dissociative nature of this transition state, namely, that the glycosidic bond is already broken and the bond to the nucleophilic group is far from established.[7,12]

The transition-state structures developed in the catalytic mechanism are stabilized by both intrinsic electronic effects and binding effects. A great component of catalysis in most glycosidases derives from noncovalent enzyme/substrate interactions that are established along the reaction pathway. Our calculations show that the structure of $TS_1$ is determined by the specific contacts established between some residues and hydroxyl groups 2, 3, 4, and 6 of the galactose species. The 2′-OH group interacts with the Glu537 nucleophilic carboxylate (1.66 Å) and the Asn460 amine groups (1.87 Å); the 3′-OH interacts with one $Mg^{2+}$ bound water molecule (1.73 Å) and with the His391 side chain (1.87 Å). The 4′-OH group establishes a hydrogen bridge with the Asp201 side-chain carboxylate group (1.62 Å), whereas the 6′-OH group interacts with one water molecule (1.72 Å) and with the His540 side chain (2.92 Å). Some of these hydrogen bridges are very short, supporting their importance in the stabilization of the transition state and explaining the catalytic effect predicted by experimental findings.[7,52] As one can see in Figures 2 and 3, all of these hydrogen interactions are shorter in the transition state than in the reactant structure, revealing their importance to the catalytic mechanism, in the spirit of the archetypal Linus Pauling catalytic concept.

These effects were also studied by comparing the kinetics of the reaction with 2,4-dinitrophenyl $\beta$-galactosides in which different OH groups of the substrate were replaced by H or F atoms.[13,53] These studies suggest that absent hydroxyl groups in positions 3, 4, or 6 increase the activation barrier of the galactosylation step by at least 4 kcal/mol. Other studies have revealed that the His391 and His540 residues play important roles in providing a transition-state stabilization through their interactions with the 3′-OH and 6′-OH groups, respectively, of the galactose molecule.[6,18] Some

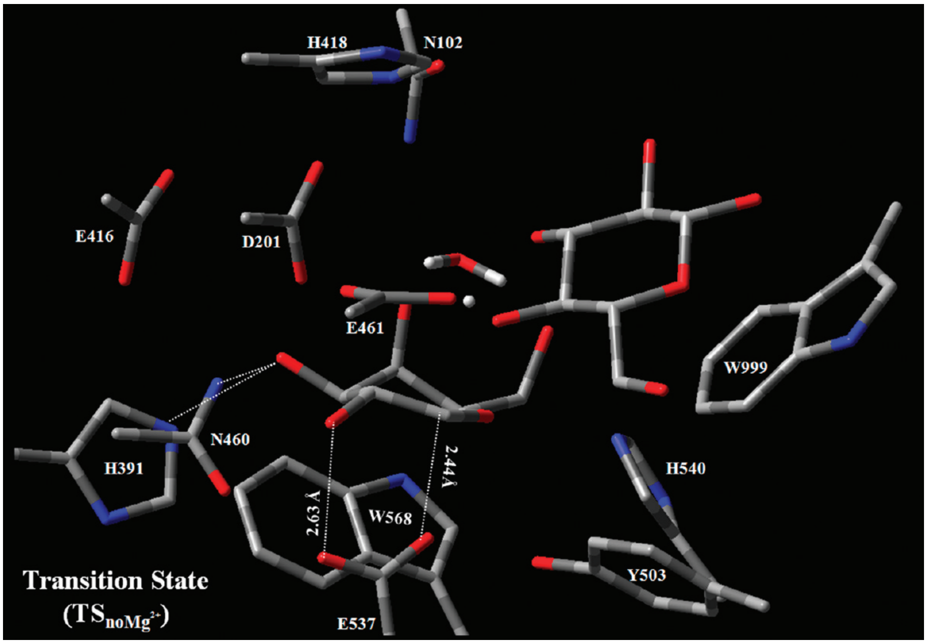


***Figure 4.*** Representation of the optimized structure of the transition state with no $Mg^{2+}$ ($TS_{noMg^{2+}}$).

studies have shown that a histidine residue equivalent to His540 is conserved in every $\beta$-galactosidase.[46] These data suggest that the interaction between this residue and the 6′-OH group is very important for the binding of the substrates and for the stabilization of the transition states in both steps of the catalytic mechanism. This study also suggests that this effect is more significant for the degalactosylation step than for the galactosylation step. The most interesting result is that the substitution of the 2′-OH group lowered the reaction rate considerably, contributing approximately 4−5 kcal/mol for TS stabilization. However, this group can reach a maximum of ca. 10 kcal/mol of transition-state stabilization for some $\beta$-glycosidases.[7,53] One can see that this hydrogen bridge has a length of 1.71 Å in the reactants, decreasing to 1.66 Å in the $TS_1$ structure, revealing its importance in the stabilization of this structure, as well as its catalytic effect on the reaction pathway.

A conformational rearrangement of the galactosyl ring is observed in the transition state, resulting in a half-chair conformation ($^4E$). Many glycosidases show this same conformational distortion in their catalytic mechanisms.[7] However, in our enzymatic model, this reorganization occurs easily because the galactosyl ring in the reactants is already distorted inside the active site. The "electron donation" from the oxygen of the ring to the anomeric carbon results in a partial double bond that facilitates a planar arrangement of the bonds around the anomeric carbon, as one can see in Figure 3. Furthermore, the mentioned hydrogen bridge established between the 2′-HO group and the Glu537 nucleophilic carboxylate group also contributes to this distortion of the galactosyl ring. According to our results, the glycosidic oxygen of the leaving group is equatorial, and hence, this distortion allows for a direct attack of the nucleophilic group to the anomeric center, creating an optimal charge distribution between the ring oxygen and the anomeric carbon for the formation of an oxocarbenium ion. It is obvious that this rearrangement facilitates the progression to $TS_1$, lowering the activation barrier. Additionally, the hydrogen bond from the proton of the Glu461 carboxylate group to the glycosidic oxygen atom comes closer to the plane of the galactosyl ring; hence, it places the glycosidic oxygen atom in an appropriate position for protonation by the proton donor. Moreover, the two catalytic acids become closer to each other, suggesting an implication for tuning the acid p$K_a$ values through the reaction cycle,[51] as will be discussed shortly. The distortion of the galactosyl ring can also modify the interactions of the other sugars with the residues present in the binding pocket. However, in $\beta$-galactosidase, the leaving group is only a glucose molecule, which establishes very few specific interactions because of its nonspecific subsite, and therefore, no substantial modifications occur in its binding pocket.

Analysis of the products of the galactosylation step indicates that a covalent galactosyl−enzyme intermediate forms and that the Glu461 residue donates its proton to the glycosidic oxygen atom, resulting in the departure of the leaving group. When the Glu537 nucleophilic group makes a covalent bond with the galactosyl group, its negative charge is neutralized. As the two carboxylic groups become closer to each other, that modification in charge decreases the p$K_a$ value of the acidic Glu461 residue, inducing proton transfer to the leaving group. This allows the latter residue to act as a base in the degalactosylation step, in a reverse mode of this first step. Some studies have shown that, in enzymes whose activities are highly dependent on existing magnesium ions, the acid catalysis that promotes leaving-group departure is facilitated by their presence.[4,6] It is well-known that $\beta$-galactosidase depends catalytically on the presence of the magnesium ion. Additionally, as the Glu461 residue coordinates with the latter, it has been suggested that it could tune the p$K_a$ value of the amino acid.[51] To better understand the key role of the magnesium ion in the reaction pathway, we repeated the mechanistic step using a similar model, but without the bivalent cation; the results are presented in the next section.

**3.3. First Mechanistic Step without the Magnesium Ion.** As mentioned, Lewis catalysis promoted by $Mg^{2+}$, involving a direct electrophilic attack on the glycosidic oxygen by that same ion, does not seems plausible because of the large distance between those two entities in all crystallographic structures. Various studies have suggested that the magnesium ion might have a dynamic role in tuning the p$K_a$ of the Glu461 residue during the reaction. However, this ability remains to be confirmed.[4,6,10,11,50,51] Other studies have suggested that the conformational changes that occurs in the protein structure along the reaction pathway are the cause for the catalytic dependency on the magnesium ion. In this way, these conformational changes would be favored differently depending on the aglycon bound.[6,11]

Analysis of our data indicates a structural reorganization around the Glu461 residue at the transition state without $Mg^{2+}$ ($TS_{noMg^{2+}}$); moreover, some differences in the specific contacts established between the protein and the galactosyl group were observed, as can be seen in Figure 4. The length of the glycosidic bond is 2.40 Å at $TS_{noMg^{2+}}$; comparison with the value obtained for $TS_1$ (2.25 Å) suggests that this mechanistic step becomes slower in the absence of the magnesium ion. Additionally, the proton of the Glu461 residue comes closer to the glycosidic oxygen atom (1.44 Å), similarly to the behavior observed for the $TS_1$ structure (1.46 Å). On the other hand, the bond distance between the nucleophilic oxygen and the anomeric carbon is shorter than in $TS_1$, having a value of 2.32 Å as compared to 2.45 Å at $TS_1$, also pointing to a later transition state. In summary, a shift toward a delayed transition state was observed upon deletion of the $Mg^{2+}$ ion. Moreover, in the model that includes the magnesium ion, the distance between this ion and the glycosidic oxygen changes from 4.70 to 4.19 Å, suggesting that the crucial role played by this cation is none other than the stabilization of the glycosidic anion in the transition-state structure. It is well-known that this negative charge is directly involved in the increase of the activation barrier values. Furthermore, in the absence of this cation, the acid/base residue (Glu461) becomes less stabilized (it was coordinated to $Mg^{2+}$), which results in its approach toward the Asn460 residue to establish a hydrogen interaction with its amine side chain (2.24 Å).

Values for the activation barrier and reaction energy were calculated for all optimized geometries with the BB1K hybrid-meta functional and the 6-311+G (2d, 2p) basis set. Various studies have suggested that the BB1K functional is the best for calculation of the activation barrier of this kind of reaction and that it overestimates these values typically by about 1.5 kcal/mol.[12,54] Figure 5 shows the activation barriers and reaction energies for both systems, with and without the magnesium ion. Comparing the activation barrier values obtained for the two reactions, 15.0 and 29.9 kcal/mol, respectively, one can see that the initial value increased significantly in the absence of the cation. Experimental data suggest that the presence of the magnesium ion increases the activity of the $\beta$-galactosidase by 5−100-fold, depending on the substrate.[16] However, taking into account the calculated activation barrier, one can assume that this mechanistic step is magnesium-dependent and that, in the absence of $Mg^{2+}$, the reaction either does not occur or occurs through a different pathway. The reproduced barrier difference with and without $Mg^{2+}$ ion is larger than the corresponding experimental value. However, all of the experimental thermodynamic and kinetic data were well reproduced with our model with the exception of the magnitude of the magnesium ion effect. The exaggerated prediction of the effect of the $Mg^{2+}$ ion might occur because the enzyme rearranges more extensively at the TS than was captured with the energy minimizations (or even MD simulations, as these rearrangements have a long time scale).

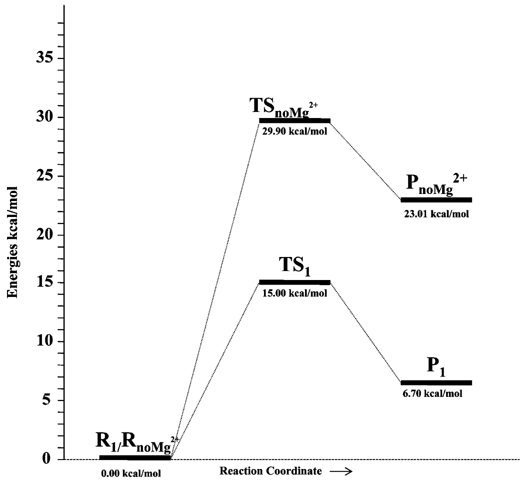


***Figure 5.*** Energetic pathway for the galactosylation mechanistic step of the hydrolysis reaction of the glycosidic linkages with and without magnesium ion.

On the other hand, the reaction energy value obtained for the model with the magnesium ion is 6.7 kcal/mol, compared to 23.0 kcal/mol without the magnesium ion. The variation in the Gibbs energy is positive, which indicates that this first mechanistic step of the net reaction of glycosidic bond hydrolysis is thermodynamically unfavorable. However, it becomes far more unfavorable in the absence of $Mg^{2+}$, emphasizing the importance of the metal ion to the catalytic cycle.

It is interesting to note that the calculated barrier for the galactosylation step without the magnesium ion is similar to that obtained in the small model with the BB1K functional. However, the Gibbs energy is much higher than that in the small-active-site model. Some interactions might contribute to the destabilization of the products in the enzyme without the $Mg^{2+}$ ion in relation to the small model, for example, an unfavorable contact of Glu416 with Glu461. The unprotonated Glu416 belongs to the coordination shell of $Mg^{2+}$ ion and was retained in the model without the magnesium ion but not included in the small model. As the system moves from the reactants to the products, a negative charge is transferred to Glu416, and the system is strongly destabilized because of the short distance between the two negative glutamate residues (ca. 4.5 Å). This is an important factor, but not the only one that justifies the difference in the reaction energy of the two systems. The sum of a very large number

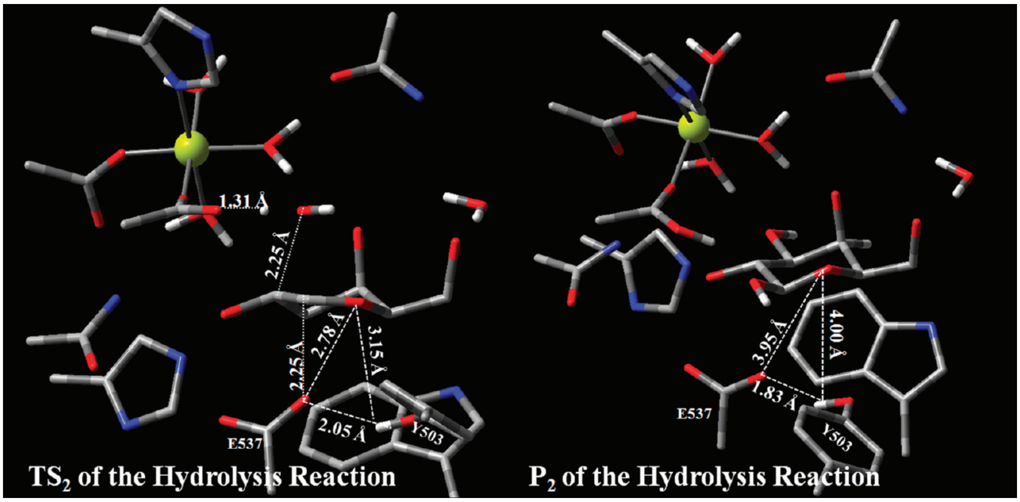


***Figure 6.*** Representation of the structures of the transition state ($TS_2$) and products ($P_2$) for the degalactosylation step.

of small interactions present in the larger model might, as a whole, increase the energy of the products, even though their effect is not trivial to pinpoint individually.

**3.4. Second Mechanistic Step: Hydrolysis Reaction.** The second mechanistic step involves the attack of the covalent galactosyl−enzyme intermediate by a water molecule. To obtain the reactants for this step, we transformed the leaving group into a water molecule, keeping the glycosidic hydroxyl and replacing the remaining one by a proton directed along the C−O bond. In the reactant structures ($R_2$), the distance between the anomeric carbon atom and the oxygen atom of the attacking water is 3.43 Å. Figure 6 shows the structure of the hydrolytic transition state ($TS_2$) in which the breaking of the bond between the anomeric center and the nucleophile group occurs. The length of this bond is 1.53 Å in the reactants and 2.25 Å at $TS_2$. Therefore, the bond might be considered as being broken at this stage. The attacking water molecule comes closer to the anomeric carbon atom, with the oxygen atom performing the attack. However, one can see that the covalent bond has not yet been established (2.25 Å) at this stage. Only in the products ($P_2$) is the bond fully formed, with a length of 1.41 Å. Additionally, one proton of the water molecule gets closer to the base Glu461 carboxylate group at the TS, with a distance of 1.31 Å. At this stage, proton transfer is ready to occur. The data also confirm the expected dissociative nature of the degalactosylation step, with the bond to the nucleophilic group already broken and the glycosidic bond not yet established.[7,45]

Furthermore, the cleavage of the nucleophilic bond transfers the negative charge back to the Glu537 carboxylate group. During this degalactosylation step, a trigonal oxocarbenium ion has formed, again stabilized by interactions between the Glu537 carboxylic side chain, the Tyr503 hydroxyl side chain, and the oxygen atom of the galactosyl ring. Figure 6 shows the distances among all of these atoms for the $TS_2$ and $P_2$ structures. Secondary kinetic isotope effects previously suggested this to be a trigonal oxocarbenium ion.[19,55] The short hydrogen bridge established with the Tyr503 hydroxyl group could facilitate the elimination of the negative nucleophile. Studies performed with a myrosinase (from *Sinapis alba*) and a xylanase (from *Bacillus circulans*) suggest that an equivalent tyrosine residue near the nucleophilic catalytic group plays a similar role in these catalytic mechanisms.[6,56,57]

One can see that, in the products, the galactosyl group adopts the lower-energy chair conformation. At the end of all reactions, the acid/base Glu461 ends up protonated, whereas the Glu537 nucleophilic residue is negatively charged; that is, the two catalytic residues are prepared for another reaction cycle.

Values for the activation barrier and reaction energy for this second step were also calculated, for the optimized geometries, using the BB1K hybrid-meta functional and the 6-311+G (2d, 2p) basis set. The values obtained were 15.5 and −9.2 kcal/mol, respectively. At the end of this step, a glucose molecule dissociates from the active site. The kinetics and the free energy profile for the dissociation step are probably very complex, but they must be included in the free energy profile for the reaction, as their thermodynamic contribution must be far from negligible. To introduce this effect, we used a simplified, but still accurate, model in which we considered only a glucose molecule surrounded by (and geometry-optimized within) two dielectric continuum solvents, one with $\varepsilon = 4$ and the other with $\varepsilon = 80$. A dielectric constant of 4 mimics a hydrophobic protein environment, whereas a value of 80 corresponds to an aqueous environment. The dissociation free energy, $\Delta G_{diss}$(Glu), corresponds to the difference between these two values (−9.09 kcal/mol).

Upon addition of $\Delta G_{diss}$(Glu), the rate-limiting activation barrier (calculated from the initial reactants) and the Gibbs energy of reaction were calculated as 15.00 and −11.53 kcal/mol, respectively (Figure 7). These values indicate that this mechanistic step is both kinetically and thermodynamically favorable. Our results are in agreement with experiment, as the value of $k_{cat}$ for the hydrolysis of lactose has been reported as 60 $s^{-1}$ (corresponding to an activation barrier value of approximately 15 kcal/mol).[16,58,59] Kempton and Withers studied the *Agrobacterium sp.* $\beta$-glucosidase using different substrates and described a relationship in which the p$K_a$ values of the leaving groups are good predictors of $k_{cat}$, as

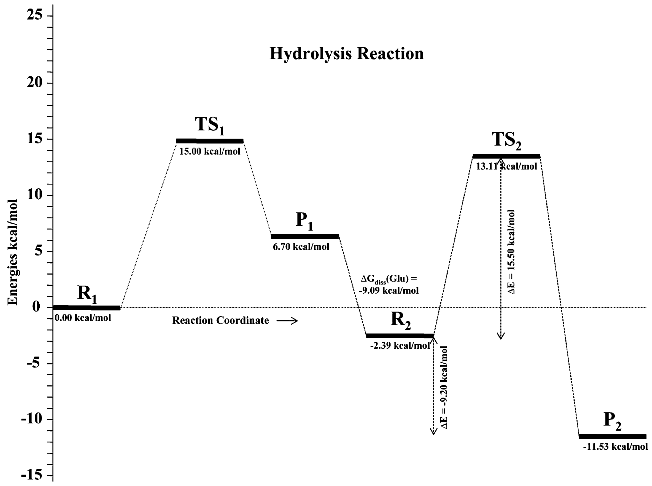


*Figure 7.* Energetic pathway for the hydrolysis reaction of the glycosidic linkages.

well as the rate-limiting step.[14] They found that the first step should be rate-limiting for leaving groups with p$K_a$ values above 8. It is well-known that lactose is a substrate with a very unfavorable leaving group. The p$K_a$ of the glycosyl oxygen is 11,[60] which indicates that the galactosylation step should be rate-limiting. However, our data indicate similar values of the activation barriers for both steps (galactosylation and degalactosylation). Therefore, although the activation energy for this degalactosylation step is higher than that for the galactosylation step, the small difference between them is smaller than the accuracy of the computational method, and therefore, this computational method is unable to confirm which one is the rate-limiting step.

**3.5. Second Mechanistic Step: Transglycosylation Reactions.** Transglycosylation reactions occur when the same leaving group, or another carbohydrate molecule, attacks the covalent galactosyl−enzyme intermediate. At the end of the first step, if the glucose molecule does not leave the active site, it can rotate and attack the intermediate with one of its several hydroxyl groups. As we have already mentioned, the glucose molecule makes few specific interactions and has substantial freedom of movement in the binding pocket, which allows for rotations and conformational transitions. On the other hand, this molecule is a poor leaving group, suggesting that transglycosylation reactions are favorable. In the case of $\beta$-galactosidase from *E. coli*, some studies have suggested that galactosyl-$\beta$(1−6)-glucose (allolactose) is the preferred transglycosylation product, showing a yield of ca. 97% over the other disaccharides.[58] Considering that the allolactose molecule is the natural *lac* operon inducer (thus strictly necessary for the $\beta$-galactosidase production), this preference is expected and is physiologically important to the *E. coli* bacterium. To better clarify the origin of the stereoselectivity of this enzyme during the transglycosylation reactions, we performed a study of the PESs for the attacks of the anomeric center on the oxygens of the different hydroxyl groups, namely, those of the $C_3$, $C_4$, and $C_6$ atoms of the glucose group. Figures 8 and 9 show the activation barriers and reaction energies obtained for the $\beta$(1−3) and $\beta$(1−6) transglycosylation reactions, respectively.

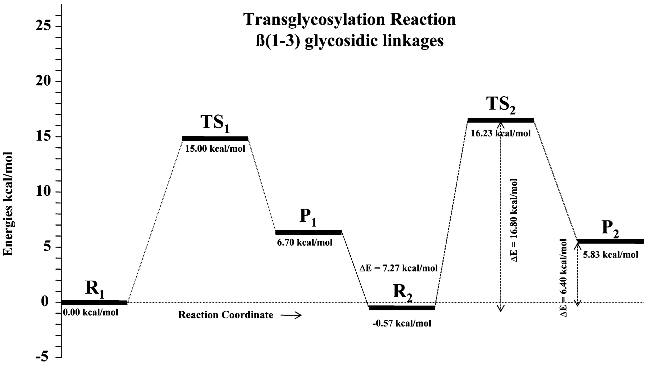


*Figure 8.* Energetic pathway for the transglycosylation reaction that produces $\beta$(1−3) glycosidic linkages.

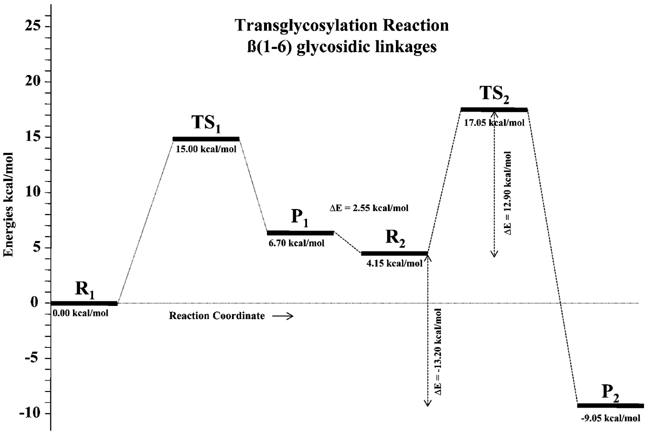


*Figure 9.* Energetic pathway for the transglycosylation reaction that produces $\beta$(1−6) glycosidic linkages.

The first step, in which the covalent galactosyl−enzyme intermediate is formed, is common to all of the reaction pathways because of the necessary cleavage of the natural lactose substrate. In the $\beta$(1−4) transglycosylation reaction, the products of the first step are also the reactants of the second step; thus, this transglycosylation pathway is a reverse mode of the galactosylation pathway (Figure SI-2 in the Supporting Information). Therefore, the activation barrier and total Gibbs energy are 15.0 and 0.0 kcal/mol, respectively.

On the other hand, the products of the cleavage step are quite different from the reactants for both $\beta$(1−3) and $\beta$(1−6) transglycosylation steps. By analyzing the products of the galactosylation step, one can verify that only the O−$C_4$ bond is directed toward the anomeric carbon atom, whereas the O−$C_3$ and O−$C_6$ bonds are pointing in other directions. Therefore, rotation and conformational adjustment in the glucosyl ring are necessary for the attack to be performed by the oxygens of hydroxyl groups 3′-HO and 6′-HO. Figure 10 shows the structures of the reactants for the $\beta$(1−3) and $\beta$(1−6) transglycosylation steps.

Comparing the energies of the $P_1$ and $R_2$ states of these transglycosylation reactions, one can see that these modifications decrease the energies values of $R_2$ by 7.27 and 2.55 kcal/mol for $\beta$(1−3) and $\beta$(1−6) glycosidic bonds, respectively, as compared to those of $\beta$(1−4) glycosidic linkages. According to these data, one can assume that the reactants' geometries for the $\beta$(1−3) transglycosylation pathway are

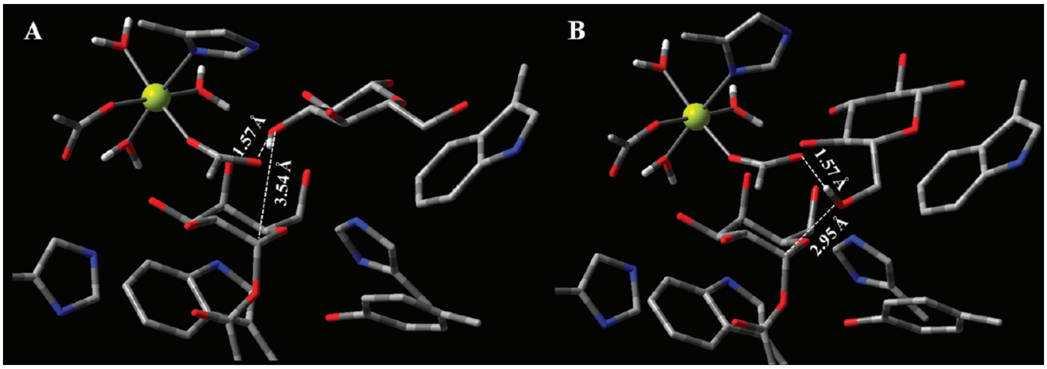


***Figure 10.*** Structures of reactants for different transglycosylation steps: (A) $\beta$(1−3) glycosidic linkages, (B) $\beta$(1−6) glycosidic linkages.

the most stable, showing the lowest energy values. However, the activation barrier and reaction energies for the $\beta$(1−3) transglycosylation reaction are 16.23 and 5.83 kcal/mol, respectively, whereas those for the $\beta$(1−6) transglycosylation reaction are 17.05 and −9.05 kcal/mol, respectively. In summary, the activation barriers for the three different pathways studied are 16.23, 15.00, and 17.05 kcal/mol for the $\beta$(1−3), $\beta$(1−4), and $\beta$(1−6) transglycosylation reactions, respectively. one can see that all barriers are similar (which seems logical, as the bonds being broken and formed are of the same type in all three reactions). Additionally, as was already mentioned, the density functional used in our calculations is known to overestimate the activation barriers.[12,54] With a small model, we confirmed that this meta-hybrid density functional overestimates the activation energy of this reaction by 1.7 kcal/mol in relation to high-level post-HF calculations.[45] Subtracting this error from the activation barrier for the $\beta$(1−6) transglycosylation reaction, we obtain the value of 15.35 kcal/mol, which is in agreement with the experimental data that proposed a $\Delta G^{\ddagger}$ value of approximately 15 kcal/mol for the allolactose production. These data suggest that the transglycosylation reactions to produce all of the aforementioned kinds of disaccharides are thermodynamically favorable. In summary, the reaction energy values for the three different reactions studied are 5.83, 0.00, and −9.05 kcal/mol for the $\beta$(1−3), $\beta$(1−4), and $\beta$(1−6) transglycosylation reactions, respectively. Concerning these results, one can see that the formation of the $\beta$(1−3) glycosidic bond is thermodynamically unfavorable, which suggests that this enzyme does not produce this kind of linkage. In contrast, the fact that the formation of the $\beta$(1−6) glycosidic bond has the most negative value of the Gibbs energy shows that this transglycosylation reaction is thermodynamically more favorable than any other, explaining the preference of the $\beta$-galactosidase for the production of the allolactose molecule. All of these data are in agreement with the experimental studies that suggested that the galactosyl-$\beta$(1−6)-glucose is the preferred transglycosylation product of this enzyme. The basis for the stereoselectivity is therefore thermodynamic, rather than kinetic, and relies on the more favorable binding of the $\beta$(1−6) product to the active site. Therefore, the enzyme promotes enrichment in $\beta$(1−6) linkages through selective stabilization of the desired product.

## 4. Conclusions

Despite the large number of studies on the glycosidase family of enzymes and their reactions mechanisms, many atomistic insights are still not fully elucidated. In the theoretical study presented here, BB1K:AMBER QM/MM calculations were performed on a large enzymatic model in order to fully understand the catalytic mechanism of the hydrolysis, its dependence on magnesium, and the origin of the stereoselectivity of the different transglycosylation reactions performed by $\beta$-galactosidase from *E. coli*.

Our analysis of the optimized structures of the reactants, transition states, and products for both galactosylation and degalactosylation steps of the mechanism confirms the dissociative nature of the transition states, as is generally accepted for glycosidases. In the first TS, the glycosidic bond is very elongated, and the nucleophilic group is still far from the anomeric carbon atom, as well as the hydrogen atom of the proton donor, which, in turn, is still bound to the acid/base residue. In the second TS, the covalent bond established between the anomeric carbon atom and the nucleophilic group is almost broken, and the attacking oxygen of the water molecule is still far from the anomeric carbon. Additionally, a proton from a water molecule comes close to the acid/base residue, but does not transfer at this stage. Furthermore, our data show the presence of a shorter hydrogen bridge between the nucleophilic group and 2′-HO of the galactosyl group (1.66 Å), as well as the ring planarization toward the half-chair conformation at the transition state. These phenomena have a crucial role in lowering the activation energy of the system and help stabilizing the nascent oxocarbenium ion.

The key role played by the critical magnesium ion in the hydrolysis catalytic mechanism of this enzyme was also studied, and we found that the activation barrier is significantly affected by the absence of this ion. In such a situation, the activation energy rises by 14.9 kcal/mol, emphasizing the necessity of this magnesium ion for the catalytic mechanism to take place. We suggest that this occurs probably through the stabilization of the leaving group by the bivalent cation.

The galactosylation step of the hydrolysis reaction is rate-limiting, having an activation barrier of 15.0 kcal/mol. Moreover, the total reaction energy is −11.5 kcal/mol. Therefore, we conclude that this catalytic reaction is kinetically and thermodynamically favorable, which is in complete agreement with the experimental data on $\beta$-galactosidade from *E. coli*. Comparing these values with those obtained with a similar small system, we verified that these values are much more favorable; therefore, the contribution of the enzymatic environment to the reaction kinetics is strictly necessary to characterize quantitatively the catalytic mechanism for this galactosidase. It was concluded that the enzyme scaffold binds the transition state better than the reactants, providing a huge catalytic effect.

Different transglycosylation reactions that produce the $\beta$(1−3), $\beta$(1−4), and $\beta$(1−6) glycosidic linkages were also studied. Comparison of the energetic values for these reactions shows that the transglycosylation reactions are all very similar from a kinetic perspective, which seems reasonable given the similarity in the bond-breaking/bond-forming processes. However, thermodynamically, they are quite dissimilar, with the transglycosylation step to make $\beta$(1−6) glycosidic bonds being significantly favored. We can conclude that this retaining $\beta$-galactosidase has a transglycosylation preference for glycosidic linkages in the order $\beta(1-6) > \beta(1-4) > \beta(1-3)$. Therefore, our data suggest that the allolactose molecule is the preferred product obtained, which is in agreement with the experimental data. According to the free energy values obtained here, the $\beta$(1−6) product should be selected to ~100%, which is very close to the 97% preference observed experimentally. The origin for the stereoselectivity was found to be thermodynamic, with the enzyme stabilizing the preferred product.

This QM/MM study allows for a complete comprehension of this catalytic mechanism with atomistic detail. As the $\beta$-galactosidase from *E. coli* is an enzyme commonly used in molecular biology research, knowledge of the different reaction pathways is crucial to the development of new chromophore substrates. Furthermore, such knowledge helps improve the efficiency of large-scale industrial design and synthesis of new inhibitors and carbohydrates for both the pharmaceutical and food industries.

**Acknowledgment.** N.F.B. thanks the Fundação para a Ciência e a Tecnologia (FCT) for a Ph.D. grant (SFRH/BD/31359/2006).

**Supporting Information Available:** Atomic point charges calculated, with a Mulliken population analysis, for all key atoms involved in bond breaking/bond formation during the scan of the reactants, transition states, and products for all reactions studied (Tables SI-I−SI-III); structure of the natural substrate lactose, energetic pathway for the transglycosylation reaction that produces $\beta$(1−4) glycosidic linkages, and all linear transit schemes for all steps studied (Figures SI-1 and SI-2 and Schemes SI-1−SI-5, respectively). This information is available free of charge via the Internet at http://pubs.acs.org/.

## References

(1) Perugino, G.; Trincone, A.; Rossi, M.; Moracci, M. *Trends Biotechnol.* **2004**, *22*, 31.

(2) Maugard, T.; Gaunt, D.; Legoy, M. D.; Besson, T. *Biotechnol. Lett.* **2003**, *25*, 623.

(3) Jakeman, D. L.; Withers, S. G. *Can. J. Chem.* **2002**, *80*, 866.

(4) Juers, D. H.; Jacobson, R. H.; Wigley, D.; Zhang, X. J.; Huber, R. E.; Tronrud, D. E.; Matthews, B. W. *Protein Sci.* **2000**, *9*, 1685.

(5) Jacobson, R. H.; Zhang, X. J.; Dubose, R. F.; Matthews, B. W. *Nature* **1994**, *369*, 761.

(6) Juers, D. H.; Heightman, T. D.; Vasella, A.; McCarter, J. D.; Mackenzie, L.; Withers, S. G.; Matthews, B. W. *Biochemistry* **2001**, *40*, 14781.

(7) Zechel, D. L.; Withers, S. G. *Acc. Chem. Res.* **2000**, *33*, 11.

(8) Koshland, D. E. *Biol. Rev. Cambridge Philos. Soc.* **1953**, *28*, 416.

(9) Crout, D. H. G.; Vic, G. *Curr. Opin. Chem. Biol.* **1998**, *2*, 98.

(10) Richard, J. P.; McCall, D. A. *Bioorg. Chem.* **2000**, *28*, 49.

(11) Sinnott, M. L.; Withers, S. G.; Viratelle, O. M. *Biochem. J.* **1978**, *175*, 539.

(12) Bras, N. F.; Moura-Tamames, S. A.; Fernandes, P. A.; Ramos, M. J. *J. Comput. Chem.* **2008**, *29*, 2565.

(13) Namchuk, M. N.; McCarter, J. D.; Becalski, A.; Andrews, T.; Withers, S. G. *J. Am. Chem. Soc.* **2000**, *122*, 1270.

(14) Kempton, J. B.; Withers, S. G. *Biochemistry* **1992**, *31*, 9961.

(15) Jahn, M.; Withers, S. G. *Biocatal. Biotransform.* **2003**, *21*, 159.

(16) Huber, R. E.; Parfett, C.; Woulfeflanagan, H.; Thompson, D. J. *Biochemistry* **1979**, *18*, 4090.

(17) Cupples, C. G.; Miller, J. H.; Huber, R. E. *J. Biol. Chem.* **1990**, *265*, 5512.

(18) Huber, R. E.; Hlede, I. Y.; Roth, N. J.; McKenzie, K. C.; Ghumman, K. K. *Biochem. Cell Biol.* **2001**, *79*, 183.

(19) Penner, R. M.; Roth, N. J.; Rob, B.; Lay, H.; Huber, R. E. *Biochem. Cell Biol.* **1999**, *77*, 229.

(20) Roth, N. J.; Rob, B.; Huber, R. E. *Biochemistry* **1998**, *37*, 10099.

(21) *Insight II*, version 2.3.0; Accelrys: San Diego, CA, 1993.

(22) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727.

(23) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. *Proteins* **2003**, *52*, 609.

(24) Bras, N. F.; Fernandes, P. A.; Ramos, M. J. *Theor. Chem. Acc.* **2009**, *122*, 283.

(25) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, H. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *Amber 8*; University of California: San Francisco, CA, 2004.

(26) Basma, M.; Sundara, S.; Calgan, D.; Vernali, T.; Woods, R. J. *J. Comput. Chem.* **2001**, *22*, 1125.

(27) Kirschner, K. N.; Woods, R. J. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10541.

(28) Kirschner, K. N.; Woods, R. J. *J. Phys. Chem. A* **2001**, *105*, 4150.

(29) Asensio, J. L.; Jimenezbarbero, J. *Biopolymers* **1995**, *35*, 55.

(30) Izaguirre, J. A.; Catarello, D. P.; Wozniak, J. M.; Skeel, R. D. *J. Chem. Phys.* **2001**, *114*, 2090.

(31) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. *Biopolymers* **1992**, *32*, 523.

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

(33) Hammonds, K. D.; Ryckaert, J. P. *Comput. Phys. Commun.* **1991**, *62*, 336.

(34) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision D.01/D.02; Gaussian, Inc.: Pittsburgh, PA, 2003.

(35) Dapprich, S.; Komaromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *J. Mol. Struct. (THEOCHEM)* **1999**, *461*, 1.

(36) Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170.

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

(38) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040.

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

(40) Lundberg, M.; Blomberg, M. R. A.; Siegbahn, P. E. M. Developing Active Site Models of ODCase—from Large Quantum Models to a QM/MM Approach. In *Orotidine Monophosphate Decarboxylase: A Mechanistic Dialogue*; Topics in Current Chemistry; Springer-Verlag: Berlin, 2004; Vol. 238, pp 79−112.

(41) Mulholland, A. J. *Chem. Cent. J.* **2007**, Vol. 1, Number 19.

(42) Carvalho, A. T. P.; Fernandes, P. A.; Ramos, M. J. *J. Phys. Chem. B* **2006**, *110*, 5758.

(43) Cioslowski, J. *J. Am. Chem. Soc.* **1989**, *111*, 8333.

(44) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715.

(45) Bras, N. F.; Fernandes, P. A.; Ramos, M. J. *J. Mol. Struct. (THEOCHEM)* 2009; doi:10.1016/j.theochem.2009.08.039.

(46) Roth, N. J.; Huber, R. E. *J. Biol. Chem.* **1996**, *271*, 14296.

(47) Deslongchamps, P. Intramolecular strategies and stereoelectronic effects − Glycosides hydrolysis revisited. *Pure and Applied Chemistry*, 1993; Vol. 65, pp 1161−1178.

(48) Fushinobu, S.; Mertz, B.; Hill, A. D.; Hidaka, M.; Kitaoka, M.; Reilly, P. J. *Carbohydr. Res.* **2008**, *343*, 1023.

(49) Cremer, D.; Pople, J. A. *J. Am. Chem. Soc.* **1975**, *97*, 1354.

(50) Loeffler, R. S. T.; Sinnott, M. L.; Sykes, B. D.; Withers, S. G. *Biochem. J.* **1979**, *177*, 145.

(51) Richard, J. P.; Huber, R. E.; Lin, S.; Heo, C.; Amyes, T. L. *Biochemistry* **1996**, *35*, 12377.

(52) McCarter, J. D.; Adam, M. J.; Withers, S. G. *Biochem. J.* **1992**, *286*, 721.

(53) Namchuk, M. N.; Withers, S. G. *Biochemistry* **1995**, *34*, 16194.

(54) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Phys. Chem. A* **2007**, *111*, 10439.

(55) Sinnott, M. L. *FEBS Lett.* **1978**, *94*, 1.

(56) Burmeister, W. P.; Cottaz, S.; Driguez, H.; Iori, R.; Palmieri, S.; Henrissat, B. *Structure* **1997**, *5*, 663.

(57) Sidhu, G.; Withers, S. G.; Nguyen, N. T.; McIntosh, L. P.; Ziser, L.; Brayer, G. D. *Biochemistry* **1999**, *38*, 5346.

(58) Huber, R. E.; Kurz, G.; Wallenfels, K. *Biochemistry* **1976**, *15*, 1994.

(59) Huber, R. E.; Gaunt, M. T.; Sept, R. L.; Babiak, M. J. *Can. J. Biochem. Cell Biol.* **1983**, *61*, 198.

(60) Ye, J. N.; Zhao, X. W.; Sun, Q. X.; Fang, Y. Z. *Mikrochim. Acta* **1998**, *128*, 119.

CT900530F

JCTC Journal of Chemical Theory and Computation

# Constant Constraint Matrix Approximation: A Robust, Parallelizable Constraint Method for Molecular Simulations

Peter Eastman*,† and Vijay S. Pande‡

*Department of Bioengineering and Department of Chemistry, Stanford University, Stanford, California 94305*

Received June 16, 2009

**Abstract:** We introduce a new algorithm, the constant constraint matrix approximation (CCMA), for constraining distances in molecular simulations. It combines the best features of many existing algorithms while avoiding their defects: it is fast and stable, can be applied to arbitrary constraint topologies, and can be efficiently implemented on modern parallel architectures. We test it on a protein with bond length and limited angle constraints and find that it requires less than one-sixth as many iterations as SHAKE to converge.

## Introduction

Rigid distance constraints are a popular method of increasing the integration step size in simulations of macromolecules. Using standard molecular force fields with no constraints, one is generally limited to a step size of about 1 fs. By constraining the lengths of bonds involving a hydrogen atom, one can increase the step size to 2 fs, thus doubling the amount of time that can be simulated in a given number of time steps. By constraining all bond lengths, as well as the most rapidly oscillating bond angles, the step size can be further increased to 4 fs.[1] Furthermore, due to the quantization of vibrational motion of bonds, constraints may be a more realistic representation of these stiff degrees of freedom than the harmonic forces conventionally used for them.[2]

Many algorithms have been suggested for implementing these constraints, but all of them have disadvantages that restrict their usefulness. The choice of which to use involves trade-offs between speed, stability, and range of applicability. For example, some algorithms are only useful for small molecules, or for short time steps, or for particular constraint topologies, or on particular computer architectures.

In this paper, we introduce a new constraint algorithm called the constant constraint matrix approximation (CCMA). It combines the best features of many existing algorithms while avoiding their disadvantages: it is fast, has good stability, can be applied to arbitrary sets of constraints, and can be efficiently implemented on a variety of modern computer architectures.

## Background

Most constraint algorithms used in molecular simulations are based on (or are equivalent to) the method of Lagrange multipliers. For each interatomic distance that is to be constrained, one defines an error function

$$\sigma_i(\{\mathbf{r}_k\}) = |\mathbf{r}_m - \mathbf{r}_n| - d_i \tag{1}$$

where $i$ is the index of the constraint, $\{\mathbf{r}_k\}$ is the set of all atomic coordinates, $\mathbf{r}_m$ and $\mathbf{r}_n$ are the positions of the two atoms whose distance is constrained, and $d_i$ is the required distance between them. One then applies a constraint force $\lambda_i(t)$ to atoms $m$ and $n$, which produces a combined displacement $\delta_i$ along the constraint direction during each time step. (Atom $m$ is displaced by $(\delta_i/m_m)/(1/m_m + 1/m_n)$, while atom $n$ is displaced by $-(\delta_i/m_n)/(1/m_m + 1/m_n)$, where $m_m$ and $m_n$ are the masses of the two atoms). The challenge at each time step is to calculate the vector of displacements $\boldsymbol{\delta}(t)$ such that $\sigma_i(\{\mathbf{r}_k\}) = 0$ for every constraint at the end of the time step.

This requires solving a system of nonlinear equations, which is typically done with an iterative algorithm such as Newton iteration:

$$\delta^{N+1} = \delta^{N} - \mathbf{J}^{-1}\boldsymbol{\sigma}^{N} \tag{2}$$

* Corresponding author e-mail: peastman@stanford.edu.

† Department of Bioengineering.

‡ Department of Chemistry.

10.1021/ct900463w © 2010 American Chemical Society
Published on Web 01/07/2010

where $\boldsymbol{\delta}^N$ is the vector of displacements calculated in the $N$th iteration, $\boldsymbol{\sigma}^N$ is the vector of constraint errors in the $N$th iteration, and $\mathbf{J}$ is the Jacobian matrix

$$\mathbf{J}_{ij} = \frac{\partial \sigma_i}{\partial \delta_j} \tag{3}$$

where $i$ and $j$ each run over all constraints in the system.

The most straightforward way to implement this is to explicitly construct the Jacobian matrix $\mathbf{J}$ and then invert it using a standard technique such as LU decomposition. This is, in fact, precisely what the M-SHAKE algorithm does.[3] The result is a stable algorithm that converges rapidly. Unfortunately, the time required to build and invert $\mathbf{J}$ increases rapidly with the number of constraints. For this reason, M-SHAKE is only useful for small molecules, not for macromolecules such as proteins and nucleic acids.

The LINCS algorithm takes a slightly different approach to performing the iteration.[4] Instead of explicitly calculating and inverting the Jacobian matrix, it represents $\mathbf{J}^{-1}$ as a power series. The usefulness of this approach depends on how quickly the series converges. For weakly connected sets of constraints, such as when only bond lengths are constrained, it converges quickly. For more strongly connected systems, such as when both bond lengths and angles are constrained, it converges more slowly, and may even fail to converge at all. For this reason, LINCS is generally only useful for bond length constraints.

Instead of accurately calculating $\mathbf{J}^{-1}$, one can instead try to approximate it with a different matrix $\mathbf{K}^{-1}$ that is easier to calculate. It can be shown that this approximation has no effect on the final result: because the displacements are uniquely determined by the requirement $\sigma_i(\{\mathbf{r}_k\}) = 0$, any convergent procedure is guaranteed to produce the same result.[5] On the other hand, the approximation will generally increase the number of iterations required and may also decrease the radius of convergence. How serious these problems are depends on how close $\mathbf{K}^{-1}$ is to $\mathbf{J}^{-1}$. The challenge is to find a matrix that is as close as possible to $\mathbf{J}^{-1}$ while still being easy to calculate.

The simplest approximation one might consider is the identity matrix. This is equivalent to assuming that all constraints are decoupled from each other, so that the force applied along one constraint has no effect on any other constrained distance. For weakly connected sets of constraints, this is actually not too bad an approximation and may produce a useful algorithm. For more strongly connected sets of constraints, however, it produces very poor convergence.

The SHAKE algorithm uses a small variation on this procedure that significantly improves its speed and stability.[5] It still computes $\delta_i$ independently for each constraint, but it processes them serially: each constraint force is calculated and the positions of its two atoms are updated before the next $\delta_i$ is calculated. As a result, each constraint implicitly sees the effect of all other constraints that were processed before it, but not those processed after it. This is equivalent to approximating $\mathbf{J}^{-1}$ using its true upper triangle, while setting all elements below the diagonal to zero. The result is significantly improved convergence at very little extra cost, which accounts for the popularity of this method.

SHAKE has an important disadvantage, however: it is an inherently serial algorithm. Each constraint must be fully processed and the atom positions updated before the next constraint can be processed. As a result, it is impossible to implement SHAKE efficiently on parallel architectures (multicore processors, graphics processing units, clusters, etc.). As parallel computing has become increasingly prevalent, the need for alternatives to SHAKE has become clear.

Another important class of constraint algorithms is ones that solve the constraint equations analytically rather than using an iterative method. The most important algorithm in this class is SETTLE, which uses an analytical solution for rigid water molecules.[6] It is both fast and extremely stable. As a result, it is clearly the method of choice for simulations involving explicit water molecules. Because it is applicable only to one very specific type of molecule, however, another algorithm must be used along with it to constrain the geometry of solute molecules.

A very different approach to implementing constraints is to work in internal coordinates.[7,8] Instead of representing the molecular conformation by the Cartesian coordinates of each atom, one instead represents it by the bond lengths and angles between atoms. It then becomes trivial to constrain those lengths and angles by keeping the corresponding coordinates fixed. This leads to a description of the system as a set of rigid bodies, each containing multiple atoms, connected by a minimal set of internal coordinates. Because the molecular force field depends on the Cartesian coordinates of atoms, it is necessary to convert positions and forces between Cartesian and internal coordinates as part of each time step. The algorithms for doing this are difficult to implement and add overhead to each time step. They also involve tree-structured computations that are difficult to parallelize efficiently. For these reasons, internal coordinates have been much less widely used than Cartesian coordinates for molecular simulations. They have the interesting property that their computational cost scales with the number of *free* degrees of freedom, in contrast to most other constraint algorithms whose cost scales with the number of *constrained* degrees of freedom. This makes internal coordinates most appropriate for highly constrained systems, such as when entire secondary structure elements or even protein domains are held rigid.

Many other constraint algorithms have been proposed, and a complete survey of them is beyond the scope of this paper. The methods described above include the most popular ones and are illustrative of the general approaches taken by many algorithms. Below, we expand on the details of our proposed approach, the constant constraint matrix approximation.

## Constant Constraint Matrix Approximation

The CCMA algorithm is based on the observation that the Jacobian matrix changes very little over the course of a simulation. All elements along the diagonal are equal to 1. Each off-diagonal element describes the coupling between two constraints. If the two constraints do not share an atom, the corresponding element is zero. If they do share an atom, it is equal to

$$\mathbf{J}_{ij} = \frac{1/m_1}{1/m_1 + 1/m_2}\cos\theta \tag{4}$$

where $m_1$ is the mass of the atom that is shared by the two constraints, $m_2$ is the mass of the other atom affected by constraint $i$, and $\theta$ is the angle between the two constraints. The atomic masses usually do not change with time. If the angle $\theta$ is itself constrained, the corresponding element of **J** is constant over the simulation. In fact, if all bond lengths and angles are constrained, then **J** is constant.

If the angle is not constrained, the element will vary with time, but usually not very much. Molecular force fields typically include a harmonic force term for each angle that restricts its motion to a narrow range. This suggests that if we construct and invert **J** once at the start of the simulation, we can reuse it for every time step and it will continue to be a good approximation. (We note in passing that this same observation was made by Weinbach and Elber, but they did not pursue it further or develop an algorithm based on it.[9])

Specifically, we construct and invert a matrix **K** that is an approximation to **J** as follows. For each element in **K**:

(1) If the angle between the two constraints is itself constrained, we calculate the value on the basis of the actual constrained angle.

(2) Otherwise, we calculate it on the basis of the equilibrium angle of the corresponding harmonic force term.

How much **K** deviates from **J** is determined by how far the unconstrained angles vary from their equilibrium values. For typical molecular force fields, these deviations are very small. In the more general case of arbitrary constrained systems, however, there might be situations where angles have more flexibility. For example, some coarse-grained lipid models use a relatively soft force term on angles that allows larger fluctuations.[10] CCMA will still work with these systems, but the number of required iterations is expected to increase as the difference between **J** and **K** increases.

When solving the constraint equations for each time step, we replace $\mathbf{J}^{-1}$ in eq 2 by $\mathbf{K}^{-1}$. This involves a matrix-vector multiplication at each iteration, which will be efficient if and only if $\mathbf{K}^{-1}$ is sufficiently sparse. **K** is very sparse, since a single atom is almost never bonded to more than four other atoms, but it does not automatically follow that $\mathbf{K}^{-1}$ is also sparse. In practice, we find that most of its elements are extremely small and can be neglected. We therefore set all elements of $\mathbf{K}^{-1}$ that fall below a cutoff to zero, yielding a sparse matrix which still is an excellent approximation to $\mathbf{J}^{-1}$.

For highly constrained systems, such as when all bond lengths and angles are constrained, care must be taken to prevent **K** from becoming singular. This happens when a rigid cluster of atoms contains more constraints than are necessary to remove all internal degrees of freedom of the cluster. For example, a methane molecule has nine internal degrees of freedom, but if one naively constrains all bond lengths and angles, this produces ten constraints. Ideally, one should identify such clusters and remove the redundant constraints. Alternatively, one can invert **K** with a method that is robust to singular matrices, such as QR decomposition or singular value decomposition.[11] This approach assumes that the redundant constraints are all consistent with each other; if the constraints are inconsistent, it is impossible to find a solution which satisfies all of them.

***Table 1.*** Average Number of Iterations Needed for the Constraint Algorithm To Converge with a Relative Tolerance of $10^{-4}$

| | 1 fs | 2 fs | 3 fs | 4 fs |
|---|---|---|---|---|
| CCMA (0.01 cutoff) | 3.09 | 4.02 | 4.64 | 5.03 |
| CCMA (0.1 cutoff) | 3.58 | 4.50 | 4.79 | 5.38 |
| SHAKE | 20.8 | 27.7 | 31.9 | 34.7 |

## Results

To test the CCMA algorithm, we incorporated it into OpenMM, a library for performing molecular simulations on graphics processing units (GPUs) and other high-performance architectures.[12] The implementation was straightforward since all elements of the algorithm (computing the vector of constraint errors, the sparse matrix−vector multiply, and updating atom positions) are easily parallelized. We also created a serial implementation to facilitate comparison with other algorithms.

We tested it by simulating the D14A variant of the lambda repressor monomer,[13,14] an 80 residue protein, in implicit solvent (Onufriev−Bashford−Case generalized Born model[15]). All bond lengths were constrained, as well as angles of the form H−X−H or H−O−X. This gives a total of 1570 constraints, none of which are redundant. Keeping all elements of $\mathbf{K}^{-1}$ whose absolute value is greater than 0.1 gives 8.1 nonzero elements per constraint, making the matrix−vector multiplies extremely fast. If we instead keep all elements greater than 0.01, there are 19.9 nonzero elements per constraint. The maximum number of nonzero elements in any row of $\mathbf{K}^{-1}$ (that is, the maximum number of other constraints that any constraint is directly affected by) is 22 with a cutoff of 0.1, or 47 with a cutoff of 0.01.

Simulations were run using time steps of 1−4 fs with both CCMA and SHAKE. Iteration was continued until all constraints were satisfied to within a relative tolerance of $10^{-4}$. All simulations used a Langevin integrator to couple the protein to a thermal bath at 300 K with a friction coefficient of 91 $\text{ps}^{-1}$.

The results are shown in Table 1. CCMA requires only a small fraction as many iterations as SHAKE. More computation is required for each iteration due to the matrix−vector multiply, but the total number of FLOPS is still much smaller. We profiled a single threaded CPU implementation of each algorithm to precisely measure the computational work required for each one. When using a 4 fs time step, 1.1% of the total CPU time is spent in the SHAKE algorithm, while CCMA with a cutoff of 0.1 takes up 0.8% of the total CPU time.

More importantly, CCMA is easily parallelized. This makes it a far more efficient algorithm than SHAKE on modern parallel architectures. Massively parallel processors such as GPUs typically have hundreds or even thousands of processing units, and all other parts of the simulation can be efficiently implemented on them.[12] SHAKE, being a single threaded algorithm, would then become more expensive than all other parts of the simulation put together. In contrast, CCMA can be efficiently parallelized and remains a small contributor to the computation time on a GPU.

The rate of convergence is only weakly affected by how many elements of $\mathbf{K}^{-1}$ we keep. Decreasing the cutoff from 0.1 to 0.01 decreases the average required iterations by 3−16%, but it also more than doubles the number of elements (and hence the cost of the matrix−vector multiply). Cutoffs much larger than 0.1, on the other hand, do significantly impact the convergence. The optimal value for this cutoff will depend on the detailed performance of a particular implementation. On a cluster, for example, there is a communication overhead for every iteration, so it is probably best to use a small value; on a multicore shared memory computer, a larger value that minimizes the total amount of computation will likely be faster.

We also studied the effect of constraint topology on the rate of convergence. We repeated the above simulations using a time step of 2 fs, but constraining only bond lengths, not any angles. In that case, the average number of iterations for SHAKE drops by a factor of 3 to 9.67, while CCMA drops to 2.56 with a cutoff of 0.01, or to 3.52 with a cutoff of 0.1. We see that CCMA is less sensitive than SHAKE to the constraint topology. This is not surprising, since the accuracy of its approximation to the Jacobian does not change significantly, whereas SHAKE uses a much less accurate approximation when constraints are highly coupled.

## Conclusions

We have developed a new constraint algorithm for use in molecular simulations. It produces very rapid convergence and has a lower overall computational cost than many popular algorithms. It can be used for arbitrary constraint topologies and works well for constraining angles as well as bond lengths. It also is easy to parallelize, making it a good choice for use on modern parallel architectures.

## Availability

The implementation reported in this paper will be made available at Simtk.org as part of the OpenMM API (http://simtk.org/home/openmm). OpenMM is designed for incorporation into molecular dynamics codes to enable execution on GPUs and other high-performance architectures.

**Acknowledgment.** The authors would like to thank Michael Sherman for useful conversations related to the design of this algorithm. This work was supported by Simbios via the NIH Roadmap for Medical Research Grant U54 GM072970.

## References

(1) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. Improving Efficiency of Large Time-Scale Molecular Dynamics Simulations of Hydrogen-Rich Systems. *J. Comput. Chem.* **1999**, *20*, 786–798.

(2) Tironi, I. G.; Brunne, R. M.; van Gunsteren, W. F. On the Relative Merits of Flexible Versus Rigid Models for Use in Computer Simulations of Molecular Liquids. *Chem. Phys. Let.* **1996**, *250*, 19–24.

(3) Kräutler, V.; van Gunsteren, W. F.; Hünenberger, P. H. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. *J. Comput. Chem.* **2001**, *22*, 501–508.

(4) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(5) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comp. Phys.* **1977**, *23*, 327–341.

(6) Miyamoto, S.; Kollman, P. A. SETTLE: An Analytical Version of the SHAKE and RATTLE Algorithm for Rigid Water Models. *J. Comput. Chem.* **1992**, *13*, 952–962.

(7) Vaidehi, N.; Jain, A.; Goddard, W. A., III. Constant Temperature Constrained Molecular Dynamics: The Newton−Euler Inverse Mass Operator Method. *J. Phys. Chem.* **1996**, *100*, 10508–10517.

(8) Schwieters, C. D.; Clore, G. M. Internal Coordinates for Molecular Dynamics and Minimization in Structure Determination and Refinement. *J. Magn. Reson.* **2001**, *152*, 288–302.

(9) Weinbach, Y.; Elber, R. Revisiting and Parallelizing SHAKE. *J. Comp. Phys.* **2005**, *209*, 193–206.

(10) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B* **2004**, *108*, 750–760.

(11) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P., *Numerical Recipes in C++*, 2nd ed.; Cambridge University Press: Cambridge, U.K., 2003.

(12) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; LeGrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. Accelerating Molecular Dynamic Simulation on Graphics Processing Units. *J. Comput. Chem.* **2009**, *30*, 864–872.

(13) Yang, W. Y.; Gruebele, M. Rate−Temperature Relationships in $\lambda$-Repressor Fragment $\lambda_{6-85}$ Folding. *Biochemistry* **2004**, *43*, 13018–13025.

(14) Yang, W. Y.; Gruebele, M. Folding $\lambda$-Repressor at Its Speed Limit. *Biophys. J.* **2004**, *87*, 596–608.

(15) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins* **2004**, *55*, 383–394.

CT900463W

JCTC Journal of Chemical Theory and Computation

# Problems with Some Current Water Models for Close Pair Interactions That Are Not Near the Minimum Energy

Eric M. Yezdimer[†,‡] and Robert H. Wood[*,§]

*Industrial Summit Technology Corporation, 250 Cheesequake Road, Parlin, New Jersey 08859, I.S.T. Corporation, 5-13-13 Ichiriyama, Otsu, Shiga 520-2153, Japan, and Department of Chemistry and Biochemistry, University of Delaware, Newark, Delaware 19716*

Received August 21, 2009

**Abstract:** The ability of an empirical, polarizable model of water to predict a thermal ensemble of molecular configurations at ambient conditions was examined using first-principle quantum mechanics. The empirical model of water selected for this evaluation was the TTM2-F model. The quantum mechanical methodology selected was the second-order Møller−Plesset model (MP2). Only pairwise interaction energies were considered. Significant deviations from the empirical model were found. Similar results were found for ad-hoc comparisons with several other common water models including the TIP3P, TIP4P, TIP4P-FQ, TIP5P, TTM2.1-F, TTM2.2-F, TTM3-F, and POL5/QZ potential models. Our results show that spatially close dimer configurations with interaction energies notably above the potential well minimum (but are still thermally accessible at ambient conditions) are the source of the largest deviations. To assist others in future water model parametrizations we report the MP2 near complete basis set limit energies for 840 water dimer configurations sampled from an approximate thermal ensemble at ambient conditions.

## 1. Introduction

This work started as an attempt to calculate the molecular radial distribution functions of bulk liquid water at room temperature using the complete basis set (CBS) limit of the second-order Møller−Plesset (MP2)[1] quantum mechanical model. Direct simulation of bulk liquid water at a density of 997 kg/m$^3$ requires a minimum of at least 64 water molecules to produce a simulation box large enough that system size effects can safely be ignored.[2] Direct quantum simulations using a basis set large enough to approximate the CBS limit for the MP2 model is not computationally feasible for this system size and molecular density, even using the largest supercomputers currently available. During this research the authors had hoped to circumvent this technical barrier by using non-Boltzman sampling.[3] The idea was to use a molecular dynamics simulation of an approximate water model to produce a thermal ensemble of molecular configurations and then reweight those configurations to yield the molecular distribution functions for the MP2 model. The approximate model chosen for this study was the TTM2-F model of Burnham and Xantheas.[4] This model, unlike a full MP2 quantum mechanical simulation, is relativity computationally inexpensive and allows a large number of independent molecular configurations for bulk liquid water to be generated. In order to reweight the distributions, each ensemble configuration was broken down into subsets of water dimer and trimer configurations. The two- and three-body configurational energies from each subset can then be calculated using the MP2 model near the CBS limit using a more reasonable amount of CPU resources. Each of the quantum energies is then compared directly to the energy predicted from the approximate model. The radial distribution functions for the approximate model can then be converted, through an energy reweighting process,[3] into the radial distribution for a water molecule whose interaction energies with the surrounding waters is given by a quantum

[*] To whom correspondence should be addressed.

[†] Industrial Summit Technology Corporation.

[‡] I.S.T. Corporation.

[§] University of Delaware.

10.1021/ct900447n © 2010 American Chemical Society
Published on Web 01/13/2010

mechanical model, while the interactions between the surrounding waters are given by the approximate model. This method avoids a large, brute force quantum mechanical simulation. This methodology has been proven successful in previous studies predicting the $g_{Na,O}$[5] and $g_{Cl,O}$[6] in high-temperature aqueous solutions. However, in the present calculations, Bootstrap error estimates showed that the 160 independent configurations sampled in this study were not nearly enough to achieve accurate radial distribution functions. Since the number of independent configurations needed is very strongly dependent on the accuracy of the model, it is easier to improve the model rather than sample thousands of configurations.

The poor result for liquid water was very surprising because the approximate model, TTM2-F, has been shown to accurately predict the MP2/CBS energies of ice clusters of 2−6 water molecules to within 0.2 kJ/mol per hydrogen bond.[4] A study of our results showed that there were many pair configurations whose interaction energies were not accurately described by the TTM2-F model. These configurations were often spatially close and far above the minimum energy configuration. Comparisons with other well-studied polarizable (TIP4P-FQ,[7] TTM2.1-F,[8] TTM3-F,[9] and POL5/QZ[10]) and nonpolarizable (TIP3P,[11,12] TIP4P,[2] and TIP5P[13]) models of water also showed poor accuracy for many of these configurations. The purpose of this paper is to report on the deficiencies of the above models. We also report all of our molecular configurations and interaction energies in the hope they will be useful in the development of new and more accurate water models.

## 2. Methods

Molecular trajectories from a NVT simulation for a system of 128 TTM2-F water molecules at room temperature were provided by George Fanourgakis and Sotiris S. Xantheas. From these configurations 160 independent configurations were sampled. To efficiently calculate $\Delta U$ for each configuration it is convenient to explicitly define the total potential energy for each configuration, $U$, as a series of single- and multiple-body interactions

$$U = \sum_i u_i^o + \sum_i (u_i - u_i^o) + \sum_{i<j} (u_{ij} - u_i - u_j) + \sum_{i<j<m} (u_{ijm} - u_{ij} - u_{im} - u_{jm}) + ... \quad (1)$$

where $u_{\{x\}}$ denotes the total potential energy contained in a set of $\{x\}$ molecules and the first, second, third, and fourth terms denote the minimum potential energy of the monomers, the distortion energy of the monomers, the pair wise potential energies ($U_2$), and the three-body potential energies ($U_3$), respectively. In this study we calculated only the pairwise potential energies, $U_2$.

Pairwise potential energies ($U_2$) were calculated using the MP2/aug-cc-pVDZ ($U_{DZ}$) and MP2/aug-cc-pVTZ ($U_{TZ}$) quantum methods. All quantum calculations were performed using the NWChem software package (version 4.7).[14] For all MP2 calculations basis set superposition error (BSSE) corrections were made using the counterpoise method.[15] Preliminary comparisons of the DZ energies with the TTM2 model showed many large differences for oxygen−oxygen distances ($R_{OO}$) less than 4.1 Å. We selected 60 dimer outliers from the first solvation shell ($R_{OO} < 4.1$) and performed additional MP2/aug-cc-pVQZ ($U_{QZ}$) and MP2/aug-cc-pV5Z ($U_{5Z}$) calculations (corrected for BSSE). We repeated these calculations using 50 additional dimer geometries chosen at random again with $R_{OO} < 4.1$ Å. The complete basis set limit for pairwise interactions ($U_{CBS}$) for each configuration was determined by fitting the following equation

$$U_{LZ} = U_{CBS} + \frac{b}{(1 + L)^4} + \frac{c}{(1 + L)^5} \quad (2)$$

where $U_{CBS}$, $b$, and $c$ are adjustable parameters. The value $L$ denotes the maximum angular momentum for the aug-cc-pVDZ, aug-cc-pVTZ, aug-cc-pVQZ, and aug-cc-pV5Z basis sets and was taken to be 2, 3, 4, and 5, respectively. Energies that were not BSSE corrected did not typically show a clear convergence progression with increased basis set size, so they were not used.

We noticed a strong correlation of both $R_{OO}$ and ($U_{TZ} - U_{DZ}$) with ($U_{TZ} - U_{CBS}$) for our sample of 110 pair configurations. Least-squares fits showed that $U_{TZ}$ could be corrected to near the CBS limit (NCBS) by the equation

$$U_{NCBS} = U_{TZ} + 0.238 + 0.280(U_{TZ} - U_{DZ}) + 0.0263(U_{TZ} - U_{DZ})^2 - 148.9R_{OO}^{-5} \quad (3)$$

where the energies are given in units of kJ/mol and the distances are in units of Angstroms. For our sample of 110 pair configurations, $U_{NCBS}$ is more accurate than $U_{5Z}$: the average and standard deviation of ($U_{CBS} - U_{NCBS}$) are 0.00 and 0.13 kJ/mol, respectively, while for ($U_{CBS} - U_{5Z}$) they are −0.245 and 0.13 kJ/mol. The maximum difference between $U_{CBS}$ and $U_{NCBS}$ was 0.34 kJ/mol. This correlation appeared highly accurate, so we used it to calculate $U_{NCBS}$ for all of our 840 pair configurations with $R_{OO} < 4.1$ Å.

During the course of this research it also became clear that an unphysical decrease in the van der Waals potential of the TTM2-F model at O−O distances of under 2.2 Å could cause unrealistic molecular configurations to be generated during a molecular dynamics simulation run. A revised form of the TTM2-F dubbed TTM2.1-F with the addition of an extra exponential repulsive term was published to correct this problem.[8] As a precaution we checked all the oxygen−oxygen distances inside our selected clusters and found that only a few were under 2.5 Å and none were under 2.4 Å. The difference in the van der Waals potential energy between the TTM2-F and TTM2.1-F potential models at the energy minimum for $R_{OO} = 2.4$ and 2.5 Å is 0.7 and 0.2 kJ/mol, respectively, giving us confidence that this deficiency in the TTM2-F model did not significantly skew our sampling of the ensemble. We then calculated the energies of the TTM2.1-F model using the same configurations selected from the TTM2-F ensemble. The results were very similar, and the energy differences between the TTM2.1-F and the MP2/CBS models remained quite large.

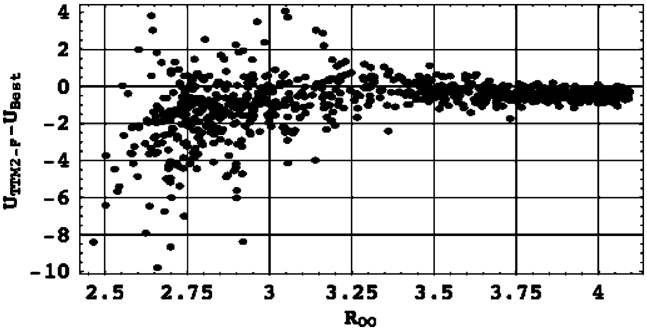


***Figure 1.*** Differences in pairwise energy between the TTM2-F model and the Best MP2 estimate ($U_{TTM2-F} - U_{Best}$) are plotted versus $R_{OO}$ (Å) for our sample of 840 pairs. $U_{Best}$ is the near CBS limit when the CBS limit is not available. Energies are given in kJ/mol.

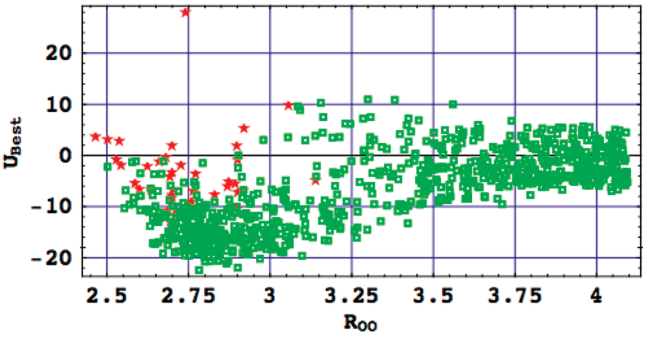


***Figure 2.*** Pairwise energies for the interaction of a central water molecule with another surrounding water molecule, $U_{Best}$, as a function of the oxygen−oxygen distance, $R_{OO}$ (Å). The 5% of points with the largest errors ($U_{TTM2-F} - U_{Best}$) are plotted as red stars. These 42 points have model energies, $U_{TTM2-F}$, that are too negative by −3.8 to −9.9 kJ/mol. Energies are given in kJ/mol.

## 3. Results and Discussions

**3.1. Comparison with the TTMX-F Models.** We then compared the models with our most accurate pairwise energy, $U_{Best}$, which was $U_{NCBS}$ when $U_{CBS}$ was not available. Please note that we calculated $u_{ij} - u_i - u_j$ for both models, and this did not include the distortion energy, $u_i - u_o$ energy.

Figure 1 shows the difference $U_{TTM2-F} - U_{Best}$ as a function of $R_{OO}$. For $R_{OO}$ greater than 3.2 Å the differences are small. For $R_{OO}$ less than 3.2 Å, the energy differences are much larger and ranged from +4 to −10 kJ/mol, with an average difference of −0.8 kJ/mol and a standard deviation of 1.5 kJ/mol. For our sample of 840 pairs with $R_{OO} < 4.1$ Å there are a few outliers that are too high in energy by 2−4 kJ/mol and many outliers that were too low in energy by 4−10 kJ/mol. Figure 2 shows $U_{Best}$ versus $R_{OO}$ for all water dimers with $R_{OO} < 4.1$. There are many near-neighbor water molecules possessing repulsive pairwise interactions that are up to 12 kJ/mol and even one as high as 28 kJ/mol. In Figure 2 the 5% of points with the largest errors are plotted as red stars. These 42 points have TTM2-F model energies that are too negative by −3.8 to −9.9 kJ/mol. Figure 2 shows that these outliers occur when $R_{OO}$ is short and $U_{Best}$ is high, that is, these molecules are close neighbors but do not form strong hydrogen bonds. The TTM2-F model does not reflect these ab initio predictions and instead describes dimer energies that are considerably lower. We suppose the source of these discrepancies could be either (1) an unknown deficiency in the equations governing the potential model or (2) a failure to appropriately represent these kinds of configurations during parametrization.

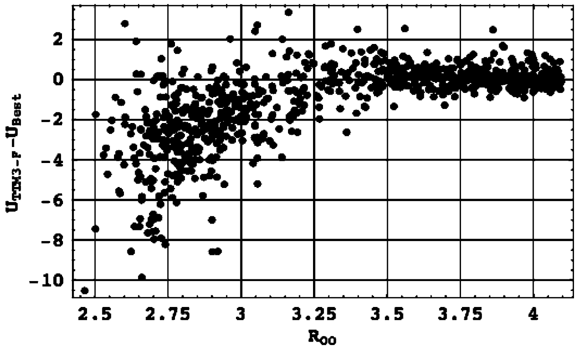


***Figure 3.*** Differences in the pairwise energy between the TTM3-F model and the Best MP2 estimate ($U_{TTM3-F} - U_{Best}$) are plotted versus $R_{OO}$ for our sample of 840 pairs. $U_{Best}$ is the near CBS limit when the CBS limit is not available. Energies are given in kJ/mol.

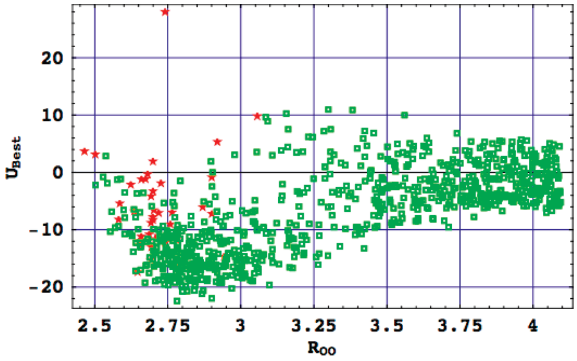


***Figure 4.*** Pairwise energies for the interaction of a central water molecule with another surrounding water molecule, $U_{Best}$, as a function of the oxygen−oxygen distance, $R_{OO}$. The 5% of points with the largest errors ($U_{TTM3-F} - U_{Best}$) are plotted as red stars. These 42 points have model energies, $U_{TTM3-F}$, that are too negative by −4.9 to −10.2 kJ/mol. Energies are given in kJ/mol.

Next we considered the TTM3-F model,[9] which is a substantial revision motivated by the failure of this and other models to reproduce the OH stretching vibrations in both water clusters and in liquid water. Using the same configurations selected from the TTM2-F ensemble the dimer energies were recalculated using the TTM3-F potential model. For our selected configurations with $R_{OO} < 4.1$ average differences between the TTM2-F and the TTM3-F models were small. The average difference is −0.23 kJ/mol, with a maximum positive difference of 2.5 kJ/mol and a maximum negative difference of −5.3 kJ/mol. Although Figures 3 and 4 show that the revisions from the TTM2-F model to the TTM3-F model did increase the number of outliers below −6 kJ/mol from 11 to 22, the 11 configurations that were the worst outliers for the TTM2-F model are still outliers in the TTM3-F model (<−6 kJ/mol) and the outliers again occur mainly for interactions between very close neighbors. Of the 22 outliers, 20 have configurations that would normally be called hydrogen bonds ($R_{OO} < 3.2$ Å and O−H−O angle > 140°). The hydrogen bond length is also very short ($1.46 < R_{OH} < 1.82$) for 17 of the 20 configurations. The other two outliers have $R_{OH}$ near 2 Å and O−H−O angles near 127°. These findings indicate there are many configurations that are both accessible at ambient temperatures and not well represented by either the TTM2-F or the TTM3-F model.

In addition to the outliers, there is a systematic bias in the results; the average differences are $U_{\text{TTM-F}} - U_{\text{Best}}$, −0.8 kJ/mol, and $U_{\text{TTM3-F}} - U_{\text{Best}}$, −1.1 kJ/mol. The systematic bias in the dimer energies is cumulative and leads to significant errors for the entire molecular configuration. Since these models have the correct total energy for liquid water at ambient conditions, a systematic and canceling error in the multibody interactions is possible.

**3.2. Comparison with Other Models.** The poor performance of the TTM2-F and TTM3-F models surprised us, so we looked at some of the standard water models to see how they performed. To this end we arbitrarily chose the TIP3P, TIP4P, TIP4P-FQ, TIP5P, and POL5/QZ models and calculated the energy for the 110 configurations for which we had MP2/CBS energies. The TIP3P (its simplest flexible form was used[12]) and TIP4P models were studied because of their prevalence in commercial molecular dynamics simulations. The TIP5P model was included because of its excellent description of $g_{OO}$ as measured by recent X-ray diffraction experiments.[13] Technically it is not fair to compare the TTM2-F dimer configurations using water potentials with a fixed internal geometry. Complete fairness however was not our purpose here as we only wanted to gain a generalized understanding as to how well other models performed for these configurations, and we settled on making an ad-hoc comparison with two slightly different methods. In one comparison, the rigid model was confined to the plane of the flexible model with oxygens superimposed and the M site on the H−O−H bisector (the A models). In the other comparison the rigid molecules were distorted with oxygens and hydrogens superimposed, and the M site was positioned on the bisector. To calculate the energy of this distorted configuration the site−site interactions between the water molecules were calculated with the usual algorithm even though the molecule was distorted (the B models). The differences between the two calculations (A and B) were small compared to the errors discussed here (see below).

We first examined the distribution of energy differences between the models. Figure 5 shows the percentile distribution of energy differences between each model and the ab initio pairwise energies ($U_{\text{Model}} - U_{\text{CBS}}$). The small uncertainties shown in Figure 5 were estimated from the difference between the results of the two methods of calculation (A and B above).

As expected, the TTM2-F and TTM3-F models were very similar and the most accurate. They have the smallest spreads of any models with a fairly negative bias. The polarizable TIP4P-FQ and POL5/QZ models show a smaller but positive bias, but the spreads are larger. Both of these models had smaller minimum errors than the TTM3-F model (the minimum errors are only about −4 kJ/mol). However, these models had much higher maximum errors (the maximum energies are 15 and 20 kJ/mol, respectively).

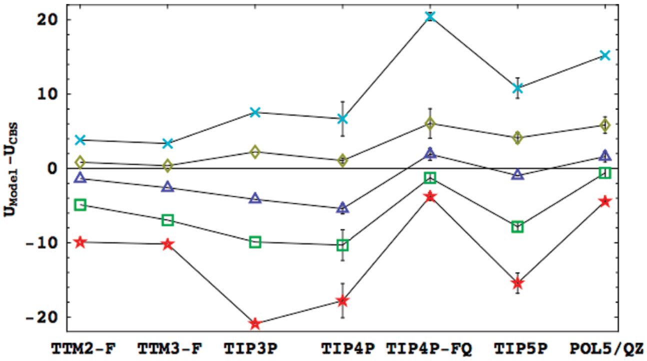


***Figure 5.*** Percentile distributions of the differences in pairwise energy ($U_{\text{Model}} - U_{\text{CBS}}$) for 110 configurations are plotted for each model. The symbols are from top to bottom: the maximum difference, the difference that is higher than 83.5% of the differences, the median difference, the difference that is higher than 16.5% of the differences, and the minimum difference. Energies are given in kJ/mol. Uncertainties less than the symbol size are not shown.

It is important to note that comparing the present ab initio pairwise energies to nonpolarizeable models, such as the TIP3P, TIP4P, and TIP5P models, is also not fair because these models were never meant to reproduce specific pairwise interactions. They instead use an "effective" pair interaction that was empirically adjusted to compensate for the average multibody interactions in liquid water at room temperature. This compensation is probably the reason for the negative biases of these models. The TIP5P model however, which is known to give an excellent description of the oxygen−oxygen radial distribution function at ambient conditions, appears to have a minimal average bias in the pairwise interactions. This is puzzling because we would expect a more negative bias because the effective pairwise potential should include a significant contribution to empirically account for attractive three-body interactions. In our opinion these three models have some very large errors that are too large to be solely due to neglecting compensations for multibody effects. As before, the largest errors were spatially close.

For each model we plotted both the error and $U_{\text{Best}}$ as a function of $R_{OO}$ as in Figures 1 and 2 to find out where the largest outliers were located. We looked at the top 11 outliers of each model. For all models the A and B comparisons were very similar. The highest 11 outliers all had $R_{OO}$ less than 3 Å with most of them being close interactions with $R_{OO}$ less than 2.75 Å. For TTM3-F, TIP3P, and TIP4P (A and B) the top 11 outliers were all negative and TIP5P had most of its outliers negative. For TIP4P-FQ and POL5/QZ the outliers were all positive. This is interesting as the TIP4P-FQ and POL5/QZ are fluctuating point charge models and treat polarization in a fundamentally different way than the TTM2-F and TTM3-F models. The values of $U_{\text{Best}}$ showed that almost all of the outliers did not have strong hydrogen bonds. We also checked whether the same configurations tended to be outliers for more than one model. We found that only 12 configurations accounted for 59% of the outliers for the models in Figure 5. Nine of these 12 configurations had large errors with both polarizable and nonpolarizable models.

## 4. Concluding Remarks

Many current models do a poor job of representing spatially close dimer interactions that are not near the minimum energy. There are substantial numbers of these "outliers" found in simulations of water at ambient conditions, so adjustments to the models are necessary. It is believed that ab initio methods are very helpful in parametrizing polarizable models for close interactions. Complete results on the 840 pairs used in this study are given in the Supporting Information, so that these pairs can be used in future modeling efforts. The main limitation of this study is that only pairwise interactions were studied. The errors found are much larger than the expected errors in the quantum method used.

**Acknowledgment.** We thank Sortiris Xantheas and George Fanourgakis for providing the results of a molecular dynamics simulation of the TTM2-F model and Harry Stern and Bruce Berne for providing the POL5/QZ dimer energies. The quantum mechanical calculations described in this paper were performed using the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research, located at Pacific Northwest National Laboratory. We thank Douglas Doren for many helpful discussions.

**Supporting Information Available:** Complete results including coordinates and energies of the 840 pairs used in this study. This material is available free of charge via the Internet at http://pubs.acs.org.

### References

(1) Møller, C.; Plesset, M. S. Note on an approximation treatment for Many-electron systems. *Phys. Rev.* **1934**, *46* (7), 618–622.

(2) Jorgensen, W. L.; Madura, J. D. Temperature and size dependence for Monte Carlo simulations of TIP4P water. *Mol. Phys.* **1985**, *56* (6), 1381–1392.

(3) Wood, R. H.; Liu, W. B.; Doren, D. J. Rapid calculation of the structures of solutions with ab initio interaction potentials. *J. Phys. Chem. A* **2002**, *106* (29), 6689–6693.

(4) Burnham, C. J.; Xantheas, S. S. Development of transferable interaction models for water. IV. A flexible all-atom polarizable potential (TTM2-F) based on geometry dependent charges derived from ab initio monomer dipole moment surface. *J. Chem. Phys.* **2002**, *116* (12), 5115–5124.

(5) Dong, H.; Liu, W.; Doren, D. J.; Wood, R. H. Structure of an accurate ab initio model of the aqueous $Na^+$ ion at high temperatures. *J. Phys. Chem. B* **2008**, *112*, 13552–13560.

(6) Dong, H.; Liu, W.; Doren, D. J.; Wood, R. H. Structure of an accurate ab initio model of the aqueous $Cl^-$ ion at high temperatures. *J. Phys. Chem. B* **2006**, *110*, 18504–18514.

(7) Rick, S. W.; Stuart, S. J.; Berne, B. J. Dynamical fluctuating charge force-fields: Application to liquid water. *J. Chem. Phys.* **1994**, *101*, 6141–6156.

(8) Fanourgakis, G. S.; Xantheas, S. S. The flexible, polarizable, thole-type interaction potential for water (TTM2-F) revisited. *J. Phys. Chem. A* **2006**, *110*, 4100–4106.

(9) Fanourgakis, G. S.; Xantheas, S. S. Development of transferable interaction potentials for water V. Extension of the flexible, polarizable, Thole-type model potential (TTM3-F, v. 3.0) to describe the vibrational spectra of water clusters and liquid water. *J. Chem. Phys.* **2008**, *128* (7), 074506, 1–11.

(10) Stern, H. A.; Rittner, F.; Berne, B. J.; Friesner, R. A. Combined fluctuating charge and polarizable dipole models: Application to a five-site water potential function. *J. Chem. Phys.* **2001**, *115* (5), 2237–2251.

(11) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(12) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Kenneth, M.; Merz, J.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of protein, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(13) Mahoney, M. W.; Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **2000**, *112* (20), 8910–8922.

(14) Aprà, E.; Windus, T. L.; Straatsma, T. P.; Bylaska, E. J.; de Jong, W.; Hirata, S.; Valiev, M.; Hackler, M.; Pollack, L.; Kowalski, K.; Harrison, R.; Dupuis, M.; Smith, D. M. A.; Nieplocha, J.; Tipparaju, V.; Krishnan, M.; Auer, A. A.; Brown, E.; Cisneros, G.; Fann, G.; Früchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R. ; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z. *NWChem: Computational Chemistry Package for Parallel Computers*, version 4.7; Pacific Northwest National Laboratory: Richland, WA, 2004.

(15) Boys, S. F.; Bernardi, F. Calculation of small molecular interactions by differences of separate total energies - some procedures with reduced errors. *Mol. Phys.* **1970**, *59* (19), 553–566.

CT900447N

JCTC Journal of Chemical Theory and Computation

# Multi-Level Ewald: A Hybrid Multigrid/Fast Fourier Transform Approach to the Electrostatic Particle-Mesh Problem

David S. Cerutti* and David A. Case

*Department of Chemistry and Chemical Biology, and BioMaPS Institute, Rutgers University, 610 Taylor Road, Piscataway, New Jersey 08854-8066*

Received October 3, 2009

**Abstract:** We present a new method for decomposing the one convolution required by standard Particle−Particle Particle-Mesh ($P^3M$) electrostatic methods into a series of convolutions over slab-shaped subregions of the original simulation cell. Most of the convolutions derive data from separate regions of the cell and can thus be computed independently via FFTs, in some cases with a small amount of zero padding so that the results of these subproblems may be reunited with minimal error. A single convolution over the entire cell is also performed, but using a much coarser mesh than the original problem would have required. This “Multi-Level Ewald” (MLE) method therefore requires moderately more FFT work plus the tasks of interpolating between different sizes of mesh and accumulating the results from neighboring subproblems, but we show that the added expense can be less than 10% of the total simulation cost. We implement MLE as an approximation to the Smooth Particle-Mesh Ewald (SPME) style of $P^3M$ and identify a number of tunable parameters in MLE. With reasonable settings pertaining to the degree of overlap between the various subproblems and the accuracy of interpolation between meshes, the errors obtained by MLE can be smaller than those obtained in molecular simulations with typical SPME settings. We compare simulations of a box of water molecules performed with MLE and SPME and show that the energy conservation, structural, and dynamical properties of the system are more affected by the accuracy of the SPME calculation itself than by the additional MLE approximation. We anticipate that the MLE method’s ability to break a single convolution into many independent subproblems will be useful for extending the parallel scaling of molecular simulations.

## 1. Introduction

Observing biochemical processes through computer simulations requires thorough equilibrium sampling of a protein or nucleic acid system with thousands of degrees of freedom. The quality of the molecular model is of utmost importance, but validation requires extensive simulations to yield precise results for properties such as equilibrium conformations,[1] crystallographic temperature factors,[2] binding energies,[3] and molecular folding rates.[4] The capabilities of molecular simulations and the models themselves therefore evolve in step with computer performance and parallel algorithm design.

The central challenge with parallel molecular dynamics algorithms is the treatment of electrostatic interactions. Because the electrostatic potential decays as the inverse distance, charged particles influence one another at long-range, implying a great deal of information sharing and potentially a great deal of algorithmic complexity, as much as $O(N^2)$ in the number of particles $N$. Particle $\leftrightharpoons$ mesh implementations of the Ewald sum,[5,6] and more generally Particle−Particle Particle-Mesh ($P^3M$) methods,[7] are popular choices for treating long-ranged electrostatic forces in

* Corresponding author phone: (732) 445-0334; fax: (732) 445-5958; e-mail: dcerutti@rci.rutgers.edu.

10.1021/ct900522g
Published on Web 12/29/2009

molecular simulations because of the favorable complexity of the algorithms, $O(M \log M)$ or $O(M)$, for a number of mesh grid points $M$ depending on the choice of Poisson solver. For commodity hardware, Poisson solvers based on the fast-Fourier transforms (FFTs) are commonly used because of their computational efficiency,[5,8] which is so great that the parallel scaling of these approaches is still limited, on most clusters, to a few hundred processors. Most molecular dynamics codes meet high scaling targets by dedicating a subset of the processors to the FFT, but the number of messages and the amount of data that must be shared can still limit the total number of processors that can be devoted to the calculation and thus the maximum speed of molecular simulations.

Some recent variations of P$^3$M[9,10] make use of real-space Poisson solvers based on finite difference or multigrid methods. These approaches offer better algorithmic complexity ($O(M)$ for the real-space methods, versus $O(M \log M)$ for the FFT-based methods) and asymptotically better interprocessor communication in parallel calculations. However, because the real-space solvers require significantly more work to map the particles' charge density to the mesh and extract forces from the mesh, their only successful application has been on specialized hardware.[9]

Other strategies for solving particle ⇋ mesh problems can be found in the broad class of Multilevel Summation methods pioneered by Brandt[11] and developed for molecular simulations by Skeel[12] and others, the Fast-Multipole Method,[13] and Adaptive P$^3$M techniques used in astrophysical gravity calculations.[14−16] These mesh refinement techniques, along with the basic P$^3$M method, may all be viewed as variations on the theme of smoothly and (in essence) isotropically splitting a long-ranged potential into short- and long-ranged components that can be accurately represented on meshes of different resolutions. Like multigrid Poisson solvers, when applied to condensed-phase systems, the mesh refinement methods also exchange computational effort for scaling benefits, but the simplicity of the algorithms and communication patterns makes these methods highly adaptable for applications on commodity hardware and general-purpose graphics processor units.

In this Article, we present an alternative method for replacing the one convolution required by traditional electrostatic P$^3$M solvers with a series of convolutions, each pertaining to a slab subdomain of the simulation cell, building up to a single convolution involving a coarsened mesh which describes the entire simulation cell. In contrast to the smooth splitting employed by other mesh refinement methods, our approach is to split the mesh-based potential sharply and anisotropically such that the individual components all contain discontinuities but can nonetheless recover a smooth potential when summed. This "Multi-Level Ewald" (MLE) approach produces the electrostatic potential in a single pass over all levels of the mesh; the extra computational effort is small.

We explore numerous strategies for manipulating the parameters of MLE scheme itself and the details of the associated particle ⇋ mesh operations to tune the accuracy of the resulting forces and energies with small amounts of overlap between adjacent slabs. Approximating the reciprocal space convolution using MLE can incur scarcely more error in the resulting particle forces than would be obtained with an equivalent P$^3$M calculation. We expect that MLE can reduce the data communication requirements of molecular dynamics simulations for modern networked computing architectures and will prove adaptable for balancing communication loads when the network connectivity is heterogeneous.

## 2. Theory

**2.1. The Problem of Computing Long-Ranged Electrostatics, the Ewald Solution, and Its Evolution into P$^3$M.** The Ewald method can be summarized as splitting the calculation of the electrostatic energy of a periodic system of point (or otherwise highly localized) charges $E^{(\mathrm{coul})}$ into a "reciprocal space" sum describing the energy of a system of spherical Gaussian charges, which has identical coordinates to the system of interest, and a "direct space" sum, which modifies the energy of the reciprocal space sum to recover the energy of the system of point charges:

$$\begin{aligned}
E^{(\mathrm{coul})} &= \frac{1}{2}\sum_{n}\sum_{i}\sum_{j\neq i}\frac{q_iq_j}{4\pi\varepsilon_0|\mathbf{n}\cdot\mathbf{L}+\mathbf{r}_{ij}|}\\
E^{(\mathrm{dir})} &\simeq \frac{1}{2}\sum_{i}\sum_{j\neq i}\frac{q_iq_j(1-\mathrm{erf}(\beta|\mathbf{r}|))}{4\pi\varepsilon_0|\mathbf{r}_{ij}|} = \sum_{i}\sum_{j\neq i}q_iq_j\theta^{(\mathrm{dir})}\\
E^{(\mathrm{rec})} &= \frac{1}{2}\sum_{n}\sum_{i}\sum_{j\neq i}\frac{q_iq_j\mathrm{erf}(\beta|\mathbf{r}|)}{4\pi\varepsilon_0|\mathbf{n}\cdot\mathbf{L}+\mathbf{r}_{ij}|} = \sum_{i}\sum_{j\neq i}q_iq_j\theta^{(\mathrm{rec})}\\
E^{(\mathrm{coul})} &= E^{(\mathrm{dir})}+E^{(\mathrm{rec})}
\end{aligned}\tag{1}$$

In these equations, $\mathbf{n}\cdot\mathbf{L}$ represents images of the unit cell throughout all space, $i$ and $j$ run over all charged particles in the system, $\mathbf{r}_{ij}$ is the distance between particles $i$ and $j$, $\varepsilon_0$ is the permittivity of free space, and $\beta$ is the "Ewald coefficient". The reciprocal and direct space sums, $E^{(\mathrm{rec})}$ and $E^{(\mathrm{dir})}$, obtain their names because each converges absolutely in Fourier (reciprocal) or real (direct) space, respectively. The splitting is necessary because a straightforward summation over many periodic images of all charges in the system will not converge absolutely.

In its original formulation, the Ewald method relies only on the positions of particles. The direct space calculation involves a loop over all particles with a nested loop over each particle's neighbors within a cutoff distance $L_{\mathrm{cut}}$ sufficient to give a convergent direct space sum. The reciprocal space calculation involves a loop over all particles with a nested loop that again involves all particles.

Splitting a potential into short- and long-ranged components is also the basis of Particle−Particle Particle-Mesh (P$^3$M) methods.[7] These similarities led Darden and colleagues to propose "Particle-Mesh Ewald"[6] as a special case of P$^3$M, which incorporated a Gaussian function for splitting the potential function. Many variants of this particular case have since been developed,[5,9,10] along with distinct approaches for optimizing the influence function that modulates the interaction of charges on the mesh.[17] In all of these methods, the basic procedure may be summarized: (1) assign charges to the density mesh $Q$ using the positions and charges of particles plus a suitable particle → mesh interpolation kernel, (2) solve the field on the mesh by convolution with the mesh-based representation of the interparticle potential function, and (3) interpolate forces on all particles given the particle positions, charges, field values, and particle → mesh interpolation kernel.

The most common motivation for using particle-mesh strategies is to exploit the convolution theorem, which states that for two sequences of numbers $f_1$ and $f_2$,

$$f_1 \bigstar f_2 = \mathscr{F}^{-1}[\mathscr{F}(f_1)\cdot\mathscr{F}(f_2)] \tag{2}$$

Above, ★ represents a convolution, $\mathscr{F}(f)$ is the (fast) discrete Fourier transform (FFT) of $f$, $\mathscr{F}^{-1}(f)$ is the inverse FFT of $f$ such that $\mathscr{F}^{-1}(\mathscr{F}(f)) = f$, and $\mathscr{F}(f_1)\cdot\mathscr{F}(f_2)$ is the element-wise product of $\mathscr{F}(f_1)$ and $\mathscr{F}(f_2)$.

The most popular variant of P$^3$M for electrostatics, Smooth Particle-Mesh Ewald (SPME),[5] makes use of cardinal B-splines[18] to map the system's charges to the mesh $Q$. An elegant derivation of the Fourier transform of the reciprocal space pair potential, $\mathscr{F}(\theta^{(\mathrm{rec})})$, is obtained by folding together Euler exponential splines in the mesh $B$ (the Fourier-space representation of B-splines are Euler splines) and the Fourier-space representation of the Gaussian charge smoothing function $W$ (the Fourier transform of a Gaussian is another Gaussian):

$$B(x, y, z) = |b(x)|\cdot|b(y)|\cdot|b(z)|$$

$$b(\eta_\alpha) = \frac{\exp(2\pi i(n-1)\eta_\alpha/g_\alpha)}{[\sum_{p=0}^{n-2} M_n(p+1)\exp(2\pi i p\eta_\alpha/g_\alpha)]} \tag{3}$$

$$W(x, y, z) = \frac{\exp(-4\pi^2|\mathbf{k}|^2\sigma^2)}{\pi V|\mathbf{k}|^2} \tag{4}$$

$$\mathscr{F}(\theta^{(\mathrm{rec})}) = B\cdot W \tag{5}$$

In the equations defining $B$ and $W$, $M_n$ represents a cardinal B-spline of order $n$, $i$ is the square root of $-1$, α is one of the mesh dimensions $x$, $y$, or $z$, $\eta_\alpha$ is a displacement in the mesh dimension α, and $g_\alpha$ is the size of the mesh in α. Also, in the equation defining $W$, $V$ is the volume of the simulation cell, $\sigma$ is the rms of the Gaussian charge smoothing function (note that $\sigma = 1/(2\beta)$), and $\mathbf{k}$ is the displacement from the origin in Fourier space. (Readers should consult the original SPME reference[5] for a detailed presentation of the derivation of this approximation to $\theta^{(\mathrm{rec})}$ and particle ⇋ mesh interpolation using B-splines. We have provided the most important definitions here because they will be important later as we develop our new method.) After $\mathscr{F}(\theta^{(\mathrm{rec})})$ has been prepared, the electrostatic potential $U^{(\mathrm{rec})}$ is computed with only two FFT operations:

$$U^{(\mathrm{rec})} = \mathscr{F}^{-1}[\mathscr{F}(Q)\cdot\mathscr{F}(\theta^{(\mathrm{rec})})] \tag{6}$$

The electrostatic potential energy of the system $E^{(\mathrm{rec})}$ may then be obtained by element-wise multiplication of the charge density $Q$ with the electrostatic potential:

$$E^{(\mathrm{rec})} = Q\cdot U^{(\mathrm{rec})} \tag{7}$$

This operation would be performed in real space and would require that a copy of the original charge density be saved before computing $U^{(\mathrm{rec})}$. To avoid this extra memory requirement, FFT-based Poisson solvers use an identity to obtain $E^{(\mathrm{rec})}$ during the element-wise multiplication in Fourier space, when the system virial is available as well. However, we emphasize the real-space expression for the energy as this will be necessary as we develop a replacement to the convolution step of P$^3$M methods in electrostatics.

**2.2. Accurate Decomposition of the Mesh-Based Sum: The MLE Method.** In the interest of improving the parallel scaling of P$^3$M methods for molecular electrostatics, we focused on improving the method in which $Q\bigstar\theta^{(\mathrm{rec})}$ is computed while preserving other aspects of the algorithm. Our approach was to split $\theta^{(\mathrm{rec})}$ into fine and coarse resolution components as shown in Figure 1. Rather than splitting the potential isotropically in terms of the absolute distance between points, however, the splitting is done anisotropically along planes perpendicular to one dimension, which we will call $\hat{x}$. The fine resolution potential $\theta_\mathrm{h}^{(\mathrm{rec})}$ exactly describes the interactions between any two mesh points separated by up to (and including) $T_\mathrm{cut}$ in $\hat{x}$, regardless of the distance between the points in the other unit cell dimensions $\hat{y}$ and $\hat{z}$. Conversely, the low resolution pair potential $\theta_\mathrm{c}^{(\mathrm{rec})}$ approximately describes the interactions of points separated by more than $T_\mathrm{cut}$ in $\hat{x}$, no matter their locations in $\hat{y}$ and $\hat{z}$. For convenience, we will refer to the set of mesh points that share the same coordinate in the $\hat{x}$ direction as a "page" of the mesh. The meshes of different resolutions used in our approximation will be referred to as different "levels" of mesh. Our convention is to call the finest resolution mesh the lowest level; the extent of the reciprocal space pair potential grows as the mesh level becomes higher.

With $\theta^{(\mathrm{rec})}$ split into two components, the convolution can be restated:

$$Q\bigstar\theta^{(\mathrm{rec})} \simeq Q\bigstar\theta_\mathrm{h}^{(\mathrm{rec})} + Q_\mathrm{c}\bigstar\theta_\mathrm{c}^{(\mathrm{rec})} = U_\mathrm{h}^{(\mathrm{rec})} + U_\mathrm{c}^{(\mathrm{rec})} \simeq U^{(\mathrm{rec})} \tag{8}$$

where $Q_\mathrm{c}$ is a coarsened charge mesh interpolated from $Q$. (Note that, while $\theta_\mathrm{h}^{(\mathrm{rec})}$ is sparse and $\theta_\mathrm{c}^{(\mathrm{rec})}$ contains a void where $\theta_\mathrm{h}^{(\mathrm{rec})}$ is nonzero, $Q$ and $Q_\mathrm{c}$ are full: every point in $Q_\mathrm{c}$ is interpolated from the appropriate points in $Q$.) This approximation does not immediately reduce the communication requirements of computing $Q\bigstar\theta^{(\mathrm{rec})}$, but because $\theta_\mathrm{h}^{(\mathrm{rec})}$ is zero in all but a narrow region $2(T_\mathrm{cut}) + 1$ pages thick, the convolution $Q\bigstar\theta_\mathrm{h}^{(\mathrm{rec})}$ can be accomplished as a series of convolutions:

$$Q\bigstar\theta_\mathrm{h}^{(\mathrm{rec})} = \sum_{i=1}^{P}[Q_i\bigstar\theta_\mathrm{h}^{(\mathrm{rec})}] \tag{9}$$

where each submesh $Q_i$ spans the simulation box in $\hat{y}$ and $\hat{z}$ and is zero-padded $2T_\mathrm{cut}$ pages in $\hat{x}$. We will refer to these submeshes as "slabs". Each of the $P$ slabs of the lowest level mesh is therefore padded by $2T_\mathrm{cut}$ pages of zeroes, and each of the series of convolutions described in eq 9 can be computed independently. The padding is done so that FFTs

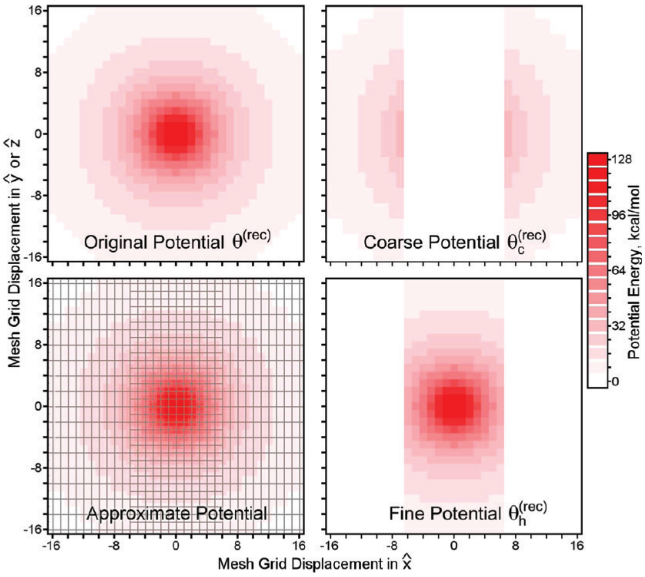


***Figure 1.*** A multilevel mesh-based approximation to the Ewald reciprocal space pair potential. In Ewald mesh calculations, the reciprocal space pair potential $\theta^{(rec)}$ can be visualized in real space. At short-range, $\theta^{(rec)} \sim \mathrm{erf}(\beta|\mathbf{r}|)/(4\pi\varepsilon_0|\mathbf{r}|)$, where $\varepsilon_0$ is the permittivity of free space and erf is the error function. $\theta^{(rec)}$ was computed for a 96 Å cubic box on a mesh of $96^3$ points; slices of the potential through the *xy* (or *xz*) planes are shown with varying intensities of red to indicate the magnitude. The color scale is deliberately coarse to make the potential isocontours apparent. $\theta^{(rec)}$ varies most rapidly along paths passing directly through the source at (0,0,0); paths that move tangentially to the source encounter much slower variations in $\theta^{(rec)}$. It is more feasible to approximate $\theta^{(rec)}$ with high- and low-resolution potentials $\theta_h^{(rec)}$ and $\theta_c^{(rec)}$ as shown, avoiding mesh $\leftrightharpoons$ mesh interpolation along vectors pointed at the source as much as possible. $\theta_c^{(rec)}$ uses double the mesh spacing along the $\hat{y}$ and $\hat{z}$ axes, as indicated by the mesh overlay in the lower left panel, but the same spacing as $\theta_h^{(rec)}$ in the $\hat{x}$ direction. $\theta_c^{(rec)}$ therefore presents a fine mesh spacing for interpolating gradients of the true potential $\theta^{(rec)}$ when the true potential varies rapidly, but presents a coarse spacing when the true potential varies slowly.

may be used for the convolution without having charges near the *yz* faces of any slab "wrap around" and erroneously influence other parts of the same slab, and so that the influence of charges near the *yz* faces of each slab will be recorded as the results of all of these convolutions over slabs are then spliced back together to accumulate the approximation to $U^{(rec)}$. The electrostatic influence of charge density in slab $Q_i$ on the neighboring slabs is recorded in its zero-padded pages. (Generally, the neighbors of $Q_i$ are $Q_{i-1}$ and $Q_{i+1}$, although $Q_P$ and $Q_1$ are neighbors due to the periodicity of the unit cell.) The basic procedure is illustrated in Figure 2. In principle, convolutions with many radially symmetric potential functions could be split in this manner, although we focus on the case of the inverse distance kernel for application to biomolecular simulations. Numerous styles of P$^3$M are also compatible with this convolution splitting procedure; we have chosen to implement it within the Smooth Particle-Mesh Ewald style described above and call the new method "Multi-Level Ewald" (MLE).

In our MLE implementation, the coarsened reciprocal space pair potential $\theta_c^{(rec)}$ is obtained simply by extracting points from $\theta^{(rec)}$ at regular intervals of the coarsening factor, $C_{yz}$, in the $\hat{y}$ and $\hat{z}$ directions:

$$\theta_c^{(rec)}(i,j,k) = \begin{cases} \theta^{(rec)}(i, C_{yz}j, C_{yz}k) & ,i > T_{cut} \text{ or } i < = g_x - T_{cut} \\ 0 & ,\text{otherwise} \end{cases} \qquad (10)$$

where $i$, $j$, and $k$ represent coordinates in the $\hat{x}$, $\hat{y}$, and $\hat{z}$ directions, respectively, and $g_x$ is the mesh size in the $\hat{x}$ direction. (Here, $k$ should not be confused with $\mathbf{k}$ used earlier to describe a vector in Fourier space.) While this approach may appear to discard much of the information present in $\theta^{(rec)}$, we will show that it can produce very high accuracy depending on the other MLE parameters. We wrote an optimization procedure to try and improve the coarsened reciprocal space pair potential mesh by using steepest descent optimization to adjust the values of $\theta_c^{(rec)}$ at individual mesh points and minimize the root mean squared (rms) error in the approximate $U^{(rec)}$. This approach could only reduce the error rate of MLE calculations by about 2% (data not shown) and was not given further consideration in these studies.

Similar to the construction of $\theta_c^{(rec)}$, $Q_c$ is interpolated from $Q$ using cardinal B-splines[18] similar to those used for particle $\leftrightharpoons$ mesh interpolation in standard SPME. However, mesh $\leftrightharpoons$ mesh interpolation is a two-dimensional process as the mesh resolution is only reduced in $\hat{y}$ and $\hat{z}$ and maintained in $\hat{x}$. Each page of the mesh $Q_c$ is interpolated from the corresponding page of $Q$. As we will show in the Results, it can be advantageous to use relatively high values of the order of mesh $\leftrightharpoons$ mesh interpolation $I^{(mm)}$ as opposed to the order of particle $\leftrightharpoons$ mesh interpolation $I^{(pm)}$. In the same way that higher values of $I^{(pm)}$ improve the accuracy of SPME calculations, higher values of $I^{(mm)}$ improve the accuracy of the MLE approximation; however, whereas the cost of an SPME calculation scales as the cube of $I^{(pm)}$ because each particle has a different alignment to the mesh, the regularity of the mesh $\leftrightharpoons$ mesh interpolation makes it separable in each dimension, and thus the operation scales merely as $I^{(mm)}$.

While $\theta_h^{(rec)}$ will typically span a small region of the simulation box, if $\theta_c^{(rec)}$ were to span the rest it might be impractical to compute $Q_c\bigstar\theta_c^{(rec)}$ as a series of convolutions. However, it is still possible to add more mesh levels by splitting $\theta_c^{(rec)}$ into its own "fine" and "coarse" resolution components $\theta_{c,1}^{(rec)}$, $\theta_{c,2}^{(rec)}$, ..., $\theta_{c,n}^{(rec)}$, in the same manner that the original $\theta^{(rec)}$ was split (see Figure 3). The most general expression of the MLE method is then:

$$Q\bigstar\theta^{(rec)} = \sum_{i=1}^{P_1}[Q_i\bigstar\theta_h^{(rec)}] + \sum_{j=2}^{L-1}\sum_{k=1}^{P_j}[Q_{c,j,k}\bigstar\theta_{c,j}^{(rec)}] + Q_{c,L}\bigstar\theta_{c,L}^{(rec)} \qquad (11)$$

The scheme above involves $L - 1$ coarsened meshes with as many distinct coarsening factors. In general, a single convolution of the highest level charge mesh with the coarsest component of the reciprocal space pair potential must be performed, involving data collected over the entire simulation cell. However, because the mesh spacing in the

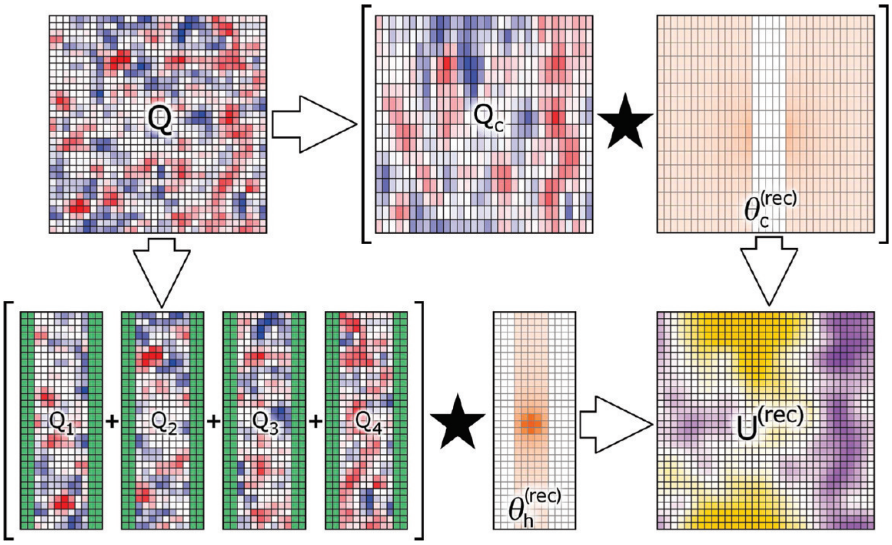


***Figure 2.*** Illustration of the Multi-Level Ewald convolution procedure. In typical Smooth Particle-Mesh Ewald (SPME) calculations, the charge mesh $Q$ is convoluted with $\theta^{(rec)}$ to arrive at the reciprocal space electrostatic potential $U^{(rec)}$. In Multi-Level Ewald (MLE), this single convolution is replaced with many smaller ones. In a basic two-level variation of MLE, the mesh $Q$ is split, with no interpolation, into multiple subregions (slabs) $Q_1$...$Q_L$ as shown. The slabs are then zero-padded, as highlighted in green in the diagram, so that they may be convoluted with the high-resolution reciprocal space pair potential $\theta_h^{(rec)}$, which is itself extracted from $\theta^{(rec)}$ without interpolation. The coarsened charge mesh $Q_c$ is interpolated from $Q$ and convoluted with the coarsened reciprocal space pair potential $\theta_c^{(rec)}$ (see Figure 1). An electrostatic potential at the resolution of the fine mesh is then interpolated from the result of $Q_c\bigstar\theta_c^{(rec)}$ to complete the approximation of $Q\bigstar\theta^{(rec)}$. This figure was made using an actual MLE calculation on a 32 Å$^3$ box of 4000 randomly distributed ions. The color scales are not given as the diagram is qualitative, but in the meshes $Q$, $Q_c$, and $Q_1$...$Q_4$ red and blue signify negative and positive charge, the intensity of orange signifies the intensity of the pair potential, and purple and gold signify negative and positive electrostatic potential in the resulting $U_c^{(rec)}$. Each colored pixel corresponds to a point in a plane cutting through the mesh in the actual MLE calculation.

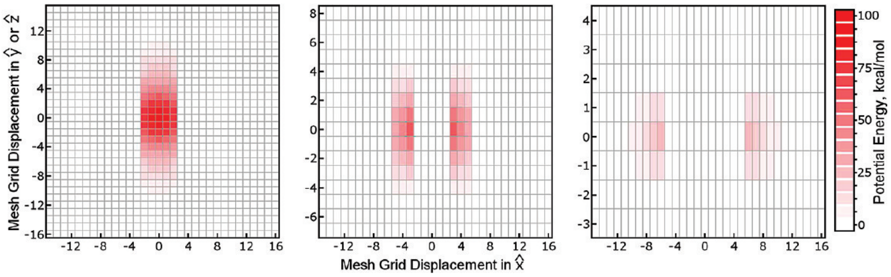


***Figure 3.*** A three-level MLE scheme. As illustrated above, the reciprocal space pair potential mesh can be split into three (or more) separate meshes, each with successively larger coarsening factors. Here, there are two coarsened meshes, with coarsening factors $C_{yz}$ of 2 and 4, respectively. In this scheme, the pair potential in the lowest level mesh extends 2 pages; slabs of the lowest level charge mesh would require 4 pages of zero-padding. The pair potential in the intermediate level mesh has $T_{cut}$ = 5, although only 6 of its pages have nonzero potential values in them (the thickness of the nonzero region of the mesh is 2 × 5 + 1 = 11 pages). While slabs of the intermediate-level charge mesh would require 10 pages of zero-padding, the intermediate level mesh is much smaller than the lowest level mesh, making such a degree of padding more economical. The color scale is not the same as that for Figure 1 because the SPME calculation this MLE scheme approximates was not the same; the diagram is intended for qualitative understanding only.

highest level charge mesh can be 2−6 times larger than the mesh spacing in $Q$, calculating $Q_{c,L}\bigstar\theta_{c,L}^{(rec)}$ is not so demanding as calculating $Q\bigstar\theta^{(rec)}$, and the communication burden is likewise reduced.

**2.3. Considerations for Constant Pressure Simulations.** It is important to note that, to make MLE run efficiently, $\theta^{(rec)}$ must be computed by taking the inverse Fourier transform of $\mathscr{F}(\theta^{(rec)})$ as is typically computed in particle-mesh methods, then extracting $\theta_h^{(rec)}$ and $\theta_c^{(rec)}$, and finally computing the Fourier transforms $\mathscr{F}(\theta_h^{(rec)})$ and $\mathscr{F}(\theta_c^{(rec)})$. This preparatory work must be done at the beginning of the simulation so that during each step of dynamics the necessary convolutions described in eq 12 or 13 can be accomplished with only two Fourier transforms each. This necessity may appear to limit the applicability of the MLE approximation to constant volume systems, where $\theta^{(rec)}$ is constant throughout the simulation. However, if the unit cell volume rescaling in constant pressure simulations is isotropic, and the Gaussian charge smoothing parameter $\sigma$ and the mesh spacing $\mu$ vary in proportion to the unit cell dimensions, the updated pair potentials $\theta_h^{(rec)*}$ and $\theta_c^{(rec)*}$ for any new unit cell volume can be obtained be simply rescaling the $\theta_h^{(rec)}$ and $\theta_c^{(rec)}$ obtained at the beginning of the simulation.

***Table 1.*** Test Cases for the Multi-Level Ewald Method[a]

| case | cell dimensions<br>($a$, $b$, $c$), Å | cell dimensions<br>($\alpha$, $\beta$, $\gamma$) | atom count |
|---|---|---|---|
| water | 31.4 × 31.4 × 31.4 | 90°, 90°, 90° | 3072 |
| streptavidin | 89.7 × 89.7 × 89.7 | 90°, 90°, 90° | 73 305 |
| protein crystal | 71.4 × 71.4 × 75.6 | 90°, 90°, 120° | 36 414 |
| glycerol solution | 69.7 × 69.7 × 89.0 | 60°, 90°, 90° | 39 808 |
| cyclooxygenase-2 | 114.8 × 114.8 × 114.8 | 109.5°, 109.5°, 109.5° | 118 833 |

[a] The cases presented here span a variety of simulation cell geometries. All systems are in the condensed phase and were pre-equilibrated by molecular dynamics simulations at constant pressure.

## 3. Methods

**3.1. The MDGX Program.** To test the Multi-Level Ewald (MLE) method, we wrote an in-house molecular dynamics program, MDGX (Molecular Dynamics with Gaussian Charges and Explicit Polarization—not all parts of the acronym are yet fulfilled, as the purpose of the program is to be a proving ground for new algorithms). Routines in MDGX are able to read AMBER topology files and produce outputs in a format like that of the SANDER module in the AMBER software package.[19] The MDGX program is able to run unconstrained molecular dynamics trajectories of systems such as a box of SPC-Fw water molecules[20] in the microcanonical (NVE) ensemble or simply compute energies and forces acting on all atoms of a system for a single set of coordinates. The MDGX program implements both Smooth Particle-Mesh Ewald (SPME) as well as our new MLE method and also offers the option of using different particle ⇋ mesh interpolation orders in different dimensions, a feature that we will show is very helpful for tuning MLE. When run with identical parameters, the SPME reciprocal space electrostatic forces computed by MDGX agree with those of sander to $1.0 \times 10^{-9}$ relative precision. MDGX links with the FFTW[21] library to perform its FFTs.

**3.2. A Matlab Ewald Calculator.** While the MDGX program is an excellent tool for testing MLE and other new variants of P$^3$M, it currently only works with orthorhombic unit cells (the reciprocal space code is actually set up to perform calculations with nonorthorhombic cells, but the direct space domain decomposition is not yet ready in this respect). The MDGX program was therefore only used for calculations involving rectangular unit cells.

Before MDGX was created, Multi-Level Ewald was discovered and verified through a set of script functions written for the Matlab software package (The MathWorks, Inc., Natick, MA). These scripts, which can easily produce forces and energies for a particular set of coordinates and charges but are not efficient enough to propagate a lengthy molecular dynamics trajectory, were used for any calculations involving nonorthorhombic unit cells. The calculator facilitates analysis of every stage of the Multi-Level Ewald process through Matlab's high-level programming language and is available from the authors on request.

**3.3. Test Systems.** As presented in Table 1, we chose a number of systems representative of those found in typical, condensed phase, biomolecular simulations. The first, a system of 1024 SPC-Fw water molecules, was used for testing energy conservation and ensemble properties of the system collected over long molecular dynamics runs. The other systems were much larger protein-in-water and protein crystal systems used in our previous study developing a different P$^3$M method.[22] Most importantly, these systems span a variety of unit cell types: as will be shown, MLE can be performed with any type of unit cell, but the geometry of the unit cell itself affects the accuracy of the MLE approximation.

**3.4. Smooth Particle-Mesh Ewald Accuracy Standards and Reference Calculations.** Standards for the accuracy of electrostatic forces in molecular simulations must be established before assessing the accuracy of MLE approximations with respect to SPME targets. Generally, we chose the accuracy obtained by the default settings of the SANDER molecular dynamics engine in the AMBER software package[19] as a reasonable level of accuracy for molecular simulations. These settings are mesh spacing $\mu$ as close to 1.0 Å as possible given a mesh size $g$ with prime factors 2, 3, 5, and possibly 7, particle ⇋ mesh interpolation order $I^{(\mathrm{pm})} = 4$, and direct sum tolerance $D_{\mathrm{tol}} = 1.0 \times 10^{-5}$ with direct space cutoff $L_{\mathrm{cut}} = 8.0$ Å leading to a Gaussian charge smoothing half width $\sigma = 1.434$ Å. They can be expected to produce electrostatic forces with a root mean squared (rms) error of about $1.0 \times 10^{-2}$ kcal/mol·Å, but the exact number varies depending on the system composition and geometry. Most SPME calculations for this work were performed with these parameters, and most modifications to the parameters were made in such a way as to conserve the overall accuracy of the calculation. For reference, very high-quality SPME calculations were performed with $\mu$ ~0.4 Å, $I^{(\mathrm{pm})} = 8$, and an identical value of $\sigma$ to ensure that the direct space and reciprocal space components of the SPME calculation could both be compared to the reference. The reference calculations produced forces convergent to within $1.0 \times 10^{-6}$ kcal/mol·Å.

## 4. Results

**4.1. Accuracy of Forces and Energies Computed with Multi-Level Ewald.** The most important products of our new Ewald reciprocal space approximation are correct reproduction of the electrostatic energy of a system of particles and correct reproduction of the gradients of that energy. As our implementation of Multi-Level Ewald (MLE) is an approximation to the Smooth Particle-Mesh Ewald (SPME) method, we performed SPME calculations with high-accuracy "reference" parameters as well as typical molecular dynamics parameters and compared MLE results to both. In typical SPME calculations, there are two sources of error to consider, arising from the direct and reciprocal space parts of the calculation, respectively. MLE uses the same direct space sum but approximates the reciprocal space sum, introducing "coarsening" errors into the electrostatics calculation. We define the coarsening errors as deviations in the MLE approximation away from the equivalent SPME calculation, where the "equivalent" SPME calculation uses the same $\sigma$, $I^{(\mathrm{pm})}$, and $\mu$ parameters as the MLE calculation and has its own degree of inaccuracy relative to the SPME reference calculation. Fundamentally, the coarsening errors are errors in the scalar values of the electrostatic potential

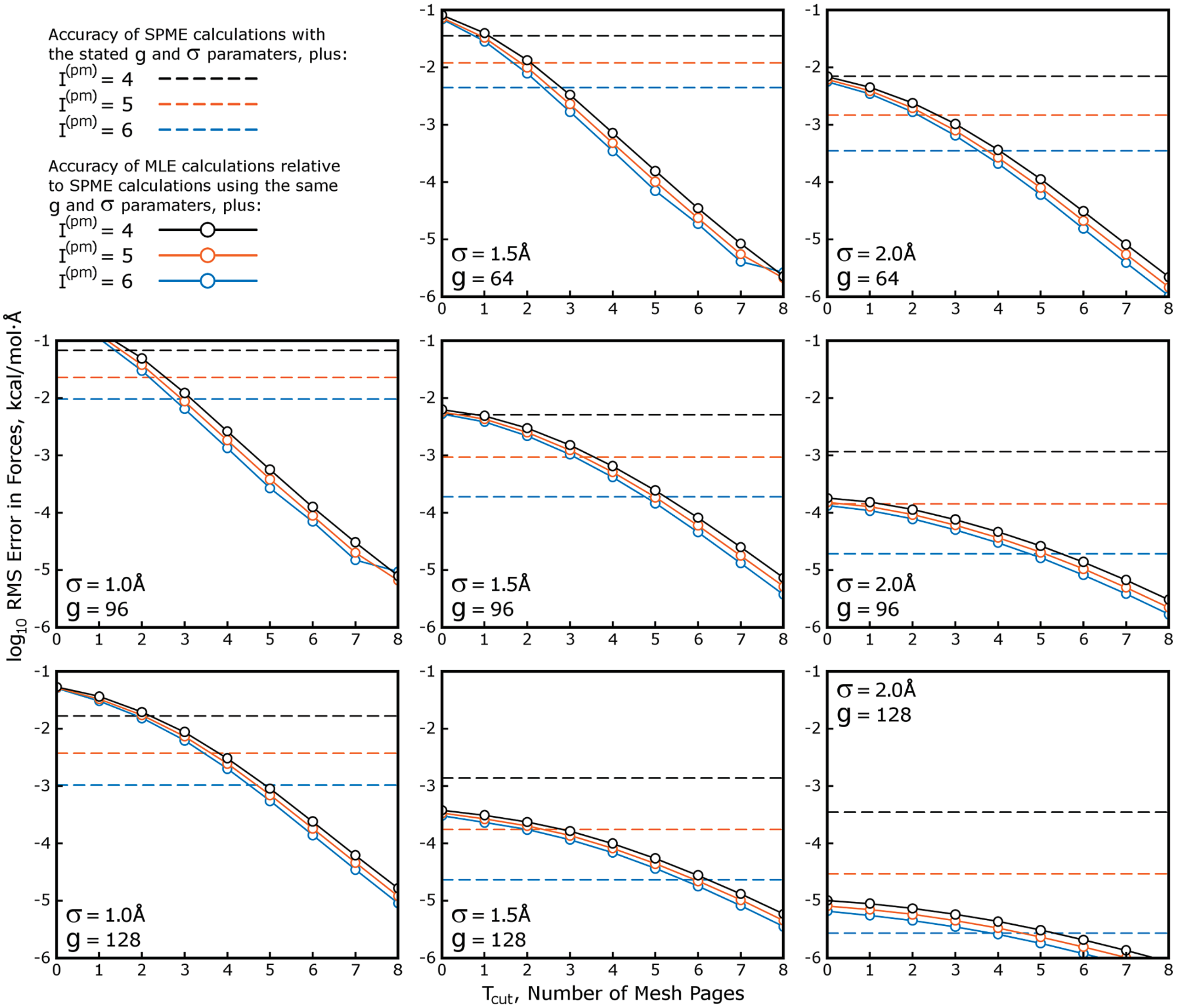


***Figure 4.*** Accuracy of MLE calculations shows dependence on the parameters of the SPME calculation that is being approximated. SPME calculations were performed on the streptavidin test case with a range of Gaussian charge smoothing half widths $\sigma$ and mesh sizes $g \times g \times g$. In practice, SPME calculations run with particle ⇋ mesh interpolation order $I^{(pm)} = 4$ require that $\sigma$ be at least 1.5× the mesh spacing $\mu$ to produce reasonable accuracy in the forces arising from the reciprocal space part; however, if $I^{(pm)}$ is set to 6, $\sigma$:$\mu$ ratios as small as 1.0 can be used. The $\sigma$:$\mu$ ratio in the center panel is roughly 1.6, and it increases moving across the panels from left to right or top to bottom. The $\sigma$ = 1.0 Å, $g$ = 64 case is omitted because it would be far too inaccurate for molecular simulations, no matter the value of $I^{(pm)}$. In each panel, horizontal dashed lines show the accuracy of SPME calculations with the stated parameters (the "equivalent" SPME calculations) relative to a high-accuracy reference calculation performed as described in Methods. Solid lines with "○" depict the accuracy of MLE calculations relative to the equivalent SPME calculations.

at points in $U^{(rec)}$ (see eq 6), which in turn imply errors in the forces on charged particles interpolated from $U^{(rec)}$. We will first focus on the errors in forces, as these are of greatest importance in molecular simulations; energies will be discussed in another section.

To quantify the coarsening errors as a function of the SPME parameters, we ran calculations on the Streptavidin test case described in Table 1 using a range of values for the SPME mesh spacing $\mu$, Gaussian charge smoothing function half width $\sigma$, and interpolation order $I^{(pm)}$. We then approximated the SPME results with MLE calculations using $C_{yz} = 2$, mesh ⇋ mesh interpolation order $I^{(mm)} = 8$, and a range of values for $T_{cut}$ ($C_{yz}$ and $I^{(mm)}$ can be varied to benefit the accuracy of MLE calculations, as we will show later, but their values were fixed for simplicity in this test). The results in Figure 4 show that the accuracy of the MLE approximation improves exponentially with $T_{cut}$ and is also very sensitive to the parameters of the equivalent SPME calculation, particularly $\sigma$ and $\mu$ and to a lesser extent $I^{(pm)}$. Although the values of $\mu$ and $\sigma$ are widely varied and not thoroughly sampled, Figure 4 establishes another important result, that MLE can be used to approximate a wide range of different SPME calculations and, without large values of $T_{cut}$, incur less error than the SPME calculation itself.

Although at first it appears that the accuracy of MLE is least sensitive to $I^{(pm)}$, this parameter can be manipulated to great advantage in MLE calculations. Of all of the commercially or academically available molecular dynamics codes, the Desmond software package[23] is, to our knowledge, the only one to permit different settings of $I^{(pm)}$ in different

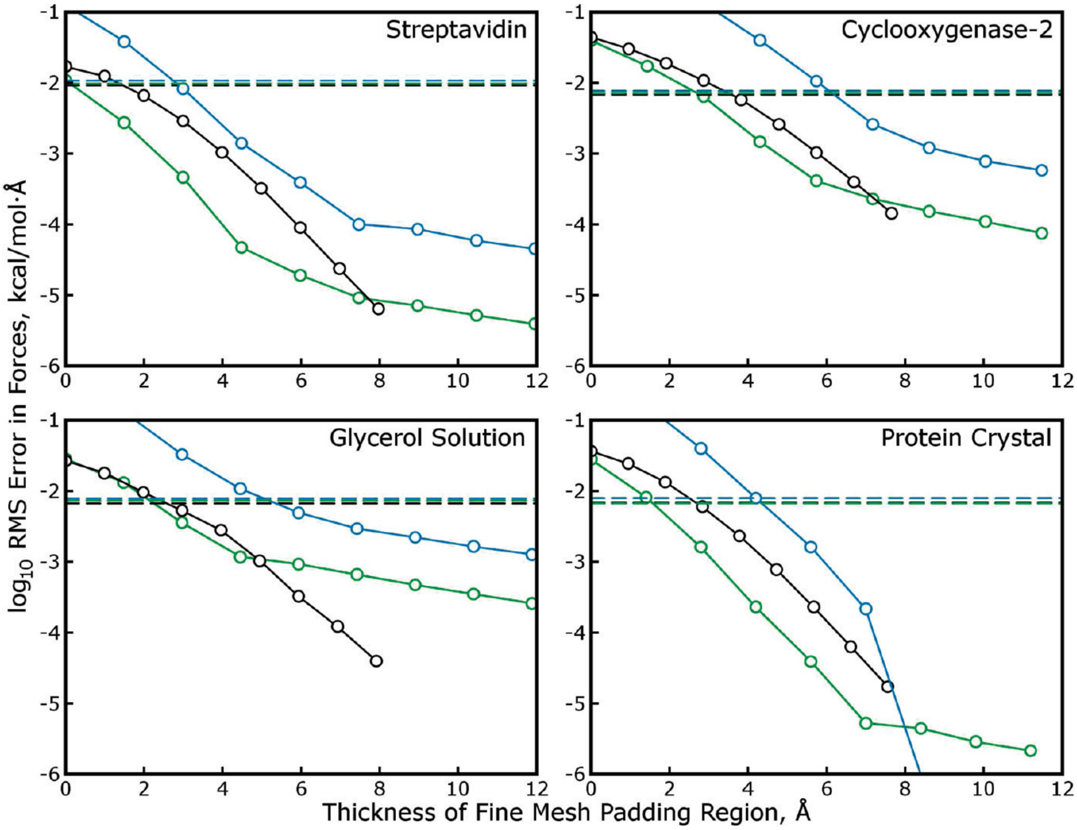


***Figure 5.*** Anisotropic mesh spacings and interpolation orders enhance the accuracy of MLE calculations. SPME calculations were performed on four of the test cases from Table 1, this time using the AMBER default parameters $\sigma \approx 1.4$ Å, $I_x^{(pm)} = I_y^{(pm)} = I_z^{(pm)} = 4$, and the smallest mesh dimensions $g_x$, $g_y$, and $g_z$ such that the $g_x$, $g_y$, and $g_z$ were multiples of 2, 3, and 5 and the mesh spacings $\mu_x$, $\mu_y$, and $\mu_z$ were less than 1.0 Å. Accurate MLE approximations of such SPME calculations are possible for all of these systems, which include monoclinic and triclinic unit cells in addition to the cubic streptavidin system. Modifying the SPME parameters by increasing $\mu_x$ to 1.5 Å and increasing $I^{(pm)}$ to compensate maintains the accuracy of the SPME calculation with a smaller amount of mesh data and can also increase the accuracy of MLE approximations. As in Figure 4, dashed lines represent the accuracy of SPME calculations relative to a high accuracy reference, and lines with "○" represent the accuracy of MLE calculations relative to SPME. Black, green, and blue lines correspond to $I^{(pm)}$ = (4,4,4), (6,4,4), and (6,6,6), respectively. Because the SPME/MLE calculations in each panel use different values of $\mu_x$, the results are plotted in terms of the physical thickness of the padding needed for each MLE approximation, $T_{cut} \times \mu$.

directions. However, we found that this is a powerful way to improve the accuracy of an MLE approximation. As we showed in previous work,[22] setting $I^{(pm)} = 6$ permits $\mu$ to be set as much as 1.5× larger than $I^{(pm)} = 4$ would allow; the result in fact applies in one, two, or all three dimensions. Strictly in terms of the number of operations, increasing $I^{(pm)}$ to 6 in only one dimension offers the most reduction in the mesh size per increase in the amount of particle ⇋ mesh work. For example, a mesh of $90^3$ points could be replaced by a mesh of $60 \times 90 \times 90$ points, at the expense of mapping particles to $6 \times 4 \times 4 = 96$ points rather than $4 \times 4 \times 4 = 64$. In contrast, setting $I^{(pm)} = 6$ in all dimensions could produce comparable accuracy in the aforementioned problem with a mesh of $60^3$ points, but at the expense of mapping all particles to 216 mesh points.

We performed additional SPME and MLE calculations on the streptavidin system, this time using the AMBER default parameters (as described in Methods) and a variation on those parameters using $I^{(pm)} = 6$ and $\mu$ approaching 1.5 Å in the $\hat{x}$ direction or in all directions. We also performed tests on other systems described in Table 1 to confirm the accuracy of MLE when applied to nonorthorhombic unit cells. All of these results are presented in Figure 5. While all of the different combinations of $I^{(pm)}$ and $\mu$ produce comparable accuracy in the SPME calculation, and while raising $I^{(pm)}$ will improve the accuracy of MLE calculations if all $\mu$ (and $\sigma$) are held fixed, increasing $\mu$ in this manner appears to be detrimental to the accuracy of the subsequent MLE approximation. However, if only $I_x^{(pm)}$ is raised and $\mu_x$ is increased accordingly, the accuracy of the subsequent MLE approximation is improved significantly in three out of the four cases. Anisotropic interpolation orders and a longer mesh spacing in the $\hat{x}$ direction therefore permit significant reductions in the number of pages $T_{cut}$ that must be computed in zero-padded FFTs and transmitted between neighboring slabs, making MLE cheaper to apply.

As can be seen in Figure 4, raising $I^{(pm)}$ is not the only way to compensate for an increase in $\mu$. Raising $\sigma$ itself is another way to maintain the critical ratio of $\sigma$ to $\mu$. Larger values of $\sigma$ are obtained by using a longer direct space cutoff $L_{cut}$; many codes[23,24] and specialized hardware for running molecular simulations[25] make use of longer values of $L_{cut}$ to reduce the size of the reciprocal space mesh. We therefore tested the accuracy of MLE calculations if larger values of

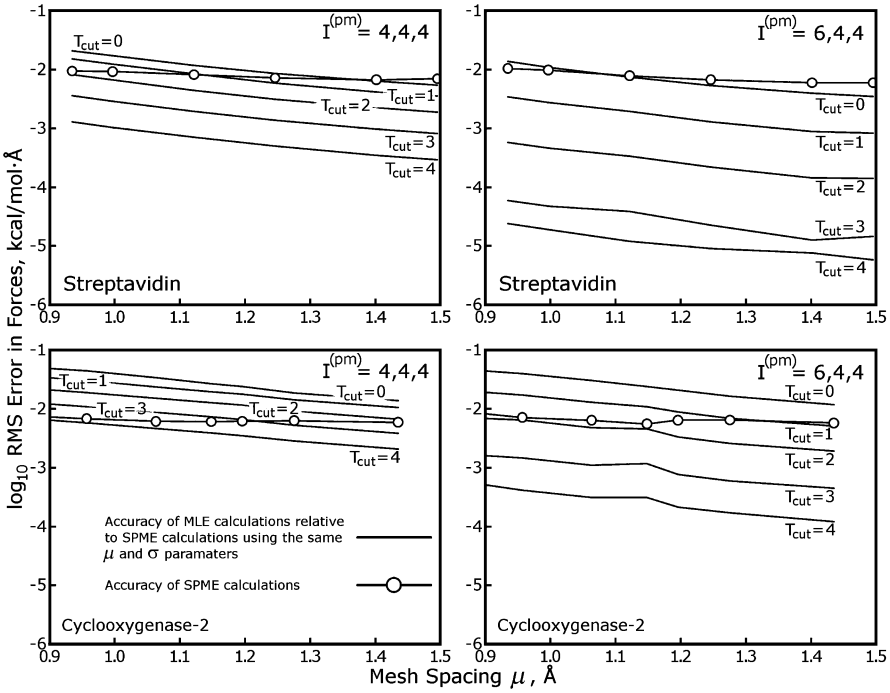


***Figure 6.*** Wider Gaussian smoothing functions enhance the accuracy of MLE. Although using a larger mesh spacing $\mu$ in conjunction with isotropic 6th order particle ⇌ mesh interpolation is detrimental to the accuracy of MLE approximations, it is possible to improve the accuracy of MLE by using a larger $\mu$ and increasing $\sigma$, the rms of the Gaussian charge smoothing function, to maintain the accuracy of the equivalent SPME calculation. The accuracy of MLE approximations for the streptavidin and cyclooxygenase-2 systems is plotted as a function of $\mu$. For each of the equivalent SPME calculations, $\sigma$ was adjusted in proportion to $\mu$ to maintain the $\sigma$ to $\mu$ ratio that would be obtained in each system by the AMBER default parameters, roughly 1.42. As shown by the solid lines with "○", this approach is also effective at conserving the accuracy of the equivalent SPME calculation, even improving it slightly as $\mu$ gets larger. The accuracies of MLE approximations improve steadily as a function of $\mu$. The inset legend in the lower left panel applies to all panels.

$\sigma$, rather than higher $I^{(pm)}$, were used in conjunction with a larger $\mu$. The results in Figure 6 stand in contrast to the results of Figure 5: the MLE approximation becomes more accurate when longer $\mu$ are used, insofar as $\sigma$ is increased accordingly. When using higher $\sigma$ and larger $\mu$, anisotropic particle ⇌ mesh interpolation is still effective at conserving the accuracy of the SPME calculation and continues to benefit the accuracy of the MLE approximation.

Noting that nonorthorhombic unit cells are detrimental to the accuracy of MLE (although only to the extent that $T_{cut}$ must be raised by 1 or 2), we tried MLE calculations with several other monoclinic unit cells, each with only one of the α, $\beta$, or $\gamma$ angles different from 90°. While we had hoped that MLE might be able to give the same accuracy in monoclinic unit cells as in orthorhombic ones if the coarsening occurred in certain dimensions with respect to the nonorthogonal unit cell vectors, the accuracy of MLE showed similar degradation no matter which angle differed from 90° (data not shown).

As mentioned in the Theory, the MLE approximation is tunable in the $I^{(mm)}$ parameter as well as in $T_{cut}$. Knowing that high values of $I^{(mm)}$ are economical in terms of the number of arithmetic operations, we tested the accuracy of MLE for orders of mesh ⇌ mesh interpolation ranging from 4 to 16. The AMBER default parameters and variants with $I^{(pm)} = 6$ were again used for this test. As shown in Figure 7, if a low order of mesh ⇌ mesh interpolation can produce accuracy on the order of the SPME reciprocal space calculation, raising $I^{(mm)}$ can improve the accuracy of an MLE approximation by an order of magnitude.

Figures 5 and 7 show that, with proper choices of $T_{cut}$ and $I^{(mm)}$ to accommodate the parameters of the equivalent SPME calculation, the coarsening errors in MLE calculations can be well below the level of reciprocal space error in the equivalent SPME calculation. However, the form of the coarsening errors themselves must be examined. As will be discussed in the following section, the reciprocal space electrostatic forces accumulate errors as a consequence of inaccuracies in the mesh $U^{(rec)}$, but because the interpolation of $U^{(rec)}$ itself is done only along certain dimensions in MLE calculations, the resulting errors in the reciprocal space electrostatic forces could be expected to be somewhat anisotropic. Figure 8 shows the magnitudes of the coarsening errors acting on individual atoms of the streptavidin system in the $\hat{x}$, $\hat{y}$, and $\hat{z}$ directions under an aggressive MLE approximation (details are given in the figure itself). As might be expected, the coarsening errors tend to be greater in the $\hat{y}$ and $\hat{z}$ directions, but only slightly: coarsening errors in the electrostatic forces also have significant components in the $\hat{x}$ direction, despite the fact that no mesh coarsening was done in $\hat{x}$. As was shown in Figure 5, use of anisotropic SPME parameters reduces the overall error in the MLE approximation, but the individual errors in forces become even more shifted toward the $\hat{y}$ and $\hat{z}$ directions. Conceptually, anisotropic errors are less desirable than isotropic ones, as they might impart the wrong energetics to interactions

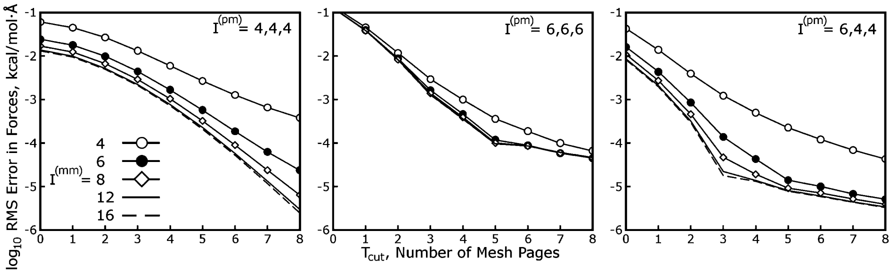


***Figure 7.*** Higher mesh ⇋ mesh interpolation orders can benefit MLE calculations. Interpolation between the finest mesh and higher level meshes in MLE is, like the particle ⇋ mesh interpolation in standard SPME calculations, based on cardinal B-splines; the grid points of the finest mesh can be thought of as particles to map onto coarser meshes. However, mesh ⇋ mesh interpolation of order $I^{(mm)}$ only occurs in only two dimensions, and, because of the regularity of the fine grid, the operations are separable in each dimension leading to $O(I^{(mm)})$ complexity and the possibly much higher orders of $I^{(mm)}$ than $I^{(pm)}$. The accuracy of MLE approximations to the SPME calculations described in Figure 5 was therefore re-evaluated as a function of $I^{(mm)}$ for the streptavidin test case. The results suggest that raising $I^{(mm)}$ is effective if the equivalent SPME calculation makes use of a high $\sigma$:$\mu$ ratio in the dimensions along which the mesh is coarsened (i.e., $\hat{y}$ and $\hat{z}$). The inset legend in the leftmost panel applies to all panels.

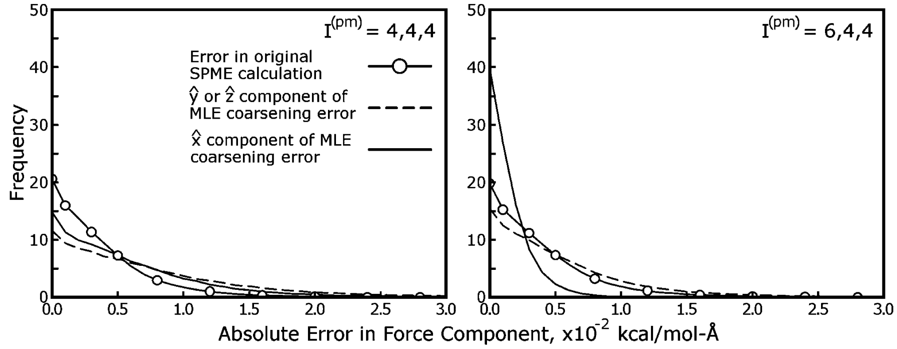


***Figure 8.*** Errors arising from the MLE approximation are anisotropic. Because the MLE approximation is applied in only two of the three unit cell dimensions, errors arising from the approximation may be larger in some dimensions than others. Electrostatic forces on atoms of the streptavidin system in Table 1 were computed using SPME and the AMBER default parameters with the particle ⇋ mesh interpolation schemes given in each panel. As shown by these histograms, the coarsening errors (differences between the equivalent SPME calculation and an aggressive MLE approximation with $T_{cut} = 0$, $C_{yz} = 2$, and $I^{(mm)} = 8$) are indeed more pronounced in the directions along which the reciprocal space mesh was coarsened, $\hat{y}$ and $\hat{z}$, even if the equivalent SPME calculation uses isotropic $\mu$ and $\sigma$ parameters. Increasing $\mu_x$ to ~1.5 Å and setting $I_x^{(pm)} = 6$ reduces errors in all directions, but the accuracy of MLE-approximated forces in the $\hat{x}$ direction shows the most improvement by far, even exceeding the accuracy of the equivalent SPME calculation. The frequencies of errors in the $\hat{y}$ and $\hat{z}$ directions were averaged and presented together as they were indistinguishable in this cubic unit cell. For comparison, we also show a histogram of the magnitudes of errors inherent in the equivalent SPME reciprocal space calculation, as judged by a high accuracy standard. The inset legend in the leftmost panel applies to both panels.

based on, for example, the orientation of a protein in the simulation cell. However, we stress that this test used an aggressive MLE approximation for demonstrative purposes and that other MLE parameters can yield errors well below those of the equivalent SPME calculation. We attempted the same test with more conservative parameters (data not shown) and found the shapes of the force error histograms to be very similar to those in Figure 5, but on an exponentially smaller scale.

**4.2. Larger Coarsening Factors and Incorporation of Multiple Higher-Level Meshes.** While there may be some advantage in being able to split the convolution $Q\star\theta^{(rec)}$ into multiple pieces (and, if $T_{cut}$ can be set as low as 0, obtain $Q_h\star\theta_h^{(rec)}$ by performing only two-dimensional FFTs, saving some FFT work and a major data transpose operation), $C_{yz}$ can be larger than 2 to reduce the size and processing requirements of the $Q_c\star\theta_c^{(rec)}$ convolution even further. The MLE scheme is also not limited to just one higher level mesh: the charge mesh $Q$ can be split into a series of meshes $Q_{c,1}$, $Q_{c,2}$, ..., $Q_{c,L}$, staged with increasing values of $C_{yz,1}$, $C_{yz,2}$, ..., $C_{yz,L}$ depending on the size of the problem.

Figure 9 shows the accuracy of Multi-Level Ewald on two of the systems in Table 1 using larger values of $C_{yz}$, demonstrating that MLE can be applied with $C_{yz}$ as high as 4−6 for 16−36-fold reductions in the amount of data present in the coarsest mesh. However, setting $T_{cut}$ to 4 or 6 could be very expensive in terms of the extra FFT work and communication cost. If eight MLE slabs were used with a mesh of 64 × 96 × 96 points with $C_{yz}$ set to 4 and $T_{cut}$ set to 5, each MLE slab would measure (64/8) + 2 × 5 = 18 points thick; the FFT work needed to compute $Q_c\star\theta_c^{(rec)}$ would be more than 16 times less than that needed to compute $Q\star\theta^{(rec)}$ in the equivalent SPME calculation, but the FFT work needed to compute the series $\sum_i Q_i\star\theta_h^{(rec)}$ would be roughly twice the original FFT burden. There would also

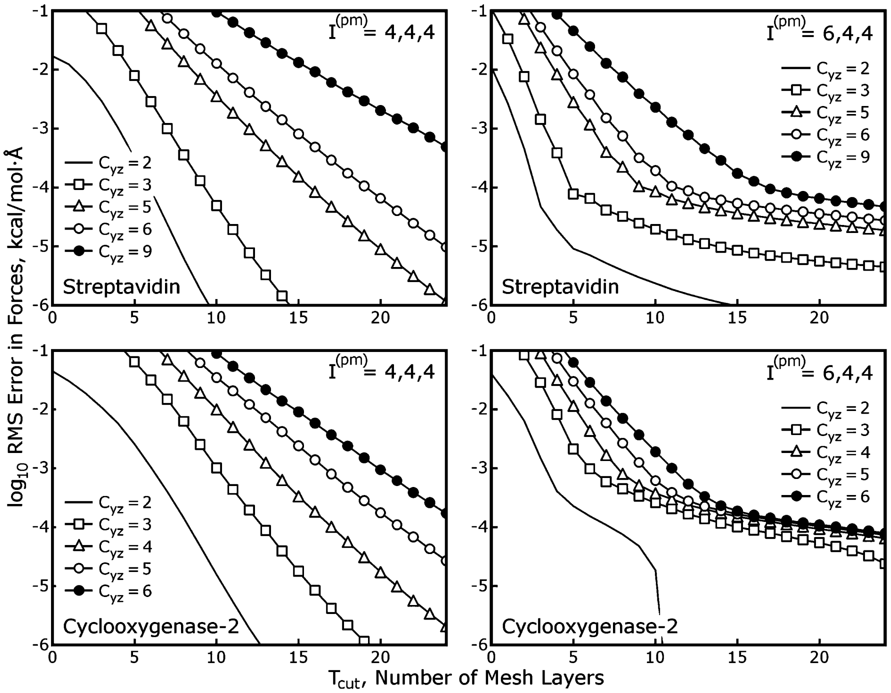


***Figure 9.*** Larger coarsening factors are available in MLE. Thus far, the results have focused on the performance of the MLE approximation for a variety of SPME calculations, emphasizing what values of $I^{(mm)}$ and $T_{cut}$ are necessary to achieve accurate results with a coarsening factor $C_{yz}$ of 2, but $C_{yz}$ is itself a tunable parameter of MLE. These plots show the accuracy of the MLE approximation for the streptavidin and cyclooxygenase-2 test cases (in cubic and triclinic unit cells, respectively) for numerous coarsening factors as shown in each diagram. The coarsening factors are limited to common factors of the mesh sizes in the $\hat{y}$ and $\hat{z}$ directions, but we do not expect this to be a serious limitation in practice. In these tests, $I^{(mm)}$ was fixed at 8. The AMBER default SPME paramaters, or the variant with anisotropic interpolation discussed in previous figures and the maint text, were used for the SPME calculations as indicated in each diagram. While larger values of $C_{yz}$ require larger values of $T_{cut}$ to produce accurate results, MLE with $C_{yz}$ as high as 6 can imply modest additional error with $T_{cut}$ as low as 7 if anisotropic particle $\rightleftharpoons$ mesh interpolation is used. As shown in Table 2, a third, intermediate mesh level, typically with $C_{yz} = 2$, is helpful for reducing the computation and communication burden of larger values of $T_{cut}$, making it possible to efficiently coarsen the reciprocal space mesh in stages.

be a considerable burden for communicating the zero-padded regions of each MLE slab.

To bridge the gap between $Q$ and $Q_c$, we introduced another mesh with an intermediate coarsening factor (i.e., $C_{yz,2} = 2$). Following the nomenclature in the Theory, we will refer to this intermediate mesh as $Q_{c,2}$ and refer to the highest level mesh, coarsened by a high value of $C_{yz,3}$, as $Q_{c,3}$. Previously, we have used $T_{cut}$ to describe the extent of the reciprocal space pair potential applied to the finest mesh $Q$ or the number of zero-padded pages in each of its slabs. When multiple coarse meshes are involved, we refer to the extent of the potential for the $n$th mesh as $T_{cut,\,n}$ and the coarsening factor for the $n$th mesh as $C_{yz,\,n}$. (In principle, for the lowest level mesh, $C_{yz,1} = 1$, and for the highest level mesh $T_{cut,\,L}$ is not defined.) In a three-mesh scheme, convoluting the lowest and intermediate level meshes $Q$ and $Q_{c,2}$ with an intermediate-ranged pair potential $\theta^{(rec)}_{c,2}$ can be accomplished as a series of convolutions over slabs as was done for $Q$ in previous MLE calculations. The slabs of the intermediate coarsened mesh, much less dense than $Q$, could be padded by a high value of $T_{cut,2}$ without adding greatly to the overall FFT computation or communication burden.

The MDGX program, but not the Matlab MLE calculator, was written to accommodate more than one level of mesh coarsening. We therefore tested the accuracy of MLE with several three-level mesh schemes on the cubic streptavidin system, as shown in Table 2. The performance of MLE in these three-level schemes is almost exactly what would be expected if the errors associated with separate two-level MLE calculations using the same parameters were combined.

**4.3. Energy Conservation and Equilibrium Properties in Simulations with Multi-Level Ewald.** The accuracy of forces obtained by the MLE approximation is encouraging, but we must still test whether the type of errors introduced by MLE, which are of a different nature than the errors in direct or reciprocal space forces arising from a standard SPME calculation, are possibly detrimental in the context of simulations. We therefore used the MDGX program to simulate a system of 1024 SPC-Fw water molecules in the microcanonical ensemble. Two different MLE schemes were used, as described in Table 3, both of them with three mesh levels. Trajectories were propagated at a 0.5 fs time step for 50 ns each, and the MLE or SPME reciprocal sums were computed at every time step to provide a stringent test of energy conservation. Coordinates were collected every 0.5 ps, and energies arising from electrostatic, Lennard-Jones, and harmonic bond and angle terms were collected every 0.05 ps.

The evolution of the total energy of the 1024 water system, simulated using each of the four methods described in Table

***Table 2.*** Multi-Level Ewald Calculations Performed with Three Mesh Levels[a]

| calc. type[b] | $\sigma$ | $I_x^{(pm)}$ | $C_{yz}$ | $T_{cut}$ | $\langle \Delta F^{(dir)} \rangle$[c] | $\langle \Delta F^{(rec)} \rangle$[d] | $\langle \Delta F^{(cor)} \rangle$[e] | $\langle \Delta F^{(ref)} \rangle$[f] |
|---|---|---|---|---|---|---|---|---|
| SPME | 1.434 | 4 | | | 7.850 | 9.288 | | 12.152 |
| MLE | 1.434 | 4 | 2 | 2 | 7.850 | 9.288 | 6.670 | 13.879 |
| MLE | 1.434 | 4 | 5 | 9 | 7.850 | 9.288 | 7.066 | 14.065 |
| MLE | 1.434 | 4 | 2,5 | 2,9 | 7.850 | 9.288 | 9.714 | 15.579 |
| SPME | 1.434 | 6 | | | 7.850 | 9.826 | | 12.598 |
| MLE | 1.434 | 6 | 2 | 1 | 7.850 | 9.826 | 2.756 | 12.890 |
| MLE | 1.434 | 6 | 5 | 5 | 7.850 | 9.826 | 2.859 | 12.924 |
| MLE | 1.434 | 6 | 2,5 | 1,5 | 7.850 | 9.826 | 3.972 | 13.209 |
| SPME | 2.151 | 4 | | | 5.492 | 7.005 | | 8.903 |
| MLE | 2.151 | 4 | 2 | 2 | 5.492 | 7.005 | 1.890 | 9.108 |
| MLE | 2.151 | 4 | 5 | 9 | 5.492 | 7.005 | 2.425 | 9.228 |
| MLE | 2.151 | 4 | 2,5 | 2,9 | 5.492 | 7.005 | 3.072 | 9.425 |

[a] All calculations in this table pertain to the streptavidin test case. Parameters for the equivalent SPME calculations included $I_y^{(pm)} = I_z^{(pm)} = 4$, $\sigma/\mu = 1.439$, $L_{cut} = 5.578\sigma$, and any parameters listed in the table pertaining to particular cases. $I^{(mm)}$ was set to 8 for all MLE calculations. The values of $\sigma$, $\mu$, and $L_{cut}$ in the first case are those obtained with the AMBER default settings for this system. All rms errors listed in this table are uncorrelated: when examined in detail, the Pearson correlation coefficients for the errors arising from distinct parts of the calculation are all smaller than 0.02. [b] Type of calculation. [c] The rms error in the direct space forces on all particles, relative to a high-accuracy SPME calculation ($\times 1.0 \times 10^{-3}$ kcal/mol·Å). [d] The rms error in the SPME reciprocal space force ($\times 1.0 \times 10^{-3}$ kcal/mol·Å). [e] The rms MLE coarsening error ($\times 1.0 \times 10^{-3}$ kcal/mol·Å). [f] Total rms error of the SPME or MLE calculation ($\times 1.0 \times 10^{-3}$ kcal/mol·Å).

***Table 3.*** Parameters Used in Long-Time Scale Simulations of SPC-Fw Water, and the Accuracy of Forces Resulting from Each Approximation[a]

| | conservative | | aggressive | |
|---|---|---|---|---|
| parameter | SPME | MLE | SPME | MLE |
| $L_{cut}$ (LJ, Å)[b] | 10.0 | 10.0 | 10.0 | 10.0 |
| $L_{cut}$ (elec, Å)[c] | 9.0 | 9.0 | 8.0 | 8.0 |
| $\sigma$, Å | 1.58 | 1.58 | 1.43 | 1.43 |
| $g$ | $(24 \times 36 \times 36)$ | $(24 \times 36 \times 36)$ | $(21 \times 32 \times 32)$ | $(21 \times 32 \times 32)$ |
| $I^{(pm)}$ | $(6 \times 4 \times 4)$ | $(6 \times 4 \times 4)$ | $(6 \times 4 \times 4)$ | $(6 \times 4 \times 4)$ |
| $I^{(mm)}$ | | 8 | | 8 |
| $C_{yz,2}$, $C_{yz,3}$ | | (2,4) | | (2,4) |
| $T_{cut,1}$, $T_{cut,2}$ | | (1,5) | | (0,4) |
| $\langle \Delta F^{(dir)} \rangle$[d] | $4.9 \times 10^{-3}$ | $4.9 \times 10^{-3}$ | $7.7 \times 10^{-3}$ | $7.7 \times 10^{-3}$ |
| $\langle \Delta F^{(rec)} \rangle$[e] | $2.3 \times 10^{-3}$ | $2.5 \times 10^{-3}$ | $8.9 \times 10^{-3}$ | $1.3 \times 10^{-2}$ |

[a] The four simulations utilize either SPME or MLE calculations for long-ranged electrostatic interactions. While all simulations make use of the same direct space cutoffs, the "conservative" simulations use roughly 60% more data in $Q$ for their SPME or MLE calculations. The "conservative" MLE scheme approximates the SPME results much more accurately than the equivalent SPME calculation obtains the true electrostatic force on each particle, as judged by a high-quality SPME calculation using $g = (96 \times 96 \times 96)$ and $I^{(pm)} = 8$. In contrast, the "aggressive" SPME scheme is somewhat less accurate, and the "aggressive" MLE scheme introduces roughly the same amount of error as the equivalent SPME scheme. Results from the simulations are presented in Figures 10 and 11. [b] Lennard-Jones potential truncation length. [c] Electrostatic direct space trucation length. [d] The rms error in direct space electrostatic forces with this approximation. [e] The rms error in reciprocal space electrostatic forces with this approximation (including MLE approximation, if applicable).

3, is shown in Figure 10. For comparison, the energy of the same system run using the "conservative" SPME parameters but a 1 fs time step is juxtaposed with the results at a 0.5 fs time step. While all of the 0.5 fs runs show some upward drift in the energy over 50 ns, it is very slow, and the time step itself clearly has a much greater impact on the energy conservation than the choice of SPME or MLE for a long-ranged electrostatics approximation. Neither MLE approximation, whether "aggressive" or "conservative", shows a visible difference when compared to the corresponding SPME calculation on the basis of energy conservation. Moreover, the differences between the two SPME methods are greater than the differences between the MLE approximations and the equivalent SPME calculations: while the axes in each panel of Figure 10 have similar scales, the energy of the system run with aggressive SPME and MLE parameters is somewhat higher and the fluctuation of the energy is noticeably larger. The reason for this increase in the recorded energy can be traced to the inaccuracies inherent in the SPME reciprocal space calculation when the $\sigma$ to $\mu$ ratio becomes smaller, as explained in the Supporting Information for our previous work on Ewald sums.[22] When run with the same aggressive SPME parameters, the MLE approximation returns similar increases in the absolute energy and fluctuations in that energy; adding more aggressive MLE parameters on top of the lower-quality SPME method does not seem to affect the results much further.

As shown in Table 4, the bulk properties of the SPC-Fw water are not significantly perturbed by any of the SPME or MLE approximations. When taken in the context of a macroscopic observable such as the heat of vaporization, the differences between the total energy of the system measured by aggressive or conservative electrostatic parameters are negligible.

We also investigated the microscopic structure of the water when simulated with each approximation. One reason for choosing the flexible SPC-Fw water model was to test whether the MLE method, which produces its lowest accuracy when computing interactions between very nearby particles, could perturb the behavior of bonded atoms. (Although electrostatic interactions are excluded between bonded atoms in most molecular force fields, this exclusion is done by computing the interaction of two Gaussian-smoothed charges at the specified distance and subtracting this from the reciprocal space sum, which necessarily computes all interactions during the mesh convolution.) As shown in Figure 11, neither the MLE approximation nor the quality of the SPME method has any significant effect on either the oxygen:hydrogen bond length, the hydrogen:hydrogen distance within each water molecule, or the radial distributions of oxygen and hydrogen atoms on different water molecules.

The timings for these single-processor MLE runs also provide an indication of how much more computational effort MLE would require over the standard SPME method. The fact that simulating the 1024 water molecules requires only 6−10% longer with MLE than with SPME indicates that most codes with well-optimized FFT routines could implement MLE without much more computational effort than SPME. Notably, while the MLE schemes are more costly in terms of FFT work, the majority of the extra cost actually comes from the mesh $\leftrightharpoons$ mesh interpolation. The FFTW

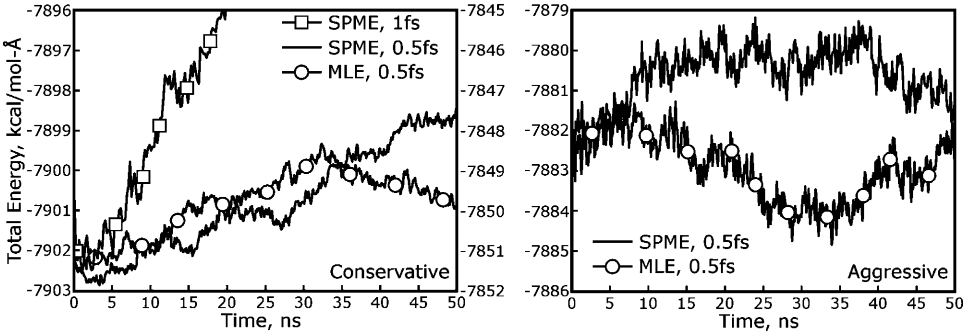


***Figure 10.*** The Multi-Level Ewald approximation yields equivalent energy conservation to traditional Smooth Particle-Mesh Ewald. Simulations of a system of 1024 SPC-Fw water molecules show that MLE is able to conserve the system's energy over 50 ns trajectories. Parameters for each simulation are given in Table 3, and the length of the time step or style of Ewald summation is given in the legend of each figure. In each simulation, the total energy of the system fluctuates because both Ewald methods entail some degree of error as the particles move relative to the mesh, and the Lennard-Jones potential is sharply truncated at 10.5 Å. The total energy is therefore plotted as a series of mean values averaged over 200 frames each. Many investigators consider the energy conservation yielded by a 1 fs time step in systems with flexible bonds to hydrogen atoms acceptable. Comparison of the results obtained with a 1.0 fs time step (results should be read from the *y*-axis on the right side of the left-hand panel) to those obtained with a 0.5 fs time step shows that the time step itself can be a more significant contributor to the upward drift of the total system energy than many of the other parameters. (With the 0.5 fs time step, the temperature drift in each simulation is only about 0.5 K over 50 ns.) As the quality of the SPME calculations decreases, the fluctuation of the total system energy increases, as evident by comparing the results for SPME simulations in each panel. The MLE approximation also reproduces these changes in the sizes of fluctuations in the total energy.

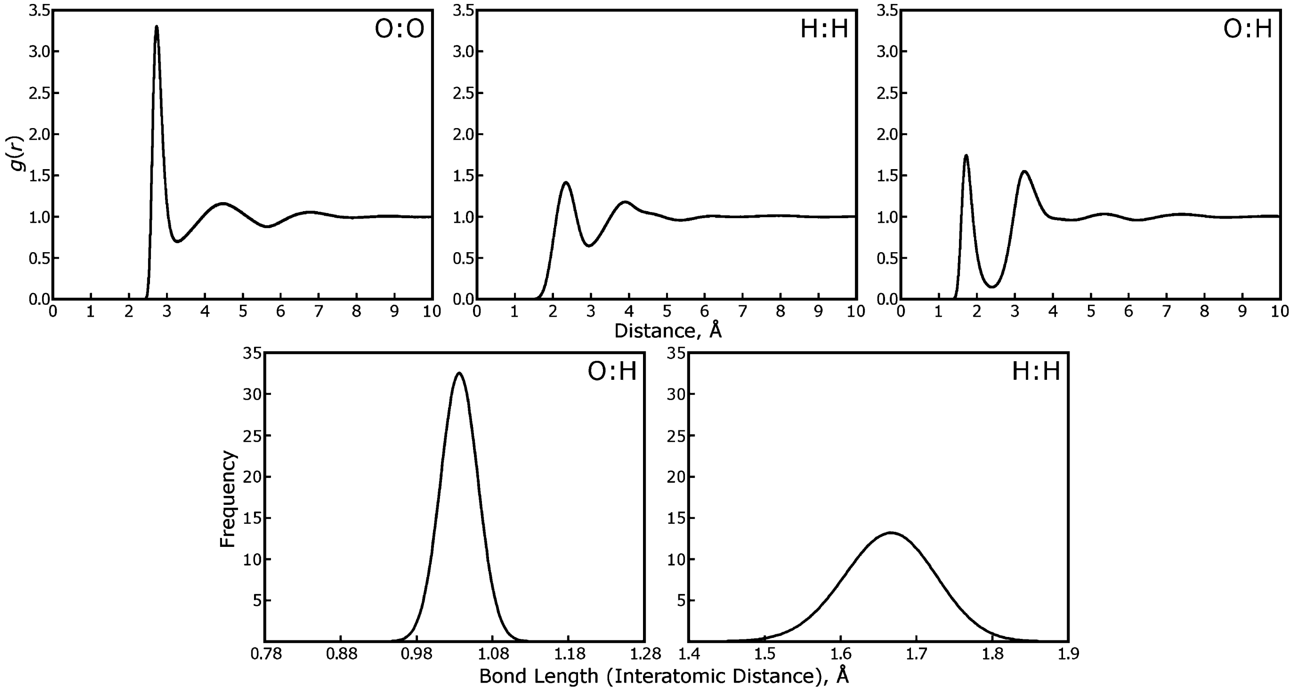


***Figure 11.*** The average structure of SPC-Fw water molecules is maintained under different Ewald approximations. Analysis of the microscopic structure of the water molecules was performed to complement the energy conservation studies presented in Figure 10. Radial distribution functions for oxygen to oxygen, hydrogen to hydrogen, and oxygen to hydrogen atoms of SPC-Fw water molecules are displayed in the top three panels. Histrograms of the oxygen−hydrogen bond length and hydrogen−hydrogen intramolecular distance are shown in the lower two panels. Results from simulations with all four of the Ewald approximations listed in Table 3 are shown with solid black lines in each plot; the distributions overlap so precisely that they are indistinguishable.

libraries used by MDGX are among the fastest available, but as shown in Table 5, many other aspects of the MDGX code are not as efficient as their counterparts in PMEMD. (We are looking into compiler optimizations that may make the difference, as we believe we have coded routines such as the particle ⇋ mesh interpolation as efficiently as possible, and there appears to be little difference between the structure of our routines and those of PMEMD.) The estimates presented in Table 5 do not include the cost of computing FFTs over zero-padded regions of each MLE slab or the possible benefits of performing numerous FFTs over small regions rather than one large FFT; instead, a single convolu-

***Table 4.*** Bulk Properties of 1024 Water Molecules Simulated in the Microcanonical Ensemble[a]

| | | conservative | | aggressive | |
|---|---|---|---|---|---|
| property | PMEMD[e] | SPME | MLE | SPME | MLE |
| $\Delta H^{vap}$, kcal/mol[b] | 10.89 ± 0.00 | 10.88 ± 0.00 | 10.88 ± 0.00 | 10.88 ± 0.00 | 10.89 ± 0.00 |
| $T$, K[c] | 289.07 ± 0.08 | 289.59 ± 0.14 | 289.65 ± 0.03 | 291.72 ± 0.13 | 291.79 ± 0.17 |
| $D$, $\times 10^{-5}$ cm $^2$/s[d] | 1.81 ± 0.02 | 1.83 ± 0.04 | 1.86 ± 0.05 | 1.85 ± 0.08 | 1.83 ± 0.07 |

[a] All values are given as averages over 12.5 ns blocks of each simulation, with standard deviations. [b] Heat of vaporization, calculated by $\Delta H^{vap} = -\langle E \rangle + RT$, where $E$ is the mean potential energy, $R$ is the gas constant, and $T$ is the mean temperature. [c] Mean temperature of the simulation. [d] Diffusion coefficient. [e] Results obtained using the PMEMD implementation of the SANDER program from the AMBER software package, running SPME with the "Aggressive" parameters, but $l^{(pm)}$ set uniformly to 4 and $g$ set uniformly to 32.

***Table 5.*** Timings for MDGX or PMEMD on a Single Processor[a]

| | conservative | | aggressive | | |
|---|---|---|---|---|---|
| routine | SPME | MLE | PMEMD | SPME | MLE |
| bonded interactions | 5.8 | 5.9 | 4.6 | 5.8 | 5.8 |
| $\Delta E^{(dir)}$, pair list | 7.7 | 7.9 | 31.3 | 7.7 | 7.8 |
| $\Delta E^{(dir)}$, interactions | 383.9 | 383.6 | 216.7 | 323.8 | 323.9 |
| $\Delta E^{(dir)}$, **total** | **391.6** | **391.6** | **248.0** | **331.5** | **331.7** |
| $\Delta E^{(rec)}$, B-splines | 14.3 | 14.4 | 2.6 | 14.2 | 14.3 |
| $\Delta E^{(rec)}$, particle → mesh | 16.2 | 16.4 | 6.0 | 15.9 | 16.1 |
| $\Delta E^{(rec)}$, convolution[b] | 2.4 | 1.2 | 11.6 | 1.7 | 0.8 |
| $\Delta E^{(rec)}$, FFT | 17.3 | 22.6 | 28.4 | 9.1 | 11.8 |
| $\Delta E^{(rec)}$, mesh → particle | 22.5 | 22.6 | 11.1 | 22.2 | 22.3 |
| mesh ⇋ mesh[c] | | 42.1 | | | 29.0 |
| $\Delta E^{(rec)}$, **total** | **72.8** | **119.3** | **59.8** | **63.1** | **94.2** |
| **total wall time** | **473.1** | **522.5** | **316.1** | **405.9** | **435.6** |

[a] 20 000 steps of dynamics were run using PMEMD or MDGX in SPME or MLE mode on an Intel Q9550 processor (Core2 architecture, 6 MB L2 cache, 2.83 GHz clock speed). Simulations made use of the parameters in Table 3; the PMEMD simulation for the "aggressive" parameters used $l^{(pm)} = 4$ and $g = 32$. Timings for different categories of calculations in the MDGX code were measured using the UNIX gettimeofday() function; those in the PMEMD code were measured with PMEMD's internal profiling functions. Because the two programs are structured differently, the exact content of each category of calculation may not match exactly; for instance, MDGX and PMEMD use different styles of pair list, and the convolution kernel is computed more rapidly in MDGX than PMEMD because PMEMD always computes virial contributions while MDGX skips them if they are not needed. Total run times for both programs are greater than the sums of all categories in each column because timings for miscellaneous routines are not listed. Standard deviations were collected over 10 trials to estimate errors in the timings; they were less than 1% across all categories, but are omitted to condense the table. [b] Multiplication of the charge mesh and reciprocal space pair potential in Fourier space. [c] Interpolation between different levels of MLE mesh.

tion over a single mesh is done at all levels of the calculation in these single-processor runs. The estimates in Table 5 also neglect some possible benefits in the case of the "aggressive" MLE parameters. Setting $T_{cut,1} = 0$ permits convolutions of the lowest level mesh to be completed with two-dimensional FFTs, saving roughly 1/3 of the FFT work for that mesh level. We are continuing to develop the MDGX program to take advantage of these optimizations.

## 5. Discussion

**5.1. Development of Multi-Level Ewald for Parallel Applications.** We do not yet have a parallel version of MDGX to run Multi-Level Ewald on many processors. However, we believe that MLE can benefit massively parallel simulations, particularly when extremely powerful, multicore nodes must be connected by comparatively weak networks, when other forms of network heterogeneity are involved, or when the problem size is very large. With MLE, there are extra communication steps as the coarse meshes must be assembled and the electrostatic potential data deposited in the zero-padded "tails" of each slab must be passed to neighboring slabs. However, because these are all local effects, the number of messages that must be passed to create the coarse meshes and contribute $U_c^{(rec)}$ to $U^{(rec)}$ is bounded, whereas the number of messages that must be passed in a convolution involving the whole $P^3M$ reciprocal space mesh grows, at best, as the square root of the number of processors.[26] For example, distributing the convolution for a 60 × 90 × 90 mesh in the streptavidin test case over six multicore nodes would require each node to transmit at least 0.65 MB (megabytes) of data (if data are transmitted in 32-bit precision). With a three-level MLE scheme placing one fine mesh slab and one intermediate mesh slab on each node and setting $C_2 = 2$, $T_{cut,1} = 0$, $C_3 = 5$, $T_{cut,2} = 5$, the total volume of data transmission between nodes could be reduced nearly 4-fold, to 0.19 MB, in the convolution step.

Several challenges remain to implementing MLE in an efficient parallel code. The most obvious is load-balancing: MLE introduces another layer of complexity for scheduling the completion of coarse mesh convolutions, plus the associated mesh ⇋ mesh interpolation. Another challenge is that, while MLE can be tuned to reduce data transmission between weakly connected processors, if many networked nodes must collaborate on each MLE slab, the original data communication problems resurface. Whereas each of $K$ nodes must pass 4 × sqrt($K$) messages in the original $P^3M$ convolution, with MLE and $P$ slabs with $K \gg P$ the number of messages is 4 × sqrt($K/P$) + $M$ (for nodes devoted to fine mesh calculations) or 4 × sqrt($K$)/$C$ + $M$ (for $K/C^2$ nodes devoted to coarse mesh calculations), where $M$ is a small constant for mesh ⇋ mesh interpolation. One possible extension of the MLE method may be helpful for the case of many (multicore) nodes collaborating on each MLE slab: subdividing the convolutions over fine mesh pages into pencils using an analogous sharp, anisotropic splitting technique, and then applying a one-dimensional coarsening to meshes spanning each MLE slab.

**5.2. Application to Multiple Time Step Algorithms.** The different mesh levels in MLE calculations may be excellent candidates for updates at different time steps, particularly because the anisotropic splitting completely captures local changes to a molecular system's electrostatics in the lower charge mesh levels. In contrast, the highest level charge mesh requires the most communication between processors devoted to disparate regions of the simulation cell, to complete the global convolution. The novel splitting approach of MLE may create its own unique types of artifacts in such simulations, however. We have shown that the sum of contributions from all mesh levels can recover a smooth potential, but this may be perturbed if the electrostatic potential of each mesh level is updated at different times. With any multiple time step method there can be subtle resonances that affect the statistical properties of the system;[27] we intend to investigate the stability and efficiency of MLE with multiple time steps in the future.

**5.3. Application to Systems With Two-Dimensional Periodicity.** While periodic boundary conditions in three dimensions have been shown to be equivalent or superior to alternative boundary conditions for many condensed-phase biomolecular simulations,[28,29] there are classes of problems, notably membrane protein simulations,[30] that produce different results if periodicity is suppressed in one dimension. Regular Ewald methods are available for imposing two-dimensional periodicity,[31,32] but they are prohibitively expensive for systems of many thousands of atoms. For larger systems, a pseudo two-dimensional periodicity may be imposed by lengthening the simulation cell in one dimension, say $\hat{x}$, confining the system to the middle of the simulation cell along $\hat{x}$ by some stochastic boundary condition or, more directly, by a physical set of walls such as sheets of platinum atoms, and then running P$^3$M calculations as usual, with three-dimensional periodicity, on the extended system.[33] This approach has been further refined by adding an electrostatic field to counteract the net dipole of the system in $\hat{x}$,[34] mirroring the way in which modified potential functions and zero-padding have been used in plasma physics and astrophysical gravity calculations.[35]

The MLE method may be suitable for systems with two-dimensional periodicity, although membrane protein simulations run in the isothermal−isobaric (NPT) ensemble tend to require anisotropic system rescaling, which MLE cannot accommodate exactly. It is likely possible to extend the MLE method to work in such cases by storing a small array of precomputed solutions of each mesh potential with different unit cell ratios and thereafter interpolating the solution for any particular time step. Other, more general, solutions to the problem of isolated boundary conditions are again found in Adaptive P$^3$M methods[14−16,36] and the Multilevel Summation method.[37,38] In all of these approaches, the advantage for isolated boundary conditions in one or more dimensions is that only the coarse mesh must be evaluated in the zero-padded, empty regions of the simulation cell.

**5.4. Diversity of Problem Decompositions for Future Machines.** In conclusion, we have shown that, for pairwise potentials that decay as the inverse distance between particles, it is feasible to subdivide the convolution in particle ⇌ mesh calculations sharply and anisotropically into many separate slabs without significantly adding to the overall cost of the calculation. This technique and the many smooth splitting approaches that already exist should be applicable to simulations on current and next-generation parallel computers, where hundreds to thousands of processors must collaborate to deliver longer simulation trajectories. It is worthwhile to develop a variety of these multilevel decompositions, as future supercomputers may come in novel architectures that offer huge advantages depending on the details of the parallel algorithm.

**Acknowledgment.** D.S.C. thanks Dr. Peter Freddolino, Professor Ross Walker, and Dr. David Hardy for helpful conversations. This work was funded by NIH grant RR12255.

## References

(1) Henzler-Wildman, K. A.; Thai, V.; Lei, M.; Ott, M.; Wolf-Watz, M.; Fenn, T.; Pozharski, E.; Wilson, M. A.; Petsko, G. A.; Karplus, M.; Hübner, C. G.; Kern, D. Intrinsic motions along an enzymatic reaction trajectory. *Nature* **2007**, *450*, 838–844.

(2) Cerutti, D. S.; Le Trong, I.; Stenkamp, R. E.; Lybrand, T. P. Simulations of a protein crystal: Explicit treatment of crystallization conditions links theory and experiment in the streptavidin system. *Biochemistry* **2008**, *47*, 12065–12077.

(3) Maruthamuthu, V.; Schulten, K.; Leckband, D. Elasticity and rupture of a multi-domain cell adhesion complex. *Biophys. J.* **2009**, *96*, 3005–3014.

(4) Freddolino, P. L.; Liu, F.; Grubele, M.; Schulten, K. Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophys. J.* **2008**, *94*, L75–L77.

(5) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. H. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(6) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An *N*log(*N*) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(7) Hockney, R. W.; Eastwood, J. Collisionless particle models. *Computer Simulation Using Particles*; Taylor and Francis Group: New York, NY, 1988; pp 260−291.

(8) Pollock, E. L.; Glosli, J. Comments on P$^3$M, FMM, and the Ewald method for large periodic Coulombic systems. *Comput. Phys. Commun.* **1996**, *95*, 93–110.

(9) Shan, Y.; Klepeis, J. L.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Gaussian Split Ewald: A fast Ewald mesh method for molecular simulation. *J. Chem. Phys.* **2005**, *122*, 054101.

(10) Sagui, C.; Darden, T. Multigrid methods for classical molecular dynamics simulations of biomolecules. *J. Chem. Phys.* **2001**, *114*, 6578–6591.

(11) Brandt, A.; Lubrecht, A. A. Multilevel matrix multiplication and fast solution of integral equations. *J. Comput. Phys.* **1990**, *90*, 348–370.

(12) Skeel, R. D.; Tezcan, I.; Hardy, D. J. Multiple grid methods for classical molecular dynamics. *J. Comput. Chem.* **2002**, *23*, 672–684.

(13) Kurzak, J.; Pettitt, B. M. Massively parallel implementation of a fast multipole method for distributed memory machines. *J. Parallel Distr. Comp.* **2005**, *65*, 870–881.

(14) Thacker, R. J.; Couchman, H. M. P. A parallel adaptive P$^3$M code with hierarchical particle reordering. *Comput. Phys. Commun.* **2006**, *174*, 540–554.

(15) Merz, H.; Pen, U.; Trac, H. Towards optimal parallel PM N-body codes: PMFAST. *New Astron.* **2005**, *10*, 393–407.

(16) Couchman, H. M. P. Mesh-refined P$^3$M: A fast adaptive *N*-body algorithm. *Astrophys. J.* **1991**, *368*, L23–L26.

(17) Sagui, C.; Darden, T. A. $P^3M$ and PME: A comparison of the two methods. In *Simulation and Theory of Electrostatic Interactions in Solution. Proceedings of the American Institute of Physics Conference, Sante Fe, NM, 1999*; Pratt, L. R., Hummer, G., Eds.; Springer: Secaucus, NJ, 2000; Vol. 492, pp 104−113.

(18) Schoenberg, I. J. *Cardinal Spline Interpolation*; Society for Industrial and Applied Mathematics; Philadelphia, PA, 1973.

(19) Case, D. A.; Cheatham, T. E., III; Darden, T. A.; Gohlke, H.; Luo, R.; Mer, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The AMBER biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

(20) Wu, Y.; Tepper, H. L.; Voth, G. A. Flexible simple point-charge water model with improved liquid-state properties. *J. Chem. Phys.* **2006**, *124*, 024503.

(21) Frigo, M.; Johnson, S. G. The design and implementation of FFTW3. *Proc. IEEE* **2005**, *93*, 216–231.

(22) Cerutti, D. S.; Duke, R. E.; Lybrand, T. P. Staggered Mesh Ewald: An extension of the Smooth Particle Mesh Ewald method adding great versatility. *J. Chem. Theory Comput.* **2009**, *5*, 2322–2238.

(23) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossv́ary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable algorithms for molecular dynamics simulations on commodity clusters. *Conference on High Performance Networking and Computing, Proceedings of the 2006 ACM/IEEE conference on Supercomputing, Tampa, FL, 2006*; Association for Computing Machinery: New York, NY, 2006; p 84.

(24) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(25) Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossvary, I.; Klepeis, J. L.; Layman, T.; McLeavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. Anton, a special-purpose machine for molecular dynamics simulation. *Special purpose to warehouse computers, Proceedings of the 34 th annual international symposium on computer architecture, San Diego, CA, May, 2007*; ACM: New York, NY, 2007; Vol. 1, pp 1−12.

(26) Sbalzarini, I. F.; Walther, J. H.; Bergdorf, M.; Hieber, S. E.; Kotsalis, E. M.; Koumoutsakos, P. PPM - A highly efficient parallel particle-mesh library for the simulation of continuum systems. *J. Comput. Phys.* **2006**, *215*, 566–588.

(27) Ma, Q.; Izaguirre, J. A.; Skeel, R. D. Nonlinear instability in multiple time stepping molecular dynamics. *Computational Sciences, Proceedings of the 2003 ACM symposium on applied computing, Melbourne, FL, March, 2003*; ACM: New York, NY, 2003; Vol. 1, pp 167−171.

(28) Freitag, S.; Chu, V.; Penzotti, J. E.; Klumb, L. A.; To, R.; Hyre, D.; Le Trong, I.; Lybrand, T. P.; Stenkamp, R. E.; Stayton, P. S. A structural snapshot of an intermediate on the streptavidin-biotin dissociation pathway. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 8384–8389.

(29) Riihimäki, E. S.; Martínez, J. M.; Kloo, L. An evaluation of non-periodic boundary condition models in molecular dynamics simulations using prion octapeptides as probes. *J. Mol. Struct. (THEOCHEM)* **2005**, *760*, 91–98.

(30) Bostick, D.; Berkowitz, M. The implementation of slab geometry for membrane-channel molecular dynamics Simulations. *Biophys. J.* **2003**, *65*, 97–107.

(31) Parry, D. E. The electrostatic potential in the surface region of an ionic crystal. *Surf. Sci.* **1975**, *49*, 433.

(32) Kawata, M.; Mikami, M. Rapid calculation of two-dimensional Ewald summation. *Chem. Phys. Lett.* **2006**, *340*, 157–164.

(33) Spohr, E. Effect of electrostatic boundary conditions and system size on the interfacial properties of water and aqueous solutions. *J. Chem. Phys.* **1997**, *107*, 6342–6348.

(34) Yeh, I.-C.; Berkowitz, M. Ewald summation for systems with slab geometry. *J. Chem. Phys.* **1999**, *111*, 3155–3162.

(35) Gelato, S.; Chernoff, D. F.; Wasserman, I. An adaptive hierarchical particle-mesh code with isolated boundary conditions. *Astrophys. J.* **1997**, *480*, 115–131.

(36) Eastwood, J. W. The Block $P^3$ M algorithm. *Comput. Phys. Commun.* **2008**, *179*, 46–50.

(37) Brandt, A.; Ilyin, V.; Makedonska, N.; Suwan, I. Multilevel summation and Monte Carlo simulations. *J. Mol. Liq.* **2006**, *127*, 37–39.

(38) Hardy, D. J.; Stone, J. E.; Schulten, K. Multilevel summation of electrostatic potentials using graphics processing units. *Parallel Comput.* **2009**, *35*, 164–177.

CT900522G

JCTC Journal of Chemical Theory and Computation

# Implementation of the CHARMM Force Field in GROMACS: Analysis of Protein Stability Effects from Correction Maps, Virtual Interaction Sites, and Water Models

Pär Bjelkmar,[†] Per Larsson,[†] Michel A. Cuendet,[‡] Berk Hess,[†] and Erik Lindahl*[,†]

*Center for Biomembrane Research, Department of Biochemistry & Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden, and Molecular Modeling Group, Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland*

Received October 16, 2009

**Abstract:** CHARMM27 is a widespread and popular force field for biomolecular simulation, and several recent algorithms such as implicit solvent models have been developed specifically for it. We have here implemented the CHARMM force field and all necessary extended functional forms in the GROMACS molecular simulation package, to make CHARMM-specific features available and to test them in combination with techniques for extended time steps, to make all major force fields available for comparison studies in GROMACS, and to test various solvent model optimizations, in particular the effect of Lennard-Jones interactions on hydrogens. The implementation has full support both for CHARMM-specific features such as multiple potentials over the same dihedral angle and the grid-based energy correction map on the $\phi$, $\psi$ protein backbone dihedrals, as well as all GROMACS features such as virtual hydrogen interaction sites that enable 5 fs time steps. The medium-to-long time effects of both the correction maps and virtual sites have been tested by performing a series of 100 ns simulations using different models for water representation, including comparisons between CHARMM and traditional TIP3P. Including the correction maps improves sampling of near native-state conformations in our systems, and to some extent it is even able to refine distorted protein conformations. Finally, we show that this accuracy is largely maintained with a new implicit solvent implementation that works with virtual interaction sites, which enables performance in excess of 250 ns/day for a 900-atom protein on a quad-core desktop computer.

## Introduction

Utilizing a force field that is as accurate as possible is one of the most important factors when using molecular dynamics to predict different macromolecular properties. The past decade has seen significant advances in force field development, and with this, many of them have become increasingly precise and accurate for predictions of nontrivial properties such as structure, dynamics, or free energies. One of the most prominent improvements in contemporary force fields has come from adjusting parameters to match quantities obtained from quantum mechanical calculations. This has, for instance, been done for the AMBER-99,[1] AMBER-03,[2] GROMOS,[3] and CHARMM22[4] force fields. The major difference between the AMBER-99 and AMBER-03 force fields, for example, includes a reparameterization of charges and dihedral angle parameters, using accurate quantum chemistry potentials together with an implicit solvent model. The idea behind this was to better mimic the internal environment of a protein, where the dielectric constant differs from either an aqueous environment or one in vacuo. Even before this, the Jorgensen group (and later Friesner) developed the OPLS-AA force field,[5,6] where a key idea was to fit parameters to

* Corresponding author e-mail: lindahl@cbr.su.se.

[†] Stockholm University.

[‡] Swiss Institute of Bioinformatics.

10.1021/ct900549r
Published on Web 01/25/2010

better reproduce the entire Ramachandran diagram for amino acids, rather than just individual dihedral potentials.

Some years ago, MacKerell et al.[7] used an analogous but more elaborate idea. The authors looked at how good CHARMM22 was at reproducing quantum chemical potential landscapes for small dipeptide fragments and characterized the deviations between the force-field-based and quantum chemistry energy landscapes. This led to the development of a grid-based energy correction map for protein backbone $\phi$ and $\psi$ dihedral angles, named CMAP, that enables almost arbitrarily smooth corrections to the Ramachandran map energy landscape. In particular, it no longer has to be a linear superposition of the constituent dihedral potentials. This correction term was implemented in the most recent version of their force field, CHARMM27.[7]

The aim of this work has been to implement these types of potentials in GROMACS, to provide an efficient (GROMACS) reference implementation of the CHARMM27 force field using them, and to assess the accuracy of force field and correction map terms when combined with the water models available in GROMACS. The present implementation supports all major CHARMM27 features, including Urey−Bradley potentials, multiple potentials over the same dihedral angle, and arbitrary correction maps for pairs of dihedrals. This support is obviously highly useful merely as another high-quality choice for large-scale parallel simulations that enables access to a wider set of parameters, and the inclusion of support for all major force fields in the GROMACS distribution will facilitate future systematic comparisons between different force fields for lipids and proteins and help determine how good these are at predicting experimental properties, to what extent they can improve protein structure, and not least how different techniques for higher performance affect various force fields.

It is also an important question to what extent these results are affected by the choice of water models. CHARMM frequently uses an extension of TIP3P with Lennard-Jones interactions also on the hydrogen atoms; while the difference from the classical TIP3P model is minimal, it can be quite significant from a performance point of view since GROMACS can use custom accelerated kernels for traditional water models. To address this, we have compared how the models affect both properties of pure water as well as protein stability assessed through rmsd (coordinate root-mean-square displacement) and dihedral angles. The rmsd is not an entirely unproblematic measure since one is comparing to a packed crystal structure with less water at a much lower temperature, but for better or worse it is still one of the most widely used quality indicators, and it is interesting to see how much force fields in general have improved compared to a decade ago.[8]

Finally, we combine CMAP and the CHARMM27 force field with a new implementation of virtual interactions better suited for all-atom force fields. The use of virtual interaction sites is a technique that goes beyond bond constraints and removes all independent hydrogen atom *degrees of freedom*—but not their interactions—by replacing hydrogens with interaction sites calculated from the heavy atoms to which they are connected.[9] Together with constraints on all bond lengths, this enables time steps of 4−5 fs without an apparent loss of accuracy even for water simulations, and possibly even larger for implicit solvent models.

The stability and accuracy of such “fast” calculation models is particularly critical in the context of protein structure refinement (improving models generated by homology modeling for example). Full protein folding requires exhaustive phase space sampling and is still obviously impossible for all but the very smallest proteins, and even then it requires months of simulation time, which makes it impossible to use in a high-throughput environment. However, it has been shown that, in some cases, with the use of efficient sampling techniques and lots of computer power, molecular dynamics simulations can help refine initially reasonable protein models toward their native state,[10] as measured by rmsd. It is worth noting that practical refinement capability is not necessarily the same as strict model accuracy. Ultimately, one would like the phase space sampling of a simulation starting from a decoy and the native structure to fully converge, but this is still not realistic with a couple of days of simulation time. In practice, the question is rather whether a force field provides an efficient funnel for refinement of structures already in the vicinity of their target. This is likely the main reason for the success of knowledge-based potentials or simple implicit solvent energy minimization techniques in structure refinement, where standard molecular dynamics simulations sometimes even deteriorate the structure.[11] Apart from the increased efficiency (and possibly lower noise) of the implicit solvent, one possible reason for this is that accurate and adequate sampling of backbone dihedral angles is key to accurately refining protein structures,[12,13] and thus applying a simple correction term only to those parameters (i.e., CMAP), while leaving the rest of the force field unaffected, is a promising option that seems to work remarkably well in our trials, even in combination with fast implicit solvent models.

## Methods

The CHARMM27 force field for proteins and lipids was ported to the force field format used by GROMACS. For most interactions, such as bonds, angles, and Lennard-Jones and Coulomb interactions, this was merely a matter of bookkeeping and converting units. The two programs for instance occasionally differ by a factor of 2 in potential energy definitions.

Some properties of the CHARMM force field are however specific to the CHARMM molecular simulation package.[14] Such features include the ability in the force field to define multiple dihedral potential terms over the same four atoms, some with a multiplicity ≥6 (and hence not representable by the Ryckaert−Bellemans function typically used in GROMACS for efficiency reasons). A new dihedral potential energy function was added in GROMACS to allow such a feature; while this does lead to a few repeated floating-point calculations when the same dihedral is calculated twice, it is quite negligible from a performance point of view. CHARMM27 also uses Urey−Bradley terms for many angles; GROMACS has had support for these internally a long time, but we now also generate them automatically with

***Table 1.*** Average Absolute Errors in Potential Energy (kJ/mol)

| GROMACS | CHARMM | <Δ*E*> | NAMD | <Δ*E*> |
|---|---|---|---|---|
| Bond | BONDs | 0.00001 | BOND | 0.00001 |
| Urey−Bradley | ANGLes+UREY-b | 0.0001 | ANGLE | 0.00002 |
| Proper+Improper Dih. | DIHEdrals+IMPRopers | 0.00003 | DIHED+IMPRP | 0.00001 |
| CMAP Dih. | CMAPs | 0.00000 | CROSS | 0.003 |
| LJ-14+LJ (SR) | VDWaals | 0.008 | VDW | 0.003 |
| Coulomb-14+Coulomb (SR) | ELEC | 0.002 | ELECT | 0.0007 |

the pdb2gmx proprocessing tool when CHARMM27 is selected, and we similarly generate the appropriate (user-selectable) CHARMM termini for polypeptides. In addition, the CMAP correction term was implemented as described by MacKerell et al.[7] The full implementation was tested by performing single-point potential energy calculations on small (four residues long) homopeptides, and energies were compared to those obtained in the native CHARMM software. This was done for all of the 20 essential amino acid residues. Then, the potential energy was split into corresponding parts (bonds, angles, electrostatics, etc.) and the absolute average error calculated in each case (Table 1).

In the original CMAP implementation, results were assessed using three proteins with PDB identifiers 1GRP, 1HIJ, and 1UBQ. To further validate and test our force field port, we were looking for a different protein with similar properties, that is, being relatively short, having a high-resolution structure, and containing both α-helix and $\beta$-sheet regions but no disulfide bridges (since those would stabilize the structure). One of the proteins that fulfilled these criteria was an IgG-binding domain from a streptococcal protein G with PDB identification 1IGD.[15] To probe the effect of the CMAP term on long timescales, as well as different water models and time steps, we used a total of 16 simulation systems, including four simulations to test refinement. The remaining 12 differ in the choice of water model (original TIP3P, CHARMM-modified TIP3P, TIP4P, implicit solvent) and the length of the time step (2 or 4 fs when using explicit solvent, 5 fs with implicit solvent). All simulations extended to 100 ns each, giving an aggregated simulation time of 1.6 $\mu$s.

**Peptide Setup.** Homopeptides with a length of four residues were generated using PyMOL.[16] In selected cases, a short steepest descent energy minimization was performed to relieve local structural strain. Potential energies were then calculated in vacuo without cutoffs to avoid any bias from different definitions of neighbor searching and interaction cutoffs between the two packages.

**System Setup.** The PDB structure 1IGD was downloaded from the Protein Data Bank,[17] and the first three, unordered, residues were deleted because their lack of a well-defined structure could potentially complicate the evaluation of the CMAP effect. The solvent water was modeled in four different ways: TIP3P and TIP4P,[18] the special CHARMM TIP3P model[4] (with LJ interaction sites also on the hydrogens), and the OBC[19] implicit solvent model, recently implemented in GROMACS. Neutral $NH_2$ and COOH termini were used, and crystal waters were retained in the systems with explicit water. The protein was placed in a rhombic dodecahedral unit cell with a minimum distance of 1.0 nm to the box edge. Steepest descent minimization was performed followed by an addition of ions to physiologically relevant levels and in order to counterbalance the protein charge, yielding a final system of about 8800 water molecules and 19 $Na^+$ and 17 $Cl^-$ ions. Finally, another steepest descent minimization was performed. Electrostatics was treated with particle-mesh Ewald (PME),[20] using a short-range cutoff of 1.2 nm, and van der Waals interactions were switched off between 1.0 to 1.2 nm. Temperatures were maintained using the thermostat of Bussi et al.[21] Periodic boundary conditions were applied, as well as isotropic pressure-coupling to a Parrinello−Rahman barostat[22] with a coupling constant of 1 ps. Simulations were run using a 2 fs (TIP3P, CHARMM TIP3P, TIP4P), 4 fs (TIP3P), or 5 fs (OBC) time step, with neighbor list updates every 20 fs.

**System Equilibration.** The systems were taken through two sets of equilibration simulations, using molecular dynamics (MD) or stochastic dynamics (SD) integration in systems with explicit and implicit solvent, respectively. First, a 2 ps simulation at 240 K was performed followed by a 1 ns simulation at the target temperature of 300 K. The constant for temperature coupling was 1.0 ps in the MD simulations, while the inverse friction constant of the SD integration was set to 91 $ps^{-1}$, in accordance with previous studies.[23] All covalent bonds were constrained to their equilibrium values by using the P-LINCS[24] algorithm, enabling a 2 fs time step in those cases, whereas in simulations with virtual sites, the time step could be pushed to 4 or 5 fs with the explicit and implicit water models, respectively.[9] For the explicit solvent simulations, all interactions were calculated as for the water system described above, while the implicit solvent simulations were run without cutoffs. Distorted structures for refinement were generated by running simulations at an elevated temperature (1500 K), starting from the energy-minimized protein. Two distorted structures were used, one with a 0.2 nm root-mean-square deviation (rmsd) of backbone heavy atoms and one with 0.3 nm. To study water structural effects, systems with 1000 previously equilibrated water molecules were simulated for 100 ns using all four different water models.

**Improved Virtual Site Construction.** Virtual interaction sites (originally called "dummies") have been available in GROMACS for almost a decade.[9] Their definition is quite straightforward: The virtual site coordinates are calculated from a set of constructing atoms every step. It takes part in the normal force evaluation, and finally the forces are spread back onto the constructing atoms by simply using the construction equations for the coordinates together with the derivative chain rule. In terms of energy conservation, the virtual site construction itself is perfect (the force is

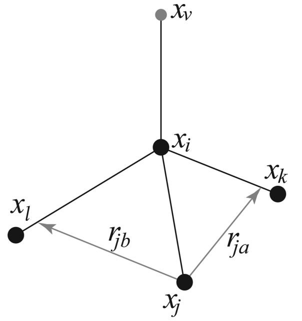


***Figure 1.*** New tetrahedral virtual site type from four atoms that is stable even when the constructing atoms *i*, *j*, *k*, and *l* are close to planar. See eq 1 for a formal definition.

exactly the derivative of the potential with respect to the constructing coordinates), but the potential for accuracy/speedup is limited by (1) whether the removed degrees of freedom were important and (2) the next fastest motion that limits the time step. Unfortunately, while the original virtual site approach was remarkably stable for force fields that only included polar hydrogens, it could occasionally cause errors with all-atom force fields. This was traced down to a particular construct (4FD of Feenstra et al.) used to build nonpolar protein $H_\alpha$ atoms from $C_\alpha$, $C_\beta$, N, and O; in the extremely rare case of the four constructing atoms being almost in a plane, this construct (but no others) can become unstable. We have corrected this by implementing a new 4FDN type. Assuming the virtual site $\mathbf{x}_v$ is connected to the tetrahedral center $\mathbf{x}_i$, which in turn is connected to $\mathbf{x}_j$, $\mathbf{x}_k$, and $\mathbf{x}_l$, the new virtual site is defined from three scalar parameters *a*, *b*, and *c* as (see Figure 1)

$$\begin{aligned} \mathbf{r}_{ja} &= a\mathbf{r}_{ik} - \mathbf{r}_{ij} = a(\mathbf{x}_k - \mathbf{x}_i) - (\mathbf{x}_j - \mathbf{x}_i) \\ \mathbf{r}_{jb} &= b\mathbf{r}_{il} - \mathbf{r}_{ij} = b(\mathbf{x}_l - \mathbf{x}_i) - (\mathbf{x}_j - \mathbf{x}_i) \\ \mathbf{r}_m &= \mathbf{r}_{ja} \times \mathbf{r}_{jb} \\ \mathbf{x}_v &= \mathbf{x}_i + c\frac{\mathbf{r}_m}{|\mathbf{r}_m|} \end{aligned} \qquad (1)$$

While this is somewhat more expensive to calculate (in particular, the analytical derivatives) due to the usage of the cross product, it is negligible for practical simulations, and this construct type has no stability issues.

## Results

For testing the correctness of the force field implementation, we compared the values for the potential energy of all amino acids to the force field included in the c33b1 release of CHARMM[14] and NAMD version 2.7b2[25] (Table 1). The CMAP implementation was also validated by comparing forces. These calculations were performed in vacuo without cutoffs, since implementation detail differences between CHARMM and GROMACS make it difficult to get exactly identical results with other setup schemes. For example, CHARMM does not necessarily work with charge groups, switch/shift functions are not the same,[26] and definitions of interaction cutoffs differ as well as rules for neighbor searching.

In analogy with MacKerell,[4] simulations with and without the CMAP correction terms were run to probe its effect. To extend their results and study the effects also on long timescales, simulations of 100 ns were performed. First, average differences between the $\phi$ and $\psi$ backbone dihedral angles in simulation were calculated relative to the crystal structure, averaged over all residues, as well as split into helical and sheet regions (Table 2). Both signed (as in the reference CMAP article) and unsigned differences were calculated, since the latter will expose systematic differences for individual residues that were canceled by the average over residues. Standard errors were estimated from autocorrelations as described by Hess;[27] while fluctuations can be large in loops and terminini, they are very small in ordered secondary structure regions, with standard errors below 2° in general. The results are consistent with MacKerell et al.; that is, these backbone dihedrals sample regions in conformational space significantly closer to those in the crystal structure when the CMAP term is applied. This is particularly obvious in the ordered secondary structure regions of helices and sheets, where these angles fluctuate around an average close to the crystal structure. Interestingly, this result is generally true for all water models and time steps investigated here. In particular, we could not detect any significant difference between the CMAP simulations that used the original TIP3P model and those with the CHARMM version of TIP3P. Even the implicit solvent simulations benefit from CMAP; the $\phi$ and $\psi$ angles are only slightly worse than with explicit water, and considerably better than the non-CMAP explicit solvent simulations.

To illustrate the behavior of these dihedral angles over time, Figure 2 shows the average $\phi$ and $\psi$ absolute differences for all residues for the frames of the 100 ns of production runs of the simulations using the TIP3P water model. Clearly, there are significant changes in the distributions of these angles on a timescale of tens of nanoseconds, which is important when considering lengthscales of refinement simulations, for example. Note that these 100 ns were preceded by 1 ns of nonrestrained equilibration simulations, during which the dihedral angles increased to the levels seen in the beginning of these graphs. As a reference, we also performed simulations with and without CMAP in a vacuum (data not shown) using 2 fs time steps, and as expected, the values of the dihedral differences and the rmsd are significantly worse in both cases. Hence, adding the CMAP term is a subtle effect, not simply overstabilizing the native state since a proper water model is needed for relevant sampling of the protein structure.

The average of the instantaneous rmsd between the crystal structure and the simulations was found to drop when the correction term was applied (Table 3). Again, the simulations using the original and CHARMM versions of the TIP3P water model were in practice indistinguishable, as well as the backbone rmsd of the two simulations with a 4 fs time step and virtual sites (with the original TIP3P). For side chains, which are generally more flexible than backbone atoms, the rmsd (of side chain heavy atoms) decreases when utilizing the virtual sites construction both with and without CMAP. One reason for this could be that the rigid interaction sites serve to make side chains slightly more stiff. Interestingly, there is no significant difference between the simula-

***Table 2.*** Average $\phi$, $\psi$ Differences (deg) from Crystal Structure

| | CHARMM27 | | | | CHARMM27 + CMAP | | | |
|---|---|---|---|---|---|---|---|---|
| system (1IGD) | $\langle\Delta\phi\rangle$ | $\langle\lvert\Delta\phi\rvert\rangle$ | $\langle\Delta\psi\rangle$ | $\langle\lvert\Delta\psi\rvert\rangle$ | $\langle\Delta\phi\rangle$ | $\langle\lvert\Delta\phi\rvert\rangle$ | $\langle\Delta\psi\rangle$ | $\langle\lvert\Delta\psi\rvert\rangle$ |
| | | | | All Residues | | | | |
| TIP3P | 0.5 | 16.5 | −10.8 | 19.8 | −1.2 | 9.3 | −4.6 | 9.8 |
| CHARMM TIP3P | 2.8 | 14.1 | −12.0 | 17.3 | −1.4 | 9.1 | −5.6 | 8.9 |
| TIP4P | 2.7 | 13.5 | −12.3 | 18.6 | −0.5 | 9.5 | −6.8 | 10.3 |
| TIP3P vsites | 2.7 | 13.4 | −16.3 | 22.4 | −2.2 | 8.9 | −3.4 | 8.5 |
| OBC | 3.6 | 14.4 | −17.2 | 21.0 | −0.9 | 12.7 | −6.4 | 15.7 |
| OBC vsites | 2.7 | 14.0 | −20.8 | 24.9 | −1.7 | 11.6 | −7.0 | 15.3 |
| | | | | Helical Residues | | | | |
| TIP3P | 4.1 | 8.0 | −3.7 | 7.0 | 0.3 | 3.4 | −0.1 | 3.5 |
| CHARMM TIP3P | 4.2 | 7.9 | −3.8 | 6.9 | 0.1 | 3.1 | −0.2 | 3.6 |
| TIP4P | 3.9 | 8.4 | −3.7 | 7.3 | 0.3 | 3.7 | 0.1 | 3.5 |
| TIP3P vsites | 4.6 | 8.2 | −4.5 | 7.8 | 0.2 | 2.7 | −0.4 | 3.4 |
| OBC | 6.3 | 8.6 | −7.4 | 9.0 | 2.7 | 5.6 | −1.9 | 4.1 |
| OBC vsites | 7.1 | 8.9 | −7.5 | 8.8 | 3.4 | 5.8 | −2.0 | 3.9 |
| | | | | Sheet Residues | | | | |
| TIP3P | 6.6 | 11.5 | −12.5 | 17.6 | −1.4 | 8.6 | −4.3 | 7.0 |
| CHARMM TIP3P | 5.3 | 10.5 | −10.2 | 16.8 | −1.6 | 8.8 | −4.1 | 6.9 |
| TIP4P | 4.9 | 8.7 | −10.5 | 17.4 | −1.4 | 9.4 | −4.5 | 7.9 |
| TIP3P vsites | 5.3 | 9.9 | −14.5 | 20.1 | −0.8 | 6.9 | −2.5 | 6.4 |
| OBC | 5.7 | 11.0 | −16.5 | 21.5 | −1.4 | 6.9 | −1.9 | 8.8 |
| OBC vsites | 5.0 | 10.5 | −18.2 | 23.1 | −0.7 | 6.4 | −2.6 | 9.5 |

tions with the implicit OBC model combined with CMAP compared to the explicit water models, not even when running with virtual sites and 5 fs time steps. Since this setup provides the highest computational performance by far, it is an interesting option for refinement. Also, with implicit solvent and virtual sites, side chains retain the same flexibility as when using only a 2 fs time step and could arguably sample the local conformational space better.

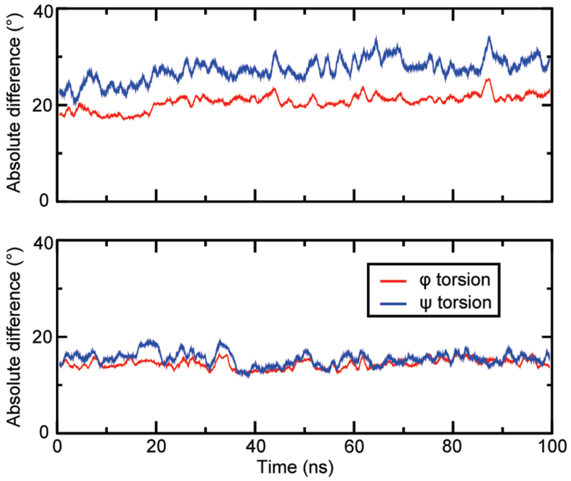


***Figure 2.*** Average absolute differences in the $\phi$ and $\psi$ backbone torsion angles compared to the corresponding angles in the protein crystal structure for the 100 ns of production runs. To decrease fluctuations, running averages have been computed from 1 ns windows. In this example, simulations were run with the TIP3P water model with and without CMAP, in the lower and upper panels, respectively. The differences in $\phi$ and $\psi$ are considerably less when applying CMAP.

***Table 3.*** Protein Stability and rmsd (nm) from Crystal Structure

| | CHARMM27 | | CHARMM27 + CMAP | |
|---|---|---|---|---|
| system (1IGD) | backbone | side-chain | backbone | side-chain |
| TIP3P | 0.12 | 0.23 | 0.09 | 0.20 |
| CHARMM TIP3P | 0.11 | 0.22 | 0.09 | 0.20 |
| TIP4P | 0.11 | 0.23 | 0.09 | 0.21 |
| TIP3P vsites | 0.12 | 0.18 | 0.09 | 0.15 |
| OBC | 0.11 | 0.23 | 0.09 | 0.20 |
| OBC vsites | 0.13 | 0.23 | 0.09 | 0.21 |

***Table 4.*** Protein G rmsd (nm) after Refinement

| | CHARMM27 + CMAP | |
|---|---|---|
| system (1IGD) | from 0.2 nm | from 0.3 nm |
| TIP3P | 0.11 | 0.28 |
| OBC vsites | 0.09 | 0.20 |

To investigate this further, we performed an elevated-temperature run in explicit water (again using the same protein, see the Methods section) to obtain structures to refine. Starting from a conformation that is initially 0.3 nm away from the crystal structure, we reach a rmsd of 0.28 nm with explicit water (calculated over the last 10 ns of a 100 ns simulation). The implicit solvent simulation does considerably better at approximately 0.20 nm. For the structure that started at 0.2 nm, we reach final values of 0.11 nm (explicit water) and 0.09 nm (implicit water; Table 4, Figure 3). While this at first sight appears to be identical to the value starting from the crystal structure, it has (unsurprisingly) not converged to cover the same phase space. When instead comparing the rmsd of *average structures*, the whole TIP3P explicit-water simulation reaches a remarkably low 0.055 nm; the last 10 ns of the explicit-water refinement starting from 0.2 nm reaches 0.09 nm, but the difference between the two is still 0.08 nm. Despite nonfully converged phase spaces, we believe this is promising for future refinement work.

Implicit solvation makes the sampling of conformational space faster as all solvent degrees of freedom are averaged out, indicating that it can move a structure far away (0.3 nm) from the conformation in the crystal structure faster

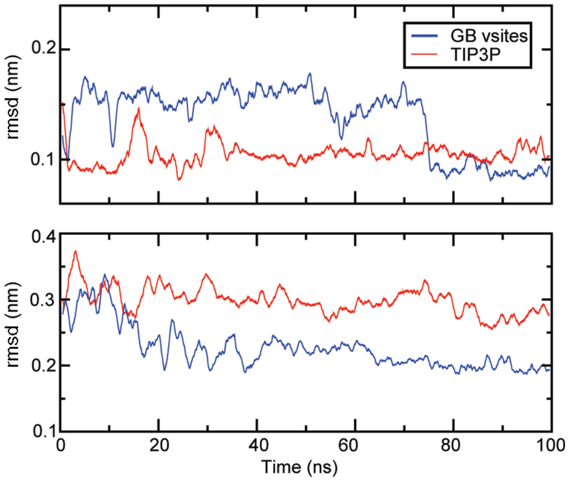


***Figure 3.*** The rmsd of refinement simulations for the 1IGD system with explicit TIP3P water molecules in red and the implicit OBC model and hydrogen virtual sites in blue. Curves were smoothened by running averages from windows of 1 ns length. The rmsd is followed for simulations starting from a protein conformation 0.2 and 0.3 nm from the crystal structure, shown in the upper and lower panels, respectively.

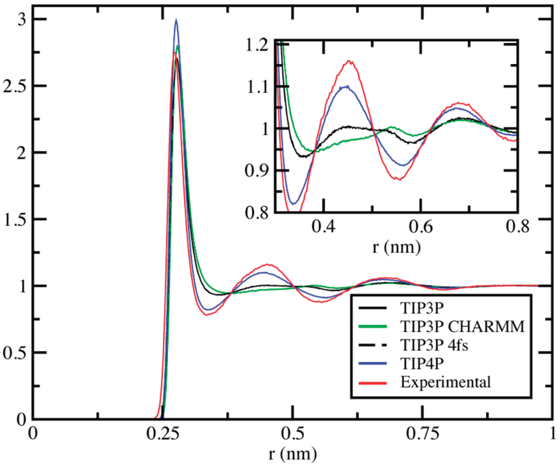


***Figure 4.*** Radial distribution functions for oxygen−oxygen distances for the different water models. TIP4P is closest to the experimental reference curve.[28] Using TIP3P with a 4 fs time step yields a RDF (dashed) that is indistinguishable from that with a 2 fs time step.

toward this state. Starting closer to the native state (0.2 nm), the protein structure is again clearly closer to the crystal conformation at the end of the simulations, but there is less difference between explicit and implicit solvent, possibly indicating that other factors, such as, for example, accurate hydrogen bonding, could be more important.

Finally, to quantify the possible differences from the various water models on the structure of water itself, and also to relate to experimental results, we calculated the oxygen−oxygen radial distribution function (RDF) from pure water simulations. As shown in Figure 4, the radial distribution functions of the two TIP3P models are very close, and both are significantly worse than TIP4P when compared to the experimental curve from Soper.[28] The average densities over the 100 ns simulations are 1001.7 ± 1.3 kg/ $m^3$ and 1014.7 ± 1.4 kg/ $m^3$ for original TIP3P and CHARMM-modified TIP3P, respectively.

## Discussion

There are no significant differences in protein backbone dihedral angles when using the CHARMM-specific TIP3P model compared to the original TIP3P with only a single Lennard-Jones interaction site, and at least for the dihedrals, the results are very close to TIP4P. In terms of long-time structural stability, using the CMAP correction term really does seem to do what it was intended to do, even on timescales where problems frequently start to show; increasing simulation time to the 100 ns scale still produces improvements consistent with those originally observed on a single nanosecond scale.[7]

Both from accuracy and computational points of view, water models can be important. CHARMM typically uses a slightly modified version of TIP3P that includes Lennard-Jones interactions on all atoms, while the original Jorgensen model[18] only does it on the water oxygen. The classical argument in the latter case is that the hydrogens are extremely small, and that the Lennard-Jones potential anyway is an approximation. Since water can account for 90% of the interactions in a typical biomolecular simulation, GROMACS includes special nonbonded kernels that can take the absence of Lennard-Jones interactions on hydrogens into account to accelerate those interactions. When first implementing the CHARMM force field in GROMACS, we considered writing similar kernels for the alternative water model.

In general, minor changes in water models can be quite important. For many features such as hydration entropies, enthalpies, or heat capacities, the choice of water model is even much more important than the rest of the force field.[29] As evident from Figure 4, both the original and CHARMM TIP3P models deviate significantly from TIP4P or the experimental results, but the differences between the two models are minimal. Both models also appear equally good at stabilizing the protein structure, so rather than implementing a separate set of kernels—which would still be slightly slower than the original ones—we would rather advocate the use of the standard Jorgensen TIP3P model, without Lennard-Jones interactions on the hydrogens (but both choices are available). This does not rule out the possibility of model differences being significant in some cases, but for critical applications, we would anyway rather recommend TIP4P, which is only 7% slower than TIP3P in GROMACS.

As with the different water models, the choice of explicit versus implicit solvent depends on the questions being addressed. If accurate dynamics or, for example, exact hydrogen bond patterns are important, explicit solvent is the natural choice. However, longer timescales can be reached with an implicit solvent, and since solvent relaxation is instantaneous, it can also provide better sampling. This can be important for structural refinement where sampling is critical—provided the lowest free energy state of the force field is close to the native state.

It is reassuring that there does not appear to be any significant difference from the use of virtual sites to extend the time step; no negative trends are visible when measuring the degree of structural divergence going from 2 to 4 fs. Even with implicit solvent and a 5 fs time step, accuracy is maintained, and the effect of the CMAP correction term is virtually the same as for the explicit solvent (Table 2). Regardless of the choice of solvent or water model, these results seem to indicate that the CHARMM force field works remarkably well with the combination of long timesteps enabled by virtual sites and implicit solvent.

It is interesting to see how the performance depends on both the solvent model and virtual interaction sites. Protein G itself consists of some 900 atoms, and with 8734 waters added, the total number of atoms exceeds 27 000 and 35 000 for TIP3P and TIP4P, respectively. Runs were performed on a 2.66 GHz quad-core Intel Nehalem, using long 1.2 nm cutoffs, PME calculated every step, and Parinello−Rahman pressure scaling, as described in the Methods section. The use of CHARMM-specific TIP3P resulted in a performance of 3.3 ns/day, while standard TIP3P reaches 5.9 ns/day and TIP4P 5.5 ns/day. Combining TIP3P with virtual sites and 4 fs steps yielded 11.3 ns/day with maintained accuracy. Finally, the (relatively expensive) OBC implicit solvent model with virtual sites and 5 fs time steps reaches a full 250 ns/day, with only marginal effects on quality.

To truly assess refinement, a more stringent test would be to take a diverse set of protein structure models, such as those produced at the biannual CASP experiment, and see whether it is possible to systematically move these closer to the native state. What we have presented here might arguably represent an easier problem, since our starting structures were produced from the native state using high-temperature simulations. The question is whether or not this means that there in some sense is an “easier” way back to the native state; that is, are “real” homology models more difficult to refine? This is a highly important question, but answering it is beyond the scope of this force field study. However, the example studied here provides promising signs that (1) refinement could indeed be possible with reasonable amounts of simulation time, (2) different solvent models might be important at different degrees of divergence from the sought native state, and (3) correction maps appear to be a universal improvement.

## Conclusions

As first observed by MacKerrel et al., the improvement from the correction maps term is significant considering the efforts that have gone into parametrizing current force fields. We find it quite striking that it seems to hold across the line in our tests, regardless of water model or even implicit solvent. The success of combining an implicit solvent model and long time steps using CHARMM27 with correction maps (a computationally very appealing combination) suggests that it is a promising setup, for example, for protein refinement simulations where the sampling efficiency is one of the limiting factors. The fact that the correction maps seem to be a close-to-universal improvement also suggests that it would be interesting to evaluate the effect of similar correction terms for other parameter sets, which is a direction we intend to pursue in the future. The implementation presented here is already available in the public GROMACS git repository (see www.gromacs.org for information) and will be officially supported as of GROMACS 4.1.

**Acknowledgment.** This work was supported by grants to E.L. from the European Research Council (209825), the Swedish Foundation for Strategic Research, and the Swedish Research Council. Supercomputing resources were provided by National Science Foundation award CNS-0619926 to Stanford University.

## References

(1) Wang, J.; Cieplak, P.; Kollman, P. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(2) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(3) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

(4) MacKerell, A. D.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

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

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

(7) MacKerell, A. D., J.; Feig, M.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1400–15, 3d.

(8) van der Spoel, D.; Lindahl, E. *J. Phys. Chem. B* **2003**, *117*, 11178–11187.

(9) Feenstra, A.; Hess, B.; Berendsen, H. *J. Comput. Chem.* **1999**, *20*, 786–798.

(10) Zhu, J.; Fan, H.; Periole, X.; Honig, B.; Mark, A. *Proteins* **2008**, *72*, 1171–1188.

(11) Chopra, G.; Summa, C.; Levitt, M. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 20239–20244.

(12) Misura, K.; Baker, D. *Proteins* **2005**, *59*, 15–29.

(13) Chen, J.; Im, W.; Brooks, C. L. *J. Comput. Chem.* **2005**, *26*, 1565–1578, 3rd.

(14) Brooks, B. R.; Bruccoleri, R. E.; B. D. Olafson, D. J. S.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(15) Derrick, J. P.; Wigley, D. B. *J. Mol. Biol.* **1994**, *243*, 906–18.

(16) DeLano, W. L. *The PyMOL Molecular Graphics System*, v1.1r1; DeLano Scientific: Palo Alto, CA, 2002.

(17) 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*, 235–42.

(18) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(19) Onufriev, A.; Bashford, D.; Case, D. *Proteins* **2004**, *55*, 383–394.

(20) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(21) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.

(22) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(23) Snow, C.; Nguyen, H.; Pande, V.; Gruebele, M. *Nature* **2002**, *420*, 102–106.

(24) Hess, B. *J. Chem. Theory Comput.* **2007**, *4*, 116–122.

(25) Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(26) Steinbach, P.; Brooks, B. *J. Comput. Chem.* **1999**, *15*, 667–683.

(27) Hess, B. *J. Chem. Phys.* **2002**, *116*, 209–217.

(28) Soper, A. *Chem. Phys.* **2000**, *258*, 121–137.

(29) Hess, B.; van der Vegt, N. *J. Phys. Chem. B* **2006**, *110*, 17616–17626.

CT900549R

JCTC Journal of Chemical Theory and Computation

# Explicit Hydrogen-Bond Potentials and Their Application to NMR Scalar Couplings in Proteins

Jing Huang and Markus Meuwly*

*Department of Chemistry, University of Basel, Klingelbergstrasse 80, 4056 Basel, Switzerland*

Received October 27, 2009

**Abstract:** Hydrogen bonds (H bonds) are fundamental for the stability, structure, and dynamics of chemically and biologically relevant systems. One of the direct means to detect H bonds in proteins is NMR spectroscopy. As H bonds are dynamic in nature, atomistic simulations offer a meaningful way to characterize and analyze properties of hydrogen bonds, provided a sufficiently accurate interaction potential is available. Here, we use explicit H-bond potentials to investigate scalar coupling constants ${}^{h3}J_{NC'}$ and characterize the conformational ensemble for increasingly accurate intermolecular potentials. By considering a range of proteins with different overall topology a general procedure to improve the hydrogen-bonding potential ("morphing potentials") based on experimental information is derived. The robustness of this approach is established through explicit simulations in full solvation and comparison with experimental results. The H-bond potentials used here lead to more directional H bonds than conventional electrostatic representations employed in molecular mechanics potentials. It is found that the optimized potentials lead to H-bond geometries in remarkable agreement with previous *ab initio* and knowledge-based approaches to H bonds in model systems and in proteins. This suggests that, by combining theory, computation, and experimental data, H-bonding potentials can be improved and are potentially useful to better study coupling, energy transfer, and allosteric communication in proteins.

## Introduction

Hydrogen bonds are ubiquitous in chemical and biological systems and are essential for the overall structure, function, and dynamics of proteins and other macromolecules.[1] The role of hydrogen bonds in protein folding,[2] the formation of secondary structural elements,[3,4] molecular recognition,[5,6] and catalysis[7,8] has been established over the past few years. A central feature of H bonds is their directionality, which cannot be easily captured by a superposition of isotropic interactions such as Coulomb interactions, as is done in customary force fields such as CHARMM, AMBER, or OPLS-AA.[9–11] In small molecules, a hydrogen bond can be characterized spectroscopically. For example, in complexes between simple ions ($HCO^+$, $HN_2^+$) and rare gas atoms (He, Ne, Ar), it is found that the fundamental infrared transitions in the electronic ground state correspond to $\Sigma-\Sigma$ transitions characteristic for linear molecules.[12–14] This can be inferred from the structure of the ro-vibrational bands (missing Q branch). Also, fitting of a model Hamiltonian[15] allows for a determination of structural constants, which in turn characterize the average geometry of the molecule. For biological macromolecules, it is more difficult to find direct measures for the directionality of H bonds and to locate the positions of the hydrogen atoms. Direct visualization through recording of a structure is impractical, as H atoms can usually not be seen in X-ray crystallography. Structure determination from NMR data, on the other hand, formulates a search problem in structure space which minimizes a cost function that involves the experimental information (usually nuclear Overhauser data) and additional physical information because experimental data are rarely sufficient to determine the three-dimensional structure of a macromolecule.[16]

One experimental signature which recently became more widely available is hydrogen bond scalar couplings, which

* Corresponding author e-mail: m.meuwly@unibas.ch.

10.1021/ct9005695 © 2010 American Chemical Society
Published on Web 01/22/2010

can be measured through NMR spectroscopy.[17,18] Scalar couplings across N−H···O═C H bonds in proteins have a typical range of about −0.2 to −1 Hz, and the measurement errors are usually less than 0.05 Hz.[19,20] Scalar $^{h3}J_{NC'}$ couplings have been observed experimentally in peptides,[21] nucleic acids,[22] and a variety of proteins.[18,23–28] Together with other NMR parameters such as relaxation times, residual chemical shift anisotropy, and dipolar couplings, $^{h3}J_{NC'}$ couplings are important in the identification of conformational dynamics taking place on the NMR time scale.[29–31] Further interesting and fundamental aspects of $^{h3}J_{NC'}$ couplings are their sensitivity to H-bonding network dynamics and cooperativity. Such effects are very difficult to probe directly through experiments, and a combined approach including atomistic simulations may prove advantageous. Earlier work established that the explicit dynamics of the solvated protein have to be taken into account to reliably calculate scalar coupling constants from molecular dynamics (MD) simulations.[32,33] This naturally paves the way to improve specific terms in empirical force fields to which the observables are sensitive. In the present case, it is the capability of a force field to correctly describe H bonds.

It has been found that $^{h3}J_{NC'}$ values can be directly correlated with H-bond geometries. Barfield proposed several empirically parametrized formulas which enable the calculation of scalar couplings from the local N−H···O═C structure.[34] As NMR spectroscopy is a time-domain method, measured scalar couplings have to be understood as time averages. From a computational point of view, molecular dynamics simulations are the method of choice for such investigations. In previous work,[32,33] a good correlation between measured $^{h3}J_{NC'}$ couplings and those derived from all-atom simulations was established by carrying out nanosecond MD simulations and averaging $^{h3}J_{NC'}$ values over entire trajectories.

Here, we combine a recently developed explicit hydrogen potential (molecular mechanics with proton transfer - MMPT)[35,36] derived from correlated quantum mechanical calculations with an established force field to characterize $^{h3}J_{NC'}$ couplings in a variety of proteins covering different folds (ubiquitin (α + β), the GB1 domain of protein G (α + β), cold-shock protein A (all β), apo-calmodulin (all α), holo-calmodulin (all α), and intestinal fatty acid binding protein (all β), see Figure 1). Compared with conventional MD studies, the deviations between calculated and experimental $^{h3}J_{NC'}$ values are notably lowered. Next, the topology of the potential energy surfaces for the H-bond potentials is modified through morphing transformations[37,38] to best describe experimentally determined couplings for three proteins. This approach is then generalized by applying it to the proteins not belonging to the training set, and very good agreement with measured coupling constants is found. Most notably, the approach pursued here leads to an average separation between the hydrogen atom and the acceptor of 1.93 Å, which agrees with a knowledge-based potential derived from 52 structures and results from electronic structure calculations.[39,40]

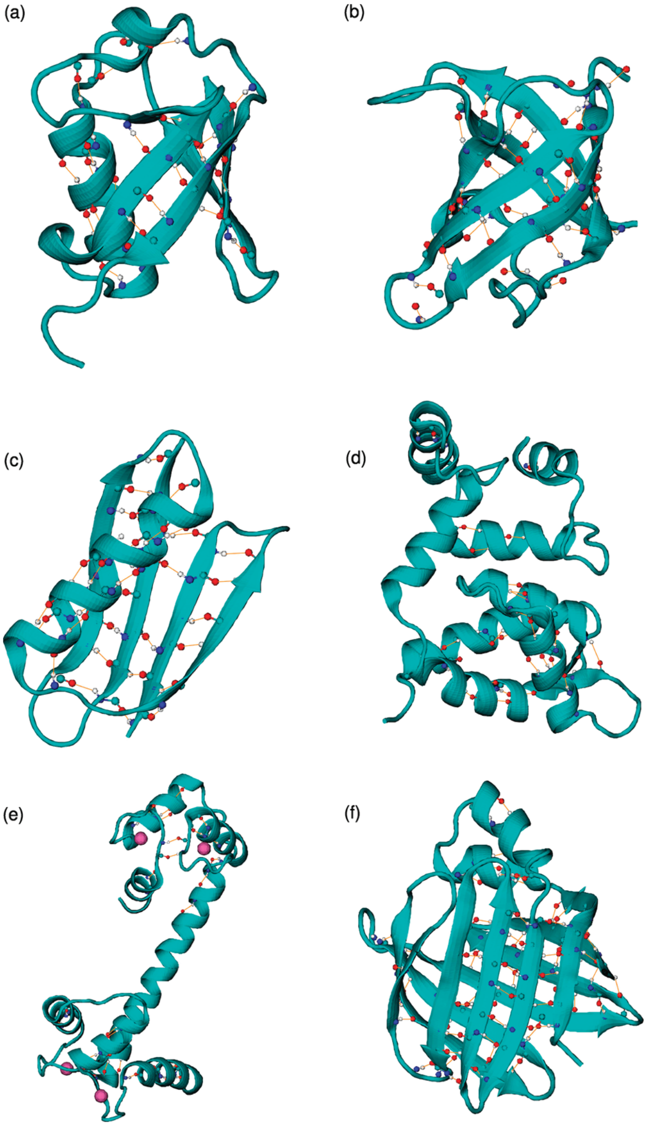


***Figure 1.*** Structure, topology, and H bonds for the six proteins investigated here. (a) Ubiquitin, (b) GB1 domain of protein G, (c) cold-shock protein A, (d) apo-calmodulin, (e) holo-calmodulin, (f) intestinal fatty acid binding protein.

## Computational Methods

**Molecular Dynamics Simulations.** All simulations were carried out with the Charmm program[41] using the CHARMM22 force field[9] and provisions for MMPT.[36] The starting structures were taken from the X-ray structures in the Protein Data Bank[42] (ubiquitin, 1ubq;[43] protein G,[44b] 2qmt;[44a] cold-shock protein A (CspA), 1mjc;[45] apo-calmodulin (apoCAM), 1qx5;[46] holo-calmodulin (holoCAM), 1cll;[47] intestinal fatty acid binding protein (IFABP), 1ifc.[48] Hydrogen atoms were generated with HBUILD,[49] and the structures were relaxed by 3000 steps of steepest descent minimization. Then, the proteins were solvated in pre-equilibrated water boxes of suitable sizes (1ubq, 65.19 Å × 52.77 Å × 49.67 Å; 2qmt, 55.88 Å × 46.56 Å × 40.36 Å; 1mjc, 52.77 Å × 52.77 Å × 46.56 Å; 1qx5, 71.40 Å × 58.98 Å × 58.98 Å; 1cll, 90.03

Å × 62.09 Å × 49.67 Å; 1ifc, 65.19 Å × 55.88 Å × 52.77 Å), and periodic boundary conditions were applied. A cutoff of 14 Å was applied to the shifted electrostatic and switched van der Waals interactions. Before free dynamics simulations, the systems were heated to 300 K and then equilibrated for $10^5$ time steps.

For conventional MD simulations, all hydrogen atoms were constrained by SHAKE,[50] whereas for simulations with MMPT, hydrogen atoms involved in ${}^{h3}J_{NC'}$ couplings were free to move and all other hydrogen atoms were treated with SHAKE. A complete list of H bonds treated by MMPT for all proteins is summarized in Supporting Information S2. In both standard MD and MMPT/MD simulations, the time step was 0.2 fs, and snapshots were taken every 0.02 ps. The hydrogen-bond coordinates were extracted from trajectories and used together with eq 1 to calculate ${}^{h3}J_{NC'}$ couplings:[32,34]

$$ {}^{h3}J_{NC'} = (-366\ \text{Hz})\exp(-3.2r_{HO'})\,[\cos^2\theta_1 - (0.47\cos^2\rho + 0.70\cos\rho + 0.11)\sin^2\theta_1] \qquad (1) $$

where $r_{HO'}$ is the distance between hydrogen and acceptor atoms, while $\theta_1$ and $\rho$ represent the H···O=C′ angle and the H···O=C′−N′ dihedral angle, respectively.

A simplified formula (eq 2) is also proposed in ref 34 and was used in a previous work:[33]

$$ {}^{h3}J_{NC'} = (-360\ \text{Hz})\exp(-3.2r_{HO'})\cos^2\theta_1 \qquad (2) $$

It captures the dominant effects of scalar couplings, while eq 1 provides a better estimate of ${}^{h3}J_{NC'}$ couplings in protein G because it accounts for the systematic difference between hydrogen bonds along the α helix and $\beta$ sheet, respectively, by including a term related to the dihedral angle $\rho$.[34] Equations 1 and 2 can provide the same accuracy as full DFT calculations[32] and have been used to calculate ${}^{h3}J_{NC'}$ couplings in different proteins.[30,32–34,51] Detailed investigations on small molecules compared the performance of DFT using VWN, BP, or PW91 functionals with results from correlated methods such as the coupled cluster singles and doubles polarization propagator approximation[52] and found that, with the exception of the HF molecule, the performance of DFT is good and provides almost quantitative spin−spin coupling constants.[53] The sensitivity to changes in the parameters of eq 2 has recently been investigated in a systematic fashion.[33] It was found that, overall, a strength factor of $\alpha = -360$ Hz and a decay of $\beta = 3.2$ Å$^{-1}$ provide a good description of most coupling constants. However, for scalar couplings in particular secondary structural elements, the values for α and $\beta$ could be optimized. As in the present work such aspects are not further pursued, eqs 1 and 2 are used, and the results are virtually identical. Generally, eq 1 leads to slightly smaller deviations between calculated and measured ${}^{h3}J_{NC'}$ values, so it was used in this work to calculate scalar couplings in all proteins except for CspA, where three backbone−side chain couplings were also included. While distinction between α-helix and $\beta$-sheet hydrogen bonds is only relevant for backbone−backbone hydrogen bonds, eq 2 has to be applied for computing these ${}^{h3}J_{NC'}$ values in CspA.

The quality of the simulations was assessed by comparing root-mean-square deviations (RMSDs) between calculated and experimental ${}^{h3}J_{NC'}$ couplings:

$$ \text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(J_i^{\text{calcd}} - J_i^{\text{exptl}})^2} \qquad (3) $$

**MMPT Potential and Morphing Transformations.** A detailed account of MMPT has been given in ref 36. Briefly, MMPT uses parametrized three-dimensional potential energy surfaces fitted to high-level *ab initio* calculations (MP2/6-311++G(d,p)) to describe the interactions within a general DH−A motif, where D is the donor, H is the hydrogen, and A is the acceptor atom. Together with a standard force field—here, CHARMM[9] is used—specific rules control how bonded interactions on the donor and acceptor side are switched on and off depending on the position of the transferring H atom (DH−A or D−HA). To adapt the overall shape of the PES to topologically similar, but energetically different, hydrogen bonding patterns—depending on the chemical environment of D and A—the PES can be “morphed”.[37,38] Morphing can be a simple coordinate scaling or a more general coordinate transformation depending on whether the purpose of the study and the experimental data justify such a more elaborate approach.

For the present case of hydrogen bonds between an amide (NH) group as the donor and the oxygen atom as the acceptor (NH···O), the MMPT potential depends on $R$ (distance between N and O), $\rho$ (relative position of H for a particular value of $R$), and $\theta$ (angle between unit vectors $\vec{R}$ and $\vec{\rho}$). The relationship between $\rho$ and the N−H distance $r$ is given by

$$ \rho = (r - r_{\min})/(R - 2r_{\min}) \qquad (4) $$

where $r_{\min} = 0.8$ Å is, in principle, arbitrary but should be sufficiently small to cover the shortest D−A separations. The angular dependence of the potential $V(R, \rho, \theta)$ is harmonic, that is, $V(R, \rho, \theta) = V_0(R, \rho) + k\theta^2$, and a typical PES along $R$ and $\rho$ is shown in Figure 2.

As mentioned above, the MMPT potentials are calculated for model systems (zeroth-order potential) and subsequently morphed to describe the situation in the actual chemical environment. Here, the asymmetric zeroth-order potential for $NH_4^+$···$OH_2$ is morphed to describe the N−H···O=C motif in proteins. Morphing is achieved by modifying the parameters and thus reshaping the MMPT potential. The original potential has a single minimum $\{R_0 = 2.71$ Å$, \rho_0 = 0.23, \theta_0 = 0°\}$ and is mapped to a new one $\{R' = R_0 + \sigma, \rho' = \rho_0 - \delta, \theta_0 = 0°\}$ where $\sigma$ and $\delta$ are positive because hydrogen bonds in proteins are weaker than in a protonated ammonia−water dimer. Since there is a one-to-one correspondence between morphing parameters $\{\sigma, \delta\}$ and PES minima $\{R', \rho'\}$, only one set—$\{R', \rho'\}$—will be used in the following. The morphed potential has its minimum energy at $\{R', \rho', 0\}$ while maintaining its overall shape, as illustrated in Figure 2.

For most X-ray structures, typically the coordinates of heavy atoms are available since only very rarely can protein crystallography resolve the positions of hydrogen atoms. Therefore, the experimental observable characterizing a

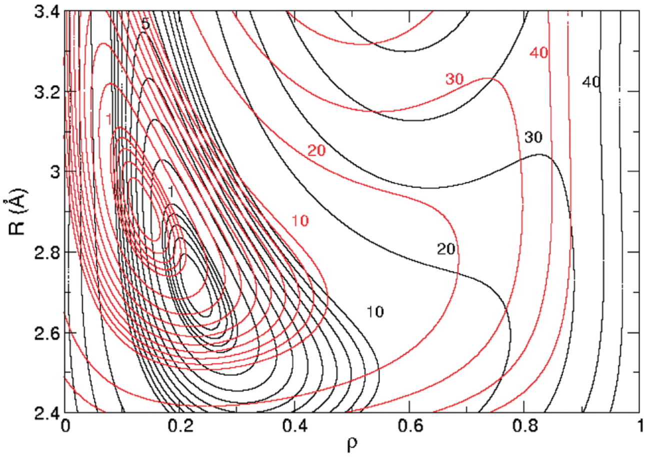


***Figure 2.*** MMPT PES for the NH−O motif and illustration of PES morphing. Black, original PES; red, morphed PES. Contour lines are drawn at intervals of 0.2 kcal/mol for energies below 1 kcal/mol, for energies between 1 and 10 kcal/mol at intervals of 1 kcal/mol, and for higher energies at intervals of 10 kcal/mol. Morphing parameters $\{R', \rho'\} = \{2.92, 0.14\}$.

hydrogen bond is the D−A distance between donor and acceptor. In the Results section, it will be shown that best value for $R'$ corresponds closely to the average D−A distances calculated from the initial X-ray structures. We also establish a relationship between optimized $R'$ and $\rho'$ (eq 5, see below). This leads to the following procedure for optimizing MMPT parameters and calculating $^{h3}J_{NC'}$ couplings by MMPT/MD simulations:

(1) From the X-ray/NMR structure, the average distance $R'$ is calculated.

(2) Compute $\rho'$ by eq 5 (see below).

(3) Morph the MMPT PES to the minima $\{R', \rho'\}$ by coordinate transformations.

(4) Carry out MD simulations with the MMPT potential, and calculate the hydrogen-bond scalar couplings according to eq 1 or 2.

## Results

**Conventional MD as Benchmarks.** Standard MD simulations 1 ns in length were first carried out for all six proteins, and the RMSDs between calculated and experimental couplings were computed as benchmarks for comparison. As shown in Supporting Information S3, $^{h3}J_{NC'}$ couplings converge well within 1 ns. Hence, the RMSD as the average over all $^{h3}J_{NC'}$ couplings is also stable during our simulation time scale; for example, the RMSDs of CspA calculated from 0.5, 1, and 1.5 ns standard MD trajectories are 0.198, 0.195, and 0.197 Hz, respectively.

We also carried out 500 ps MD simulations with CMAP for ubiquitin and CspA. CMAP is an extension of the CHARMM force field and has recently been shown to obtain a more accurate description of the peptide backbone.[54] By including grid-based energy correction maps and empirical corrections, this approach yields improved dynamical and structural properties of proteins in various simulations.[55,56] However, applied to the present simulations of $^{h3}J_{NC'}$ couplings for ubiquitin and CspA, results are very similar to simulations without CMAP, as illustrated in Supporting Information S4.

**MD Simulations with MMPT.** The zeroth-order MMPT PES is suitable to describe a N−H···O bond in $NH_4^+{-}H_2O$ and will not be directly applicable to hydrogen bonding in proteins. Therefore, it is expected that MD simulations using the unmorphed MMPT potential are unsuited for quantitative work, and large deviations between observed and calculated $^{h3}J_{NC'}$ couplings should be found, as illustrated in Figure 3. When different morphing parameters are used, the MMPT potentials will have different minimum energy geometries $\{R', \rho'\}$ and lead to different scalar couplings, which is also shown in Figure 3.

The correlation between morphing parameters and RMSDs has been investigated for ubiquitin, CspA, and protein G. First, short (20 ps) test trajectories were run to locate suitable morphing parameters, and then 100 ps MD simulations were carried out on a fine grid ($\Delta = 0.01$ Å) of $\{R', \rho'\}$ and analyzed. For combinations $\{R', \rho'\}$ with low RMSDs, simulations were continued to 500 ps. Longer trajectories (1 ns) were run for ubiquitin (morphing parameters {2.92, 0.14}), protein G ({2.95, 0.16}), and CspA ({2.96, 0.16}), and RMSDs were calculated and are summarized in Table 1, together with results obtained from standard MD simulations. As an illustration, a detailed comparison between measured and calculated $^{h3}J_{NC'}$ in CspA from standard MD and MMPT/MD simulations, and the squared deviations for each individual hydrogen-bond coupling, are shown in Figure 4. By adopting MMPT PES as the explicit hydrogen-bond potential, the correlation between calculated $^{h3}J_{NC'}$ couplings and experimental data has been enhanced for most hydrogen bonds, especially those with large deviations ($|J_{calcd} - J_{exptl}| > 0.3$ Hz). The range of scalar couplings calculated from MMPT/MD simulations, however, is narrower than that from standard MD simulations. For convergence of most scalar couplings, a total of 500 ps is typically sufficient for MD simulations with the MMPT potential (see Supporting Information S3).

**Application of the Morphed Potentials.** After establishing that morphed MMPT potentials lead to improved agreement between calculated and experimental scalar coupling constants compared to those of a conventional force field (Table 1), potential morphing is used to further improve scalar coupling constants starting from X-ray and NMR structures. This is done for the three proteins studied in the previous section: ubiquitin, CspA, and protein G. As might be suspected, somewhat different coordinate transformations are most suitable to best describe the scalar couplings in the three different proteins (a summary of the relationship between RMSD and different MMPT PESs is given in the Supporting Information S5). Due to the nonlinearity between parameters ($R'$, $\rho'$) for the H-bond potentials and the calculated RMSDs between calculated and measured $^{h3}J_{NC'}$ couplings, there is no simple, detectable relationship between the two. However, it is found that deviations are generally small around a certain $\{R', \rho'\}$ combination, and these values are summarized in Table 2. The average N−O distances computed from the initial structure are also reported and are

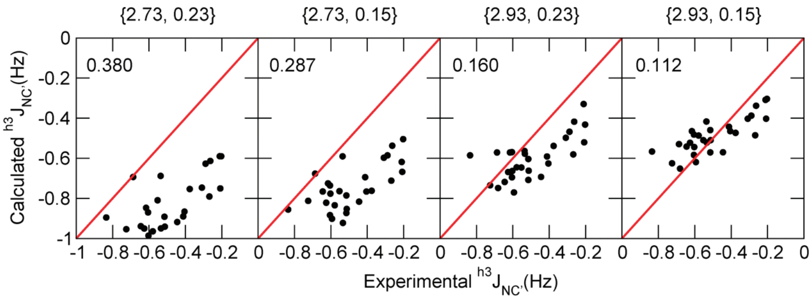


***Figure 3.*** Calculated scalar couplings from 0.2 ns of MMPT/MD simulations compared with experimental data in ubiquitin for different morphing parameters.

***Table 1.*** RMSDs of Ubiquitin, CspA, and Protein G Calculated from 1 ns Trajectory

| | ubiquitin | CspA | protein G |
|---|---|---|---|
| standard MD | 0.142 | 0.195 | 0.134 |
| MD/MMPT | 0.118[a] | 0.123[b] | 0.130[c] |

[a] Morphing parameters {2.92,0.14}. [b] Morphing parameters {2.96,0.18}. [c] Morphing parameters {2.95,0.16}.

close to $R'$. With the use of relationship 4, the actual hydrogen bondlength $r'_{OH}$ is found to be almost identical for all three proteins, namely, 1.93 Å with an average of 1.931 Å ± 0.002 Å. It is worthwhile mentioning that this value is reminiscent of the hydrogen-bond geometry parameter $\delta_{HA}$ calculated from a statistical analysis of 52 proteins.[40] In the following, potential morphing for MMPT PESs is further investigated such that the additional constraint $r'_{OH} = 1.93$ Å is fulfilled:

$$\rho' = \frac{R' - 2.73}{R' - 1.6} \qquad (5)$$

This equation directly relates the two morphing parameters.

To test the procedure, it was applied to apo-CAM, holo-CAM, and IFABP, which were not part of the training set. The 500 ps MMPT/MD simulations were carried out with MMPT PES minima $\{R', \rho'\}$ found above, and scalar couplings were calculated. RMSDs between calculated and experimentally measured $^{h3}J_{NC'}$ couplings are summarized in Table 3. Compared with results from 1 ns standard MD simulations, considerably better agreement is achieved for all six proteins we investigated.

In all previous MMPT/MD simulations, only hydrogen bonds corresponding to experimentally measured scalar couplings are treated with the explicit hydrogen-bond potential. It would be interesting to test whether $^{h3}J_{NC'}$ can be predicted by MMPT/MD simulations without knowing which couplings can be observed in E.COSY experiments. Visual Molecular Dynamics (VMD)[57] has been used to assign hydrogen bonds in ubiquitin and CspA with a distance cutoff of 3.5 Å and an angle cutoff of 40°. In both proteins, more hydrogen bonds are found with this criterion (see Supporting Information S6), but not all of the previously assigned hydrogen bonds are covered. MD simulations with all of these hydrogen bonds treated by MMPT were carried out. $^{h3}J_{NC'}$ couplings were calculated from 500 ps trajectories and compared to experimental values. The RMSDs (0.122 and 0.160 Hz) are not as good as previous MMPT/MD results (0.116 and 0.140 Hz) but are still significant improvements over standard MD simulations (0.140 and 0.195 Hz).

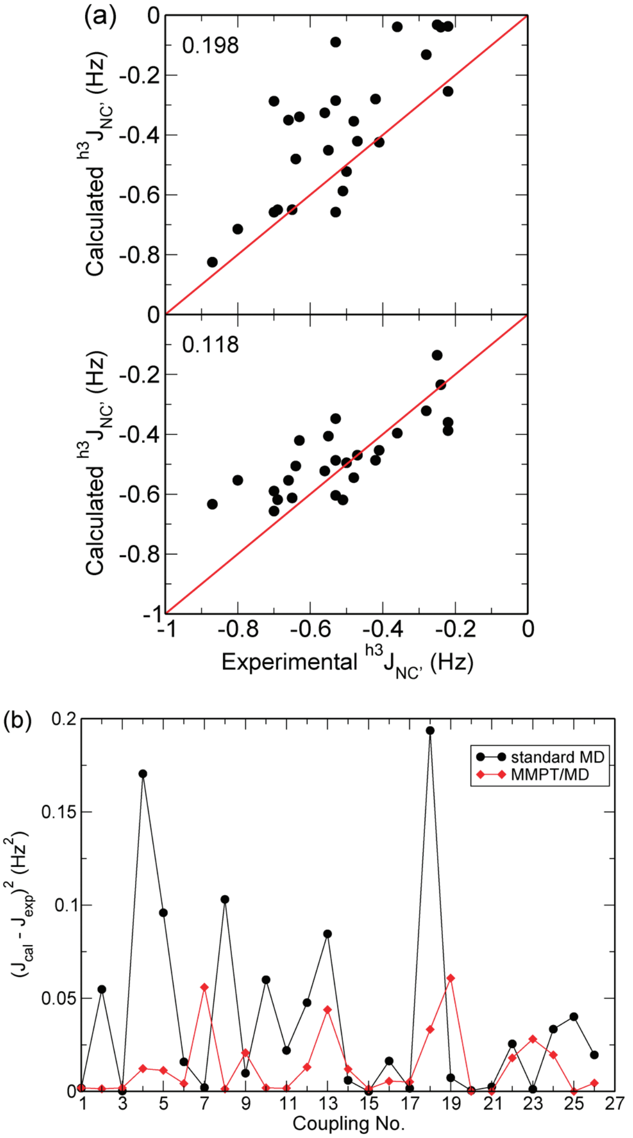


***Figure 4.*** Comparison between scalar couplings calculated by standard MD simulations and MMPT/MD simulations for cold-shock protein A. (a) Comparisons of calculated and experimental $^{h3}J_{NC'}$ couplings. (b) Squared deviations for hydrogen bond scalar couplings.

**Characterization of the Conformational Ensemble.** Once suitable morphing parameters are available, MMPT/MD can also be used to characterize the conformational

***Table 2.*** Overview of PES Morphing Parameters for Ubiquitin, CspA, and Protein G

| | $R'$, Å | $\rho'$ | $r'_{OH}$, Å | $\langle R_{X\text{-ray}}\rangle$, Å |
|---|---|---|---|---|
| 1ubq | 2.925 | 0.145 | 1.933 | 2.921 |
| 1mjc | 2.960 | 0.170 | 1.929 | 2.958 |
| 2qmt | 2.945 | 0.160 | 1.930 | 2.942 |

ensemble starting from the X-ray structure. The conformational ensembles generated by MD simulations with and without the MMPT potential are investigated through the distance between hydrogen and acceptor atoms and the angle at the acceptor atom. The respective density distributions ($r_{NO}$, $\theta_{HOC}$) from 500 ps simulations for 29 hydrogen bonds in ubiquitin are shown in Figures 5 and 6. Using the MMPT potential for the hydrogen bonds in proteins leads to slightly shorter donor−acceptor distances, more pronounced directionality of the H bonds, and significant reductions in the fluctuations of $r_{NO}$ and $\theta_{HOC}$. This is also observed in protein G and CspA (data not shown) and can be explained by the fact that the MMPT potential is stronger and more directional than a conventional superposition of Coulomb terms. Thus, the H bonds are more restricted in the conformational space, which also leads to better stability and convergence of ${}^{h3}J_{NC'}$ couplings calculated from MMPT/MD compared to standard MD simulations (Supporting Information S3).

The protein dynamics based on using MMPT as an explicit hydrogen-bond potential are also investigated by calculating the root-mean-square fluctuations (or B-factors; Figure 7) and 2D cross-correlation maps (Figure 8). Generally, using the MMPT potential leads to rigidification of the protein, which is consistent with previous efforts to better describe hydrogen bonds in proteins.[58] The cross-correlation maps of ubiquitin show that most correlated motions are caused by hydrogen-bonding structures in the protein. Cross-correlation maps computed from MMPT/MD and standard MD simulations show similar dynamical features, while the comparison indicates that the MMPT potential enhances the correlations between hydrogen-bonding residues.

## Discussion

In this work, we present a general method for deriving quantitative potential energy surfaces for H-bonding motifs and demonstrate their ability to accurately calculate scalar couplings across hydrogen bonds in proteins from atomistic simulations. Compared with standard MD simulations, RMSDs between calculated and experimental ${}^{h3}J_{NC'}$ couplings have been reduced in all six proteins investigated (Table 3). The ${}^{h3}J_{NC'}$ couplings can be calculated with an average deviation of 0.14 Hz by MMPT/MD simulation. Better agreement between calculated and experimental values are observed for all different secondary structures (Table 4), while the most significant improvements are found in loop regions. As has been noted previously,[59] current molecular mechanics force fields perform most poorly in the loop regions in proteins.

Our calculations are based on a force field treating hydrogen bonds explicitly. The MMPT potential, originally developed to investigate proton transfer reactions, has been shown to be adequate for describing hydrogen bonds in proteins by simple PES morphing techniques. This is consistent with the well-known fact that hydrogen bonds can be regarded as incipient or "frozen stage" proton transfer reactions.[1] It is possible that more sophisticated PES morphing strategies, a more realistic angular dependence (e.g., $V(R, \rho, \theta) = \Sigma_n V_n(R, \rho)P_n(\cos\,\theta)$, where $P_n$ are Legendre polynomials), or different MMPT parametrizations for H bonds in different secondary structure elements will lead to additional improvements.

The results presented here are based on an average treatment of H bonds in proteins, which means that the same MMPT potential is used for all hydrogen bonds in a certain protein. This is reflected by the fact that the PES morphing parameter $R'$ corresponds to the average D−A distance from the X-ray structures. However, hydrogen bonds in different chemical environments exhibit different strengths so describing them with environment-specific parametrizations is a possibility for improvement. In fact, this has been previously found to be the case when hydrogen bonds in different secondary structures (α helices, $\beta$ sheets, and loops) were investigated separately.[33]

On the basis of a detailed study of correlation between PES morphing parameters and RMSDs in three proteins (ubiquitin, CspA, and protein G), we propose a generic procedure whereby, starting from X-ray structures, the PES is morphed to a minimum ($R'_{NO}$, $r'_{OH}$, $\theta'_{HNO}$). Here, $R'_{NO}$ equals the average N−O distance in the X-ray structure, $r'_{OH} = 1.93$ Å and $\theta'_{HNO} = 0$. Such an approach enables us to reliably calculate ${}^{h3}J_{NC'}$ couplings, and it has been applied to a set of six proteins. Due to the nonlinear relationship between the morphing parameters, the dynamics in proteins, and the calculated RMSDs for scalar couplings, morphing parameters $\{R', \rho'\}$ may not always yield the minimal RMSD between calculated and observed couplings. For example, in CspA, the morphing parameters lead to a RMSD of 0.14 Hz, while the minima {2.96,0.16} yield 0.12 Hz. However, differences are small, and both parameter sets are significant improvements over results from standard MD simulations (0.20 Hz), given that experimental errors are usually smaller than 0.05 Hz.[19,20]

In previous work relating NMR observables and MD simulations, biased simulations with an additional restraining penalty function have been used.[51] In this approach, ${}^{h3}J_{NC'}$ couplings were taken as input information and different dynamical ensembles were generated, which enables the determination of accurate geometries and energetics of hydrogen bonds in the native states of proteins. Here, a different approach is pursued. Instead of biasing simulations, the intermolecular interactions are represented more accurately by explicitly including potentials describing H bonds. The dynamical ensemble for the two methods is comparable in that narrower distributions of hydrogen-bond lengths and more restrictions for hydrogen-bond angles are found. Because scalar coupling constants directly characterize the geometries of H bonds, it is tempting to suggest that better quantitative agreement between calculated and experimentally measured ${}^{h3}J_{NC'}$ couplings also reflects a better description of the conformational ensemble of the protein.

***Table 3.*** Comparison of RMSDs from Conventional MD Simulations and MMPT/MD Simulations with the Morphed Potentials

| | | ubiquitin | CspA | protein G | apoCAM | holoCAM | IFABP |
|---|---|---|---|---|---|---|---|
| RMSD (standard MD) | | 0.142 | 0.195 | 0.134 | 0.204 | 0.203 | 0.175 |
| morphing parameters | $R'$ | 2.921 | 2.958 | 2.942 | 2.986 | 2.937 | 2.946 |
| | $\rho'$ | 0.144 | 0.168 | 0.158 | 0.185 | 0.155 | 0.161 |
| RMSD (MD/MMPT) | | 0.116 | 0.140 | 0.134 | 0.144 | 0.142 | 0.164 |

As the results show, the procedure pursued here is generally applicable and leads to appreciable improvement for all proteins investigated, and predictions for observables can be attempted. This is, in general, not possible with biased simulations for which the bias introduced is only valid for the particular protein under investigation and is not easily transferred to a different protein.

Hydrogen-bonding dynamics between standard MD and MMPT/MD simulations have been also compared in this work. Stronger hydrogen bonding, shorter hydrogen-bond lengths, and more pronounced directionality have been observed in MMPT/MD simulations. This agrees with a statistical analysis of X-ray structures which yields $\delta_{HA} =$ 1.93 Å, which is identical to the separation found here and close to results from electronic structure calculations (1.94−1.97 Å). Furthermore, the average NHO angle from all simulations is 166°, which compares with values between 155° and 162° from electronic structure calculations, and 175° from the knowledge-based potential. Analysis of the protein dynamics shows that the MMPT potential rigidifies the entire protein and leads to stronger correlation between residues coupled by hydrogen bonds. This suggests that using explicit hydrogen-bond potentials shifts the conformational ensemble sampled in MD simulation toward the experimentally measured one.[40]

Here, we showed that an explicit, three-dimensional hydrogen-bond potential leads to—sometimes considerably—improved calculation of hydrogen bond scalar couplings from explicit atomistic simulations in full solvation for six proteins with different folds. A general computational strategy is formulated which employs the coordinates from (high-resolution) X-ray structures and leads to suitably

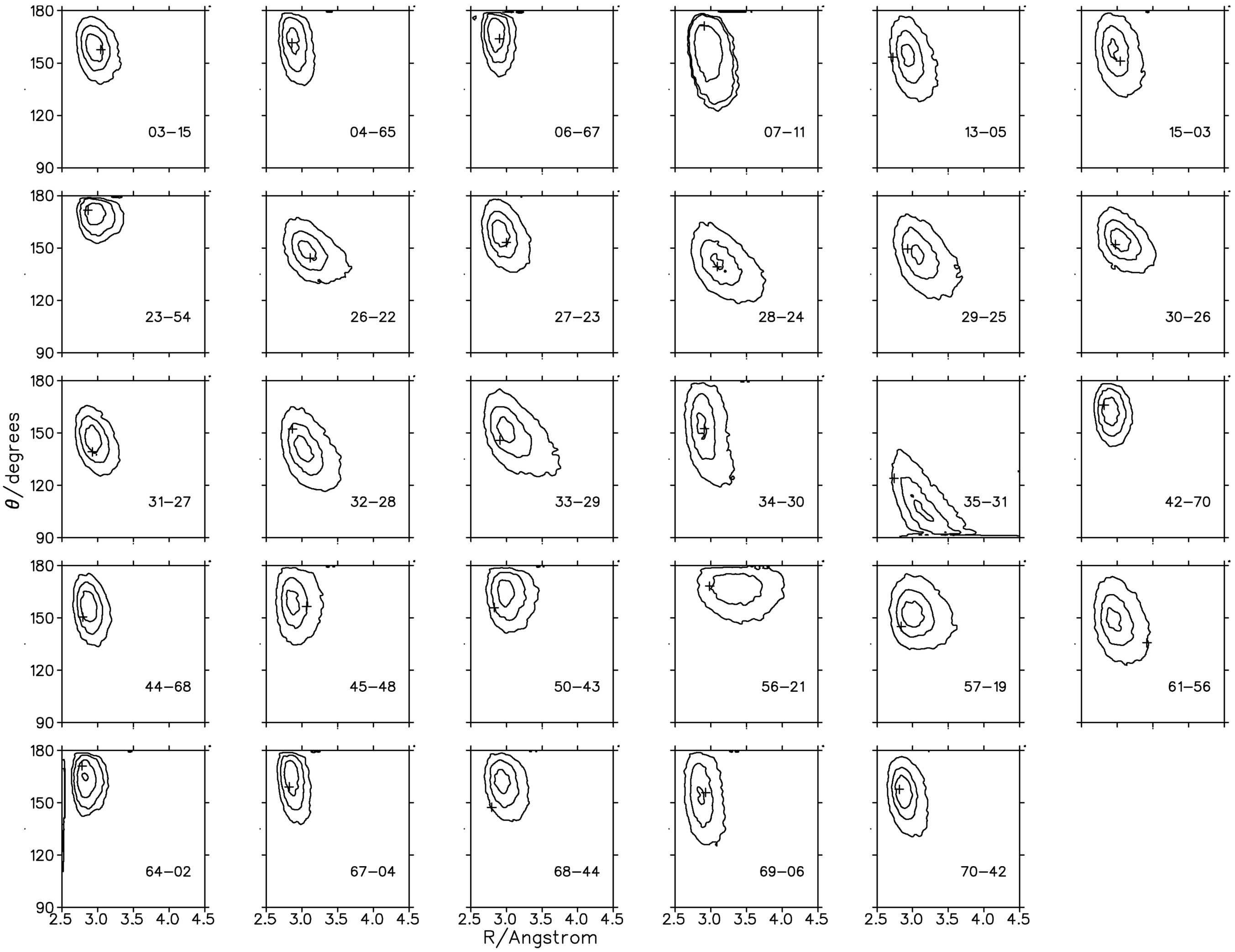


***Figure 5.*** Distributions of hydrogen-bond geometries ($r_{NO}$, $\theta_{HOC}$) populated during 500 ps standard MD simulation for 29 H bonds in ubiquitin.

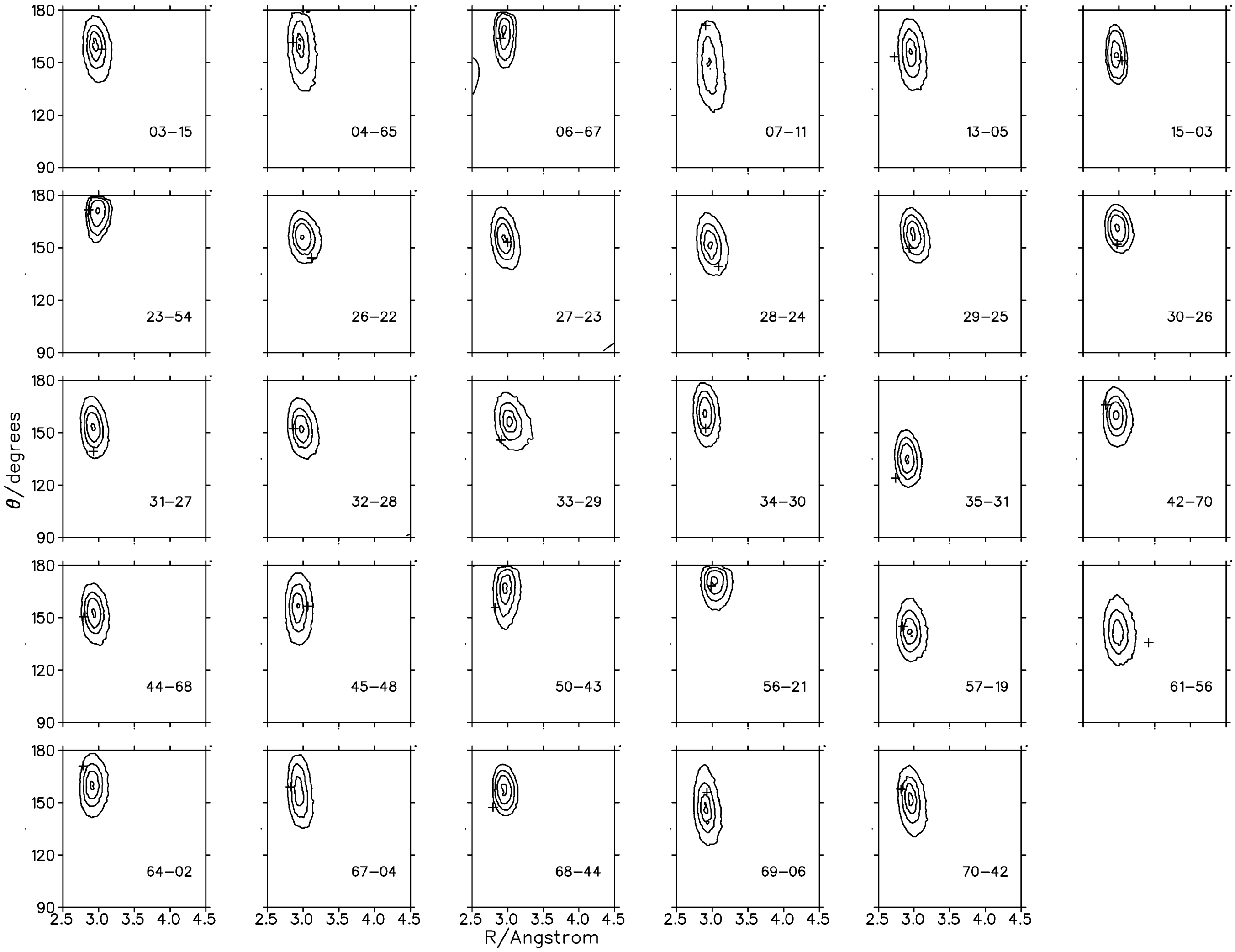


***Figure 6.*** Distributions of hydrogen-bond geometries ($r_{NO}$, $\theta_{HOC}$) populated during 500 ps MMPT/MD simulation for 29 H bonds in ubiquitin.

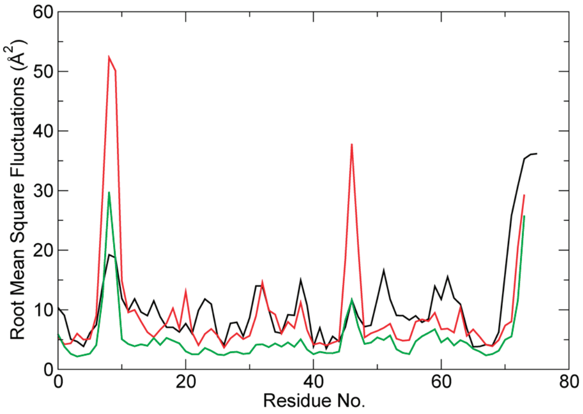


***Figure 7.*** Root mean square fluctuations of backbone atoms calculated from 1 ns standard MD (red) and MMPT/MD (green) simulations of ubiquitin. Experimental B factors are plotted on the black line.

morphed H-bonding potential that can be used to investigate the nuclear dynamics in proteins. It is further illustrated that hydrogen-bonding potentials which lead to better agreement between calculated and measured $^{h3}J_{NC'}$ couplings are those with physically meaningful (PES morphing) parameters. This opens the possibility to further improve force fields by combining NMR data and atomistic simulations, which is of particular relevance in characterizing conformational ensembles and in studies of signal transduction in proteins. Recently, a detailed analysis of the signaling pathway of rhodopsin led to the proposition that signals in proteins can be conducted through salt bridges and hydrogen bonds because they are more directional and the residues involved can act as molecular switches.[60] For such studies, which will most likely be intensified in the near future due to the fundamental interest in unraveling the means by which signaling occurs at a molecular level, accurate H-bonding potentials will be particularly important. The additional computational effort involved in using MMPT is minimal because, instead of a few harmonic potentials (conventional force field), the same number of anharmonic (Morse) terms have to be evaluated. What currently limits the standard use of MMPT is the fact that a time step of $\Delta t \approx 0.2$ fs is used to propagate the equations of motion. However, multi-time-step procedures are being considered which will largely circumvent this problem. As has been shown in a recent study on CO relaxation in myoglobin, conventional force fields based on harmonic bonded potentials which accurately describe vibrational spectra can be inappropriate when

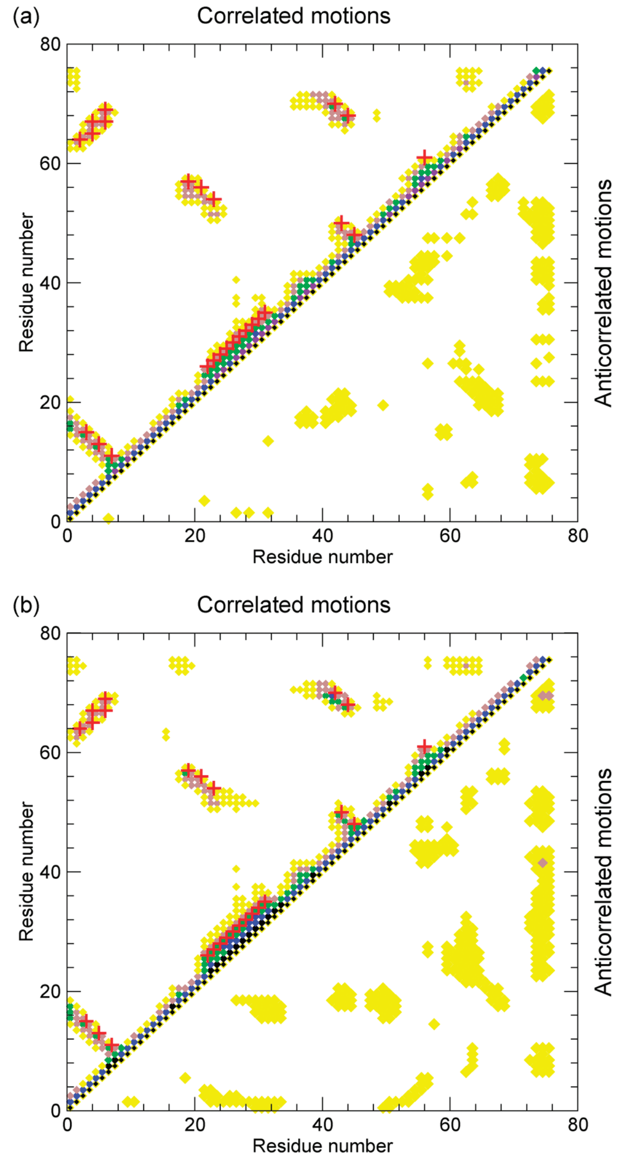


***Figure 8.*** Dynamical cross-correlation maps for ubiquitin calculated from the 1 ns (a) standard MD and (b) MMPT/MD simulations. Positive cross-correlated coefficients are collected in the upper-left triangle and negative ones in the lower-right triangle. Only cross-correlation coefficients $C_{ij}$ larger than 0.20 are shown. The intensity is represented as follows: yellow squares, $0.2 < C_{ij} < 0.35$; brown squares, $0.35 < C_{ij} < 0.5$; green squares, $0.5 < C_{ij} < 0.65$; blue squares, $0.65 < C_{ij} < 0.8$; and black squares, $0.8 < C_{ij} < 1.0$. H bonds with nonvanishing scalar couplings are marked by the red plus.

***Table 4.*** Summary of RMSDs Calculated by Standard MD Simulations and MMPT/MD Simulations for All Couplings and for Couplings in Particular Secondary Structural Elements

| | all | α helix | β sheet | loop |
|---|---|---|---|---|
| number of H bonds | 214 | 86 | 105 | 23 |
| standard | 0.174 | 0.168 | 0.169 | 0.213 |
| MMPT/MD | 0.137 | 0.128 | 0.145 | 0.128 |

considering energy transfer between vibrational modes with widely separated frequencies.[61] Thus, when energy transfer between modes is studied, details of the interaction potentials may become important. The fundamental role of H bonds, the sensitivity of ${}^{h3}J_{NC'}$ couplings to their dynamics, and the possibility to compute couplings from meaningful atomistic simulations provide an ideal stage to further develop and extend the range and applicability of simulations.

**Acknowledgment.** The authors gratefully acknowledge financial support from the Swiss National Science Foundation under grant 200021-117810.

**Supporting Information Available:** Tables of all hydrogen bonds, scalar couplings and RMSDs, a comparison of different starting structures for protein G, a comparison of MD simulation with and without CMAP correction, and figures of ${}^{h3}J_{NC'}$ convergence. This information is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Steiner, T. *Angew. Chem., Int. Ed.* **2002**, *41*, 48–76.

(2) Dill, K. A. *Biochemistry* **1990**, *29*, 7133–7155.

(3) Deechongkit, S.; Nguyen, H.; Powers, E. T.; Dawson, P. E.; Gruebele, M.; Kelly, J. W. *Nature* **2004**, *430*, 101–105.

(4) Wang, M.; Wales, T. E.; Fitzgerald, M. C. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 2600–2604.

(5) Gray, M.; Cuello, A. O.; Cooke, G.; Rotello, V. M. *J. Am. Chem. Soc.* **2003**, *125*, 7882–7888.

(6) Aruksankunwong, O.; Hannongbua, S.; Wolschann, P. *J. Mol. Struct.* **2006**, *790*, 174–182.

(7) Taylor, M. S.; Jacobsen, E. N. *Angew. Chem., Int. Ed.* **2006**, *45*, 1520–1543.

(8) Huang, Y.; Unni, A. K.; Thadani, A. N.; Rawal, V. H. *Nature* **2003**, *424*, 146.

(9) MacKerell, A. D.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(10) Ponder, J.; Case, D. *Adv. Protein Chem.* **2003**, *66*, 27–85.

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

(12) Nizkorodov, S. A.; Dopfer, O.; Ruchti, T.; Meuwly, M.; Maier, J. P.; Bieske, E. J. *J. Phys. Chem.* **1995**, *99*, 17118–17129.

(13) Meuwly, M.; Nizkorodov, S. A.; Maier, J. P.; Bieske, E. J. *J. Chem. Phys.* **1995**, *104*, 3876–3885.

(14) Nizkorodov, S.; Dopfer, O.; Meuwly, M.; Bieske, E.; Maier, J. *J. Chem. Phys.* **1996**, *105*, 1770–1777.

(15) Watson, J. K. G. *Mol. Phys.* **1968**, *15*, 479–490.

(16) Nilges, M.; Bernard, A.; Bardiaux, B.; Malliavin, T.; Habeck, M.; Rieping, W. *Structure* **2008**, *16*, 1305–1312.

(17) Dingley, A. J.; Grzesiek, S. *J. Am. Chem. Soc.* **1998**, *120*, 8293–8297.

(18) Cornilescu, G.; Ramirez, B. E.; Frank, M. K.; Clore, G. M.; Gronenborn, A. M.; Bax, A. *J. Am. Chem. Soc.* **1999**, *121*, 6275–6279.

(19) Grzesieka, S.; Cordiera, F.; Jaravinea, V.; Barfield, M. *Prog. Nucl. Magn. Reson. Spectrosc.* **2004**, *45*, 275–300.

(20) Alkorta, I.; Elguero, J.; Denisov, G. S. *Magn. Reson. Chem.* **2008**, *46*, 599–624.

(21) Eberstadt, M.; Mierke, D. F.; Kock, M.; Kessler, H. *Helv. Chim. Acta* **1992**, *75*, 2583–2592.

(22) Dingley, A. J.; Masse, J. E.; Feigon, J.; Grzesiek, S. *J. Biomol. NMR* **2000**, *16*, 279–289.

(23) Cordier, F.; Grzesiek, S. *J. Am. Chem. Soc.* **1999**, *121*, 1601–1602.

(24) Ahn, H.-C.; Juranic, N.; Macura, S.; Markley, J. L. *J. Am. Chem. Soc.* **2006**, *128*, 4398–4404.

(25) Markwick, P. R. L.; Sprangers, R.; Sattler, M. *J. Am. Chem. Soc.* **2003**, *125*, 644–645.

(26) Alexandrescu, A. T.; Snyder, D. R.; Abildgaard, F. *Protein Sci.* **2001**, *10*, 1856–1868.

(27) Juranic, N.; Atanasova, E.; Streiff, J. H.; Macura, S.; Prendergast, F. G. *Protein Sci.* **2007**, *16*, 1329–1337.

(28) Juranic, N.; Moncrieffe, M. C.; Liki, V. A.; Prendergast, F. G.; Macura, S. *J. Am. Chem. Soc.* **2002**, *124*, 14221–14226.

(29) Eberstadt, M.; Gemmecker, G.; Mierke, D. F.; Kessler, H. *Angew. Chem., Int. Ed.* **1995**, *34*, 1671–1695.

(30) Bouvignies, G.; Bernado, P.; Meier, S.; Cho, K.; Grzesiek, S.; Bruschweiler, R.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13885–13890.

(31) Grzesiek, S.; Sass, H.-J. *Curr. Opin. Struct. Biol.* **2009**, *19*, 585–595.

(32) Sass, H.-J.; Schmid, F. F.-F.; Grzesiek, S. *J. Am. Chem. Soc.* **2007**, *129*, 5898–5903.

(33) Schmid, F. F.-F.; Meuwly, M. *J. Chem. Theory Comput.* **2008**, *4*, 1949–1958.

(34) Barfield, M. *J. Am. Chem. Soc.* **2002**, *124*, 4158–4168.

(35) Lammers, S.; Meuwly, M. *J. Phys. Chem. A* **2007**, *111*, 1638–1647.

(36) Lammers, S.; Lutz, S.; Meuwly, M. *J. Comput. Chem.* **2008**, *29*, 1048–1063.

(37) Bowman, J. M.; Gazdy, B. *J. Chem. Phys.* **1991**, *94*, 816–817.

(38) Meuwly, M.; Hutson, J. M. *J. Chem. Phys.* **1999**, *110*, 8338–8347.

(39) Kortemme, T.; Morozov, A. V.; Baker, D. *J. Mol. Biol.* **2003**, *326*, 1239–1259.

(40) Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6946–6951.

(41) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(42) 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.

(43) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1987**, *194*, 531–544.

(44) (a) Schmidt, H. L. F.; Sperling, L. J.; Gao, Y. G.; Wylie, B. J.; Boettcher, J. M.; Wilson, S. R.; Rienstra, C. M. *J. Phys. Chem. B* **2007**, *111*, 14362–14369. (b) 1pga and 2igd are also used as initial structures, and the results are compared and discussed in the Supporting Information S1.

(45) Schindelin, H.; Jiang, W.; Inouye, M.; Heinemann, U. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 5119–5123.

(46) Schumacher, M.; Crum, M.; Miller, M. *Structure* **2004**, *12*, 849–860.

(47) Chattopadhyaya, R.; Meador, W. E.; Means, A. R.; Quiocho, F. A. *J. Mol. Biol.* **1992**, *228*, 1177–1192.

(48) Scapin, G.; Gordon, J.; Sacchettini, J. *J. Biol. Chem.* **1992**, *267*, 4253–4269.

(49) Brunger, A. T.; Karplus, M. *Proteins* **1988**, *4*, 148–156.

(50) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comp. Phys.* **1977**, *23*, 327–341.

(51) Gsponer, J.; Hopearuoho, H.; Cavalli, C. M.; Dobson, A.; Vendruscolo, M. *J. Am. Chem. Soc.* **2006**, *128*, 15127–15135.

(52) Raynes, W. T.; Geertsen, J.; Oddershede, J. *Int. J. Quantum Chem.* **1994**, *52*, 153–163.

(53) Malkin, V. G.; Malkina, O. L.; Salahub, D. R. *Chem. Phys. Lett.* **1994**, *221*, 91–99.

(54) Alexander, D.; Mackerell, J.; Feig, M.; Charles, L.; Brooks, I. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

(55) Buck, M.; Bouguet-Bonnet, S.; Pastor, R. W.; Alexander, D.; MacKerell, J. *Biophys. J.* **2006**, *90*, L36–L38.

(56) Bastug, T.; Kuyucak, S. *Biophys. J.* **2009**, *96*, 4006–4012.

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

(58) Ji, C.; Mei, Y.; Zhang, J. Z. H. *Biophys. J.* **2008**, *95*, 1080–1088.

(59) Case, D. A. *Acc. Chem. Res.* **2002**, *35*, 325–331.

(60) Kong, Y.; Karplus, M. *Structure* **2007**, *15*, 611–623.

(61) Devereux, M.; Meuwly, M. *J. Phys. Chem. B* **2009**, *113*, 13061–13070.

CT9005695

JCTC Journal of Chemical Theory and Computation

# Performance of the Empirical Dispersion Corrections to Density Functional Theory: Thermodynamics of Hydrocarbon Isomerizations and Olefin Monomer Insertion Reactions

Grigory A. Shamov,*[†,‡] Peter H. M. Budzelaar,[†] and Georg Schreckenbach*[†]

*Department of Chemistry, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada, and Dutch Polymer Institute (DPI), P.O. Box 902, 5600 AX Eindhoven, The Netherlands*

Received September 28, 2009

**Abstract:** Most of the commonly used approximate density functionals have systematic errors in the description of the stability of hydrocarbons. This poses a challenge for the realistic modeling of reactions involving hydrocarbons, such as olefin polymerization. Practical remedies have been proposed, including the application to usual black-box DFT of additional empirical correction $CR^{-6}$ terms for the van der Waals interaction (termed DFT-D), or introducing additional pseudopotentials that introduce some medium-to-long-range attraction (C-Pot). In this Article, we use the DFT-D scheme as realized in our BOptimize package to evaluate the performance of a range of commonly used DFT functionals (combinations of xPBE, B88, OPTX with LYP and cPBE GGAs and hybrids) for the modeling of the thermodynamics of reactions of the growth of common polyolefins. We also review and reproduce some of the previously done benchmarks in the area: alkane branching and relative stability of $C_{12}H_{12}$ and $C_{10}H_{16}$ isomers. In addition to the common DFT methods, computations with correlated wave function methods (MP2) and the new functionals B97-D and M06-L were performed. The performance of the special density functionals B97-D and M06-L is, in general, similar to the best DFT-D corrected regular functionals (BPBE-D and PBE-D). The results show that (1) the DFT-D correction is sufficient to describe alkane branching, but its performance depends on the parametrization; (2) inclusion of the correction is essential for a proper description of the thermodynamics of reactions of polymer growth; and (3) not all approximate density functionals perform effectively for the description of hydrocarbons even with the correction. The C-Pot method for the B3LYP functional shows quantitatively correct results for our test cases. The enthalpies of hydrocarbon reactions were analyzed in terms of the repulsion characteristics of a given DFT method. PBE is the least repulsive, while OLYP is the most. However, there are cases where the failure of a DFT method cannot be correlated with its repulsive character. A striking example is the performance of B3LYP and BLYP for caged molecules with small carbocycles, such as the [$D_{3d}$]-octahedrane. The stability of [$D_{3d}$]-octahedrane is underestimated by the B3LYP, BLYP, and B97-D functionals, but not by DFT methods that contain either B88 exchange or LYP correlation functionals separately. While DFT-D cannot amend the performance of the former functionals for the octahedrane, C-Pot for B3LYP does.

## Introduction

Density functional theory (DFT) methods are the most frequently used tools in today's theoretical chemists' inventory. After the introduction of the generalized gradient approximation (GGA),[1,2] and, later, the advent of hybrid density functionals,[3] it became a black-box computational method for the majority of chemists.

However, recently some flaws in the commonly used black-box density functional methods were exposed. These failures came as a surprise in part because the methods have

* Corresponding author e-mail: gas5x@yahoo.com (G.A.S.); schrecke@cc.umanitoba.ca (G.S.).

[†] University of Manitoba.

[‡] Dutch Polymer Institute.

10.1021/ct9005135 © 2010 American Chemical Society
Published on Web 01/22/2010

been thoroughly tested on sets of small molecules like those included in the G1, G2[4,5] benchmark sets. In recent years, progress of both computer hardware and DFT software (such as resolution of identity fitting techniques[6−8]) allowed for the routine treatment of systems containing up to a few hundred atoms. It turned out that density functionals (DFs) that did well for the G1 set, most notably B3LYP, can systematically fail to describe larger systems. By now, extensive literature exists on benchmarking of the performance of different DFs for a range of problems.[9−23]

One of the problematic areas for DFT is the description of the stability of hydrocarbons.[24] Schleyer and co-workers did several tests comparing isomeric polycyclic hydrocarbons.[14] Schreiner ascribed errors in calculations of isodesmic reactions to the failure of DFT to describe "protobranching" (1,3 attractive interactions) for linear hydrocarbons.[16,17] Grimme studied alkane branching as a testing ground for his methods.[12,25] These studies show that the majority of modern DFs has problems with the description of both intermolecular and intramolecular interactions in hydrocarbons. An example of the former problem is the failure of most GGA and hybrid functionals to predict the structure of the benzene shifted stacking dimer. Intramolecular issues appear when comparing branched and linear alkanes (*n*-octane vs 2,2,3,3-tetramethyl-butane). Often the simple local density approximation functional performs better than either GGAs or hybrids, giving, unlike the latter, qualitatively correct results.[10] The variability in the performance of different DF methods even led to the proposal of multilayer QM/QM computations employing several different DFs for different parts of the model system,[26] picking a functional that performs best for each particular type of interaction within the system.

A few explanations of these deficiencies and remedies for them have been proposed. On the basis of the analysis of MP2 pair correlation energy in branched hydrocarbons, Grimme coined the term "medium-range correlation energy" for the interaction energy that is not described well with GGA.[12] Yang et al.[27] traced problems of common DFs to the delocalization errors. They also noted, comparing potential energy curves of interaction between two methane molecules, that most GGAs (especially with the Becke88[1] exchange) are too repulsive as compared to CCSD(T) at shorter intermolecular distances. They suggested that for modeling of the formation of highly branched or polycyclic molecules, functionals with minimal delocalization errors such as PBE1 should be chosen.

Grimme, in his 2004 work,[28] proposed a simple empirical $CR^{-6}$ correction term (with coefficients obtained from ab initio calculations) for the long-range van der Waals (vdW) interactions, together with a damping function that turns it off at shorter distances. He applied this correction to several hydrocarbon test cases and stated that the correction alone is not enough to describe energy differences for the *n*-octane branching for GGA functionals such as PBE and BLYP. One of Grimme's proposals was to use his double-hybrid functionals containing some MP2 corrections (see ref 13 for leading references); together with the dispersion corrections (but not without them) they describe alkane branching very successfully. Later, however, he revised his dispersion correction approach.[29] Following him, we call this revised approach DFT-D hereafter. In the 2006 paper, Grimme also introduced a new parametrization of the power-series B97-type[30] density functional,[29] B97-D that, together with the dispersion correction, gave the correct sign for the alkane branching. However, simple DFs were not thoroughly tested at that time with the new DFT-D parametrization (or at least this was not reported in the paper). The widely cited (for example, see refs 16 and 31) result, based on the old DFT-D parametrization, that BLYP-D and PBE-D are insufficient for the description of nonbonded interactions in alkanes, does not necessarily hold with the new DFT-D parameter set.

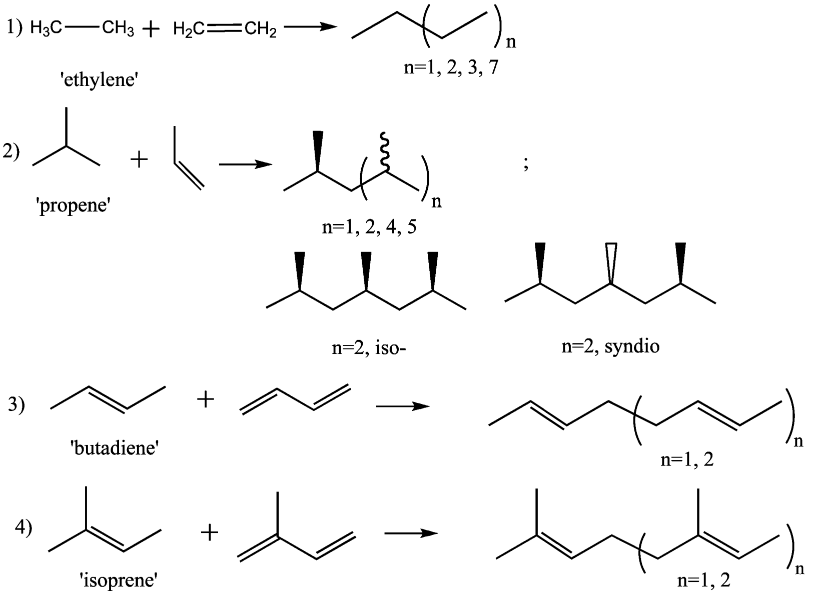
It soon became obvious that the character of the damping function is at least as important as the choice of the dispersion coefficients, for it acts in the crucial medium-range region. Head-Gordon proposed to use a power-twelve term instead of an exponential in the denominator of the damping function, otherwise following Grimme's approach.[29] Ducere and Cavallo introduced a parametrization of the damping function for intermolecular interactions of DNA base pairs.[32] Recently, Cornimonbeauf and co-workers reparametrized the DFT-D damping function of the latter work to accurately describe both intra- and intermolecular molecular interactions in alkanes.[33]

An alternative to the addition of an empirical correction is modification of core−core interactions within the DFT method itself. One of the computationally cheapest and most readily available methods for it is to modify (or add) a pseudopotential that would introduce some medium and long-distance attractive interactions to the energy, thus emulating the dispersion. The pseudopotential is usually parametrized to reproduce structures of a training set of intermolecular complexes. First introduced for the solid state computational chemistry for plane-wave computations by Röthlisberger et al.,[34,35] it was extended by DiLabio to molecular Gaussian basis set computations, and named the C-Pot method (because they only used a pseudopotential on carbon).[36] The C-Pot method was shown to help to describe intermolecular interactions of some condensed aromatics dimers such as coronene and graphene sheets. We note that the pseudopotentials are both basis set and density-functional dependent and have to be parametrized by molecular calculations. Thus, they are "less ab initio" than Grimme's DFT-D method that is based on atomic parameters and has only the global scaling factor for each density functional to be fitted.

While the ad-hoc $CR^{-6}$ dispersion correction of DFT-D can (as we will show later) indeed be very beneficial for energies of hydrocarbons, it can be problematic in cases where there can be strong changes in atomic oxidation states and hybridization. A more sophisticated method for the representation of the dispersion corrections to DFT (see refs 21, 37 for examples and leading references) might be free from these shortcomings. However, at present such methods are not widely available; also usually they rely on fitting to molecular data, which are not always available.

Modeling of organometallic catalysis (which was one of the first and the most remarkable successes of DFT) with a realistic model system should include adequate representation

***Scheme 1.*** Polymer Growth (Monomer Insertion) Reactions

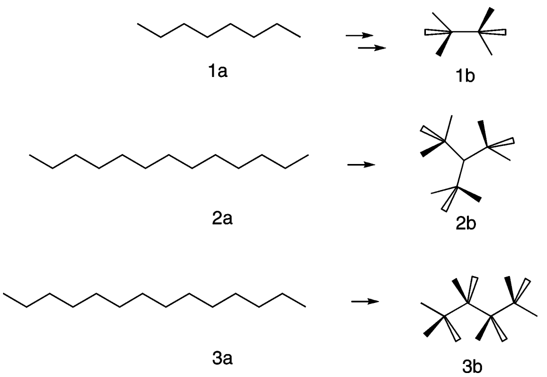


of the catalysts' ligands, weak complexes of reactants with the catalyst, and, for the case of polymerizations, accurate modeling of the growing chain. For example, we have shown in the case of an expanded porphyrin complex[38] that the inclusion of peripheral alkyl substituents in the model complex is important for the description of the complex's electronic and geometric structures. Many of the popular organometallic ligands (a good example is penta-methyl-cyclopentadienyl) are hydrocarbons themselves and/or have multiple hydrocarbon peripheral substituents that might interact with each other, with other ligands, as well as with reactants during the course of reactions. Thus, if we want to treat realistic model systems with DFT (which is usually the only affordable choice), we need to assess the accuracy of the method.

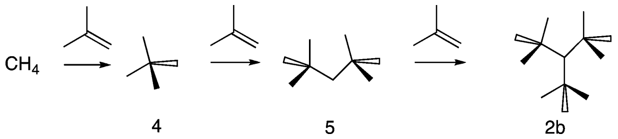
In the present study, we apply Grimme's dispersion correction together with selected commonly used density functionals to larger hydrocarbon systems, considering the accuracy of predictions of the thermodynamics of the olefin polymerization. In addition, within the same approach, we study other model systems from the literature, such as alkane branching and polycyclic hydrocarbon isomerizations. We compare our results to computational data from the literature, by reproducing selected cases from refs 12, 14, 17, 28, 29, and to experimental results. Where the latter are unavailable, we will use ab initio correlated methods (MP2). The systems under study are shown in Schemes 1−5. Unlike several previous studies dedicated to DFT performance, we aim to model not just isodesmic or homodesmotic reactions (which can indeed be used for obtaining valuable thermodynamical information as well as insights into method performance; see, for example, the recent work of ref 39), but systems of more practical, chemical relevance.

Our primary aim is to assess whether a black-box density functional method using a common, readily available density functional together with the simple DFT-D dispersion corrections is reliable for the description of larger hydrocarbon systems including systems with different degrees of sterical strain. Other options, besides black-box DFs and Grimme's DFT-D, were used for comparison. Specifically, we have tried the newly developed "dispersion-aware" density functionals B97-D and M06-L. The C-Pot method by DiLabio was also investigated. Others have already studied some of the test systems we use

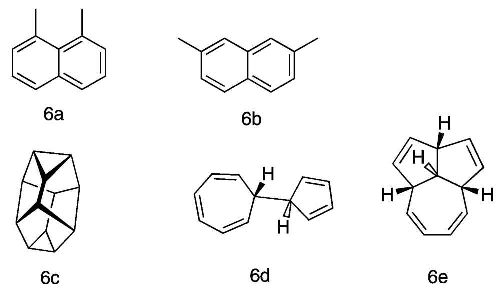
***Scheme 2.*** Alkane Branching Reactions



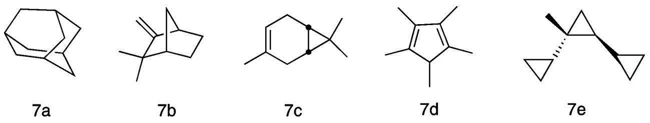
***Scheme 3.*** Addition of *t*-Bu Groups to Methane



***Scheme 4.*** Isomers of $C_{12}H_{12}$ Hydrocarbons



***Scheme 5.*** Isomers of $C_{10}H_{16}$ Hydrocarbons



individually, but as far as we know they have not been systematically considered together.

In general, our results show that common GGA and hybrid density functionals yield rather poor results for olefin polymerization energies. Properties like the energies of alkane branching are similarly predicted incorrectly. The performance of a particular DF for olefin polymerization or alkane branching can be traced to its repulsive character; the "harder" the functional is, the larger is the error. In these cases, the functionals can be improved by applying the DFT-D correction that fixes over-repulsiveness or under-attractiveness. Generally, the DFT-D correction is also beneficial for the description of other hydrocarbon isomerizations. The alternative pseudopotential modification method C-Pot is shown to yield improved results for B3LYP.

## Computational Methods

In the DFT-D approach, the attractive $C_nR^{-n}$ correction terms for the "dispersion interaction" are simply added to the total DFT energy. To avoid undesirable interference of the correction terms with the DFT energy at short distances, a damping function is used. The correction by Grimme[29] we use in the present work uses only $C_6R^{-6}$ terms and has the form of eqs 1 and 2:

$$E_{\text{total}} = E_{\text{DFT}} + E_{\text{disp}} \tag{1}$$

$$E_{\text{disp}} = -s \sum_{i=1,N}^{N} \sum_{j=1,i-1}^{N} F_{\text{damp}}(R_{ij}) C_6^{ij} R_{ij}^{-6} \tag{2}$$

The global scaling factor $s$ is optimized for each functional. In Grimme's 2006 method,[29] the atomic van der Waals coefficients $C_6{}^i$ are obtained from atomic properties as $C_6{}^i = 0.05NI\alpha$. Here, $I$ is the ionization potential, $\alpha$ is the static polarizability of the atom, and $N$ is the number of electrons in the next noble gas in the elements' period. Coefficients $C_6{}^{ij}$ were then set to the geometric mean of $C_6{}^i$ and $C_6{}^j$. The damping function used is

$$F_{\text{damp}}(R_{ij}) = \frac{1}{1 + e^{-a(R_{ij}/R_0^{ij} - 1)}} \tag{3}$$

with $a = 20$; atomic radii $R_0{}^{ij}$ are based on DFT computations for elements up to Xe with a scaling factor of 1.1 applied. Thanks to the universality and ease of use of this approach, the DFT-D correction in its Grimme-2006 form has been implemented in recent releases of many popular quantum-chemistry codes, such as, to name a few, ADF, GAMESS-US, and Gaussian 09.

Some exchange functionals are known to be less repulsive than others, and thus the DFT-D correction might overcompensate for the dispersion if used together with "too attractive" functionals. Grimme in 2006[29] and Head-Gordon in 2008[40] developed their parametrizations of the Becke-1997[30] power-series GGA specifically parametrized including DFT-D corrections, thus avoiding the dispersion double-counting that is otherwise handled by the scaling factor $s$ only. Head-Gordon also used an inverse power-twelve damping function instead of Grimme's exponential one (see the Supporting Information). The choices of parametrization of the global scaling coefficient $s$ versus scaling of the $R_0$ atomic parameters within the Grimme-2006 scheme were systematically explored in the recent work of Baldridge.[41]

Because the form of the correction does not depend on the underlying quantum-chemical method, it is straightforward to code it as an external routine, thus avoiding possible limitations of any given QM program package. We have implemented the DFT-D corrections for energies, gradients, and second derivatives of the energy in the stand-alone optimizer code BOptimize.[42] The code is now interfaced to many DFT and ab initio packages. Most of the results presented here were obtained with it.

While there are many approximate density functionals, the emphasis of this work is on practically available and computationally efficient solutions that are or have the potential to be widely used. For that reason, we chose to test the local functional VWN5[43] (with Slater exchange), and the GGA functionals BLYP,[1,2] PBE,[44] and its modifications by Adamo and Barone MPBE,[45] BPBE,[1,44] and OLYP.[2,46] All of these are local GGA functionals, and as such allow for the most efficient use of Coulomb and exchange fitting techniques. For comparison, the two popular hybrid functionals B3LYP[3] and PBE1[47] were also included; to specifically test the effect of increasing the amount of exact exchange in the latter, the HFPBE combination with PBE correlation and 100% exact exchange was also tried in selected cases. Calculations using the functionals listed above were performed with the Priroda code version 6 with nonrelativistic all-electron general-contracted Gaussian basis sets.[8,48,49] For DFT calculations with Priroda, the L11 basis set ((6s,2p)/[2s,1p] for H, (10s7p3d)/[4s3p1d] for C, roughly corresponds to the popular cc-CVDZ basis) was used along with the corresponding auxiliary fitting set. RI-MP2 calculations were performed in the larger L2 set ((8s4p2d)/[3s2p1d] for H, (12s8p4d2f)/[4s3p2d1f] for C, roughly corresponds to the popular cc-pVTZ basis).

Geometries were fully optimized at the DFT or MP2 levels of theory. For GGA DFTs, analytical second derivatives were computed. All of the GGA optimized structures had no negative Hessian eigenvalues; because hybrid DFT and MP2 optimizations were started from them, we expect them to be minima as well.

The DFT-D dispersion corrections were added by the external optimizer BOptimize with scaling factors as recommended by Grimme: 0.75 for PBE, 1.25 for BLYP, and 1.05 for B3LYP. For other functionals where Grimme's recommendations were not available, we used scaling factors as follows: 1.0 for MPBE, 1.25 for OLYP (like for BLYP), and 1.05 for BPBE (recommended for BP86, which is similar to BPBE). Finally, for the PBE1 hybrid functional, we used 0.75, the same scaling factor as for the pure GGA PBE. We have tried both single-point DFT-D calculations on geometries optimized with the same density functional, and reoptimization of geometries with the corrections applied. Because the latter did not bring significant changes in relative energies (a fraction of a kcal/mol only), we did not pursue reoptimizations any further, and all of the DFT-D energies were computed as single-point energy calculations.

The different methods that we employed (pure GGAs, hybrid DFTs, and MP2) produce frequencies of different quality. To obtain the thermal corrections required to calculate enthalpies, one usually scales down frequencies obtained for the hybrid DFs or MP2, while pure GGAs are known to systematically give lower frequencies than other methods. In the present study, we did not attempt to be very accurate in determining these enthalpy corrections, mainly because obtaining numerical second derivatives for hybrid DFs and MP2 is time-consuming with the codes we have used (Priroda and GAMESS-US). For most of the processes under scrutiny (isomerizations and polymer growth reactions), the effects of the thermal corrections on reaction enthalpies are likely to cancel out. Thus, for all MP2 and hybrid GGA calculations, we took enthalpy corrections from the corresponding PBE results. For DFT-D results, for GGA we took the corrections from the corresponding GGA DFT run.

For comparison, two other density functionals were included in our set of methods: the B97-D[29] functional by Grimme that is specially parametrized for use with DFT-D corrections, and the highly parametrized meta-GGA M06-L functional by Zhao and Truhlar.[50] For these functionals, we have used single-point calculations with the GAMESS-US code,[51] version Feb. 2009, on the RIMP2/L2 optimized geometries from the Priroda code. The L11 basis set with the same exponents and coefficients as those used in Priroda, but imported in the segmented contracted form native to GAMESS, was used for B97-D and M06-L computations (we note that due to differences in contraction scheme, total energies computed by GAMESS-US cannot be directly compared against Priroda results; the former have systematically lower total energies). Enthalpy corrections were always taken from the Priroda PBE/L11 results.

Finally, we have explored the performance of the C-Pot method by DiLabio, taking parameters for the carbon pseudopotentials from his work.[36] We have applied the PBE1 and B3LYP density functionals with the 6-31G+(d,p) basis set in single-point calculations on the Priroda MP2/L2 optimized geometries, using the Gaussian 03 program package.[52] Again, enthalpy corrections were taken from the Priroda PBE/L11 results.

Experimental heats of formations of hydrocarbons for calculation of the "experimental" reaction enthalpies used in this work were obtained from the NIST database.[53]

## Results and Discussion

**Polymerization Reactions.** The quality of the modeling of normal alkanes has been the subject of numerous previous studies, some of which were concerned with iso- and homodesmotic reactions of their formation.[14,23,33,39,54,55] Schleyer introduced the term "proto-branching energy" for one of the sources of error in these calculations.[55] It refers to attractive 1,3 interactions between methylene groups that show up when one considers the formation of the normal hydrocarbon from C1 and C2 fragments. Because polymerization is similar in this respect, one large molecule assembles from the smaller units, such that new 1,3 and 1,4 interactions arise, dispersion interactions might be of importance for it as well.

We chose to consider enthalpies of addition (i.e., formal insertion of an alkene monomer into a terminal C−H bond of corresponding substrate) for four common alkene and alkadiene polymerization reactions. The reactions are shown in Scheme 1. There is only a limited number of experimental heats of formation for larger alkanes and alkenes available. Indeed, the determination of heats of formation for these species is a goal of several computational approaches, from Bensons' increment-based calculations[56] to semiempirical tight-binding methods[22,57] and parametrized group-equivalent approaches.[23] It is interesting in itself to see whether DFT-D methods are a viable alternative. Recent work[58] has shown that for a large number of hydrocarbons and small organic molecules the B3LYP method does not have a significant edge over the semiempirical PDDG/PM3 NDO method; tight-binding methods[22,57] also have shown reasonably good performance.

A direct comparison with the experiment is possible only in the case of ethene oligomers (for $n = 1$, 2, 3, and 6) and for the addition of a single propene into isobutane ($n = 1$, "syndio-"; this is the most stable conformer).[59] Therefore, we will compare our DFT and DFT-D calculations mainly to the MP2 results. We note that the MP2 method, especially in larger basis sets, is known to overestimate the dispersion energy. Parameterizations of the Moeller−Plesset method, such as scaled opposite spin (SCS)-MP2[60] and SCS-MP3,[61] MP2.5 (with scaled MP3 correlation energy contribution),[62] have been suggested to show better performance. However, in the present study, we use the nonparameterized, vanilla MP2 method (with its known, systematic errors) and compare the MP2 results with experimental data where available.

While comparing olefin monomer insertion energies computed with different functionals, we expect that, at least within the GGA family, we will have a similar quality of computed "intrinsic" energy differences, that is, of breaking of a monomer's C−C $\pi$-bond and formation of new C−C and C−H bonds in the corresponding insertion product. Thus, the differences in energies might be related to the quality of modeling of intramolecular nonbonding interactions. This assumption should be justified for the reactions in Scheme 1 (and Scheme 3 as well) because the molecules involved are not highly polar, nor do they possess a high degree of conjugation; as such, they are not problematic cases for GGA DFT.

The computed enthalpies are relegated to the Supporting Information, Tables S1 and S2. For the ethene and propene systems, the insertion enthalpies per monomer are similar (differ by a fraction of kcal/mol) for up to $n = 6$ (ethene) or 5 (propene) for all computational methods studied. Per-monomer enthalpies for the butadiene and isoprene systems are also similar for $n = 1$ and 2. Thus, to present our results

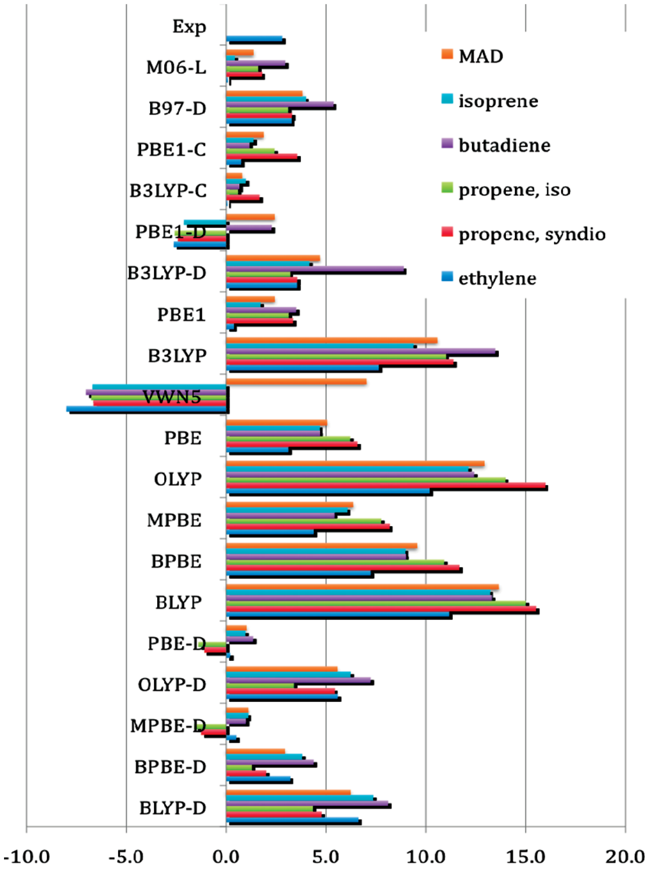


***Figure 1.*** Differences between enthalpies of monomer insertion reactions (averaged per monomer; see the text) computed by density functionals and MP2, kcal/mol. For the ethene system, the experimental enthalpy difference with respect to the MP2 result is shown as well. MAD denotes the mean average deviation from MP2 over all polymer systems.

in a more compact form, we have averaged the per-monomer insertion energies for all calculated olefin insertion systems. The differences between DFT and MP2 computed enthalpies, averaged per monomer, are shown as a bar diagram in Figure 1. The mean average deviation from MP2 for different functionals over the polymer species studied is also shown (bars labeled as MAD). For the ethene system, the experimental difference from MP2 is plotted in Figure 1 as well (labeled exp.). MP2 overestimates the ethene monomer insertion enthalpy by −2.8 kcal/mol. The best agreement with experiment, within 1 kcal/mol, is shown by the BPBE-D, BLYP-D, unmodified PBE, and B97-D functionals. The worst performance is obtained with BLYP, OLYP (strong underestimation of the enthalpy), and with VWN5 (strong overestimation of it). Most of the unmodified density functionals underestimate the insertion enthalpy. MPBE-D, PBE-D, M06-L, unmodified PBE1, and two C-Pot-modified functionals, PBE1-C and B3LYP-C, yield ethene insertion enthalpies close to the MP2/L2 values.

There are no experimental heats of formation available for oligomers of our other model systems, propene, butadiene, and isoprene. Hence, we used the computed MP2/L2 results as "the truth" in these cases. One has to keep in mind, however, that the ethene MP2 results above show a tendency for overbinding, and thus the "best" results for monomer insertion enthalpies should probably be the ones that underbind somewhat relative to MP2.

The isoprene and propene oligomers introduce branching, as compared to the polyethene. The influence of branching will also be studied in the section on alkane branching that follows, for test systems for which comparison with experiment is possible. Considering the polymerization model systems individually, we see that for propylene (both syndio- and isotactic) deviations from MP2 for uncorrected GGAs are higher than for ethene; the effect of the DFT-D correction was the largest for the propene system. Ethene and butadiene insertion that does not lead to a branched hydrocarbon product, and isoprene insertion in which the distance between the methyl side groups in the product is large, show somewhat lower sensitivity.

The VWN5 density functional systematically overestimates the monomer insertion enthalpies. All of the common GGA and hybrid density functionals underestimate the enthalpies, with errors particularly large in the cases of OLYP, BLYP, and B3LYP. Among the uncorrected DFT functionals, the hybrid PBE1 performs best. One can see that inclusion of the DFT-D correction to GGA DFT always decreases the errors. However, for OLYP-D and BLYP-D, they remain rather large. PBE-D and MPBE-D give results closest to the MP2 values. BPBE-D is slightly underbinding as compared to MP2, but, keeping in mind that MP2 itself is likely to somewhat overestimate the insertion enthalpies, we could say that this functional is probably one of the best. The corrected B3LYP-D functional still underestimates insertion exothermicities; for the PBE1-D, they are systematically overestimated, so that the MAD is similar to the uncorrected PBE1. Using the C-Pot pseudopotential method for B3LYP-C yields results very close to those of MP2/L2, with PBE1-C also having a small but slightly higher difference.

**Alkane Branching and Insertion of Crowded Alkanes.** The inability of most commonly used GGA and hybrid density functionals to accurately describe the relative stabilization of branched alkanes with respect to their linear isomers has received much attention.[12,17,29] For instance, experimentally the normal octane **1a** is less stable than 2,2,3,3-tetra-methyl-butane **1b** (Scheme 2), while GGA and hybrid density functionals predict the opposite. In the original DFT-D paper, Grimme[28] concluded that the dispersion correction (in its original parametrization) alone is not sufficient to result in the correct order of isomers **1a** and **1b**, for either the PBE or the BLYP functionals. On the basis of that, he concluded that not only "dispersion" is missing in density functional calculation but also a different effect that he named "medium-range correlation". The result was frequently cited afterward.[16,17]

In his latest revision of the DFT-D, Grimme tested his new B97-D functional including DFT-D terms and found that it indeed does predict the correct order of the octane isomers.[29] It is interesting to see what would Grimme's new DFT-D parametrization do for the case of octane branching with more common DFs like PBE or BLYP and their hybrids. In a recent paper, Cornimonbeauf and Schleyer[33] found that, in their version of the parametrization of the dispersion correction, this particular case (as well as other alkane

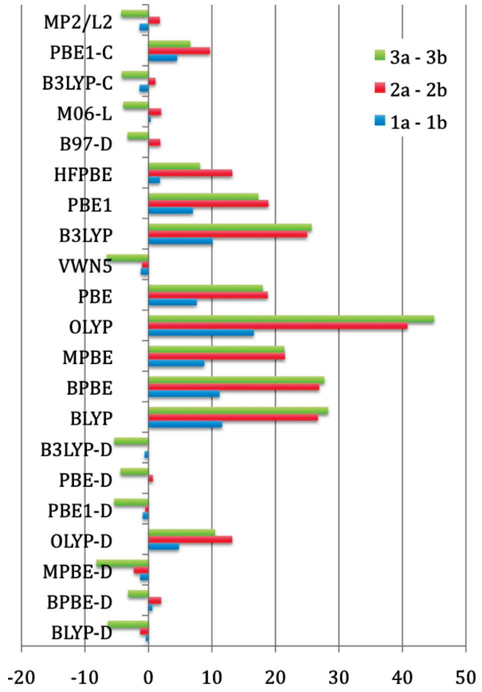


***Figure 2.*** Differences between experimental and calculated alkane branching enthalpies, kcal/mol.

isomerizations) could indeed be described with dispersion-corrected PBE.

It is known that, while for small alkanes the branched isomers are the most stable, for larger alkanes effects of the sterical strain in the highly branched isomers can prevail over the branching stabilization, making linear or other less crowded alkane isomers more stable than the highly sterically crowded branched ones. To see if the DFT-D correction can predict the right trends for these cases as well, we included two isomers of tetradecane, **3a** and **3b** (Scheme 2), for which experimental heats of formation are known,[63] to our test set. The experimental paper contains data[63] for another alkane **2b** with a different branching pattern as well. Hence, we considered its isomerization from the linear tridecane **2a**, too.

We have also designed, on the basis of the availability of experimental heats of formation, another model system with stepwise increase of the sterical crowding (Scheme 3), the addition of isobutene to methane, yielding neopentane **4**, di-$^t$Bu-methane **5**, and finally tris-$^t$Bu-methane **2b**. This system has properties of both increased product branching and monomer insertion. Thus, it is analogous to the polyolefin model systems from the previous section of the paper. Because experimental enthalpies for its products are available, we can assess our methods, including MP2, directly against them. The resulting calculated enthalpies of the processes are shown in Figures 2 and 3 (as differences from the experimental values); the numbers are collected in Tables 1 and 2.

In agreement with literature data, common GGAs and hybrids (BLYP, BPBE, MPBE, OLYP, PBE, B3LYP, PBE1) yield the wrong order of isomers **1a** and **1b**, predicting the linear isomer **1a** to be the most stable. OLYP, BLYP, and B3LYP functionals have the largest errors, while for PBE-type functionals the errors are smaller. Only VWN5 and MP2 predicted the correct order (but see the discussion below on the B97-D and M06-L functionals). With the new version of Grimme's DFT-D correction, all GGAs except OLYP-D gave the correct order of the octane isomers, as did the corrected hybrids B3LYP-D and PBE1-D. Thus, with the "new" DFT-D parametrization, the energy order of the octane isomers **1a** and **2a** can be described even with standard, widely used functionals. Interestingly, an increase in the amount of exact exchange in the uncorrected PBE functional (from pure GGA to PBE1 with 25% to HFPBE with 100% of the exact exchange) decreased the error. The latter functional even predicted correctly the order of the octane isomers, although the error in it is still larger than for DFT-D corrected PBE.

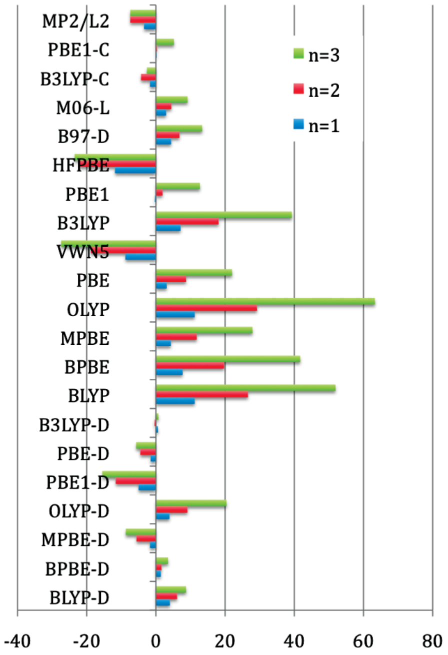


***Figure 3.*** Differences between experimental and calculated enthalpies of stepwise isobutene to methane insertion reactions, kcal/mol. The value *n* is the number of isobutene units to be inserted.

For larger alkanes, application of the DFT-D correction leads to a significant decrease in the error as well. Again, the largest discrepancy among the corrected GGA functionals was found in the case of OLYP. However, for the tetradecanes **3a** and **3b**, most of the dispersion-corrected GGAs, as well as the MP2 method, overestimate the stability of the branched isomer.

The C-Pot corrected hybrid functionals, B3LYP-C and PBE1-C, also have significantly lower errors than the uncorrected ones. The B3LYP-C functional shows very good performance for alkane branching, similar to that of B3LYP-D; however, PBE1-C somewhat underestimates the enthalpies of the branched isomers and produces a qualitatively incorrect result for the octanes.

Figure 3 and Table 2 show the results for the methane crowding system of Scheme 3. Just like for the polymerization reactions, GGAs and hybrids strongly underestimate the exothermicity of the addition, while VWN5 strongly overestimates it. MP2 also overestimates insertion exothermicities, but not as bad as VWN5. Among uncorrected

***Table 1.*** Calculated Alkane Branching Enthalpies, Shown as Differences from the Experimental Enthalpies, kcal/mol[a]

| | BLYP | BPBE | MPBE | OLYP | PBE | VWN5 | B3LYP | PBE1 | HFPBE | B97-D | M06-L | MP2/L2 | exp. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1a** → **1b** | 11.6<br>*–0.4* | 11.2<br>*0.6* | 8.8<br>*–1.3* | 16.6<br>*4.8* | 7.6<br>*0.0* | –1.2 | 10.1<br>*–0.6*<br>–1.4 | 7.0<br>*–0.9*<br>4.5 | 1.8 | *0.0* | 0.3 | –1.4 | –4.2 |
| **2a** → **2b** | 26.7<br>*–1.3* | 26.9<br>*2.0* | 21.5<br>*–2.3* | 40.8<br>*13.2* | 18.8<br>*0.7* | –1.0 | 25.0<br>*–0.1*<br>1.1 | 18.9<br>*0.5*<br>9.7 | 13.2 | *1.9* | 2.0 | 1.8 | 18.3 |
| **3a** → **3b** | 28.3<br>*–6.4* | 27.7<br>*–3.2* | 21.4<br>*–8.2* | 45.0<br>*10.5* | 18.0<br>*–4.4* | –6.6 | 25.7<br>*–5.4*<br>–4.2 | 17.3<br>–5.4<br>6.6 | 8.1 | *–3.3* | –4.0 | –4.3 | 20.0 |

[a] Enthalpies with DFT-D correction applied are shown in italics; the ones obtained with the C-Pot method in the 6-31+G(d,p) basis are underlined.

***Table 2.*** Calculated Methane *t*-Bu Saturation Enthalpies for *n* Isobutene Reagents, Shown as Differences from the Experimental Enthalpies, kcal/mol (See Scheme 3 for Description)[a]

| *n* | BLYP | BPBE | MPBE | OLYP | PBE | VWN5 | B3LYP | PBE1 | HFPBE | B97-D | M06-L | MP2/L2 | exp. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 11.2<br>*4.0* | 7.7<br>*1.4* | 4.3<br>*–1.7* | 11.2<br>*3.9* | 3.1<br>*–1.5* | –8.8 | 7.1<br>*0.6*<br>–1.7 | –0.3<br>*–5.0*<br>0.2 | –11.8 | *4.4* | 2.9 | –3.4 | –18.0 |
| 2 | 26.6<br>*6.1* | 19.7<br>*1.6* | 11.8<br>*–5.6* | 29.2<br>*9.1* | 8.7<br>*–4.5* | –18.8 | 18.1<br>*–0.4*<br>–4.3 | 1.9<br>*–11.6*<br>0.3 | –22.1 | *6.8* | 4.4 | –7.4 | –31.2 |
| 3 | 51.9<br>*8.7* | 41.7<br>*3.5* | 27.9<br>*–8.7* | 63.3<br>*20.4* | 22.0<br>*–5.7* | –27.4 | 39.3<br>*0.7*<br>–2.6 | 12.7<br>*–15.5*<br>5.2 | –23.5 | *13.4* | 9.1 | –7.4 | –25.4 |

[a] Enthalpies with DFT-D correction applied are shown in italics; the ones obtained with the C-Pot method in the 6-31+G(d,p) basis are underlined.

functionals, the largest error is shown by OLYP, followed by the BLYP and B3LYP functionals, while PBE-like functionals perform better. HFPBE shows strong overestimation, due to the obvious importance of a description of changes of the exchange-correlation energy during the reaction, for which pure Hartree–Fock exchange is inadequate. The PBE1 hybrid was the best one among the uncorrected density functionals.

Applying the dispersion correction improves the agreement with experiment, with the best performance shown by BPBE-D. PBE-D and MPBE-D overcorrect; for OLYP-D the correction is not sufficient, although it improves the results. Among the hybrids, B3LYP-D performs very well with errors under 1.0 kcal/mol. PBE1-D yields significantly overestimated exothermicities. C-Pot corrected B3LYP-C shows good performance with some overbinding, while PBE1-C somewhat underestimates the exothermicity.

We have also tested another definition of the damping function for the DFT-D method, with Head-Gordon's power-twelve function.[40] In Table S3, we provide results for the functionals that performed best for DFT-D, BPBE and PBE, with corrections using eq 3, and no DFT-D correction at all, for the alkane branching (Scheme 2) and methane crowding (Scheme 3) test systems. One can see that usage of Head-Gordon's power twelve damping function leads to undercorrection; although it reduces the errors of GGA DFs, they are still much larger that those obtained with Grimme's exponential damping function. The reason is that the power twelve function turns on too soon, as compared to the much steeper, exponential form of eq 3.

**Potential Energy Surfaces.** Scans of the intermolecular potential energy surface (PES) are usually performed around minima on the PES. Yang et al.[27] have studied the repulsive part of the methane dimer interaction curve and found that various density functionals differ significantly in that respect. They showed Becke88 exchange to be more repulsive than PBE exchange, and BLYP is more repulsive than PBE; all of these functionals are more repulsive than CCSD(T). Very recently, Becke published another study that considered interaction curves of noble gas dimers.[21] Similar observations were made for common exchange and correlation density functionals.

To rationalize our results described above, we performed potential energy scans for the collision of methane with neopentane (along the line of a C–C bond of the neopentane molecule, from the side of the latter's three methyl groups). For DFT and MP2 methods, we did fully relaxed PES scans, gradually decreasing the constrained carbon–carbon distance from 6 to 2 Å.

Figure 4 shows the energy differences between the MP2 potential energy scan curve and the curves given by other methods. The VWN5 functional consistently overbinds, as compared to MP2. This is a well-known property of the local density approximation functional.[64] The most "repulsive" density functional among those studied is OLYP, followed by BLYP and B3LYP. The PBE-containing functionals PBE, PBE1, and HFPBE give the closest resemblance to MP2 among the uncorrected functionals. The DFT-D correction plays a significant role, decreasing the repulsive character of the density functional. DFT-D for PBE makes it slightly overbind as compared to MP2. Interestingly, after addition of C-Pot pseudopotentials, the potential energy curves of PBE1-C and B3LYP-C became very close to each other; that is, with C-Pot B3LYP gets less repulsive but PBE1 gets more repulsive as compared to the uncorrected functionals.

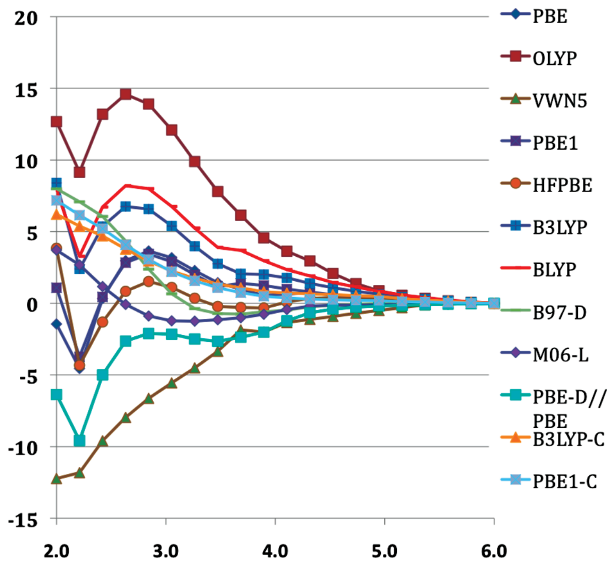


***Figure 4.*** Relaxed PES scans for neopentane−methane interactions, energies relative to MP2, kcal/mol. The energies at a C−C distance of 6.0 Å were taken as zero for each method (distance in Å).

The character of repulsiveness of the density functionals parallels their performance for polymer growth and alkane branching systems described above. The most repulsive OLYP functional gives the larger errors; Becke and LYP containing functionals are over-repulsive as well, while PBE-D gives good but slightly overbound results.

The role and character of the intramolecular interations in alkanes were subject to a prolonged discussion in the literature. Originally, Pfitzer and Catalano[65] attributed the factors that describe alkane stabilities to electron correlation (dispersion) that leads to attractive 1,4 interactions in alkanes. Their reasoning was based on an incremental calculation scheme for the heats of formation. Gronert attributed the stabilities of alkanes, alkyl radicals, and alkenes to 1,3 repulsive interactions, again based on an incremental scheme.[66] Schleyer et al.[55] as well as others[67] discussed the stabilization due to the "protobranching" effect, that is, 1,3 attractive interactions in hydrocarbons. In a very recent paper by Rogers et al.,[68] another model was proposed that was able to describe hydrocarbon stabilities by a scheme that does not include any interactions as its terms, but counts only increments for hydrocarbons' hydrogens, of eight types depending on their chemical environment. The authors of the latter work point out that a model that fits the experimental data is not necessarily a proof of the reality of the "effects" it is based upon. In 2008, Estrada,[69] based on his topological model and semiempirical tight-binding calculations, showed that the ratio of 1,3 interactions to the total of all 1,2, 1,3, 1,4 interactions determines the relative stabilities of alkane isomers. His conclusions support Schleyers view of stabilizing 1,3 protobranching interactions; also, he has shown the importance of taking into account changes in the carbon nature due to its environment, which is absent in the "1,3 repulsion" model of Gronert.

Computationally, we can see from our and others' potential energy scans for alkanes and noble gases that the problem of most common DFT methods is overestimation of repulsive interactions (or underestimation of attractive ones) at medium-range distances. The accuracy of the treatment of the interactions at these distances is crucial because it includes the 1,3 and 1,4 interactions that play an important role in hydrocarbons, as discussed above. As we have shown above, with a proper parametrization, the DFT-D approach is sufficient to make at least some of the commonly used GGA and hybrid functionals quantitatively accurate, that is, give the correct order of energies of branched and linear alkanes and the olefin monomer insertion enthalpies. It seems that the alternative approach of modifying pseudopotentials (the C-Pot method) works equally well, at least for some parametrizations (B3LYP-C). If we decompose the DFT-D energy corrections for our test molecules into H−H, C−C, and C−H components, we can see that the latter are largest in magnitude and contribute most to the DFT-D energy differences for the alkane branching and methane crowding tests (summarized in Table S4 of the Supporting Information). In the alternative C-Pot approach, only pseudopotentials for carbon are introduced; however, it somehow still works, at least for the B3LYP-C case, even for our hydrogen-rich test molecules from Schemes 1−3.

**Performance of the "Dispersion-Aware" Functionals B97-D and M06-L.** The B97-D functional was parametrized together with the DFT-D correction by Grimme.[29] Thus, it could avoid potential problems with overbinding due to dispersion being present both in the correction and in the density functional. The M06-L functional by Zhao and Truhlar is one of the highly parametrized meta-GGA functionals developed by their group, and one that is specifically recommended for main group compounds[50] (other variants of M06 family functionals exist, with different amounts of the exact exchange; we chose the semilocal M06-L functional for the performance reasons). In this study, we will apply these two specialized density functionals to the set of problems studied above. To save computational time, we took MP2 optimized geometries from Priroda calculations and computed single-point M06-L and B97-D energies for them. For the PES scans, MP2-optimized geometries of the points along the scans were taken as well.

The PES scans (Figure 4) show that both of these functionals perform well at intermediate distances; their energies are closer to the MP2 ones than those of the uncorrected GGAs and hybrids. They do not overbind as much as PBE-D does. Also, for intermediate distances they do not differ significantly from each other. However, at short ranges, B97-D rapidly becomes repulsive (reaching values closer to BLYP than to MP2), while M06-L remains "softer".

For the polyolefin growth reactions studied (as shown in Figure 1), M06-L shows results closer to those of BPBE-D (between it and MPBE-D); B97-D results are a bit farther from the latter, falling between BPBE-D and BLYP-D. Considering that MP2 is likely to have some overbinding, the performance of the "harder" B97-D for these systems might in fact be good.

For the alkane branching processes (Scheme 2), these functionals behave as follows. Both M06-L and B97-D predict the correct sign of the **1a**, **1b** energy difference (preference for the branched octane) and very small absolute

***Table 3.*** Calculated MP2/L2 Energies of $C_{12}H_{12}$ Hydrocarbon Isomerizations, and Differences between MP2 and DFT Enthalpies, kcal/mol[a]

| reaction | BLYP | BPBE | MPBE | OLYP | PBE | VWN5 | B3LYP | PBE1 | B97-D | M06-L | MP2/L2 | exp. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **6a → 6b** | –2.5 | –2.6 | –1.9 | –4.4 | –1.6 | 0.4 | –2.4 | –1.9 |  | 0.1 | –6.5 | –7.0 |
|  | *0.6* | *0.1* | *0.7* | *–1.4* | *0.3* |  | *0.3* | *0.1* | *0.0* |  |  |  |
|  |  |  |  |  |  |  | –0.4 | –0.4 |  |  |  |  |
| **6c → 6a** | –35.2 | –5.6 | –5.2 | –4.2 | –3.4 | 13.3 | –24.2 | 8.6 |  | –12.0 | –36.3 |  |
|  | *–30.5* | *–1.6* | *–1.5* | *0.4* | *–0.6* |  | *–20.3* | *11.3* | *–29.4* |  |  |  |
|  |  |  |  |  |  |  | 3.2 | –4.1 |  |  |  |  |
| **6c → 6d** | –30.8 | –7.8 | –7.4 | –6.6 | –5.9 | 7.6 | –21.5 | 4.1 |  | –7.0 | 40.2 |  |
|  | *–28.1* | *–5.5* | *–5.3* | *–4.1* | *–4.3* |  | *–19.4* | *5.5* | *–26.5* |  |  |  |
|  |  |  |  |  |  |  | 2.7 | –5.8 |  |  |  |  |
| **6c → 6e** | –38.9 | –10.1 | –9.2 | –10.1 | –6.9 | 12.9 | –26.6 | 6.1 |  | –9.9 | 27.9 |  |
|  | *–34.2* | *–6.1* | *–5.4* | *–5.5* | *–4.2* |  | *–22.7* | *8.7* | *–32.9* |  |  |  |
|  |  |  |  |  |  |  | 4.5 | –4.7 |  |  |  |  |

[a] Enthalpies with DFT-D correction applied are shown in italics; the ones obtained with the C-Pot method in the 6-31+G(d,p) basis are underlined.

errors. The agreement for the *n*-tetradecane **3a** to the branched isomer **3b** energy is better than for most of the common DFT-D functionals (PBE-D, BLYP-D, MPBE-D, and B3LYP-D), but still shows some overestimation of the stability of the branched isomer.

For the stepwise isobutene to methane insertions leading to **2b**, which have increasing steric crowding (Scheme 3), both functionals give small errors that increase with increasing branching; for M06-L the error is slightly smaller than that for B97-D. Although the discrepancies with experiment for these two functionals are larger than for the best "standard" functional (BPBE-D), both of the specialized functionals avoid the overbinding due to the dispersion correction that PBE-D and MPBE-D demonstrate (Figure 3).

**Hydrocarbon Isomerization and Cyclization Reactions.** Recently, a caged $C_{12}H_{12}$ compound **6c** (Scheme 4) containing two cyclopropane moieties linked by aliphatic bridges, named as [$D_{3d}$]-octahedrane by its creators, has been synthesized. It was crowned by its creators as "the most stable $(CH)_{12}$ hydrocarbon".[70] However, subsequent studies[14,17] have shown that quite large discrepancies exist between relative energies of compound **6c** and other $(CH)_{12}$ isomers computed by density functional methods (such as B3LYP) and MP2. This system thus appears to be an interesting test case. We will apply our set of functionals with and without DFT-D corrections, and also perform MP2 calculations with a more reliable basis and with full optimization of geometries, on selected $C_{12}H_{12}$ isomers (Scheme 4). There are experimental heats of formation[53] for the dimethyl-naphthalenes **6a** and **6b**, which also happen to be isomers of $C_{12}H_{12}$. We then consider several structures for which computations were previously reported: the compound **6c**, a "good", low-lying isomer **6d** (compound **31** from ref 17) and a randomly chosen other isomer **6e** (compound **21** from ref 17).

Calculated isomerization energies between these $C_{12}H_{12}$ hydrocarbons are presented in Table 3. First, we note that compound **6c**, according to any of the pure local, GGA, hybrid DFT, and MP2 results, is less stable than either of the dimethyl-naphthalenes. To our knowledge, there is no experimental heat of formation of the octahedrane published. The previously published CCSD(T)/cc-pVDZ//MP2 value by Schreiner et al. for the energy of the reaction **6c** to **6d** is +25 kcal/mol;[17] however, the basis set used is fairly small. One of the most recent higher-level results is a complete basis set extrapolated CCSD(T) result by Csonka et al., of +20.5 kcal/mol.[71] Comparison of these values to our MP2/L2 energy of +30.4 kcal/mol shows that, probably, our MP2 energy is slightly too high.

Second, for the isomerization of the naphthalene isomer **6a** to **6b**, inclusion of the DFT-D correction improves the DFT (both pure GGA and hybrid) results, as compared to the experimental value (−7 kcal/mol, from the NIST database[53]). Both M06-L and B97-D functionals yield perfect agreement with experiment. The local density approximation yields the smallest error as compared to uncorrected GGA functionals, but with the opposite sign, which is typical, as we have seen above. This points to the importance of the inclusion of dispersion for the treatment of interacting $CH_3$ groups.

The influence of the DFT-D corrections on the results of other reactions is also beneficial (as it decreases the differences with MP2) but rather small. For the C-Pot methods, B3LYP-C and PBE1-C, the discrepancies also become smaller than for the uncorrected hybrid functionals.

Third, and most interesting, there is a very pronounced difference between the results from the BLYP and B3LYP density functionals, on one hand, and all other methods, on the other, for isomerization energies involving compound **6c**. Both BLYP-containing methods strongly underestimate its stability relative to the naphthalenes and compounds **6d**, **6e**. The failure of BLYP and B3LYP was already reported by Schreiner et al.[17] Here, we note that it cannot be ascribed to either the Becke-88 exchange or the Lee−Yang−Parr correlation functionals alone, because other combinations we have considered that contain either of them (BPBE and OLYP) yield results that are in agreement with other GGA methods and are close to the MP2 values. The result also cannot be ascribed to the Becke88 + LYP functional being too repulsive, because the OLYP functional that is the most repulsive functional studied (see the PES discussion above) does not have such large errors as BLYP.

Inclusion of the exact exchange in B3LYP shifts the results in the right direction, but these results are still very far from the correct ones. For the PBE functional, inclusion of exact exchange (PBE1) makes the results worse, overstabilizing compound **6c**. It seems that, for the molecule **6c** in case of

***Table 4.*** Enthalpies of Reactions for $C_3H_6$ and $C_{10}H_{16}$ Hydrocarbons, Shown as Differences from the Experimental Enthalpies, kcal/mol (See Scheme 5 for the Description)[a]

| reaction | BLYP | BPBE | MPBE | OLYP | PBE | VWN5 | B3LYP | PBE1 | HFPBE | B97-D | M06-L | MP2/L2 | exp. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **7a** → **7b** | −4.6 | −0.8 | 0.1 | −3.0 | 1.1 | 9.5 | −0.6 | 5.2 | 17.4 | | −0.7 | 4.1 | 25.3 |
| | *−1.6* | *1.7* | *2.4* | *0.0* | *2.8* | | *1.9* | *6.8* | | *−1.9* | | | |
| | | | | | | | 6.3 | 5.4 | | | | | |
| **7a** → **7c** | −11.2 | −5.9 | −4.3 | −9.4 | −3.0 | 7.5 | −5.7 | 2.9 | 21.3 | | −7.8 | 5.4 | 36.6 |
| | *−4.8* | *−0.3* | *0.9* | *−2.9* | *0.9* | | *−0.2* | *6.8* | | *−6.8* | | | |
| | | | | | | | 6.7 | 1.1 | | | | | |
| **7a** → **7d** | −20.1 | −11.3 | −9.2 | −15.6 | −7.4 | 7.1 | −12.2 | 1.1 | 27.2 | | −10.7 | 2.3 | 26.2 |
| | *−10.8* | *−3.2* | *−1.6* | *−6.2* | *−1.7* | | *−4.1* | *6.8* | | *−12.1* | | | |
| | | | | | | | 3.2 | −1.5 | | | | | |
| **7a** → **7e** | −11.8 | −10.7 | −8.3 | −17.1 | −6.5 | 6.6 | −6.5 | −0.8 | 18.1 | | −15.3 | 2.4 | 75.0 |
| | *−4.3* | *−4.2* | *−2.2* | *−9.4* | *−2.0* | | *0.0* | *3.7* | | *−9.6* | | | |
| | | | | | | | 8.9 | −1.0 | | | | | |
| *n*-$C_3H_6$ → **8** | 7.6 | 2.6 | 2.6 | 2.4 | 2.2 | −0.8 | 5.8 | −3.5 | 0.6 | | −2.9 | 1.3 | 7.9 |
| | *5.0* | *0.2* | *0.0* | *0.0* | *−0.1* | | *5.4* | *0.3* | | *2.1* | | | |
| | | | | | | | −0.1 | −1.5 | | | | | |

[a] Enthalpies with DFT-D correction applied are shown in italics; the ones obtained with the C-Pot method in the 6-31+G(d,p) basis are underlined.

the combination of the Becke88 and Lee−Yang−Parr functionals, some error cancellation, which exists for other functionals, does not work.

The [$D_{3d}$]-octahedrane test is also the case where B97-D and M06-L give dissimilar results (Table 3). The enthalpies of the processes involving the octahedrane calculated by M06-L are closer to the ones given by most of the GGA functionals (BPBE, MPBE, etc.). At the same time, B97-D shows very large differences, placing it next to the BLYP and B3LYP density functionals.

While application of the DFT-D correction to all of the reactions in Scheme 4 (Table 3) somewhat improves agreement between the DFT and MP2 results for all GGAs and B3LYP (but not for PBE1), it cannot completely fix the performance of BLYP and B3LYP. One would think that the failure of the two latter functionals might have nothing to do with the over-repulsiveness of these DFs that is corrected by DFT-D. However, the C-Pot method B3LYP-C decreases the differences between MP2 and B3LYP quite dramatically. Combining C-Pot with PBE1 (to give PBE1-C) changes the energy difference in the opposite direction as compared to PBE1-D.

These observations require some rationalization. Molecule **6c** is a polycyclic molecule; besides that, some of its cycles are three-membered cyclopropane rings that are known to be chemically different from alkanes and alkenes. Either property might lead to the failure of the BLYP combination. Previously, poor performance of B3LYP was reported for terpene cyclizations and isomerization reactions by Matsuda and co-workers.[18]

Thus, we have added isomerization reactions of selected $C_{10}H_{16}$ hydrocarbons for which experimental heats of formation are available: terpenes (camphene and 3-carene), adamantane, pentamethyl-cyclopentadiene, and methyl-dicyclopropyl-cyclopropane (Scheme 5). Also, we have calculated the simple propene to cyclopropane cyclization reactions with all our methods. (We note that, for the smaller carbocycles, the experimental values have some discrepancies. There are two heats of formation for cyclopropane on the NIST website, leading to either 4.5 or 7.9 kcal/mol for the experimental cyclization enthalpy; we took the latter value to be "correct".)

Results are assembled in Table 4. With respect to the propene to cyclopropane cyclization, BLYP and B3LYP show a difference, predicting a higher endothermicity of the reaction as compared to the other density functionals and MP2; M06-L and PBE1 overestimate the stability of cyclopropane. An underestimation of cyclization enthalpies by B3LYP was noted previously in the work by Matsuda.[18] OLYP, being more repulsive than BLYP according to our PES scans, does not underestimate the stability of cyclopropane as strongly and neither does BPBE. The picture is similar to that observed for the octahedrane; however, while differences of Becke + LYP methods per cyclopropene ring are clearly present, they are unlikely to be responsible for all of the discrepancies of the octahedrane, which are much larger that just twice the difference of cyclopropane enthalpy.

Studying the isomerizations of one of the $C_{10}H_{16}$ isomers, adamantane **7a**, to various other isomers (caged molecule, the camphene **7b**, a molecule rich with terminal methyl groups **7d**, and hydrocarbons containing cyclopropane fragments, the 3-carene **7c** and **7e**; see Scheme 5) also allows us to probe the importance of the dispersion interactions and the peculiarities of hydrocarbons with cages and small rings. The results are collected in Table 4. The general picture for the common DFs is as usual; DFT-D improves the energies for all of them, except for PBE1, which is the best among uncorrected functionals. The C-Pot method usually has somewhat larger errors in case of B3LYP-C but smaller ones for PBE1-C.

For the isomerization of adamantane to pentamethylcyclopentadiene **7d**, which has numerous 1,4 interactions between its five methyl groups, PBE-D, MPBE-D, and PBE1 perform best, while errors for the "repulsive" BLYP and OLYP functionals are large and cannot be fixed entirely by the dispersion correction. As in the previous similar cases, B3LYP-D performs significantly better than the uncorrected B3LYP, while errors for PBE1, which already had a good performance by itself, are increased by the DFT-D approach.

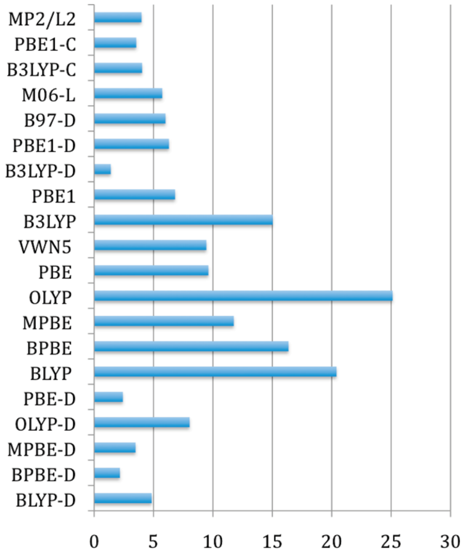


***Figure 5.*** Mean absolute deviations between computed and experimental enthalpies for the set of isomerization reactions from Schemes 2, 3, and 5, in kcal/mol.

The C-Pot corrected PBE1-C and B3LYP-C methods both show better agreement with the experiment.

M06-L and B97-D show fairly large differences with the experiment for reactions **7a** to **7c**−**e**, overestimating stabilities of the latter with respect to adamantane. For the oligocyclopropane **7e**, the discrepancy for M06-L is the largest; for pentamethyl-cyclopentadiene **7d**, both functionals overestimate its stability by over 10 kcal/mol. We note that MP2 shows a rather good performance, as compared to experiment, and that most of the functionals we used, in their DFT-D corrected form, also performed well.

Thus, it seems that the large deviation of the energy of molecule **6c** computed with BLYP, B3LYP, and B97-D as compared to other methods is really a failure of the former density functionals. We cannot trace this failure to the presence of any particular molecular fragment. Grimme's DFT-D corrections do not depend on the hybridization of the atoms; thus, one could possibly expect the $R_0$ and $C_6$ parameters for cyclopropane atoms to be slightly different from those of other types of carbon. However, this does not seem to be a problem for many of the other reactions that we have considered: the polymerization processes that involve $C^{sp3}$ to $C^{sp2}$ changes, and the isomerizations of $C_{10}H_{16}$ that contain many different types of the carbon atoms, including the cyclopropane ones. Thus, we feel that slight changes in the $C_6R^{-6}$ parameters can be ruled out as being the problem.

To summarize our results, we made an average estimation of the performance of the methods used in this work. We have computed the mean absolute deviations (MAD) of the calculated enthalpies over two sets of isomerization reactions (those on Schemes 2 and 5) and one set of monomer insertions (Scheme 3) where experimental data are available. The MAD values are shown in Figure 5. The corrected functionals B3LYP-D, BPBE-D, and PBE-D possess the smallest MADs over these sets, lower than 2.5 kcal/mol. (Note, however, that the sets do not include the octahedrane case, where B3LYP-D has larger differences with MP2 method and other functionals.) This indicates the importance of the dispersion corrections. Most of the other functionals that include dispersion corrections in the form of DFT-D or C-Pot perform accetably, with MADs under 5 kcal/mol. The exceptions are PBE1-D, which shows strong overattractiveness, and thus the MAD are not significantly lower than those of uncorrected PBE1, and OLYP-D, which is still over-repulsive despite the corrections. The specialized B97-D functional and the highly parametrized M06-L perform well for polymerization and alkane branching, but show larger errors for $C_{10}H_{16}$ hydrocarbon isomerizations; thus, their MADs are slightly over 5 kcal/mol. Among the uncorrected density functionals, PBE1 shows the best performance, and OLYP the worst.

## Conclusions

We have systematically investigated the performance of DFT on several selected test cases, olefin polymerization thermodynamics, alkane branching, and isomerization of various cyclic/caged hydrocarbons such as $C_{12}H_{12}$ and $C_{10}H_{16}$. Our approach was to combine the DFT benchmarking of interesting systems where experiment is not available against MP2 methods, with a comparison of both MP2 and DFT results against experimental ones where the latter are available. This was done specifically for systems that were chosen to be analogous to our systems of interest.

The results show that commonly used GGA and hybrid functionals perform poorly in many cases: all of them strongly and systematically underestimate the enthalpies of olefin monomer insertions; the stabilities of branched alkanes and hydrocarbons with many 1,4 methyl−methyl contacts are also strongly underestimated by these methods.

The application of the DFT-D corrections as proposed by Grimme improves the results considerably for most of the functionals and model systems we have tested. The best qualitative agreement for our test cases can be obtained by using the BPBE-D and PBE-D GGA functionals and the hybrid B3LYP-D (although the latter has problems describing the highly caged compound **6c**). Among uncorrected density functionals, PBE1 has shown the best performance, while OLYP and BLYP were usually the worst.

Other approaches, like the heavily parametrized M06-L functional, the B97-D functional, and the parametrized pseudopotential C-Pot method, also in general perform reasonably well for the cases dominated by dispersion.

The performance of density functionals was analyzed by considering the repulsive part of the alkane−alkane potential energy surfaces. There is a similarity between the over-repulsive character of a density functional (the most repulsive being OLYP, the least being PBE) and its performance for olefin insertion and alkane branching enthalpies. The GGAs and hybrids that we have studied are over-repulsive as compared to MP2, and the local VWN5 is overattractive. The introduction of the DFT-D correction amends the over-repulsiveness of the density functional; PBE-D shows slight overattractiveness as compared to MP2. Thus, the DFT-D correction with the proper parametrization and choice of the density functional is sufficient (despite previous reports[16]) for a qualitative description of intramolecular nonbonding interactions within hydrocarbons.

In cases where application of the correction is not desirable, one might test for the importance of the dispersion interactions by using the least and the most repulsive functionals (PBE or PBE1 and OLYP, correspondingly).

Not every problem of DFT can be solved by the simple DFT-D correction. For the case of the caged compound **6c**, BLYP, B3LYP, and B97-D show large errors, which are not directly related to the repulsive character of the DFs, because the most repulsive functional (OLYP) does not have this problem. These errors cannot be corrected by DFT-D but, in the case of B3LYP-C, have been successfully fixed by the C-Pot approach.

**Acknowledgment.** This work is part of the Research Programme and was performed with financial support from the Dutch Polymer Institute, Eindhoven, The Netherlands, Project #641. We thank Dr. D. N. Laikov for providing us with the Priroda code. Part of the computations was performed on high-performance computational facilities of the Canadian WestGrid consortium, which is funded in part by the Canada Foundation for Innovation, Alberta Innovation and Science, BC Advanced Education, and the participating research institutions.

**Supporting Information Available:** Tables with calculated enthalpies of polymer growth reactions, and their deviations used to create Figures 1; table with comparison of two damping functions; table with components of the DFT-D energy differences for the selected reactions; and table with selected results with the power-twelve damping function. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

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

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

(3) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(4) Pople, J. A.; Headgordon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. *J. Chem. Phys.* **1989**, *90*, 5622.

(5) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys.* **1991**, *94*, 7221.

(6) Dunlap, B. I. *J. Mol. Struct. (THEOCHEM)* **2000**, *529*, 37.

(7) Eichkorn, K.; Treutler, O.; Ohm, H.; Haser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283.

(8) Laikov, D. N. *Chem. Phys. Lett.* **1997**, *281*, 151.

(9) Swart, M. *J. Chem. Theory Comput.* **2008**, *4*, 2057.

(10) Swart, M.; van der Wijst, T.; Guerra, C. F.; Bickelhaupt, F. M. *J. Mol. Modell.* **2007**, *13*, 1245.

(11) van der Wijst, T.; Guerra, C. F.; Swart, M.; Bickelhaupt, F. M. *Chem. Phys. Lett.* **2006**, *426*, 415.

(12) Grimme, S. *Angew. Chem., Int. Ed.* **2006**, *45*, 4460.

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

(14) Wodrich, M. D.; Corminboeuf, C.; Schreiner, P. R.; Fokin, A. A.; Schleyer, P. V. *Org. Lett.* **2007**, *9*, 1851.

(15) Grimme, S.; Steinmetz, M.; Korth, M. *J. Chem. Theory Comput.* **2007**, *3*, 42.

(16) Schreiner, P. R. *Angew. Chem., Int. Ed.* **2007**, *46*, 4217.

(17) Schreiner, P. R.; Fokin, A. A.; Pascal, R. A.; de Meijere, A. *Org. Lett.* **2006**, *8*, 3635.

(18) Matsuda, S. P. T.; Wilson, W. K.; Xiong, Q. B. *Org. Biomol. Chem.* **2006**, *4*, 530.

(19) Pieniazek, S. N.; Clemente, F. R.; Houk, K. N. *Angew. Chem., Int. Ed.* **2008**, *47*, 7746.

(20) Brittain, D. R. B.; Lin, C. Y.; Gilbert, A. T. B.; Izgorodina, E. I.; Gill, P. M. W.; Coote, M. L. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1138.

(21) Kannemann, F. O.; Becke, A. D. *J. Chem. Theory Comput.* **2009**, *5*, 719.

(22) Voityuk, A. A. *J. Chem. Theory Comput.* **2008**, *4*, 1877.

(23) Tu, C. Y.; Guo, W. H.; Hu, C. H. *J. Phys. Chem. A* **2008**, *112*, 117.

(24) Check, C. E.; Gilbert, T. M. *J. Org. Chem.* **2005**, *70*, 9828.

(25) Grimme, S.; Steinmetz, M.; Korth, M. *J. Org. Chem.* **2007**, *72*, 2118.

(26) Swart, M.; Bickelhaupt, F. M. *J. Comput. Chem.* **2008**, *29*, 724.

(27) Johnson, E. R.; Mori-Sanchez, P.; Cohen, A. J.; Yang, W. T. *J. Chem. Phys.* **2008**, *129*, 204112.

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

(29) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.

(30) Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554.

(31) Swart, M.; Sola, M.; Bickelhaupt, F. M. *J. Chem. Phys.* **2009**, *131*, 094103.

(32) Ducere, J. M.; Cavallo, L. *J. Phys. Chem. B* **2007**, *111*, 13124.

(33) Wodrich, M. D.; Jana, D. F.; Schleyer, P. V.; Corminboeuf, C. *J. Phys. Chem. A* **2008**, *112*, 11495.

(34) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 4.

(35) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *J. Chem. Phys.* **2005**, *122*, 6.

(36) Mackie, I. D.; DiLabio, G. A. *J. Phys. Chem. A* **2008**, *112*, 10968.

(37) Maerzke, K. A.; Murdachaew, G.; Mundy, C. J.; Schenter, G. K.; Siepmann, J. I. *J. Phys. Chem. A* **2009**, *113*, 2075.

(38) Shamov, G. A.; Schreckenbach, G. *Inorg. Chem.* **2008**, *47*, 805.

(39) Wheeler, S. E.; Houk, K. N.; Schleyer, P. V. R.; Allen, W. D. *J. Am. Chem. Soc.* **2009**, *131*, 2547.

(40) Chai, J. D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.

(41) Peverati, R.; Baldridge, K. K. *J. Chem. Theory Comput.* **2008**, *4*, 2030.

(42) Budzelaar, P. H. M. *J. Comput. Chem.* **2007**, *28*, 2226.

(43) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.

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

(45) Adamo, C.; Barone, V. *J. Chem. Phys.* **2002**, *116*, 5933.

(46) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403.

(47) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.

(48) Laikov, D. N.; Ustynyuk, Y. A. *Russ. Chem. Bull.* **2005**, *54*, 820.

(49) Laikov, D. N. *Chem. Phys. Lett.* **2005**, *416*, 116.

(50) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.

(51) 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.

(52) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2004.

(53) Afeefy, H. Y.; Liebman, J. F.; Stein, S. E. In *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; Lindstrom, P. J., Mallard, W. G., Eds.; National Institute of Standards and Technology: Gaithersburg, MD; http://webbook.nist.gov (retrieved Feb 15, 2009).

(54) Izgorodina, E. I.; Coote, M. L.; Radom, L. *J. Phys. Chem. A* **2005**, *109*, 7558.

(55) Wodrich, M. D.; Wannere, C. S.; Mo, Y.; Jarowski, P. D.; Houk, K. N.; Schleyer, P. V. R. *Chem.-Eur. J.* **2007**, *13*, 7731.

(56) Benson, S. W. *Thermochemical Kinetics*; Wiley: New York, 1968.

(57) Voityuk, A. A. *Chem. Phys. Lett.* **2006**, *433*, 216.

(58) Tirado-Rives, J.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2008**, *4*, 297.

(59) There is a large number of conformers possible for the polypropylene isomers. We did not attempt for location of the global minima; instead, “syndio” and “iso” oligomers were constructed as to retain similarity between conformers with subsequent monomer insertions.

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

(61) Grimme, S. *J. Comput. Chem.* **2003**, *24*, 1529.

(62) Pitonak, M.; Neogrady, P.; Cerny, J.; Grimme, S.; Hobza, P. *ChemPhysChem* **2009**, *10*, 282.

(63) Verevkin, S. P.; Nolke, M.; Beckhaus, H. D.; Ruchardt, C. *J. Org. Chem.* **1997**, *62*, 4683.

(64) Koch, W.; Holthausen, M. C. *A Chemist’s Guide to Density Functional Theory*; Wiley Verlag Chemie: New York, 2000.

(65) Pitzer, K. S.; Catalano, E. *J. Am. Chem. Soc.* **1956**, *78*, 4844.

(66) Gronert, S. *J. Org. Chem.* **2006**, *71*, 1209.

(67) Mitoraj, M.; Zhu, H. J.; Michalak, A.; Ziegler, T. *J. Org. Chem.* **2006**, *71*, 9208.

(68) Zavitsas, A. A.; Matsunaga, N.; Rogers, D. W. *J. Phys. Chem. A* **2008**, *112*, 5734.

(69) Estrada, E. *Chem. Phys. Lett.* **2008**, *463*, 422.

(70) de Meijere, A.; Lee, C. H.; Kuznetsov, M. A.; Gusev, D. V.; Kozhushkov, S. I.; Fokin, A. A.; Schreiner, P. R. *Chem.-Eur. J.* **2005**, *11*, 6175.

(71) Csonka, G. I.; Ruzsinszky, A.; Perdew, J. P.; Grimme, S. *J. Chem. Theory Comput.* **2008**, *4*, 888.

CT9005135

JCTC Journal of Chemical Theory and Computation

# Computation of Nonretarded London Dispersion Coefficients and Hamaker Constants of Copper Phthalocyanine

Yan Zhao,* Hou T. Ng, and Eric Hanson

*Commercial Print Engine Lab, HP Laboratories, Hewlett-Packard Co., 1501 Page Mill Road, Palo Alto, California 94304*

Jiannan Dong, David S. Corti, and Elias I. Franses

*School of Chemical Engineering, Purdue University, 480 Stadium Mall Drive, West Lafayette, Indiana 47907-2100*

Received August 14, 2009

**Abstract:** A time-dependent density functional theory (TDDFT) scheme has been validated for predictions of the dispersion coefficients of five molecules ($H_2O$, $NH_3$, $CO_2$, $C_6H_6$, and pentane) and for predictions of the static dipole polarizabilities of three organometallic compounds ($TiCl_4$, $OsO_4$, and $Ge(CH_3)_4$). The convergence of grid spacing has been examined, and two types of pseudopotentials and 13 density functionals have been tested. The nonretarded Hamaker constants $A_{11}$ are calculated by employing a semiempirical parameter $a$ along with the standard Hamaker constant equation. The parameter $a$ is optimized against six accurate Hamaker constants obtained from the full Lifshitz theory. The dispersion coefficients of copper phthalocyanine CuPc and CuPc–$SO_3H$ are then computed. Using the theoretical densities of $\rho_1 = 1.63$ and 1.62 g/cm$^3$, the Hamaker constants $A_{11}$ of crystalline α-CuPc and β-CuPc are found to be $14.73 \times 10^{-20}$ and $14.66 \times 10^{-20}$ J, respectively. Using the experimentally derived density of $\rho_1 = 1.56$ g/cm$^3$ for a commercially available β-CuPc (nanoparticles of ~90 nm hydrodynamic diameter), $A_{11} = 13.52 \times 10^{-20}$ J is found. Its corresponding effective Hamaker constant in water ($A_{121}$) is calculated to be $3.07 \times 10^{-20}$ J. All computed $A_{11}$ values for CuPc are noted to be higher than those reported previously.

## 1. Introduction

van der Waals interactions play key roles in numerous physical phenomena and applications, such as in crystal packing, colloidal stability, interfacial adhesion, self-assembly, molecular recognition, protein folding, nucleobases stacking, drug intercalation, solvation, supramolecular chemistry, pigment dispersion, and capillarity of liquids. The London dispersion forces are the major component of the long-range interparticle forces between nanoparticles or between colloidal particles. Those forces are described to first order by the macroscopic Hamaker constant.[1] This material-dependent quantity is difficult to measure experimentally.[2,3] Although the Hamaker constant can be calculated from Lifshitz's continuum theory,[4] it requires detailed molecular and macroscopic information on the dielectric or optical properties of the material over a wide frequency range. These methods can pose a challenge for estimating Hamaker constants for many materials. Simplified models based on more recent extensions of the Lifshitz theory have been proposed.[5−10]

In this Article, a computational approach based on time-dependent density functional theory (TDDFT) is benchmarked for predictions of the London dispersion coefficients ($C_{11}$) of five molecules ($H_2O$, $NH_3$, $CO_2$, $C_6H_6$, and pentane) and for predictions of the static dipole polarizabilities of three organometallic compounds ($TiCl_4$, $OsO_4$, and $Ge(CH_3)_4$). The

* Corresponding author e-mail: yan.zhao3@hp.com.

10.1021/ct900431a © 2010 American Chemical Society
Published on Web 01/06/2010

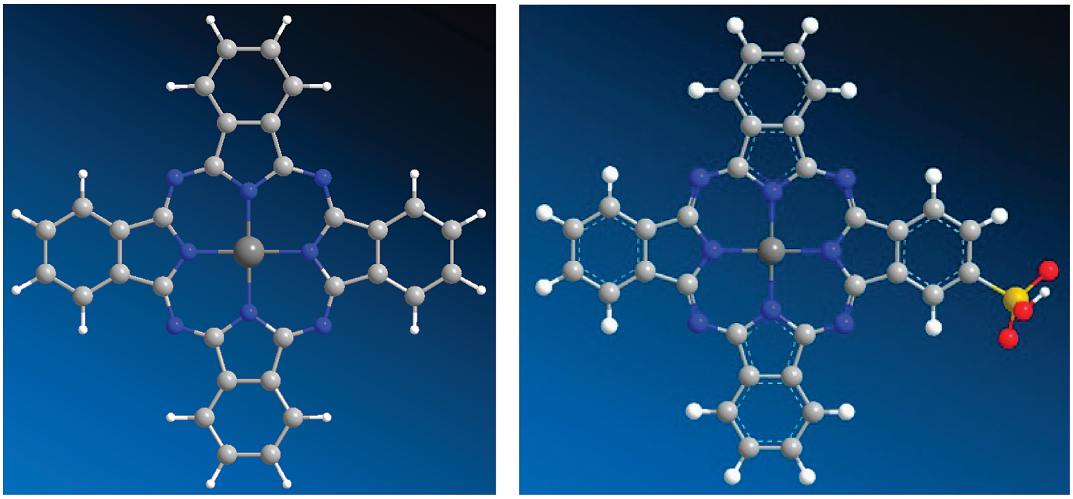


***Figure 1.*** Molecular structures of CuPc ($C_{32}H_{16}CuN_8$, mol wt = 576.1) and CuPc−$SO_3$H ($C_{32}H_{16}CuN_8SO_3$, mol wt = 656.1).

validated TDDFT scheme is then employed to calculate the dispersion coefficients for copper phthalocyanine (CuPc), and for monosulphonated CuPc (CuPc−$SO_3$H), which are important for certain ink pigments. The molecular structures of CuPc and CuPc−$SO_3$H are shown in Figure 1.

A semiempirical model based on the original Hamaker equation is then developed to calculate the Hamaker constant from the London dispersion coefficients and the densities of the particles. This model is used to predict values of the Hamaker constants of CuPc pigments.

In the present Article, the conventional notations $A_{11}$ and $A_{12}$ are used to denote Hamaker constants in vacuum for two like- and unlike-particles, while $C_{11}$ and $C_{12}$ are used to denote their London molecular dispersion coefficients. Two experimentally derived Hamaker constants $A_{11}$ for CuPc particles have been reported in the literature.[11,12] Among them, the value of $A_{11} = 0.2 \times 10^{-20}$ J[11] was estimated from certain mechanical strength properties by using some assumptions regarding (i) the detailed structure of a CuPc powder and the relationship of $A_{11}$ to its mechanical properties and (ii) molecular distances of adjacent particles in the powder at close contact. This value is an order of magnitude lower than the values of ca. (4−7) × $10^{-20}$ J calculated or measured for many organic materials. Because CuPc contains Cu and benzene rings, its density is higher than those of hydrocarbons, and hence the value of $A_{11}$ for CuPc should be higher than 7 × $10^{-20}$ J. For this reason, the accuracy of the above value of 0.2 × $10^{-20}$ J is questionable. A higher value of $A_{11} = 3.7 \times 10^{-20}$ J for CuPc green (i.e., of chlorinated CuPc) was reported recently in ref 12, although it still appears to be too low for the similar reasons detailed above.

Hence, there is a need to obtain more accurate values of $A_{11}$ (for CuPc), which would be useful for ink dispersion stability studies. In the present study, a TDDFT scheme is benchmarked for predictions of the dispersion coefficients of five molecules ($H_2O$, $NH_3$, $CO_2$, $C_6H_6$, and pentane) and for predictions of the static dipole polarizabilities of three organometallic compounds ($TiCl_4$, $OsO_4$, and $Ge(CH_3)_4$). Next, this validated TDDFT scheme was used to determine the London dispersion coefficients ($C_{11}$) for the CuPc molecule. Because certain CuPc particles are stabilized by chemically attached sulfonate groups ($SO_3$H) at their surface, the $C_{11}$ value for CuPc−$SO_3$H is also computed. The $A_{11}$ values for two commonly available CuPc crystal polymorphs, α-CuPc and β-CuPc, are then calculated using the ideal crystal densities, as they are estimated from the crystal lattice parameters. In addition, the experimentally derived density of some commercial β-CuPc particles, stabilized by $SO_3$H, is used to obtain another estimate for $A_{11}$ (β-CuPc). Our approach yields values of $A_{11}$ that range from ca. 13 to 15 × $10^{-20}$ J. While no reliable direct experimental data are available, the above predicted values are nonetheless thought to be more plausible and in turn more accurate than the previous estimates.

## 2. Theory and Computational Methods

**2.1. London Dispersion Parameters and Hamaker Constants.** In 1930, London[13] performed a quantum mechanical analysis based on perturbation theory to predict the long-range dispersion interaction potential energy $E_{dis}$, between two atoms or molecules, 1 and 2, which is of the form:

$$E_{\text{dis}} = -\frac{C_{12}}{r^6} \qquad (1)$$

where $r$ is the separating distance between the atomic or molecular centers, and $C_{12}$ is the dispersion coefficient as defined by London. Hamaker[1] integrated the interaction potential energies, based on the additivity concept as proposed by London, to calculate the total interaction energy between two macroscopic bodies (or particles, each consisting of either molecule 1 or 2):

$$E_{12} = -\int_{V_1}\int_{V_2} \frac{\rho_1\rho_2 C_{12}\,\mathrm{d}V_1\,\mathrm{d}V_2}{r^6} \qquad (2)$$

where $\rho_1$ and $\rho_2$ are the number densities, and $V_1$ and $V_2$ are the volumes of particles 1 and 2, respectively. Given the above equation, the Hamaker constant is defined as

$$A_{12} \equiv \pi^2 C_{12}\rho_1\rho_2 \tag{3}$$

The remaining terms in eq 2 produce a purely geometrical term upon integration.

The derivation that leads to the above relation is based on the following assumptions taken into consideration: (A) additivity: a pairwise summation of the individual contributions provides the total interaction; (B) continuous medium: integration over the volumes of the interacting bodies replaces the pairwise summation; (C) uniform material properties: each $\rho$ and $C_{11}$ are considered to be uniform over the total volume of the interacting bodies; and (D) medium: the above interaction is in a vacuum.

If the particles 1 are in a medium consisting of material 2, the following Hamaker constants are defined: in a vacuum, $A_{11}$, $A_{22}$, and $A_{12}$ as above and for two particles, both of type 1, located in a medium 2 (particle 2), $A_{121}$. An approximate estimation of $A_{121}$ has been proposed:[9]

$$A_{121} \approx A_{11} + A_{22} - 2A_{12} \tag{4}$$

and it is often assumed, based somewhat on London's theory and the assumption of $C_{12} \approx (C_{11}C_{22})^{1/2}$, that

$$A_{12} = \sqrt{A_{11}A_{22}} \tag{5}$$

It then follows that

$$A_{121} \approx (\sqrt{A_{11}} - \sqrt{A_{12}})^2 \tag{6}$$

The above results are also only applicable to nonretarded van der Waals interactions,[14] which require that the interparticle distances are smaller than about 0.1 $\mu$m.[15]

In 1956, Lifshitz[16] developed a macroscopic continuum theory for Hamaker constants based on quantum electrodynamics and quantum field theory. The Lifshitz theory requires data on the complex dielectric constants of each material at all frequencies. Because of a relativistic effect, the Hamaker constant is distance-dependent, or "retarded", beyond about 0.1 $\mu$m.[15,17]

In the present study, only the nonretarded London dispersion coefficients are considered. An efficient computational model based on the time-dependent density functional theory (TDDFT) for $C_{12}$[18,19] has been chosen as described below. Moreover, a modified form of eq 3 was used to calculate the nonretarded Hamaker constants from $C_{12}$:

$$A_{12} = a\pi^2 C_{12}\rho_1\rho_2 \tag{7}$$

where $a$ is an empirical parameter, devised here to account for the shortcomings of assumptions A, B, and C mentioned above. In particular, assumption A ignores the many-body effects in the condensed phase. Assumption B ignores the intrinsic discontinuous nature of the materials, and assumption C does not take into account the possible nonuniform density variations of the macroscopic particles. In addition, the parameter $a$ also accounts for the errors in the TDDFT calculations of dispersion coefficients (see section 4.4).

### 2.2. Time-Dependent DFT for Computing $C_{12}$.

The nonretarded dispersion coefficient for molecules 1 and 2, averaged over all possible orientations, is given by the Casimir−Polder relation:[20]

$$C_{12} = \frac{3}{\pi}\int_0^\infty \mathrm{d}u\, \alpha^1(i\varpi)\alpha^2(i\varpi) \tag{8}$$

where $\alpha^X(i\varpi)$ is the trace of the dipole polarizability tensor of molecule $X$ (=1 or 2) evaluated at the imaginary frequency $i\omega$. The function $\alpha^X(i\varpi)$ can be calculated by a TDDFT time propagation scheme as reported in ref 18 and implemented in the OCTOPUS code.[21]

To evaluate the Casimir−Polder integral for the dispersion coefficients in eq 8, the polarizabilities were calculated at the imaginary frequencies from the Gauss−Legendre integration schemes.

### 2.3. Benchmark Data.

It is important to validate the accuracy of the TDDFT scheme described in section 2.2 for predictions of dispersion coefficients. The dispersion coefficients ($C_{11}$) of five molecules, $H_2O$, $NH_3$, $CO_2$, $C_6H_6$, and pentane, have been used as a benchmark data. The reference values for these five molecules have been taken from the results of the dipole oscillator strength distribution (DOSD) method of Meath and co-workers.[22−26]

Because CuPc is an organometallic compound, it is desirable to include some organometallic compounds in the benchmark set. However, we are unable to find any accurate dispersion coefficients for organometallic compounds from the literature. We have chosen to use the static dipole polarizabilities of three organometallic compounds, $TiCl_4$, $OsO_4$, and $Ge(CH_3)_4$, to benchmark the quality of the employed methods. The reference static dipole polarizability for $OsO_4$ is taken from a collision-induced light scattering experiment of Hohm and Maroulis,[27] and the reference value for $TiCl_4$ is from a combined experimental and theoretical study by the same authors.[28] The reference polarizability for $Ge(CH_3)_4$ is taken from a recent collision-induced light scattering experiment of Maroulis and Hohm.[29]

We collected accurate Hamaker constants for six compounds ($H_2O$, pentane, decane, hexadecane, polystyrene, and poly(methyl methacrylate)) to obtain the parameter $a$ in eq 7, and the reference Hamaker constants are taken from the accurate Lifshitz theory calculations by Hough and White.[5]

### 2.4. Computational Details.

The molecular geometries of the selected chemical species were optimized with the M06-L density functional[30] and the 6-31+G(d,p) basis set.[31] The OCTOPUS program has been employed for the TDDFT propagation calculations of the dispersion coefficients and polarizabilities as described in section 2.2.

OCTOPUS is a pseudopotential real space DFT program, in which electrons are described quantum-mechanically with DFT or TDDFT, nuclei are described classically as point particles, and interactions between electrons and nuclei are described with the pseudopotential approximation. We examined the convergence of the grid spacing with two pseudopotentials; one is the Hartwigsen−Goedecker−Hutter (HGH) type of pseudopotentials,[32] which are relativistic separable dual-space Gaussian pseudopotentials. We also examined the FHI pseudopotentials developed by Fuchs and Scheffler[33] at Fritz-Haber-Institut.

We have tested the performance of 13 density functionals for the calculation of dispersion coefficients and static dipole

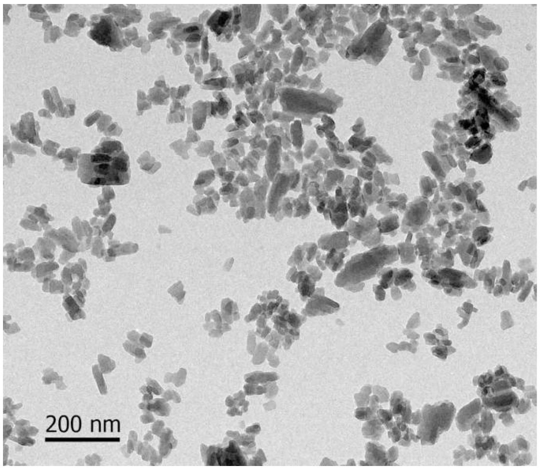


***Figure 2.*** TEM image of $\beta$-CuPc pigment particles dried on a holey carbon TEM grid.

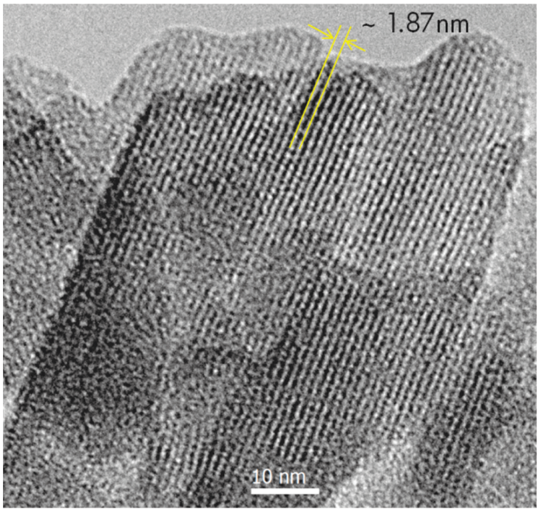


***Figure 3.*** High-resolution TEM image of a $\beta$-CuPc pigment particle. Distinctive lattice fringes with a spacing of ~1.9 nm can be clearly observed.

polarizabilities. The tested density functionals include the SPZ local spin density approximation (LSDA),[34-36] six generalized gradient approximations (GGAs), PW91,[37] PBE,[38] HCTH,[39] RPBE,[40] WC, and XLYP,[41] and six hybrid GGAs (B3PW91,[42] B3LYP,[43] B3P86,[43] PBE0,[44] O3LYP,[45,46] and X3LYP[41]).

## 3. Experimental Characterization of the CuPc Pigment Particles

The CuPc particles were obtained from Cabot Corp. (MA) as a 10 wt % stable dispersion in water and were used as received. They consist of pure CuPc stabilized by chemically attached sulfonate groups. The particles were characterized with high-resolution (HR) TEM. As shown in Figure 2, their morphology is tubular-like or cubical-like. These particles are crystalline. Distinctive lattice fringes, extending up to the edges of the pigment particles, were observed, with a spacing of ~1.9 nm, which corresponds to the lattice constant of the $\beta$-CuPc structure (Figure 3). The thickness of the sulfonate groups is likely to be less than 0.5 nm on each side. Hence, the particles are more than 99% $\beta$-CuPc and less than 1% CuPc-$SO_3H$.

***Table 1.*** Convergence of the Gauss–Legendre Quadrature Scheme for the Calculation of the Dispersion Coefficient ($C_{11}$) of Benzene[a]

| $N^b$ | $C_{11}$ (au) |
|---|---|
| 4 | 1800.5 |
| 6 | 1796.0 |
| 8 | 1796.0 |
| 10 | 1796.0 |

[a] TDDFT calculations with the SPZ local density functional, using the Hartwigsen–Goedecker–Hutter (HGH) pseudopotentials,[32] and a grid spacing of 0.25 Å. [b] $N$ denotes the number of imaginary frequencies ($i\omega$ in eq 8) for which the dynamic polarizabilities have been calculated, using the Gauss–Legendre integration scheme.

Standard dynamic light scattering measurements performed with a Brookhaven ZetaPALS dynamic light scattering instrument at a wavelength of 659 nm, which has a BI-9000AT digital autocorrelator at a scattering angle of 90°, revealed an average hydrodynamic diameter of about 90 ± 3 nm.

The ideal, or theoretical, particle densities for α- and $\beta$-CuPc were calculated on the basis of the crystal lattice parameters as follows:[26] (A) For α-CuPc, $a = 25.92$ Å, $b = 3.79$ Å, $c = 23.92$ Å, $\beta = 90°$, $\rho_1 = 1.63$ g/cm$^3$. (B) For $\beta$-CuPc, $a = 19.407$ Å, $b = 4.79$ Å, $c = 14.628$ Å, $\beta = 120°$, $\rho_1 = 1.62$ g/cm$^3$.

The density of the actual CuPc nanoparticles was also measured experimentally with the following method. A given mass $m_T$ of a dispersion had a volume $V_T$ at 25 °C and a dry weight $m_p$, as determined by drying in an oven at 50 °C for 3 days. Next, the particle density $\rho_p$ was determined with the following equation and the measured values of $m_T$, $m_p$, and $V_T$:

$$V_T = \frac{m_p}{\rho_p} + \frac{m_T - m_p}{\rho_w} \tag{9}$$

where $\rho_w$ is the literature density of water at 25 °C. The volumes of the particles and the water are assumed to be additive, because the particles are dispersed as a separate phase. The weight fraction of the particles was found to be 0.1005 ± 0.0004, which compares well with the nominal value of 0.10. The particle density was found to be $\rho_p = 1.56 \pm 0.03$ g/cm$^3$ (average of $n = 3$ measurements). This value is about 4% smaller than the ideal value above. The reasons for the small discrepancy are probably: (a) the presence of crystal imperfections or voids in the particles; (b) not accounting for the surface sulfonate groups and associated counterions; and (c) other experimental errors. A 4% discrepancy results in an 8% discrepancy in the value of $A_{11}$. Values of $A_{11}$ for both values of the densities are reported.

## 4. Results and Discussion

**4.1. Gauss–Legendre Integration for the Casimir–Polder Equation.** To determine a minimum number of data points while maintaining an accurate evaluation of the Casimir–Polder integral (eq 8), we benchmarked the Gauss–Legendre

***Table 2.*** Convergence of Dispersion Coefficients and Dipole Polarizabilities with Grid Spacing[a]

| | | $C_{11}$ (au) | | | | | static dipole polarizability α (au) | | |
|---|---|---|---|---|---|---|---|---|---|
| | grid spacing (Å) | $H_2O$ | $NH_3$ | $CO_2$ | $C_6H_6$ | $C_5H_{10}$ | $TiCl_4$ | $OsO_4$ | $Ge(CH_3)_4$ |
| HGH[b] | 0.20 | 50.4 | 93.4 | 162.1 | 1797 | 2000 | 99.0 | 50.9 | 87.6 |
| | 0.25 | 50.8 | 93.2 | 163.2 | 1796 | 2000 | 99.0 | 50.6 | 87.6 |
| | 0.30 | 52.6 | 92.3 | 167.0 | 1795 | 2000 | 99.1 | 50.1 | 87.6 |
| | 0.35 | 60.4 | 92.8 | 182.7 | 1796 | 1997 | 99.3 | 47.7 | 88.2 |
| FHI[c] | 0.20 | 50.1 | 92.6 | 161.8 | 1784 | 1978 | 97.8 | 50.6 | 87.2 |
| | 0.25 | 50.1 | 92.6 | 161.7 | 1781 | 1984 | 97.8 | 50.6 | 87.5 |
| | 0.30 | 49.7 | 92.3 | 161.4 | 1786 | 1987 | 98.2 | 50.6 | 87.3 |
| | 0.35 | 49.8 | 92.8 | 160.4 | 1785 | 1986 | 98.2 | 50.5 | 87.4 |

[a] All calculations employed the SPZ LSDA density functional. [b] The Hartwigsen−Goedecker−Hutter (HGH) type of pseudopotentials.[32] [c] The FHI pseudopotentials developed by Fuchs and Scheffler[33] at Fritz-Haber-Institut.

***Table 3.*** Performance of Density Functionals for Dispersion Coefficients and Static Dipole Polarizabilities with the HGH Pseudopotentials

| | $C_{11}$ (au) | | | | | | static dipole polarizability α (au) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| method | $H_2O$ | $NH_3$ | $CO_2$ | $C_6H_6$ | $C_5H_{10}$ | MAPE[a] | $TiCl_4$ | $OsO_4$ | $Ge(CH_3)_4$ | MAPE[b] | AMAPE[c] |
| best estimate[d] | 45.3 | 89.0 | 158.7 | 1723 | 1905 | | 101.4 | 51.0 | 83.2 | | |
| SPZ | 50.8 | 93.2 | 163.2 | 1796 | 2000 | 5.8 | 99.0 | 50.6 | 87.6 | 2.8 | 4.3 |
| XLYP | 42.5 | 76.2 | 143.2 | 1547 | 1649 | 10.8 | 89.6 | 47.5 | 78.2 | 8.1 | 9.5 |
| PW91 | 41.0 | 73.8 | 139.1 | 1512 | 1633 | 13.1 | 88.4 | 46.6 | 77.4 | 9.4 | 11.3 |
| PBE | 40.7 | 73.2 | 138.6 | 1505 | 1630 | 13.5 | 88.1 | 46.5 | 77.1 | 9.8 | 11.6 |
| X3LYP | 39.6 | 75.9 | 138.0 | 1523 | 1626 | 13.3 | 87.0 | 46.3 | 76.7 | 10.4 | 11.9 |
| WC | 40.2 | 72.4 | 137.5 | 1499 | 1633 | 14.1 | 88.1 | 46.1 | 76.9 | 10.1 | 12.1 |
| B3LYP | 41.3 | 74.9 | 137.0 | 1509 | 1606 | 13.3 | 86.6 | 46.1 | 76.0 | 10.9 | 12.1 |
| RPBE | 38.9 | 73.4 | 138.7 | 1502 | 1617 | 14.4 | 87.7 | 46.6 | 77.0 | 9.8 | 12.1 |
| HCTH | 40.5 | 73.4 | 135.4 | 1486 | 1607 | 14.4 | 87.0 | 46.2 | 76.8 | 10.4 | 12.4 |
| PBE0 | 39.8 | 72.6 | 132.9 | 1475 | 1587 | 15.6 | 85.2 | 45.2 | 75.0 | 12.4 | 14.0 |
| B3P86 | 37.6 | 72.1 | 132.5 | 1470 | 1578 | 16.9 | 85.3 | 45.2 | 74.6 | 12.6 | 14.7 |
| O3LYP | 39.1 | 70.9 | 130.7 | 1445 | 1542 | 17.4 | 84.8 | 45.2 | 74.4 | 12.8 | 15.1 |
| B3PW91 | 37.1 | 71.0 | 131.9 | 1458 | 1564 | 17.7 | 84.9 | 45.1 | 74.3 | 12.9 | 15.3 |

[a] MAPE is the mean absolute percentage error for the dispersion coefficients, which is calculated as

$$\text{MAPE} = \sum_{i=1}^{5} \frac{|C_{11}^{\text{cal},i} - C_{11}^{\text{best est.},i}|}{C_{11}^{\text{best est.},i}} \times 100\%/5$$

[b] MAPE is the mean absolute percentage error for the dipole polarizabilities, which is calculated as

$$\text{MAPE} = \sum_{i=1}^{3} \frac{|\alpha_i^{\text{cal}} - \alpha_i^{\text{best est.}}|}{\alpha_i^{\text{best est.}}} \times 100\%/3$$

[c] AMAPE is the average of the two MAPE. [d] The best estimates of the dispersion coefficients ($C_{11}$) are taken from the DSOD results of Meath and co-workers.[22−26] The reference static dipole polarizability for $OsO_4$ is taken from a collision-induced light scattering experimental study by Hohm and Maroullis,[27] and the reference value for $TiCl_4$ is from a combined experimental and theoretical study by the same authors.[28] The refrence polarizability for $Ge(CH_3)_4$ is taken from a recent collision-induced light scattering experiment of Maroullis and Hohm.[29]

integration scheme against the calculation of the dispersion coefficient of benzene. As shown in Table 1, a four-digit accuracy of the dispersion coefficient was found with the use of the six-point Gauss−Legendre quadrature scheme. We therefore used the six-point Gauss−Legendre integration in all of our calculations.

**4.2. Convergence for Grid Spacing.** The OCTOPUS code[21] performs a TDDFT calculation on a real space mesh. Therefore, it is important to examine the convergence of dispersion coefficients and polarizabilities with the grid spacing. Table 2 gives the results for different grid spacings with the SPZ functional[35,36] and two pseudopotentials (HGH[32] and FHI[33]). As shown in Table 2, with the grid spacing of 0.25 Å, the dispersion coefficients and static dipole polarizabilities are converged to better than 1% for both pseudopotentials. Therefore, the grid spacing of 0.25 Å is used to test the performance of different density functionals.

**4.3. Benchmarking Density Functionals for Dispersion Coefficients and Static Dipole Polarizabilities.** The performance of 13 density functionals for predictions of dispersion coefficients and static dipole polarizabilities is presented in Table 3 (with the HGH pseudopotentials[32]) and Table 4 (with the FHI pseudopotentials[33]). Two statistical errors are tabulated in both tables; mean absolute percentage error (MAPE) is a measure of accuracy of different functionals, whereas AMAPE is an average of the MAPEs for dispersion coefficients and for static dipole polarizabilities.

As shown in Tables 3 and 4, the SPZ functional gives the best performance for dispersion coefficients and static polarizabilities for both HGH and FHI pseudopotentials.

***Table 4.*** Performance of Density Functionals for Dispersion Coefficients and Static Dipole Polarizabilities with the FHI Pseudopotentials

| | $C_{11}$ (au) | | | | | | static dipole polarizability $\alpha$ (au) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| method | $H_2O$ | $NH_3$ | $CO_2$ | $C_6H_6$ | $C_5H_{10}$ | MAPE[a] | $TiCl_4$ | $OsO_4$ | $Ge(CH_3)_4$ | MAPE[b] | AMAPE[c] |
| best estimate[d] | 45.3 | 89.0 | 158.7 | 1723 | 1905 | | 101.4 | 51.0 | 83.2 | | |
| SPZ | 50.1 | 93.2 | 163.2 | 1796 | 2000 | 5.5 | 98.1 | 50.6 | 87.5 | 3.0 | 4.3 |
| XLYP | 41.6 | 76.2 | 143.2 | 1547 | 1649 | 11.2 | 88.8 | 47.3 | 77.8 | 8.7 | 9.9 |
| PW91 | 40.2 | 73.8 | 139.1 | 1512 | 1633 | 13.4 | 87.7 | 46.5 | 77.1 | 9.9 | 11.7 |
| PBE | 40.0 | 73.2 | 138.6 | 1505 | 1630 | 13.8 | 87.4 | 46.4 | 76.7 | 10.2 | 12.0 |
| RPBE | 40.2 | 73.4 | 138.7 | 1502 | 1617 | 13.8 | 87.1 | 46.5 | 76.6 | 10.3 | 12.1 |
| X3LYP | 38.9 | 75.9 | 138.0 | 1523 | 1626 | 13.6 | 86.2 | 46.1 | 76.4 | 10.9 | 12.3 |
| WC | 39.6 | 72.4 | 137.5 | 1499 | 1633 | 14.4 | 87.3 | 46.1 | 76.7 | 10.5 | 12.4 |
| HCTH | 39.4 | 73.4 | 135.4 | 1486 | 1607 | 14.9 | 86.4 | 46.1 | 76.3 | 10.9 | 12.9 |
| B3LYP | 38.5 | 74.9 | 137.0 | 1509 | 1606 | 14.5 | 85.8 | 45.9 | 75.7 | 11.4 | 13.0 |
| PBE0 | 36.9 | 72.6 | 132.9 | 1475 | 1587 | 16.9 | 84.4 | 45.1 | 74.8 | 12.8 | 14.8 |
| B3P86 | 36.8 | 72.1 | 132.5 | 1470 | 1578 | 17.2 | 84.5 | 45.0 | 74.3 | 13.0 | 15.1 |
| B3PW91 | 36.6 | 71.0 | 131.9 | 1458 | 1564 | 17.9 | 84.1 | 44.9 | 74.0 | 13.3 | 15.6 |
| O3LYP | 36.9 | 70.9 | 130.7 | 1445 | 1542 | 18.4 | 84.1 | 45.1 | 74.1 | 13.2 | 15.8 |

[a] MAPE is the mean absolute percentage error for the dispersion coefficients, which is calculated as

$$\mathrm{MAPE} = \sum_{i=1}^{5} \frac{|C_{11}^{\mathrm{cal},i} - C_{11}^{\mathrm{best\ est.},i}|}{C_{11}^{\mathrm{best\ est.},i}} \times 100\%/5$$

[b] MAPE is the mean absolute percentage error for the dipole polarizabilities, which is calculated as

$$\mathrm{MAPE} = \sum_{i=1}^{3} \frac{|\alpha_i^{\mathrm{cal}} - \alpha_i^{\mathrm{best\ est.}}|}{\alpha_i^{\mathrm{best\ est.}}} \times 100\%/3$$

[c] AMAPE is the average of the two MAPE. [d] The best estimates of the dispersion coefficients ($C_{11}$) are taken from the DSOD results of Meath and co-workers.[22−26] The reference static dipole polarizability for $OsO_4$ is taken from a collision-induced light scattering experimental study by Hohm and Maroullis,[27] and the reference value for $TiCl_4$ is from a combined experimental and theoretical study by the same authors.[28] The refrence polarizability for $Ge(CH_3)_4$ is taken from a recent collision-induced light scattering experiment of Maroullis and Hohm.[29]

***Table 5.*** Comparison of Predicted $A_{11}$ ($\times 10^{-20}$ J) Values to the "Best Estimates" of Hough and White

| material | $C_{11}$ (au)[a] | $\rho_1$ (g/cm$^3$) | $A_{11}$ best estimate[b] | $A_{11}$ $a = 1.0$ | $A_{11}$ $a = 0.6883$ | $A_{11}$ $a = 0.6815$ |
|---|---|---|---|---|---|---|
| $H_2O$ | 50.8 | 1.000 | 3.70 | 5.38 | 3.70 | 3.66 |
| pentane | 2000 | 0.626 | 3.75 | 5.16 | 3.55 | 3.52 |
| decane | 7649 | 0.730 | 4.72 | 6.89 | 4.74 | 4.70 |
| hexadecane | 19 368 | 0.770 | 5.23 | 7.67 | 5.28 | 5.23 |
| polystyrene | 3599 | 1.050 | 6.58 | 10.11 | 6.95 | 6.89 |
| PMMA | 2324 | 1.190 | 7.11 | 11.32 | 7.79 | 7.71 |
| MAPE[c] | | | | 48.08 | 3.66 | 3.48 |

[a] All dispersion coefficients are calculated with the SPZ functional and the HGH pseudopotentials using a grid spacing of 0.25 Å. [b] Taken from Hough and White.[5] [c] MAPE is the mean absolute percentage error for Hamaker constants, which is calculated as

$$\mathrm{MAPE} = \sum_{i=1}^{6} \frac{|A_{11}^{\mathrm{cal},i} - A_{11}^{\mathrm{best\ est.},i}|}{A_{11}^{\mathrm{best\ est.},i}} \times 100\%/6$$

However, with the FHI pseudopotentials, we encountered some self-consistent field convergence problems for CuPc, so we have chosen to use the SPZ functional and the HGH pseudopotentials with a grid spacing of 0.25 Å for the calculation of Hamaker constants.

**4.4. Estimation of the Parameter *a* in Eq 7.** We employed eq 7 to compute Hamaker constants. Because the TDDFT calculated $C_{11}$ values are used in eq 7 to compute the Hamaker constant, the parameter $a$ explicitly corrects the error in the TDDFT calculation as well as the deficiencies of the three assumptions mentioned in section 2.1. One way to determine the parameter $a$ is to use the accurate Hamaker constant of water determined by Hough and White as the "standard". Using the value of $A_{11} = 3.7 \times 10^{-20}$ J and the calculated $C_{11}$ for $H_2O$, it is found that $a = 0.6883$. Another way of determining the parameter $a$ is to minimize the MAPE for a benchmark set of Hamaker constants of $H_2O$, pentane, decane, hexadecane, polystyrene, and PMMA. The value for the optimized parameter $a$ is found to be 0.6815 by this minimization. It is encouraging that this optimized parameter differs insignificantly from the value of 0.6883 determined by using the Hamaker constant of $H_2O$.

Table 5 lists the Hamaker constants calculated by using $a = 1.0$ (the same as the original Hamaker eq 3), $a = 0.6883$ (determined from the $A_{11}$ of $H_2O$), and $a = 0.6815$ (optimized against the benchmark set of six Hamaker constants). As

***Table 6.*** Dynamic Dipole Polarizabilities $\alpha(i\omega)$ (au) and Dispersion Coefficients $C_{11}$(au)[a]

| | $\alpha(i\omega)$[b] | | | | | | |
|---|---|---|---|---|---|---|---|
| molecule | $\omega = 0.01048$ | $\omega = 0.06118$ | $\omega = 0.18441$ | $\omega = 0.48804$ | $\omega = 1.47101$ | $\omega = 8.58488$ | $C_{11}$ |
| CuPc | 598.31 | 528.15 | 373.25 | 201.66 | 54.49 | 2.40 | 78 926 |
| CuPc−SO$_3$H | 646.10 | 570.64 | 405.02 | 220.58 | 60.25 | 2.69 | 93 224 |

[a] The dynamic polarizabilities and dispersion coefficients in this table are calculated with the SPZ functional and the HGH pseudopotentials using a grid spacing of 0.25 Å. [b] The values of $\omega$ are from the six-point Gauss−Legendre integration scheme.

***Table 7.*** Calculated Values of $A_{11}$ and $A_{121}$ for CuPc Using $a = 0.6815$ in Eq 7

| particle | $\rho_1$ (g/cm$^3$) | $A_{11} \times 10^{-20}$ (J)[a] | $A_{121} \times 10^{-20}$ (J) | comment |
|---|---|---|---|---|
| α-CuPc | 1.63 | 14.73 | 3.66 | ideal density |
| β-CuPc | 1.62 | 14.66 | 3.63 | ideal density |
| β-CuPc | 1.56 | 13.52 | 3.07 | measured density |

[a] The Hamaker constants are calculated with the $C_{11}$ of CuPc in Table 6.

shown in Table 5, the original Hamaker equation ($a = 1$) gives large errors as compared to the Lifshitz theory. The MAPE for $a = 1.0$ is about 50%, which is 13 times larger than the MAPEs of $a = 0.6883$ or 0.6815. Table 5 also shows that using $a = 0.6815$ is slightly more accurate than $a = 0.6883$, as shown by its smaller MAPE. Therefore, $a = 0.6815$ is used for the computation of Hamaker constants of the CuPc pigments.

**4.5. Computation of Dispersion Coefficients and Hamaker Constants for CuPc.** Using the SPZ functional and the HGH pseudopotentials, the dynamic dipole polarizabilities and dispersion coefficients for CuPc and CuPc−SO$_3$H have been calculated, and the results are shown in Table 6. The plots of the dynamic dipole polarizabilities are given in the Supporting Information.

For densities of α-CuPc and β-CuPc shown in section 3, the Hamaker constants for CuPc nanoparticles were determined in vacuum ($A_{11}$) and in water ($A_{121}$), and the results are listed in Table 7.

These $A_{11}$ values are noticeably larger than the previously reported values of 0.2 and $3.7 \times 10^{-20}$ J (see section 1). By being larger than $A_{11}$ of polystyrene, the values computed here seem to be more reasonable, because the density of CuPc particles is much greater than that of polystyrene.

The effect of the surface sulfonate groups on the total value of $A_{11}$ is expected to be small. Using the values of $C_{11}$ for CuPc and CuPc−SO$_3$H the estimated volume fraction of CuPcSO$_3$H should alter the computed value of $A_{11}$ by no more than 1−2%. Such a correction may be more important for much smaller CuPc nanoparticles ($d < 10$ nm), but less important for larger nanoparticles ($d > 200$ nm). Because the relative uncertainties in the particle density, the $C_{11}$ computation, and the used value of $a$ are larger than 2%, the effect of the SO$_3$H groups on $A_{11}$ of the 90 nm β-CuPc particles will be ignored.

The predicted values of $A_{121}$ are also listed in Table 7. The relative uncertainty of $A_{121}$ is larger than that of $A_{11}$. These values look to be plausible estimates of the Hamaker constants, which can be used as input into the DLVO theory[47,48] for estimating colloidal stability. Having these values for $A_{11}$ ((13−15) $\times 10^{-20}$ J), rather than the 0.2 and $3.7 \times 10^{-20}$ values, makes a big difference in the computation of $A_{121}$. If the value of $A_{11} = 3.7 \times 10^{-20}$ J were used, the resulting $A_{121}$ value would be predicted to be zero, and this would have a big impact on the predictions of the DLVO theory.

**4.6. Limitations of the Proposed Model.** As shown in the previous sections, a semiempirical parameter $a$ is used to connect the Hamaker theory to the more accurate Lifshitz theory. The advantage of this model is that the experimental inputs for eq 7 are the densities of the particles, which are much easier to obtain than the experimental inputs for the Lifshitz theory, that requires detailed molecular and macroscopic information on the dielectric and/or optical properties of the material over a wide frequency range.[5] However, one limitation of the model proposed in the present study is that the training set of the parameter $a$ includes only the Hamaker constants of water, hydrocarbons, and polymers (PMMA and polystyrene). The transferability of the parameter $a$ may be a concern for other type of materials. We expect that the accuracy of our model may be degraded for ionic crystals or metal clusters due to the presence of a large amount of complicated electrostatic and screening many-body interactions in these materials. Another limitation is that our model needs a well-defined building block (for $C_{11}$) of the particle. This is not a problem for H$_2$O, hydrocarbons, crystals, or hompolymers, but a well-defined building block for a random copolymer or for a protein needs extra effort.

## 5. Conclusions

A TDDFT method has been benchmarked for the computation of the London dispersion coefficients and static dipole polarizabilities. The Hamaker constants for nonretarded van der Waals interactions were calculated from $C_{11}$, the densities of particles, and an empirical correction to the original Hamaker equation. The value of this empirical parameter $a$ was determined to be 0.6815, by optimizing against a benchmark set of six accurate Hamaker constants. Using this procedure resulted in an MAPE of 3.5% for the predictions of $A_{11}$ in the benchmark set.

After the methods for determining $C_{11}$ and $A_{11}$ were benchmarked, the dynamic dipole polarizabilities and $C_{11}$ for two target molecules, CuPc and CuPc−SO$_3$H, were computed. The Hamaker constants for α-CuPc and β-CuPc particles, which are important in pigment dispersions, were predicted to be in the range from 13 to 15 $\times 10^{-20}$ J. Such values are much larger than the available literature values of $0.2 \times 10^{-20}$ and $3.7 \times 10^{-20}$ J, which were inferred indirectly from certain previously published experiments. Overall, the new $A_{11}$ value for CuPc seems to be a reasonably rigorous and accurate estimate that is more in line with our expectations of the values of the Hamaker constant for similar organically based compounds. It can be argued that the previously reported estimates of $A_{11}$ for CuPc are too low. While the current results suggest that the present method yields reliable predictions, more extensive tests are needed. In such tests, additional estimates of $C_{11}$ and $A_{11}$ may be calculated and compared to other reliable literature data or predictions.

**Acknowledgment.** This work was supported by HP Laboratories, Hewlett-Packard Co. We thank the developers of the OCTOPUS code for making this excellent computer program freely available to the scientific community. We are grateful to Mr. Eric Wu for providing computational resources for this work.

**Supporting Information Available:** Cartesian coordinates and the plots of the dynamic polarizabilities of CuPc molecules. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Hamaker, H. C. *Physica* **1937**, *4*, 1058.

(2) Das, S.; Sreeram, P. A.; Raychaudhuri, A. K. *Nanotechnology* **2007**, *18*, 35501.

(3) Farmakis, L.; Lioris, N.; Kohadima, A.; Karaiskakis, G. *J. Chromatogr., A* **2006**, *1137*, 231.

(4) Dzyaloshinskii, I. E.; Lifshitz, E. M.; Pitaevskii, L. P. *Adv. Phys.* **1961**, 10.

(5) Hough, D. B.; White, L. R. *Adv. Colloid Interface Sci.* **1980**, *14*, 8.

(6) Bowen, W. R.; Jenner, F. *Adv. Colloid Interface Sci.* **1995**, *56*, 201.

(7) Ackler, H. D.; French, R. H.; Chiang, Y. M. *J. Colloid Interface Sci.* **1996**, *179*, 460.

(8) Bergström, L. *Adv. Colloid Interface Sci.* **1997**, *70*, 125.

(9) French, R. H. *J. Am. Ceram. Soc.* **2000**, *83*, 2117.

(10) Kim, H.-Y.; Sofo, J. O.; Velegol, D.; Cole, M. W.; Lucas, A. A. *Langmuir* **2007**, *23*, 1735.

(11) Li, Q.; Feke, D. L.; Manas-Zloczower, I. *Powder Technol.* **1997**, *92*, 17.

(12) Hui, D.; Nawaz, M.; Morris, D. P.; Edwards, M. R.; Saunders, B. R. *J. Colloid Interface Sci.* **2008**, *324*, 110.

(13) London, F. *Z. Physik* **1930**, *60*, 245.

(14) Kaplan, I. G. *Intermolecular Interactions*; Wiley: Hoboken, NJ, 2006; p 72.

(15) Calbi, M. M.; Gatica, S. M.; Velegol, D.; Cole, M. W. *Phys. Rev. A* **2003**, *67*, 33201.

(16) Lifshitz, E. M. *Sov. Phys.* **1956**, *2*, 73.

(17) Parsegian, V. A. In *Physical Chemistry: Enriching Topics from Colloid and Surface Science*; Olphen, H. V., Mysels, K. J., Eds.; Theorex: La Jolla, CA, 1975; p 27.

(18) Marques, M. A. L.; Castro, A.; Malloci, G.; Mulas, G.; Botti, S. *J. Chem. Phys.* **2007**, *127*, 14107.

(19) Botti, S.; Castro, A.; Andrade, X.; Rubio, A.; Marques, M. A. L. *Phys. Rev. B* **2008**, *78*, 35333.

(20) Casimir, H. B. G.; Polder, D. *Phys. Rev.* **1948**, *73*, 360.

(21) Castro, A.; Marques, M. A. L.; Appel, H.; Oliveira, M.; Rozzi, C.; Andrade, X.; Lorenzen, F.; Gross, E. K. U.; Rubio, A. *Phys. Status Solidi B* **2006**, *243*, 2465.

(22) Margoliash, D. J.; Meath, W. J. *J. Chem. Phys.* **1978**, *68*, 1426.

(23) Jhanwar, B. L.; Meath, W. *J. Chem. Phys.* **1982**, *67*, 185.

(24) Jhanwar, B. L.; Meath, W. J. *Mol. Phys.* **1980**, *41*, 1061.

(25) Kumar, A.; Meath, W. J. *Mol. Phys.* **1992**, *75*, 311.

(26) Wu, Q.; Yang, W. *J. Chem. Phys.* **2002**, *116*, 515.

(27) Hohm, U.; Maroulis, G. *J. Chem. Phys.* **2004**, *121*, 104118.

(28) Hohm, U.; Maroulis, G. *J. Chem. Phys.* **2006**, *124*, 124312.

(29) Hohm, U.; Maroulis, G. *Phys. Rev. B* **2007**, *76*, 32504.

(30) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.

(31) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986; p 125.

(32) Hartwigsen, C.; Goedecker, S.; Hutter, J. *Phys. Rev. B* **1998**, *58*, 3641.

(33) Fuchs, M.; Scheffler, M. *Comput. Phys. Commun.* **1999**, *119*, 67.

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

(35) Slater, J. C. *Quantum Theory of Molecular and Solids. Vol. 4: The Self-Consistent Field for Molecular and Solids*; McGraw-Hill: New York, 1974; p 136.

(36) Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981**, *23*, 5048.

(37) Perdew, J. P. In *Electronic Structure of Solids '91*; Ziesche, P., Eschig, H., Eds.; Akademie Verlag: Berlin, 1991; p 11.

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

(39) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 6264.

(40) Hammer, B.; Hansen, L. B.; Norskov, J. K. *Phys. Rev. B* **1999**, *59*, 7413.

(41) Xu, X.; Goddard, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2673.

(42) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

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

(44) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158.

(45) Hoe, W.-M.; Cohen, A. J.; Handy, N. C. *Chem. Phys. Lett.* **2001**, *341*, 319.

(46) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403.

(47) Derjaguin, B.; Landau, L. *Acta Physico Chemica URSS* **1941**, *14*, 633.

(48) Verwey, E. J. W.; Overbeek, J. T. G. *Theory of the Stability of Lyophobic Colloids*; Elsevier: Amsterdam, 1948; p 100.

CT900431A

JCTC Journal of Chemical Theory and Computation

# Bulk and Surface Properties of Rutile $TiO_2$ from Self-Consistent-Charge Density Functional Tight Binding

H. Fox,[†] K. E. Newman,[‡] W. F. Schneider,[§] and S. A. Corcelli*,[†]

*Department of Chemistry and Biochemistry, Department of Physics, Department of Chemical and Biomolecular Engineering, and Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana 46556*

Received October 2, 2009

**Abstract:** Bulk rutile $TiO_2$ and its (110) surface have been investigated with a computationally efficient semiempirical tight binding method: self-consistent-charge density functional tight binding (SCC-DFTB). Comparisons of energetic, mechanical, and electronic properties are made to density functional theory (DFT) and to experiment to characterize the accuracy of SCC-DFTB for bulk rutile $TiO_2$ and $TiO_2$(110). Despite the fact that the SCC-DFTB parameters for Ti, Ti−Ti, and Ti−O were developed in the context of small biologically relevant Ti containing compounds, SCC-DFTB predicts many properties of bulk $TiO_2$ and the $TiO_2$(110) surface with accuracy similar to local and gradient-corrected DFT. In particular, SCC-DFTB predicts a direct band gap of $TiO_2$ of 2.46 eV, which is in better agreement with experiment, 3.06 eV, than DFT utilizing the local density approximation (LDA), 2.0 eV. SCC-DFTB also performs similar in terms of accuracy as LDA-DFT for the phonon frequencies of the bulk lattice and for the relaxed geometry of the $TiO_2$(110) surface. SCC-DFTB does, however, overestimate the surface energy of $TiO_2$(110) compared to LDA-DFT. Nevertheless, the overall accuracy of SCC-DFTB, which is substantially more computationally efficient than DFT, is encouraging for bulk rutile $TiO_2$ and $TiO_2$(110).

## I. Introduction

Self-consistent-charge density functional tight binding (SCC-DFTB) is a powerful semiempirical tight binding method, which retains much of the physics of density functional theory (DFT) at a significantly reduced computational cost.[1−5] SCC-DFTB utilizes an optimized linear combination of atomic orbitals basis set and attains its computational efficiency in part by precomputing and tabulating all of the necessary one- and two-center matrix elements and overlap integrals. Charge transfer and polarizability are introduced by allowing fluctuations in individual atomic charge densities.The reduced computational cost of SCC-DFTB relative to plane-wave DFT offers the opportunity to investigate the properties of complex interfaces that are inaccessible with other methods (e.g., catalysis at solid−liquid boundaries). Like any emerging computational methodology, it is important that its accuracy be fully evaluated before proceeding to more complex applications. Previous studies have successfully utilized SCC-DFTB to study materials[5] including zinc oxide and sulfide[6] and silicon dioxide and carbide[7,8] and the interaction of graphite surfaces with water,[9] as well as biological systems,[10−25] protonated water clusters,[26] and liquid water.[27]

SCC-DFTB has not, however, been fully evaluated for titania ($TiO_2$). Two previous studies have used DFTB (without the SCC correction) to investigate the electronic properties of $TiO_2$ nanostructures.[28,29] While these studies illustrated the breadth of applicability of DFTB, neither addressed properties of bulk $TiO_2$ nor was their focus on validation of the DFTB methodology. A more recent study by Luschtinetz et al. used SCC-DFTB to investigate the adsorption of phosphonic acid on the (101) surface of anatase $TiO_2$ and the (110) of rutile $TiO_2$.[30] These authors reported structural properties of bulk rutile $TiO_2$ and $TiO_2$(110)

* To whom correspondence should be addressed. E-mail: scorcell@nd.edu.

[†] Department of Chemistry and Biochemistry.

[‡] Department of Physics.

[§] Department of Chemical and Biomolecular Engineering and Department of Chemistry and Biochemistry.

10.1021/ct900665a
Published on Web 01/22/2010

predicted with SCC-DFTB, including lattice parameters of bulk rutile $TiO_2$ and atomic displacements normal to the optimized $TiO_2(110)$ surface. This paper aims to provide a more comprehensive characterization of SCC-DFTB for bulk rutile $TiO_2$ and $TiO_2(110)$ by considering a much more expansive set of properties, including the electronic band structures and vibrational properties of the materials. Also, in this paper we demonstrate that publicly available[31] SCC-DFTB parameters developed and evaluated for titanium atoms in biological contexts[32] are transferable to bulk rutile $TiO_2$ and the $TiO_2(110)$ surface. This is important because it establishes the broad applicability of SCC-DFTB for Ti-containing compounds and materials without recourse to reparameterization for each new problem.

Titanium dioxide is thought of as a prototypical metal oxide and has been extensively studied experimentally and theoretically due to its many industrial applications, principally as a white pigment and in heterogeneous catalysis. Titania and zirconia ($ZrO_2$) are unique among transition metal oxides because they are stable in aqueous solution and are thus particularly important for aqueous radiation and photochemistry. For example, $TiO_2$ is a model system for the photocatalytic disproportionation of water into hydrogen and oxygen gases using solar light.[33−43] The dearth in our understanding of the factors that influence the thermodynamics, kinetics, and mechanisms of chemical reactions at liquid−solid interfaces presents a fundamental impediment to the rational design of improved low-temperature catalysts. Accurate computer simulations of reactivity at liquid−solid interfaces would offer tremendous insight and guidance for improvement of these systems, and due to the large number of atoms involved in such a calculation, an alternative to density functional theory that could capture the science at less computational cost would be very beneficial.

A review by Diebold provides a broad introduction to the physical and chemical properties of bulk $TiO_2$ and its surfaces.[44] The (110) surface is the most stable rutile surface, and its structure has been the focus of numerous theoretical and computational investigations. The energetics and structure of the rutile $TiO_2(110)$ surface have been investigated with DFT using various plane-wave[45−47] and atomic orbital[48−50] implementations. Experimentally, a recent low-energy electron diffraction study by Lindsay et al.[51] has found good agreement with theoretical studies of the (110) surface, compared to the surface X-ray diffraction study by Charlton et al.[52] We will show in this paper that titanium dioxide is described well by SCC-DFTB relative to DFT results using the local density approximation (LDA)[53] or generalized gradient approximation (GGA).[54,55] We will first give a brief introduction to the SCC-DFTB method in section II; section III includes calculations of bulk rutile $TiO_2$ structural, electronic and vibrational properties; section IV repeats this for the (110) surface; and finally, concluding remarks are found in section V.

## II. SCC-DFTB Methodology

There are a number of features of the DFTB method (without the SCC correction) that are responsible for its computational efficiency and transferability. Most central to the latter is that the total energy of the system is expressed within a tight-binding (TB) formalism, where the matrix elements are precomputed and tabulated:[3,10]

$$E^{\mathrm{DFTB}} = \sum_i \sum_{\mu\nu} c_\mu^i c_\nu^i H_{\mu\nu}^{(0)} + E_{\mathrm{rep}} \tag{1}$$

The sum $i$ is over occupied Kohn−Sham orbitals, $\Psi^i(\mathbf{r})$, which have been expanded as a linear combination of atomic orbitals (LCAO) in a basis of confined local pseudoatomic orbitals for the valence electrons, $\phi_\nu(\mathbf{r} - \mathbf{R}_\alpha)$

$$\Psi^i(\mathbf{r}) = \sum_\nu c_\nu^i \phi_\nu(\mathbf{r} - \mathbf{R}_\alpha) \tag{2}$$

where $c_\nu^i$ are expansion coefficients, $\mathbf{r}$ is the position of the electron, and $\mathbf{R}_\alpha$ is the position of nucleus α. The local atomic orbitals are specific to each atom type and are determined by performing a DFT calculation on the isolated atom with an extra electrostatic potential term that confines the electron density to regions that are close to the nucleus, $(r/r_0)^M$. This confining potential contains two parameters, $r_0$ and $M$, that could in principle be optimized. However, in practice $M = 2$ and $r_0 \approx 2r_{\mathrm{cov}}$, where $r_{\mathrm{cov}}$ is the covalent radius of the atom, are found to be reasonable.

By invoking a two-center approximation, the nonzero Hamiltonian matrix elements, $H_{\mu\nu}^{(0)}$, appearing in eq 1 can be conveniently expressed as

$$H_{\mu\nu}^{(0)} = \begin{cases} \varepsilon_\mu^{\text{neutral free atom}} & \alpha = \beta \\ \langle \phi_\mu^\alpha | \hat{T} + \hat{V}_0^\alpha + \hat{V}_0^\beta | \phi_\nu^\beta \rangle & \alpha \neq \beta \end{cases} \tag{3}$$

where $\hat{T}$ is the single-electron kinetic energy operator, $\varepsilon_\mu^{\text{neutral free atom}}$ are the Kohn−Sham orbital eigenvalues of the neutral free atom, and the additional superscripts, α and $\beta$, denote orbitals on atomic sites. $\hat{V}_0^\alpha$ is the effective one-electron Kohn−Sham potential for the compressed reference density on atom α. $\hat{V}_0^\alpha$ is evaluated without the confining potential $(r/r_0)^M$ but using the self-consistently determined confined reference density on atom α. Precomputing and tabulating these matrix elements as a function of the interatomic separation, $R_{\alpha\beta} = |\mathbf{R}_\alpha - \mathbf{R}_\beta|$, results in a significant computational savings.

By applying the variational principle to eq 1, a set of algebraic equations is obtained

$$\sum_\nu c_\nu^i (H_{\mu\nu}^{(0)} - \varepsilon_i S_{\mu\nu}) = 0 \tag{4}$$

whose solution provides the Kohn−Sham orbital coefficients, $c_\nu^i$, and eigenvalues, $\varepsilon_i$. In eq 4, $S_{\mu\nu} = \langle \phi_\mu | \phi_\nu \rangle$ represents the overlap between local pseudoatomic orbitals. Once the coefficients $c_\nu^i$ have been determined, the total energy of the system can be evaluated using eq 1, where $E_{\mathrm{rep}}$ denotes the total core−core repulsion energy. $E_{\mathrm{rep}}$ is calculated from short-ranged purely repulsive pair-potentials, $V_{\mathrm{rep}}(R_{\alpha\beta})$, as follows:

$$E_{\mathrm{rep}} = \sum_\alpha \sum_{\beta > \alpha} V_{\mathrm{rep}}(R_{\alpha\beta}) \tag{5}$$

The effective pair-potentials, $V_{\mathrm{rep}}(R_{\alpha\beta})$, are obtained by empirically fitting $E^{\mathrm{DFTB}}$ to the corresponding total energy, $E^{\mathrm{DFT}}$, computed with DFT for a set of configurations of an appropriately chosen reference system. Determining $V_{\mathrm{rep}}(R_{\alpha\beta})$ empirically has both advantages and disadvantages. From a theoretical perspective, it makes the DFTB method not entirely ab initio and, in principle, less transferable. However, the empiricism improves the accuracy of DFTB by incorporating information from higher-level DFT calculations, and from a practical perspective, the DFTB method has proven itself as an accurate and computationally efficient alternative to self-consistent DFT calculations.

An exciting advantage of the DFTB formalism is that it can be systematically improved by incorporating the effects of charge transfer (i.e., electronic polarizability); the effects of charge transfer are especially important for the description of bonding between atoms of different polarity and the electronic structure of metal oxides. In the self-consistent-charge (SCC) extension of DFTB, the matrix elements in eq 3 are rigorously corrected (through second order in perturbation theory) to take into account small fluctuations in the atomic charge density[3,56]

$$E^{(2)} = \frac{1}{2}\sum_{\alpha}\sum_{\beta}\Delta q_{\alpha}\Delta q_{\beta}\gamma_{\alpha\beta} \tag{6}$$

where the $\Delta q_{\alpha} = q_{\alpha} - q_{\alpha}^{(0)}$ are atomic charge fluctuations. $\gamma_{\alpha\beta} = \gamma_{\alpha\beta}(U_{\alpha},U_{\beta},R_{\alpha\beta})$ is a function that has been derived analytically by Elstner et al.[3] to arrive at the fairly simple, but approximate, expression (using atomic units)

$$\gamma_{\alpha\beta} = \frac{1}{R_{\alpha\beta}} - f(\tau_{\alpha}, \tau_{\beta}, R_{\alpha\beta}) \tag{7}$$

where $f(\tau_{\alpha},\tau_{\beta},R_{\alpha\beta})$ is an exponentially decaying short-ranged function, and $\tau_{\alpha} = 16/5U_{\alpha}$ is given in terms of the Hubbard parameter for atom α, $U_{\alpha}$. Physically, $U_{\alpha}$ is approximately twice the chemical hardness of the atom and can be readily computed by calculating the derivative of the highest occupied atomic orbital with respect to its occupation number using DFT.

Although the accuracy of the DFTB method is significantly increased by incorporating the role of charge transfer, the method becomes more computationally expensive because the atomic charges in eq 6 must be determined self-consistently. However, since the charges do not vary significantly between molecular dynamics time steps, few iterations are typically needed to achieve convergence. Furthermore, since all of the terms in the Hamiltonian matrix elements are still precomputed and tabulated, the SCC-DFTB method remains computationally efficient. Several recent papers have proposed further extension of the DFTB method by incorporating higher-order fluctuations in the charge density.[10,57]

SCC-DFTB calculations were performed using the program DFTB+.[58] All precomputed matrix elements are held in Slater−Koster files, downloaded from http://www.dftb.org. The mio parameter set was used for O−O interactions[3] and the trans3d parameter set for Ti−O, Ti−Ti interactions.[32]

## III. Bulk Rutile $TiO_2$

**A. Crystal Geometry and Bulk Modulus.** The unit cell of titanium dioxide is tetragonal, with six atoms per unit cell. The geometry of the unit cell is described in terms of two lattice parameters, $a$ and $c$, in addition to the internal parameter $u$. Using SCC-DFTB we calculated the equilibrium lattice parameters to be $a = 4.70$ Å and $c = 2.92$ Å and the internal parameter to be $u = 0.300$ Å, in agreement to within 3% of the experimental values of $a = 4.58$ Å, $c = 2.95$ Å, and $u = 0.305$ Å[59] and other theoretical calculations.[60] Our results for the equilibrium lattice parameters are also consistent with the results of Luschtinetz et al.[30] ($a = 4.61$ Å, $c = 2.97$ Å, and $u = 0.302$ Å) using different SCC-DFTB parameters for bulk rutile $TiO_2$. The calculated rutile structure is stable to symmetry-breaking distortions. SCC-DFTB predicts that the titanium and oxygen atoms have atomic charges of $0.814e$ and $-0.407e$, respectively. The bulk modulus, $K$, of the material was found by varying the volume of the unit cell and taking the second derivative of the energy $U$:

$$K = V\left.\frac{\partial^2 U}{\partial V^2}\right|_0 \tag{8}$$

where the subscript zero denotes that the derivatives are to be evaluated at the equilibrium cell geometry. SCC-DFTB yielded a bulk modulus of 243 GPa, compared to an experimental value of 210 GPa.[59] DFT calculations predict a bulk modulus in slightly better agreement with experiment. With the LDA functional the bulk modulus is overestimated (230 GPa), whereas with the PBE[54] functional it is underestimated (194 GPa).[60]

**B. Electronic Structure.** Titanium dioxide is a wide-band gap semiconductor with relatively strong ionicity, being of advantage in photoassisted dissociation of molecules. The chemical bonding is governed by the interaction between oxygen 2p and titanium 3d states, compared to other common semiconductors such as ZnO and GaAs. Figure 1 shows the band structure and density of states of $TiO_2$ calculated using SCC-DFTB, where the zero of energy has been taken as the top of the valence band. The band structure was obtained by self-consistently converging charges with a tolerance of $1 \times 10^{-5}$ on a $4 \times 4 \times 8$ Monkhorst−Pack mesh[61] in **k**-space. These charges were then utilized in a nonself-consistent-charge calculation of the energy along the high-symmetry directions of the Brillouin zone. The lower valence bands, of mostly O 2s character, and upper valence bands, from hybridization between O 2*p* and Ti 3*d* states, compare qualitatively well with LDA-DFT[62] and PBE-DFT[63] results; however both the upper and lower valence band widths are underestimated. Table 1 compares numerical properties of the SCC-DFTB bulk rutile $TiO_2$ electronic structure with experiment and with DFT calculations using the LDA and PBE functionals. A direct band gap of 2.46 eV is predicted by SCC-DFTB, which agrees better with the experimental[64,65] value of 3.06 eV than LDA-DFT (2.0 eV)[62] and PBE-DFT (1.88 eV).[63] It should be noted that the LDA-DFT and PBE-DFT band structure calculations were performed using different basis sets. The biggest difference between the SCC-

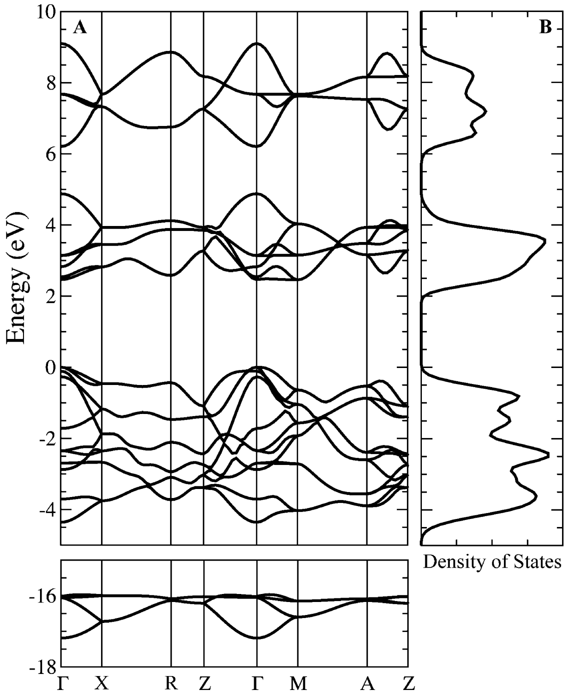


***Figure 1.*** (A) Electronic band structure along high-symmetry directions of the irreducible Brillouin zone (Γ−X−R−Z−Γ−M−A−Z). For an illustration of the Brillouin zone of $TiO_2$, see, for example, Glassford and Chelikowsky.[62] (B) Electronic density of states.

***Table 1.*** Electronic Structure of Bulk Rutile $TiO_2$

| property | SCC-DFTB | PBE-DFT[a] | LDA-DFT[b] | experiment |
|---|---|---|---|---|
| lower valence bandwidth (eV) | 1.22 | 1.79 | 1.8 | 1.9[c] |
| upper valence bandwidth (eV) | 4.36 | 5.69 | 5.7 | 5.4[c] |
| lower valence/ conduction band separation at Γ (eV) | 18.5 | 18.13 | 17 | 16 − 18[d] |
| band gap (eV) | 2.46 | 1.88 | 2.0 | 3.06[c,e] |

[a] Reference 63. [b] Reference 62. [c] Reference 64. [d] Reference 86. [e] Reference 65.

DFTB and LDA-DFT band structures is in the conduction bands, where there is a gap of 1.2 eV at around 5 eV. The LDA-DFT results do find that the lower conduction bands can be divided into two distinct groups, corresponding to $t_{2g}$ and $e_g$ d orbital states of Ti, which are separated over the entire Brillouin zone, except for a small overlap at Γ. This overlap is absent in the SCC-DFTB band structure and is also absent in other empirical tight-binding studies.[66]

**C. Vibrational Spectra.** The lattice dynamics of bulk rutile titanium dioxide have been shown to be very important in determining its technological properties. For example, it has an exceptionally high static dielectric constant along the *c*-direction, which increases as the temperature is lowered. This has been explained in terms of the transverse optic $A_{2u}$ mode, which is only 173 $cm^{-1}$ at room temperature and shifts 36 $cm^{-1}$ to the red at lower temperatures (142 $cm^{-1}$ at 4 K).[67] This softness is not observed in isostructural materials like tin dioxide or germanium dioxide (465 and 455 $cm^{-1}$, respectively).

We calculated the vibrational frequencies of bulk titania at the Γ point using, first, the dynamical matrix method, where the frequencies are equal to the eigenvalues of the matrix $D_{ij}$ defined by

$$D_{ij} = \frac{-1}{\sqrt{m_i m_j}} \frac{\partial F_j}{\partial x_i}\bigg|_0 = \frac{1}{\sqrt{m_i m_j}} \frac{\partial^2 E}{\partial x_i \partial x_j}\bigg|_0 \tag{9}$$

where $m_i$ is the mass of atom *i*, $x_i$ is the displacement of atom *i* from its equilibrium position, and the 0 subscript denotes that the derivatives are to be performed with all atoms in their equilibrium positions. $D_{ij}$ is a $3N \times 3N$ matrix, where *N* here is 6, corresponding to the six atoms in the unit cell, and the eigenvectors of the matrix correspond to the displacements of the unit cell atoms in each of the various vibrational modes. The positions of the atoms in the unit cell were optimized until all forces were less than $10^{-4}$ eV/Å, and then a displacement of 0.01 Å was made in each of the 3*N* degrees of freedom. The derivative of the force was then calculated using a finite difference approximation. Second, we computed vibrational frequencies of selected modes using the frozen phonon method, which models a single mode directly and extracts a frequency from the energy found as a function of displacement. This method can readily expose any instabilities or anharmonicity, but it requires knowledge of the symmetry and displacement vector of a nondegenerate mode of interest. In contrast, the dynamical matrix method does not require any a priori information regarding the vibrational modes. When computing vibrational frequencies with the frozen phonon method, we used three displacements of 0.01 Å in the positive and negative directions to map the potential energy (−0.03, −0.02, −0.01, 0.00, 0.01, 0.02, 0.03 Å), which was then fit to a quadratic form to obtain the vibrational frequency of the mode.

Shown in Table 2 are bulk rutile $TiO_2$ vibrational frequencies computed with SCC-DFTB using both the dynamical matrix and the frozen phonon methods (frozen phonon frequencies were computed for four selected modes: $B_{2g}$, $A_{1g}$, $E_g$, and $A_{2u}$). For comparison we also show vibrational frequencies measured with several experimental techniques, including neutron scattering,[67] Raman scattering,[68] and infrared absorption.[69] Table 2 also contains bulk rutile $TiO_2$ vibrational frequencies reported by Sikora for LDA-DFT,[70] which are consistent and therefore representative of earlier DFT calculations,[60,71] and PBE-DFT results reported by Montanari and Harrison.[60] Four modes ($A_{2u}$, $E_u^1$, $E_u^2$, and $E_u^3$) that are strongly perturbed by an interaction with the electric fields created by long wavelength vibrations of the crystal [i.e., modes that undergo longitudinal optical-transverse optical (LO-TO) splitting] are omitted from Table 2. Correcting for LO-TO splitting to make a meaningful comparison to experiment requires a more sophisticated analysis and is inconsequential to validating the accuracy of SCC-DFTB for $TiO_2$. The agreement with experiment of the SCC-DFTB vibrational frequencies is generally impressive. For the higher frequency modes (>400 $cm^{-1}$) the root-mean-squared (rms) deviation of the calculated frequencies compared to the neutron scattering measurements is slightly better for SCC-DFTB than for LDA-DFT and PBE-DFT (16.4 $cm^{-1}$ compared to 17.1 and 39.3 $cm^{-1}$). The lower frequency modes are significantly more challenging to compute correctly. For example, PBE-DFT predicts that the $A_{2u}$ mode

***Table 2.*** Calculated and Measured Bulk $TiO_2$ Phonon Frequencies (in $cm^{-1}$)

| | SCC-DFTB | | | | | |
|---|---|---|---|---|---|---|
| mode | dynamical matrix | frozen phonon | neutron scattering[67] | IR[69] and Raman[68] | PBE-DFT[60] | LDA-DFT[70] |
| $B_{2g}$ | 822 | 823 | 825 | 827 | 774 | 801 |
| $A_{1g}$ | 577 | 583 | 610 | 612 | 566 | 615 |
| $E_u^3$ | 505 | | 494 | 500 | 469 | 498 |
| $E_g$ | 447 | 448 | 445 | 447 | 429 | 472 |
| $B_{1u}^2$ | 417 | | 406 | inactive | 358 | 417 |
| $A_{2g}$ | 406 | | inactive | inactive | 424 | 413 |
| $E_u^2$ | 388 | | inactive | 388 | 354 | 393 |
| $A_{2u}$ | 235 | 186 | 173 | 167 | 86.3*i* | 191 |
| $E_u^1$ | 201 | | 189 | 183 | 124 | 144 |
| $B_{1u}^1$ | 150 | | 113 | inactive | 79 | 118 |
| $B_{1g}$ | 113 | | 142 | 143 | 154 | 132 |

is imaginary,[60] whereas SCC-DFTB and LDA-DFTB both predict that this mode is stable. Although SCC-DFTB is less accurate than LDA-DFT for the lowest frequency modes, the overall level of quantitative accuracy is impressive considering that the SCC-DFTB calculations are substantially faster than DFT.

In addition, we calculated an approximation to the infrared absorption spectrum using molecular dynamics (MD) simulations. The infrared absorption spectrum, $I(\omega)$, is proportional to the Fourier transform of the quantum mechanical electric dipole moment time-correlation function (TCF):[72,73]

$$I(\omega) \sim \int_{-\infty}^{\infty} \mathrm{d}t\, \mathrm{e}^{-i\omega t} \frac{\mathrm{Tr}[\mathrm{e}^{-\beta\hat{H}}\hat{\mu}(0)\,\hat{\mu}(t)]}{\mathrm{Tr}[\mathrm{e}^{-\beta\hat{H}}]} \qquad (10)$$

where $\hat{\mu}$ denotes the electric dipole moment operator of the system, $\hat{H}$ is the Hamiltonian operator for the nuclear and electronic degrees of freedom of the system, $\beta$ is the inverse temperature times Boltzmann's constant ($\beta = 1/k_BT$), and Tr denotes a quantum mechanical trace over all degrees of freedom. Since eq 10 is generally computationally intractable, we utilized the classical dipole approximation[72-75] to replace the quantum mechanical electric dipole moment TCF by its classical analog

$$I(\omega) \sim Q(\omega) \int_{-\infty}^{\infty} \mathrm{d}t\, \mathrm{e}^{-i\omega t} \langle \mu(0)\mu(t)\rangle \qquad (11)$$

making it amenable to calculation within MD simulations. In eq 11 $\mu$ is the classical electric dipole moment of the system, the angular brackets represent a classical equilibrium ensemble average, and $Q(\omega)$ is a frequency dependent correction factor to compensate approximately for replacing a quantum mechanical TCF with a classical TCF.[76] In the present application, where we are primarily interested in peak locations of relatively low frequency vibrations (<850 $cm^{-1}$), we will assume $Q(\omega) = 1$. Since even computing the classical electric dipole TCF is nontrivial because it tends to be a slowly converging quantity, we instead invoked one further common approximation and calculated the Fourier transform of the velocity TCF

$$I_v(\omega) \sim \int_{-\infty}^{\infty} \mathrm{d}t\, \mathrm{e}^{-i\omega t} \langle v(0)v(t)\rangle \qquad (12)$$

where $v$ is the $3N$-dimensional velocity vector. This last approximation has the disadvantage that the magnitude of the absorption intensity will not necessarily be accurate compared with experiment, but the location of the peaks in the spectrum should be maintained. For a further discussion of the approximations required to arrive at eq 12, see, for example, the discussion in the appendix of Lobaugh and Voth.[77] The advantage of eq 12 is that the phonon spectrum can readily be computed from MD simulations. In our MD simulations the system was first equilibrated for 1 ps in the NVT ensemble using an Andersen thermostat and for 1 ps without the theormostat.[78] The simulation then proceeded in the NVE ensemble, where the average temperature was observed to remain stable at the targeted temperature throughout the simulation. Production runs were between 20 and 50 ps with a time step of 1 fs. A slowly decaying exponential function was applied to the velocity TCF before it was Fourier transformed to force it smoothly to zero at long time. The spectra reported were completely insensitive to the details of this exponential function.

Figure 2 shows $I_v(\omega)$ at 30 and 300 K, where molecular dynamics has been performed on the six-atom unit cell. Restricting the size of the system to the primitive unit cell limits the spectrum to only reflect frequencies at the $\Gamma$ point, thus ignoring dispersion effects. Temperature broadening of the peaks due to anharmonicity[74] is observed at 300 K. There is generally excellent agreement between the location of peaks in $I_v(\omega)$ compared to those computed with the dynamical matrix and/or frozen phonon methods. For example, the peak at 821 $cm^{-1}$ corresponds to the highest frequency $B_{2g}$ mode calculated as 822 and 823 $cm^{-1}$ with the dynamical matrix and frozen phonon methods, respectively. Figure 3 shows $I_v(\omega)$ at 30 and 300 K, where MD has been performed on a 3 × 3 × 3 supercell containing 162 atoms. The temperature broadening becomes more pronounced and more modes appear due to dispersion, as the supercell effectively includes more points in the Brillouin zone.

## IV. $TiO_2$(110) Surface

**A. Surface Geometry.** The most stable termination of titania is the stoichiometric (110) surface, shown in Figure 4. The $TiO_2$(110) surface has a plane containing two titanium atoms (one 6-fold and one 5-fold coordinated) and two oxygen atoms per unit cell. One oxygen resides 1.3 Å above this plane (the bridging oxygen), and another oxygen sits similarly below the plane. These six atoms constitute the

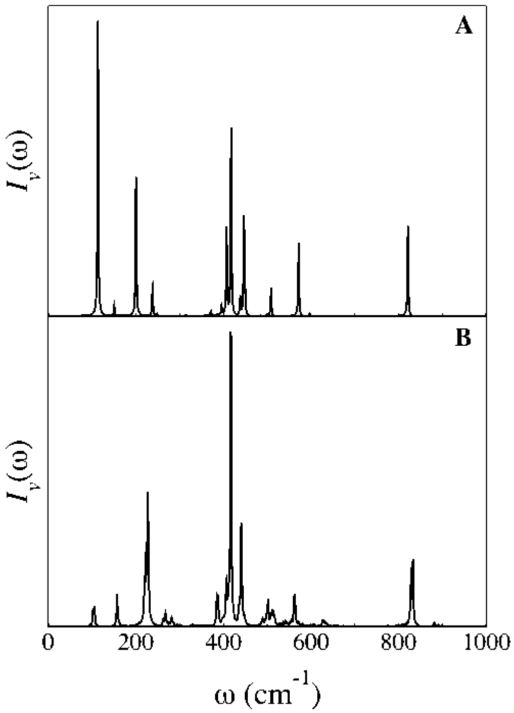


***Figure 2.*** Phonon frequency spectrum, $I_v(\omega)$, calculated from the velocity time autocorrelation function, eq 12, at (A) 30 K and (B) 300 K for the six-atom rutile $TiO_2$ unit cell.

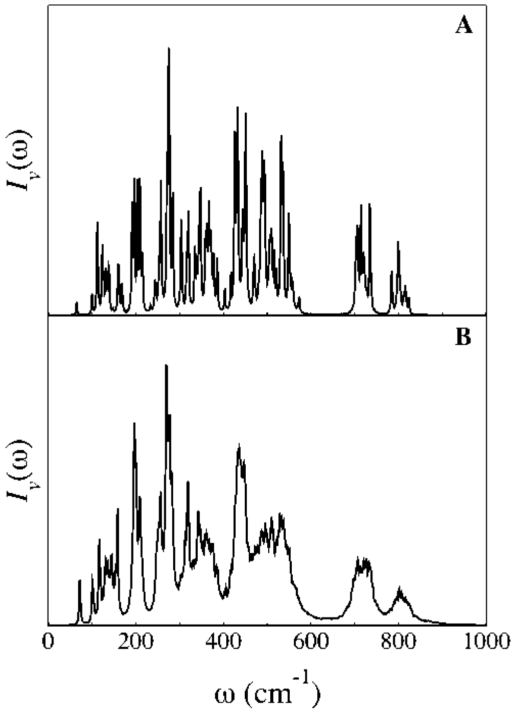


***Figure 3.*** Phonon frequency spectrum, $I_v(\omega)$, calculated from the velocity time autocorrelation function, eq 12, at (A) 30 K and (B) 300 K for the a 162-atom 3 × 3 × 3 supercell of rutile $TiO_2$.

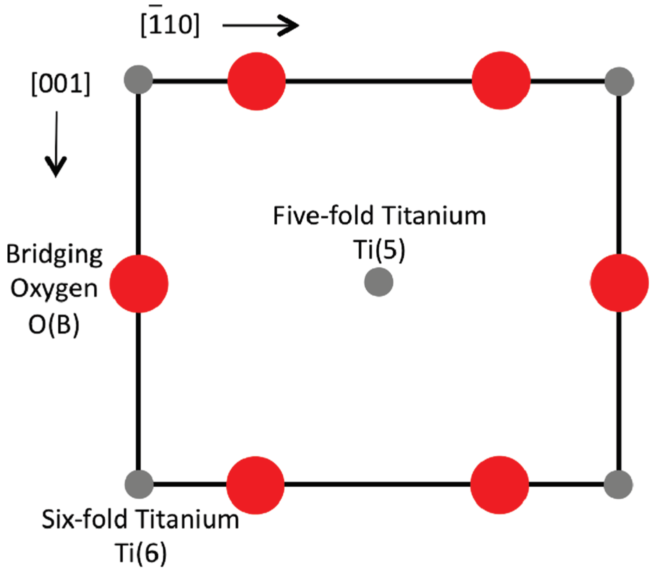


***Figure 4.*** Schematic view of the idealized $TiO_2(110)$ surface. Oxygen atoms are shown in red and titanium atoms in gray. All of the atoms are in the same plane, except for the bridging oxygens, O(B), which lie above the plane.

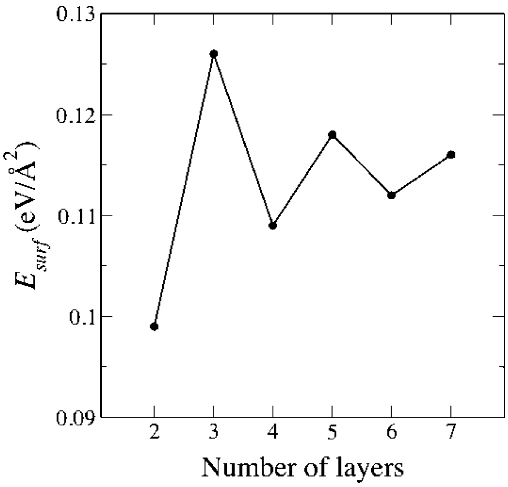


***Figure 5.*** Convergence of the surface energy, $E_{surf}$, with the number of layers in the slab.

top layer, and lower layers are offset in the [$\bar{1}$10] direction by half a cell length. The first DFT studies of $TiO_2(110)$ found that relaxations of the surface atoms from their bulk-terminated positions were substantial and responsible for a large reduction in calculated surface energies.[79,80] Many studies since have observed a strong dependence of relaxations and surface energy on slab thickness.[45,48,50,81] Figure 5 shows the convergence of the surface energy, $E_{surf} = (E_{slab} - E_{bulk})/2A$ ($E_{slab}$ is the energy of the slab, $E_{bulk}$ is the energy of an equivalent quantity of the bulk, and $A$ is the surface area) with an increasing number of layers. It fluctuates about the converged value due to the structural difference between slabs with even and odd numbers of layers. Odd-layered slabs have a symmetry plane through the central layer, while this symmetry is absent in even-layered slabs. For sufficiently large slabs, these differences should cease to matter. The surface energy is reasonably well converged at seven layers, with a value of 0.116 eV/Å$^2$. Some recent DFT studies agree that the surface energy is approximately 0.056 eV/Å$^2$.[50,81] However, the exact value of $E_{surf}$ depends strongly on the method and exchange-correlation functional used, dropping down to 0.030 eV/Å$^2$ with the PBE[54] functional.[82] Nevertheless, SCC-DFTB appears to overestimate the surface energy substantially.

***Table 3.*** Displacements of Surface Atoms, Perpendicular to the Surface, in Å as a Function of the Number of Layers in the Slab, *N*

| | $E_{surf}$ (eV/Å$^2$) | Ti(6) | Ti(5) | O(IP) | O(B) |
|---|---|---|---|---|---|
| $N = 4$ | 0.109 | 0.30 | −0.22 | 0.14 | 0.07 |
| $N = 5$ | 0.118 | 0.20 | −0.16 | 0.16 | −0.02 |
| $N = 6$ | 0.112 | 0.28 | −0.19 | 0.16 | 0.05 |
| $N = 7$ | 0.116 | 0.23 | −0.17 | 0.16 | 0.01 |
| previous SCC-DFTB[30] | | 0.27 | −0.11 | 0.22 | 0.07 |
| PBE-DFT ($N = 8$) | 0.029 | 0.35 | −0.09 | 0.27 | 0.16 |
| experiment[51] | | 0.25 | −0.19 | 0.27 | 0.10 |

***Table 4.*** Atomic Charges in Units of *e* of Surface and Internal Layer Atoms for the $N = 7$ $TiO_2$(110) Slab[a]

| | Ti(6) | Ti(5) | O(IP) | O(B) | O(I) |
|---|---|---|---|---|---|
| surface layer | 0.76 | 0.86 | −0.40 | −0.35 | −0.45 |
| middle layer | 0.82 | 0.81 | −0.41 | −0.39 | −0.39 |

[a] Charges in the bulk are 0.814*e* for Ti and −0.407*e* for O. O(I) corresponds to the subsurface oxygen directly under the bridging oxygen in the six-atom surface unit.

Table 3 lists the displacements of the surface atoms perpendicular to the surface. These atoms relax inward or outward from their bulk terminated positions, and the number of layers in the slab can greatly affect the size of these movements. A seven-layer slab, using SCC-DFTB, is reasonably well converged, and the displacements agree well with PBE-DFT calculations. Although some debate remains about the exact relaxations at the (110) surface,[83] the LEED results[51] shown in the Table 3 are representative and are well reproduced by our SCC-DFTB calculations. The atomic displacements normal to the surface are also consistent with the previous SCC-DFTB calculations (with different parameters) of Luschtinetz et al.[30] also shown in Table 3. Table 4 gives the atomic charges of the surface and middle layer atoms for the seven layer slab. Compared with the bulk, the 5-fold titanium has increased in charge, while the 6-fold titanium and bridging oxygen have decreased in charge. The charges on the middle layer agree with bulk charges to within 0.02*e*, confirming that the slab is large enough for the interior to behave like the bulk material.

**B. $TiO_2$(110) Electronic Structure.** Previous studies have widely agreed that no surface electronic states are observed or predicted for $TiO_2$(110).[44] Figure 6 shows the electronic density of states for a four-layer slab of $TiO_2$(110) containing 24 atoms calculated with SCC-DFTB, compared to that of the bulk. The band structure was obtained by self-consistently converging charges with a tolerance of $1 \times 10^{-5}$ on a $2 \times 4 \times 1$ Monkhorst−Pack mesh[61] in **k**-space. These charges were then utilized in a non-self-consistent-charge calculation of the energy along the high symmetry directions of the Brillouin zone. As expected, the presence of the stoichiometric surfaces in the supercell does not have an appreciable effect on the electronic structure; the valence and lower conduction bands have a similar number of peaks in terms of location and magnitude as the bulk, and the direct band gap of 2.46 eV is maintained. The gap between the $t_{2g}$ and $e_g$ conduction bands at around 5 eV shows the largest difference, decreasing from a gap of approximately 1 eV to less than 0.5 eV.

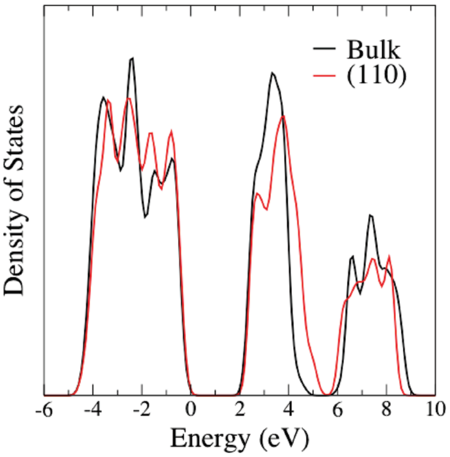


***Figure 6.*** Comparison of the electronic density of states between bulk $TiO_2$ and a slab of $TiO_2$(110). The number of atoms in both systems is 24. The zero is taken as the Fermi energy of the bulk system.

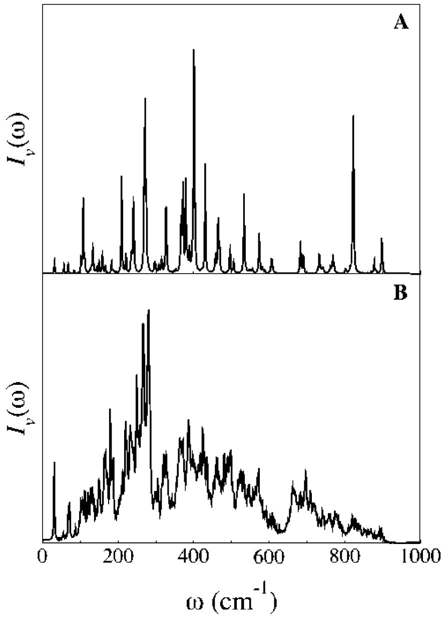


***Figure 7.*** Phonon frequency spectra, $I_v(\omega)$, calculated from the velocity time autocorrelation function, eq 12. (A) 30 K with a $TiO_2$ (110) slab containing 24 atoms and (B) 300 K with a $TiO_2$(110) slab containing 96 atoms.

**C. $TiO_2$(110) Vibrational Spectra.** Two experimental measurements of the (110) phonon spectrum have appeared in the literature using electron energy loss spectroscopy (EELS)[84] and high-resolution EELS (HREELS).[85] The HREELS study found three bands at approximately 365, 445, and 755 cm$^{-1}$, which were consistent with the earlier EELS measurements. Figure 7 shows the phonon frequency spectrum, $I_v(\omega)$, calculated as the Fourier transform of the velocity TCF (eq 12) of (A) a 30 ps NVE MD simulation $TiO_2$ (110) slab containing 24 atoms equilibrated to 30 K and (B) a 20 ps MD simulation of a $TiO_2$(110) slab containing 96 atoms at 300 K. The spectra retain some of the features of the bulk phonon spectra (Figures 2 and 3), particularly the characteristic gap around 640 cm$^{-1}$. A direct comparison of the calculated spectra to experiment is not possible without a detailed model for intensities of the surface modes as measured with HREELS. We also performed

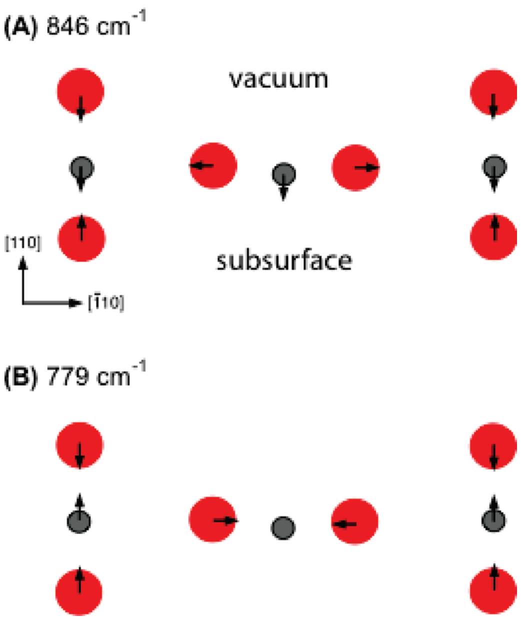


***Figure 8.*** Schematic illustration of two surface vibration modes found using the dynamical matrix method for $TiO_2$(110). The view is perpendicular to the plane of the surface, and oxygen atoms are shown in red and titanium atoms in gray.

a dynamical matrix calculation on a subset of a 24-atom slab of $TiO_2$(110), corresponding to the six atoms of one surface. The two highest frequency modes (846 and 779 cm$^{-1}$) involve motions of the bridging oxygen, 6-fold coordinated titanium, and a subsurface oxygen perpendicular to the plane of the (110) surface, as well as motions of the in-plane oxygen atoms parallel to the surface (Figure 8). Another mode of interest involves the opposite motion of the in-plane oxygens and 5-fold Ti atoms (428 cm$^{-1}$).

## V. Conclusions

SCC-DFTB is a computationally efficient semiempirical tight-binding method, which has been successfully applied to crystals and crystal surfaces, as well as biological systems. Because it is relatively inexpensive compared with conventional DFT methods, it is possible to simulate systems containing large numbers of atoms or to perform long molecular dynamics simulations. Titanium dioxide has been extensively studied by DFT, but cell sizes have mostly been limited to contain under 100 atoms, and molecular dynamics run lengths are restricted to a few tens of picoseconds. DFTB+ molecular dynamics simulations based on SCC-DFTB forces are hundreds of times faster than equivalent VASP calculations based on DFT forces. Although SCC-DFTB parameters have not yet been parametrized specifically for titania, we have demonstrated that existing parameters for titanium in biological systems reproduce properties of the material well. A band gap of 2.46 eV is predicted, closer to experiment than DFT results. Other electronic properties and vibrational mode frequencies also agree well with experiment and theory. Relaxations of (110) surface atoms are well-reproduced, and although the surface energy is overestimated by perhaps a factor of 2, this property has been seen to be strongly method-dependent and is therefore not necessarily of large concern. On the basis of these encouraging results, computationally efficient investigations of more complex phenomena involving the $TiO_2$ surface (e.g., reactions of adsorbate molecules) with SCC-DFTB are justified and are presently being pursued.

**Acknowledgment.** The authors are grateful for support from the University of Notre Dame Energy Center and the University of Notre Dame Center for Research Computing.

### References

(1) Seifert, G. *J. Phys. Chem. A* **2007**, *111*, 5609.

(2) Seifert, G.; Porezag, D.; Frauenheim, T. *Int. J. Quantum Chem.* **1996**, *58*, 185.

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

(4) Porezag, D.; Frauenheim, T.; Kohler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, *51*, 12947.

(5) Frauenheim, T.; Seifert, G.; Elstner, M.; Niehaus, T.; Kohler, C.; Amkreutz, M.; Sternberg, M.; Hajnal, Z.; Di Carlo, A.; Suhai, S. *J. Phys.-Condens. Mat.* **2002**, *14*, 3015.

(6) Moreira, N. H.; Dolgonos, G.; Aradi, B.; da Rosa, A. L.; Frauenheim, T. *J. Chem. Theory Comput.* **2009**, *5*, 605.

(7) Rauls, E.; Elsner, J.; Gutierrez, R.; Frauenheim, T. *Solid State Commun.* **1999**, *111*, 459.

(8) Kohler, C.; Frauenheim, T. *Surf. Sci.* **2006**, *600*, 453.

(9) Lin, C. S.; Zhang, R. Q.; Lee, S. T.; Elstner, M.; Frauenheim, T.; Wan, L. J. *J. Phys. Chem. B* **2005**, *109*, 14183.

(10) Elstner, M. *Theor. Chem. Acc.* **2006**, *116*, 316.

(11) Elstner, M.; Frauenheim, T.; Kaxiras, E.; Seifert, G.; Suhai, S. *Phys. Status Solidi B* **2000**, *217*, 357.

(12) Elstner, M.; Jalkanen, K. J.; Knapp-Mohammady, M.; Frauenheim, T.; Suhai, S. *Chem. Phys.* **2000**, *256*, 15.

(13) Elstner, M.; Jalkanen, K. J.; Knapp-Mohammady, M.; Frauenheim, T.; Suhai, S. *Chem. Phys.* **2001**, *263*, 203.

(14) Hu, H.; Elstner, M.; Hermans, J. *Proteins* **2003**, *50*, 451.

(15) Jalkanen, K. J.; Elstner, M.; Suhai, S. *J. Mol. Struct. THEOCHEM* **2004**, *675*, 61.

(16) Kumar, A.; Elstner, M.; Suhai, S. *Int. J. Quantum Chem.* **2003**, *95*, 44.

(17) Kumar, A.; Knapp-Mohammady, M.; Mishra, P. C.; Suhai, S. *J. Comput. Chem.* **2004**, *25*, 1047.

(18) Liu, H. Y.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Hermans, J.; Yang, W. T. *Proteins* **2001**, *44*, 484.

(19) Okada, T.; Sugihara, M.; Bondar, A. N.; Elstner, M.; Entel, P.; Buss, V. *J. Mol. Biol.* **2004**, *342*, 571.

(20) Reha, D.; Kabelac, M.; Ryjacek, F.; Sponer, J.; Sponer, J. E.; Elstner, M.; Suhai, S.; Hobza, P. *J. Am. Chem. Soc.* **2002**, *124*, 3366.

(21) Shishkin, O. V.; Elstner, M.; Frauenheim, T.; Suhai, S. *Int. J. Mol. Sci.* **2003**, *4*, 537.

(22) Shishkin, O. V.; Gorb, L.; Luzanov, A. V.; Elstner, M.; Suhai, S.; Leszczynski, J. *J. Mol. Struct. THEOCHEM* **2003**, *625*, 295.

(23) Sugihara, M.; Buss, V.; Entel, P.; Elstner, M.; Frauenheim, T. *Biochemistry* **2002**, *41*, 15259.

(24) Wanko, M.; Hoffmann, M.; Strodel, P.; Koslowski, A.; Thiel, W.; Neese, F.; Frauenheim, T.; Elstner, M. *J. Phys. Chem. B* **2005**, *109*, 3606.

(25) Zhou, H. Y.; Tajkhorshid, E.; Frauenheim, T.; Suhai, S.; Elstner, M. *Chem. Phys.* **2002**, *277*, 91.

(26) Yu, H. B.; Cui, Q. *J. Chem. Phys.* **2007**, *127*, 234504.

(27) Hu, H.; Lu, Z. Y.; Elstner, M.; Hermans, J.; Yang, W. T. *J. Phys. Chem. A* **2007**, *111*, 5685.

(28) Enyashin, A. N.; Ivanovskii, A. L. *J. Mol. Struct. THEOCHEM* **2006**, *766*, 15.

(29) Enyashin, A. N.; Seifert, G. *Phys. Status Solidi B* **2005**, *242*, 1361.

(30) Luschtinetz, R.; Frenzel, J.; Milek, T.; Seifert, G. *J. Phys. Chem. C* **2009**, *113*, 5730.

(31) http://www.dftb.org.

(32) Zheng, G.; Witek, H. A.; Bobadova-Parvanova, P.; Irle, S.; Musaev, G.; Prabhakar, R.; Morokuma, K.; Lundberg, M.; Elstner, M.; Kohler, C.; Frauenheim, T. *J. Chem. Theory Comput.* **2007**, *3*, 1349.

(33) Fox, M. A.; Dulay, M. T. *Chem. Rev.* **1993**, *93*, 341.

(34) Fujishima, A.; Honda, K. *Nature* **1972**, *238*, 37.

(35) Fujishima, A.; Rao, T. N.; Tryk, D. A. *J. Photochem. Photobiol. C* **2000**, *1*, 1.

(36) Harada, H.; Ueda, T. *Nouv. J. Chim.* **1984**, *8*, 123.

(37) Hoffmann, M. R.; Martin, S. T.; Choi, W. Y.; Bahnemann, D. W. *Chem. Rev.* **1995**, *95*, 69.

(38) Kamat, P. V. *Chem. Rev.* **1993**, *93*, 267.

(39) Kawai, M.; Naito, S.; Tamaru, K.; Kawai, T. *Chem. Phys. Lett.* **1983**, *98*, 377.

(40) Kawai, T.; Sakata, T. *J. Chem. Soc. Chem. Commun.* **1980**, 694.

(41) Linsebigler, A. L.; Lu, G. Q.; Yates, J. T. *Chem. Rev.* **1995**, *95*, 735.

(42) Sakata, T. *J. Photochem.* **1985**, *29*, 205.

(43) Sakata, T.; Kawai, T. *Chem. Phys. Lett.* **1981**, *80*, 341.

(44) Diebold, U. *Surf. Sci. Rep.* **2003**, *48*, 53.

(45) Bates, S. P.; Kresse, G.; Gillan, M. J. *Surf. Sci.* **1997**, *385*, 386.

(46) Harrison, N. M.; Wang, X. G.; Muscat, J.; Scheffler, M. *Faraday Discuss.* **1999**, *114*, 305.

(47) Lindan, P. J. D.; Harrison, N. M.; Gillan, M. J.; White, J. A. *Phys. Rev. B* **1997**, *55*, 15919.

(48) Bredow, T.; Giordano, L.; Cinquini, F.; Pacchioni, G. *Phys. Rev. B* **2004**, *70*, 035419.

(49) Muscat, J.; Harrison, N. M.; Thornton, G. *Phys. Rev. B* **1999**, *59*, 2320.

(50) Swamy, V.; Muscat, J.; Gale, J. D.; Harrison, N. M. *Surf. Sci.* **2002**, *504*, 115.

(51) Lindsay, R.; Wander, A.; Ernst, A.; Montanari, B.; Thornton, G.; Harrison, N. M. *Phys. Rev. Lett.* **2005**, *94*, 246102.

(52) Charlton, G.; Howes, P. B.; Nicklin, C. L.; Steadman, P.; Taylor, J. S. G.; Muryn, C. A.; Harte, S. P.; Mercer, J.; McGrath, R.; Norman, D.; Turner, T. S.; Thornton, G. *Phys. Rev. Lett.* **1997**, *78*, 495.

(53) Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981**, *23*, 5048.

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

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

(56) Foulkes, W. M. C.; Haydock, R. *Phys. Rev. B* **1989**, *39*, 12520.

(57) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H. B.; Ghosh, N.; Prat-Resina, X.; Konig, P.; Li, G. H.; Xu, D. G.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, *110*, 6458.

(58) Aradi, B.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007**, *111*, 5678.

(59) Burdett, J. K.; Hughbanks, T.; Miller, G. J.; Richardson, J. W.; Smith, J. V. *J. Am. Chem. Soc.* **1987**, *109*, 3639.

(60) Montanari, B.; Harrison, N. M. *Chem. Phys. Lett.* **2002**, *364*, 528.

(61) Monkhorst, H. J.; Pack, J. D. *Phys. Rev. B* **1976**, *13*, 5188.

(62) Glassford, K. M.; Chelikowsky, J. R. *Phys. Rev. B* **1992**, *46*, 1284.

(63) Labat, F.; Baranek, P.; Domain, C.; Minot, C.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 154703.

(64) Kowalczyk, S. P.; McFeely, F. R.; Ley, L.; Gritsyna, V. T.; Shirley, D. A. *Solid State Commun.* **1977**, *23*, 161.

(65) Pascual, J.; Camassel, J.; Mathieu, H. *Phys. Rev. B* **1978**, *18*, 5606.

(66) Vos, K. *J. Phys. C Solid State* **1977**, *10*, 3917.

(67) Trayler, J. G.; Smith, H. G.; Nicklow, R. M.; Wilkinso, Mk. *Phys. Rev. B* **1971**, *3*, 3457.

(68) Porto, S. P. S.; Fleury, P. A.; Damen, T. C. *Phys. Rev.* **1967**, *154*, 522.

(69) Eagles, D. M. *J. Phys. Chem. Solids* **1964**, *25*, 1243.

(70) Sikora, R. *J. Phys. Chem. Solids* **2005**, *66*, 1069.

(71) Lee, C.; Ghosez, P.; Gonze, X. *Phys. Rev. B* **1994**, *50*, 13379.

(72) Gordon, R. G. *J. Chem. Phys.* **1965**, *43*, 1307.

(73) McQuarrie, D. A. *Statistical Mechanics*; University Science Books: Sausalito, CA, 2000.

(74) Schmidt, J. R.; Corcelli, S. A. *J. Chem. Phys.* **2008**, *128*, 184504.

(75) Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1981**, *74*, 4872.

(76) Egorov, S. A.; Everitt, K. F.; Skinner, J. L. *J. Phys. Chem. A* **1999**, *103*, 9494.

(77) Lobaugh, J.; Voth, G. A. *J. Chem. Phys.* **1997**, *106*, 2400.

(78) Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384.

(79) Ramamoorthy, M.; Kingsmith, R. D.; Vanderbilt, D. *Phys. Rev. B* **1994**, *49*, 7709.

(80) Ramamoorthy, M.; Vanderbilt, D.; Kingsmith, R. D. *Phys. Rev. B* **1994**, *49*, 16721.

(81) Hameeuw, K. J.; Cantele, G.; Ninno, D.; Trani, F.; Iadonisi, G. *J. Chem. Phys.* **2006**, *124*, 024708.

(82) Kiejna, A.; Pabisiak, T.; Gao, S. W. *J. Phys.-Condens. Mat.* **2006**, *18*, 4207.

(83) Pang, C. L.; Lindsay, R.; Thorton, G. *Chem. Soc. Rev.* **2008**, *37*, 2328.

(84) Rocker, G.; Schaefer, J. A.; Gopel, W. *Phys. Rev. B* **1984**, *30*, 3704.

(85) Henderson, M. A. *Surf. Sci.* **1996**, *355*, 151.

(86) Knotek, M. L.; Feibelman, P. J. *Phys. Rev. Lett.* **1978**, *40*, 964.

CT900665A

JCTC Journal of Chemical Theory and Computation

# Diffusion of Hydrides in Palladium Nanoclusters. A Ring-Polymer Molecular Dynamics Study of Quantum Finite Size Effects

F. Calvo* and D. Costa†

*Laboratoire de Spectrométrie Ionique et Moléculaire (LASIM), Université Claude Bernard Lyon 1 and Centre National de la Recherche Scientifique (CNRS) UMR 5579, Bât. A. Kastler, 43 Boulevard du 11 Novembre 1918, F69622 Villeurbanne, France*

Received October 20, 2009

**Abstract:** The diffusion kinetics of hydrogen in bulk palladium and in Pd nanoclusters containing up to 512 atoms has been theoretically investigated at 3% loading using ring-polymer molecular dynamics simulations. The electronic ground-state energy surfaces are modeled using an explicit many-body potential fitted to reproduce the properties of bulk palladium and palladium hydrides. The diffusion constant, calculated by integration of the velocity autocorrelation function, shows Arrhenius behavior with inverse temperature. In addition, both the prefactor and activation energy are found to exhibit approximately linear variations with inverse cluster radius for sizes exceeding 128 Pd atoms. Vibrational delocalization generally enhances diffusion, this effect being stronger in clusters than in bulk. An inherent structure analysis from the positions of the centroids was used to characterize the diffusion mechanisms. Quantum effects lead to not only a higher coordination of hydrogen atoms both in bulk (fcc) palladium and in clusters but also favor further softening of the outer layers.

## 1. Introduction

Upon absorption of hydrogen, bulk palladium changes its mechanical and thermodynamical properties to a significant extent.[1−3] The two hydrides exhibited by this metal, namely the $\alpha$ phase at low concentration and the $\beta$ phase at high concentration, should be considered as being similar to a solid solution and a defective NaCl rocksalt structure, respectively. They are separated from each other by a so-called miscibility gap associated with a phase transformation.[1] The amount of hydrogen that can be naturally absorbed is particularly high and reaches around 70% at saturation concentration,[1,4] which makes palladium a model system for hydrogen storage.[5] Palladium surfaces have interest of their own in the field of catalysis, with applications such as olefin hydrogenation or ammonia synthesis.[6] The catalytic efficiency of palladium can be magnified by further reducing the dimensionality and by studying clusters or nanoparticles, due to their higher surface/volume ratio. The higher sorption ability has been demonstrated by Huang and co-workers[7] for Pd nanoparticles smaller than 10 nm in diameter and supported on silica surfaces. More recently, Rather and co-workers[8] even found that an hyperstoichiometric concentration of 1.12 could be reached for surfactant-stabilized Pd nanoparticles of comparable sizes.

Nanometer palladium particles interacting with hydrogen have also been investigated for their structural and sorption properties.[9−15] Hydrogen-induced transitions between cubic and icosahedral clusters have been reported experimentally[10] and studied theoretically,[13] which could open some possible ways of controlling the nanoparticle shape by varying the external hydrogen pressure. The lattice expansion of Pd nanoparticles upon hydrogen absorption has been characterized by diffraction (X-ray and synchrotron radiation) techniques,[10,11,14] and one application to hydrogen sensor through tunneling has been proposed by van Lith and co-workers.[16] The interplay between the lattice size and the miscibility gap has been shown to depend on the presence

* Corresponding author. Telephone: 33 4 72 44 83 14. E-mail: fcalvo@lasim.univ-lyon1.fr.

† Present address: Groupe Métallurgie, MMC, EDF, Les Renardières, F-77818 Moret-sur-Loing, France.

10.1021/ct900554r
Published on Web 01/13/2010

of encapsulating surfactant molecules.[12,14] Very recently, Di Vece and co-workers[15] found that hydrogen adsorption can lead to Ostwald ripening of a film assembly of Pd nanoparticles at room temperature. These last examples further emphasize the important role of the surface on the hydrogenation process.

The mechanism for hydrogen diffusion in bulk palladium has been identified as a hopping motion between octahedral sites through intermediate tetrahedral sites,[17,18] and the variations of the diffusion constant $D(T)$ generally show Arrhenius behavior with inverse temperature:

$$D(T) \simeq D_0 \exp\left[-\frac{A}{k_B T}\right] \tag{1}$$

where the prefactor $D_0$ accounts for a typical attempt frequency, and $A$ is an activation energy. The Arrhenius behavior also holds for deuterium, despite this isotope is known to occupy preferentially tetrahedral sites due to less favorable zero-point energy correction.[18−20] Most theoretical works aimed at characterizing the diffusion constant have so far relied on harmonic transition-state theory (TST),[21,22] which explicitly employs such an Arrhenius form. This approach, together with conflicting experimental measurements, has been criticized in the recent review by Jewell and Davis.[23] The static TST seems justified by the complexity of the overall diffusion process, since it can operate using accurate energetics obtained with methods that explicitly account for electronic structure and includes quantum tunneling effects using dedicated models.[24,25] Harmonic TST is also useful for studying hydrogen diffusion on free Pd surfaces[26,27] but cannot deal in itself with nonperiodic systems, such as nanoparticles. One extension of TST, recently applied by Hao and Sholl[28] to diffusion in amorphous $Fe_3B$, consists of sampling local minima connected by transition states and performing a kinetic Monte Carlo simulation of the hydrogen diffusion process. This approach allows multiple hydrogen atoms to be dealt with simultaneously and can also be carried out at first-principle levels using modern molecular dynamics techniques. However, it still neglects the motion of palladium atoms, which may be particularly important near the free surfaces of nanoparticles.

Diffusion constants for hydrogen motion in fcc palladium have also been calculated by direct molecular dynamics simulations by various authors,[17,29,30] who used explicit potentials. The molecular dynamics approach treats hydrogen atoms as classical particles because a full quantum treatment is practically unfeasible for this problem. Wavepacket simulations could be performed for $H_2$ molecules interacting with free Pd surfaces,[31,32] but this method is limited to very few degrees of freedom and, in particular, also assumes host Pd atoms to remain fixed. Semiclassical techniques, such as those based on the path-integral representation of quantum mechanics,[33] are a powerful alternative which can address the aforementioned limitations. Classical and quantum Monte Carlo simulations have been performed by Chen and co-workers,[34] who investigated the stable structures, the adsorption sites, and the temperature effects in small Pd−H clusters. Path-integral methods have also been used by Forsythe and Makri to study diffusion of hydrogen and deuterium in crystalline silicon.[35]

In the present work, the diffusion of hydrogen in palladium has been investigated using quantum ring-polymer molecular dynamics (RPMD).[36] The RPMD method provides exact results for the quantum dynamics in the limit of harmonic systems or in short time scales[38] and is accurate for the velocity autocorrelation function up to a leading error of $\mathcal{O}(t^6)$, which is better than the corresponding error of the centroid molecular dynamics of Cao and Voth[37] that scales as $\mathcal{O}(t^4)$. Since its introduction, the RPMD method has been applied to several quantum diffusion processes,[39,40] particularly for the similar problem of hydrogen impurities in condensed water,[41] and we use it here for both fcc palladium and Pd nanoparticles. Our main motivation is to characterize the extent of finite size effects on the diffusion constant of hydrogen in palladium, following our previous investigations on the structural and dynamical properties of Pd−H clusters.[13,42] Our main result is that, similar to bulk palladium, the diffusion of hydrogen in Pd nanoclusters follows Arrhenius behavior but with prefactors and activation energies that strongly depend on the number of Pd atoms. Above some size, both quantities are found to vary approximately linearly with inverse cluster radius, a well-known manifestation of cluster size effects in the scaling regime.[43]

The article is organized as follows: In Section II, we briefly review the potential energy surface chosen to model the interaction among Pd and H atoms and give some details about our implementation of the RPMD method for the present systems. The results on the diffusion constants are given and discussed in Section III, where scaling laws relating the Arrhenius parameters to the cluster size are proposed. In this section, we also attempt to analyze the role of quantum vibrational effects on the diffusion mechanisms by looking at the instantaneous inherent structures, thus getting insight into the nature of adsorption sites in both bulk and finite Pd systems. A summary and some perspectives finally end the paper in Section IV.

## II. Methods

Our computational study relies on the recently introduced ring-polymer molecular dynamics technique.[36] Because the systems we deal with are rather large (up to several hundreds of atoms), we could not afford an explicit description of the electronic structure of palladium hydrides and turned instead to semiempirical many-body potentials. Several groups have employed models of similar complexity in previous work on Pd−H systems in both bulk[17,29,30,44] and nanoscale[9,13,34,42,45] forms.

**A. Potential.** The many-body alloy (MBA) model of Zhong and co-workers[46] served as a template for the present investigation. This potential is based on the second moment approximation to the electronic density of states in the tight-binding model and expresses the cohesion energy of the $N$-atom system with configuration $\mathbf{R} = \{x_i, y_i, z_i\}$ as

$$V(\mathbf{R}) = \sum_{i<j} \varepsilon_{ij} \exp\left[-p_{ij}\left(\frac{r_{ij}}{r_{ij}^0} - 1\right)\right] - \sum_i \left\{\sum_{j\neq i} \zeta_{ij}^2 \exp\left[-2q_{ij}\left(\frac{r_{ij}}{r_{ij}^0} - 1\right)\right]\right\}^{1/2} \qquad (2)$$

where $\varepsilon$, $\zeta$, $p$, $q$, and $r^0$ are 3 × 5 parameters defined for all pairs of elements Pd−Pd, H−H, and Pd−H. These parameters were originally optimized to reproduce mechanical properties of bulk palladium hydrides and of some energetic properties of H and $H_2$ adsorbed on ⟨001⟩ and ⟨110⟩ Pd surfaces, all computed using density functional theory (DFT) within the local density approximation.[46] As shown in a previous contribution,[13] this original potential does not perform so well for hydrides absorbed in fcc palladium, at least when compared with more sophisticated recent calculations.[22] A much better agreement could be reached simply by borrowing the Pd−Pd parameters from another work by Rey and co-workers,[47] employing a similar expression as eq 2 but for pure palladium. The new set of combined parameters are given in Table I.

The new potential predicts the correct lattice size (3.89 Å) and the cohesion energy (3.91 eV/atom) for pure fcc palladium at 0 K. The most significant improvement over the initial MBA model of Zhong et al. lies in the binding energy of hydrogen in fcc palladium, which changes from −0.79 to −0.15 eV with the new parameters (octahedral site), after taking zero-point energy corrections into account at the harmonic level.[13] The latter value is in good agreement with the DFT result (GGA level) of −0.16 eV obtained by Kamakoti and Sholl.[22]

The above potential was first used to locate relevant structures for the Pd−H clusters to be used subsequently as initial configurations for the molecular dynamics trajectories. Global optimization by Monte Carlo plus by minimization was carried out, starting with large icosahedral clusters of pure palladium and inserting several hydrogen atoms at random locations before locally optimizing the geometry. In the present work, a fixed low hydrogen concentration of 3% was kept for a better comparison with available data.[17] Our bulk reference system is $Pd_{256}H_8$, which was treated using periodic boundary conditions in the minimum image convention and without truncating the interactions. Cubic boxes of volume $V = L^3$ were taken to depend on temperature in order to mimic the thermal expansion naturally occurring in experimental palladium hydride under fixed pressure. We chose the same linear dependence for the lattice spacing $a(T)$ as the authors of ref 17 for the same system, namely $a(T) = a(0)[1 + \alpha T]$, with $a(0) = 3.89$ Å and $\alpha = 2.1 \times 10^{-5}$ K$^{-1}$ in the temperature range 500−1 000 K. This expression for $a(T)$, which accounts reasonably well for the measured expansion of the Pd lattice upon absorbing low amounts of hydrogen,[2] gives the $Pd_{256}H_8$ system a box length of $L(T) \approx L(0)[1 + \alpha T]$ with $L(0) = 8.300$ Å. In our simulations of the bulk system, the hydrogen atoms were initially placed in octahedral sites distant from each other (see Figure 1) for both quantum and classical simulations.

***Table I.*** Parameters of the Many-Body Potential Defined by eq 2, as Used in This Work[a]

| pair type | $\varepsilon$ (eV) | $\zeta$ (eV) | $p$ | $q$ | $r^0$ (Å) |
|---|---|---|---|---|---|
| Pd−Pd | 0.17375 | 1.70769 | 10.8874 | 3.75433 | 2.748 |
| H−H | 0.1601 | 0.9093 | 5.28 | 3.22 | 2.3 |
| Pd−H | 0.6794 | 2.5831 | 5.5 | 2.2 | 1.769 |

[a] The Pd−Pd parameters are taken from ref 47, while Pd−H and H−H parameters are those proposed by Zhong et al. (ref 46).

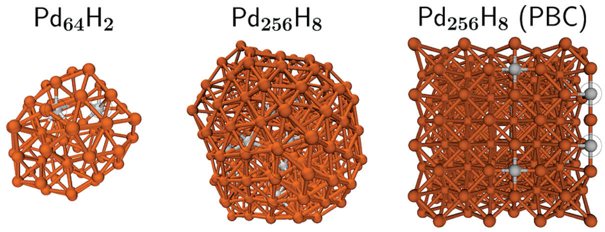


***Figure 1.*** (Color online) Stable configurations used for the initial conditions of the simulations. Left: $Pd_{64}H_2$ cluster; middle: $Pd_{256}H_8$ cluster; and right: $Pd_{256}H_8$ with fcc lattice and with periodic boundary conditions (PBC). Some hydrogen atoms are highlighted with black circles.

In order to unravel size effects in nanoscale Pd−H systems, the clusters were chosen at the same hydrogen concentration as the bulk, and we studied $Pd_{64}H_2$, $Pd_{128}H_4$, $Pd_{256}H_8$, and $Pd_{512}H_{16}$ without imposing periodic boundaries. Our search for stable structures consisted of $10^3$ basin-hopping moves, which lead to icosahedral shape with subsurface absorbed hydrogens, two examples of which are shown in Figure 1. While our global optimization was rather limited, the structures obtained only served as initial configurations for the molecular dynamics trajectories, which were conducted at significantly high temperatures $T \geq 400$ K in order to detect some diffusion under the relatively short simulated times. Under these conditions, the global minimum structure is not so relevant, especially since many isomers differing in the hydrogen sites have comparable energies.[13]

**B. Ring-Polymer Molecular Dynamics.** The diffusion kinetics of hydrogen in Pd was studied using molecular dynamics simulations. Quantum vibrational effects were included with the ring-polymer molecular dynamics method developed by Craig and Manolopoulos.[36] It is not the purpose of the present paper to review this method in detail or how it compares with the related centroid molecular dynamics technique of Cao and Voth,[37] so we will only briefly describe its main features of relevance to the present problem.

RPMD is based on the so-called primitive path-integral representation of the quantum partition function. In the RPMD method, each classical atom is described by a number $M$ of "beads" or monomers that act as imaginary time slices along the thermal path. These monomers interact successively through effective harmonic bonds, in such a way that the dynamics of the system is ruled by the following Hamiltonian:[36]

$$H(\{\mathbf{R}_i, \mathbf{P}_i\}) = \sum_{i=1}^{M} \sum_{\alpha \in \text{atoms}} \frac{\vec{p}_{i,\alpha}^{\,2}}{2m_{i,\alpha}} + \sum_{i=1}^{M} \sum_{\alpha \in \text{atoms}} \frac{m_{i,\alpha} M}{2\beta^2\hbar^2} \|\vec{r}_{i,\alpha} - \vec{r}_{i+1,\alpha}\|^2 + \frac{1}{M}\sum_{i=1}^{M} V(\mathbf{R}_i) \qquad (3)$$

In the previous equation, we have denoted $\mathbf{R}_i$ and $\mathbf{P}_i$ as the set of positions and the associated momenta of atoms that belong to the replica $i$, $1 \leq i \leq M$. The vector $\vec{r}_{i,\alpha}$ and associated momentum $\vec{p}_{i,\alpha}$ refer to atom α among replica $i$. In eq 3, we have implicitly used the cyclic condition $\vec{r}_{M+1,\alpha} = \vec{r}_{1,\alpha}$ for all α. In RPMD, the atomic mass $m_{i,\alpha}$ is taken just as the physical mass $m_\alpha$, which is the main practical difference with the partially adiabatic version of the centroid molecular dynamics method.[36,48]

Solving the equations of motion for the above Hamiltonian was shown by Craig and Manolopoulos to yield correct quantum dynamics for the centroids $\bar{\mathbf{R}} = \{\bar{\mathbf{r}}_\alpha\}$:

$$\bar{\mathbf{r}}_\alpha(t) = \frac{1}{M}\sum_{i=1}^{M} \vec{r}_{i,\alpha} \tag{4}$$

in the $t \rightarrow 0$ and harmonic $V(\mathbf{R})$ limits.[36,38] In practice, the equations are solved in the normal mode representation, which diagonalizes the harmonic part of eq 3. We denote $\mathbf{K}$ as the matrix with elements $\mathbf{K}_{ij} = 2\delta_{ij} - \delta_{i,j-1} - \delta_{i,j+1}$ (and the cyclic condition) as well as the unitary eigenvector matrix $\mathbf{U}$, which diagonalizes $\mathbf{K}$ as $\mathbf{U}^T\mathbf{K}\mathbf{U} = \mathrm{diag}(\lambda_i)$, $\lambda_i$ being the corresponding eigenvalues. Under the linear transformation from the Cartesian coordinates $\{\vec{r}_{i,\alpha}\}$ to the normal modes $\{\vec{a}_{k,\alpha}\}$:

$$\vec{a}_{k,\alpha} = \frac{1}{\sqrt{M}}\sum_{i=1}^{M} U_{ki}\vec{r}_{i,\alpha} \tag{5}$$

the Hamiltonian becomes decoupled

$$\begin{aligned} H(\{\mathbf{A}_k, \mathbf{\Pi}_k\}) &= \sum_{k=1}^{M}\sum_{\alpha\in\text{atoms}} \frac{\vec{\pi}_{k,\alpha}^2}{2m_\alpha} \\ &+ \sum_{k=1}^{M}\sum_{\alpha\in\text{atoms}} \frac{m_\alpha M}{2\beta^2\hbar^2}\lambda_k \vec{a}_{k,\alpha}^2 \\ &+ \frac{1}{M}\sum_{i=1}^{M} V\left(\left\{\sqrt{M}\sum_{k=1}^{M} U_{ik}\vec{a}_{k,\alpha}\right\}\right) \end{aligned} \tag{6}$$

where we have denoted $\vec{\pi}_{k,\alpha} = m_\alpha\sqrt{M}\mathrm{d}\vec{a}_{k,\alpha}/\mathrm{d}t$ as the momentum of atom α among replica $k$. This normal mode expression is especially useful for propagating the equations of motion because the decoupling of harmonic bonds allows their analytical integration using the reference system propagation algorithm,[49] hence, no loss in time step duration with respect to the classical case $M = 1$.

An important issue of path-integral molecular dynamics, especially relevant when simulating finite-size systems, relates to thermostating. The effective potential of eqs 3 and 6 explicitly depends on some inverse temperature $\beta$, however, the RPMD dynamics is Newtonian. To address these difficulties, Craig and Manolopoulos originally advocated that the atomic momenta be periodically redrawn from the Maxwell−Boltzmann distribution.[36] In the present work, we have followed a different strategy by coupling each normal mode vector $\vec{a}_{k,\alpha}$, carrying 3 degrees of freedom, to a separate Nosé-Hoover thermostat. This simulation is only carried out to generate proper initial conditions pertaining to the canonical ensemble at a fixed temperature, from which the Hamilton equations of motion of the RPMD method are solved without coupling with the thermostat.[50] In the case of finite systems, the angular momentum of the centroids motion is not conserved during the thermostated equilibration stage, which may result in some undesired global rotation of the system during the subsequent RPMD propagation. This rotation can be suppressed by adding some extra angular velocity that exactly compensates the current angular momentum, a method recently used by Witt and co-workers.[51]

Diffusion processes have been characterized by evaluating the Kubo-transformed position and the velocity autocorrelation functions and by monitoring the mean square displacement of the centroids and the diffusion coefficient $D$. Within the RPMD framework, and neglecting exchange effects, the quantum mechanical expression for $D$ is approximated by the Green−Kubo relation[39,40] involving the centroids velocities:

$$D \simeq \frac{1}{3}\int_0^\infty \langle\bar{\mathbf{v}}(t) \times \bar{\mathbf{v}}(0)\rangle \mathrm{d}t \tag{7}$$

where the average is taken over the different hydrogen atoms and multiple time origins, $\bar{\mathbf{v}}(t)$ being expressed for atom α as

$$\bar{\mathbf{v}}_\alpha(t) = \frac{1}{M}\sum_{i=1}^{M} \frac{\vec{p}_{i,\alpha}}{m_\alpha} \tag{8}$$

Likewise, the mean square displacement $\langle\vec{r}^2\rangle(t)$ is approximated from the centroids position $\bar{\mathbf{r}}_\alpha$, eq 4, which is also obtained from the normal mode $\vec{a}_{k,\alpha}$ corresponding to $\lambda_k = 1$. This centroid approximation reverts to the common classical expressions in the case of $M = 1$ and is exact in the limit of short times or harmonic potentials.[36,38]

For bulk systems, the calculated diffusion constant is known to depend on the size of the simulation box, and several authors[52,53] have shown that the asymptotic (infinitely large box) diffusion constant should include a hydrodynamic finite box correction proportional to $k_BT/\eta L$, where $\eta$ is the viscosity. We have not tried to calculate the parameter $\eta$ from simulations, but due to these corrections, the values for the diffusion constants of all bulk samples should be meant as lower bounds.

The hydrodynamic corrections for the bulk system are due to the spurious interactions between particles and their periodic images, thus, they do not have a counterpart in clusters. However, and strictly speaking, the diffusion constant should be zero for any finite system at any temperature because the mean square displacement cannot grow arbitrarily large in a restricted volume. The same problem arises for confined solvated systems, and Berne and co-workers have proposed methods based on fluctuating boundary conditions to deal with such situations.[54,55] Following Beck and Marchioro,[56] the time scale for calculating the velocity autocorrelation function was chosen sufficiently long with respect to the vibrational period but not exceedingly long, in order to avoid the saturation regime of the mean square displacement.

**C. Numerical Details.** The simulations for the periodic systems have been performed in the 500−1 000 K temperature range by 100 K steps and for all clusters in the 400−600 K range by 50 K steps. The first thermalization stage was carried out with thermostated RPMD trajectories employing a short time step of 0.05 fs, for a duration of 2 × $10^5$ steps (10 ps). The number of beads was taken either as $M = 1$ (classical case) or $M = 24$ to account for quantum delocalization. Intermediate values were tried also to assess the convergence at this higher value (see below). From the last 5 ps of the canonical simulations, 100 initial conditions were saved every 50 fs for further propagation without the thermostats. Each RPMD trajectory was then carried out for 20 ps with time step of 0.2 fs. Along these trajectories, the mean square displacement of centroid positions and the velocity autocorrelation functions were accumulated over 10 ps long time windows before final averaging over all trajectories.

In the case of clusters, these various MD trajectories sometimes had to be repeated, especially at high temperatures due to the spontaneous desorption of $H_2$ molecules into vacuum. Such dissociation events turned out to take place even more often for quantum simulations but would have hindered the calculation of the diffusion constant by contributing dominantly to the autocorrelation functions.

In addition to diffusion observables, we monitored the atomistic mechanisms of diffusion by periodically quenching the system to either its nearest classical local minimum or its inherent structure. Starting from the centroids positions $\bar{\mathbf{R}} = \{\bar{\mathbf{r}}_\alpha\}$, the potential energy $V$ was locally minimized by a conjugate gradient. The set of structures obtained with this scanning procedure were eventually analyzed in terms of hydrogen coordination inside palladium sites.

## III. Results

We first consider the influence of the Trotter number $M$ on the centroids trajectories in the case of the periodic system $Pd_{256}H_8$ at intermediate temperature $T = 800$ K. The optimal value for $M$ should depend first on the system as well as on the temperature, more beads being required to describe the increasingly broad nuclear wave functions as $T$ decreases. We were not able to carry out reliably converged simulations for the largest 528-atom cluster with $M > 24$ due to the frequent hydrogen desorptions at temperatures $T \geq 500$ K that required restarting the trajectories.[57]

The time variations of the mean square displacement of hydrogen atoms are represented in Figure 2 for different values of $M$ ranging from 1 to 24. At the temperature considered here, hydrogen exhibits some clear diffusion manifested on the positive slope of $\langle\bar{r}^2\rangle(t)$ for times longer than about 3 ps. The classical and quantum mean square displacements show different behaviors, with the classical motion slightly less diffusive by about 16%. However, looking at the diffusion constants obtained from the integrated velocity autocorrelation function, this factor reduces to 14% for $M = 16$, instead of 24, and to 9% for $M = 8$. The discrepancy between the results obtained for $M = 16$ and $M = 24$ is rather small. Additional simulations for the smallest system $Pd_{64}H_2$ at 400 K using $M = 48$ indicate a further increase of the diffusion rate with respect to the $M = 24$ results, however, the effect is marginal and lies below 2%. Based on these observations and keeping computational feasibility into account, we believe that the error associated with employing the fixed value of $M = 24$ for the Trotter discretization number in our simulations should not exceed 5% in the estimated diffusion constants.

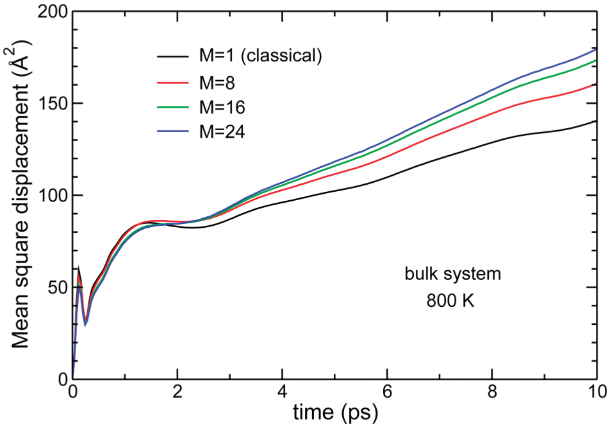


***Figure 2.*** Mean square displacement of hydrogen atoms in $Pd_{256}$ with periodic boundary conditions, as a function of time for different numbers of beads $M$ in the RPMD simulations. The temperature is 800 K.

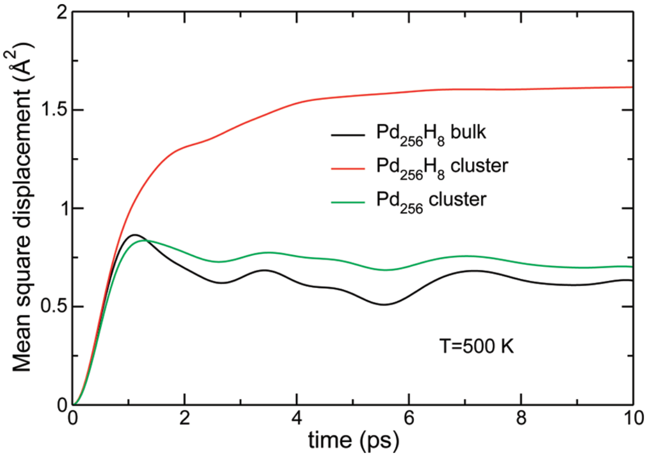


***Figure 3.*** Mean square displacement of palladium atoms in several systems, as a function of time, as obtained from the simulations at 500 K. Classical molecular dynamics were used for the pure Pd cluster.

Hydrogen absorbed in bulk palladium is known to alter its thermomechanical properties,[1−3] and we have investigated the palladium motion in both periodic and finite systems. Figure 3 shows the mean square displacement of Pd atoms for the bare and hydrogenated 256-atom systems, as a function of time and temperature at $T = 500$ K. At this temperature, the bare palladium cluster is essentially solid-like,[13] and the atoms vibrate around their equilibrium positions. The dynamics of bulk palladium hydride is also poorly diffusive as far as the Pd atoms are concerned. In contrast, Pd atoms in the free $Pd_{256}H_8$ cluster remarkably exhibit some diffusion. The variations of the atom-resolved displacement indices, together with direct visual inspection, indicate that the cluster is softer near the hydrides and is even partially melted at the surface. This agrees with our previous findings in classical Pd−H nanoclusters[13] and is also consistent with previous studies by Grönbeck and co-workers, who performed classical molecular dynamics simulations with the original MBA potential.[45] The melted surface and the relatively more rigid core result from some

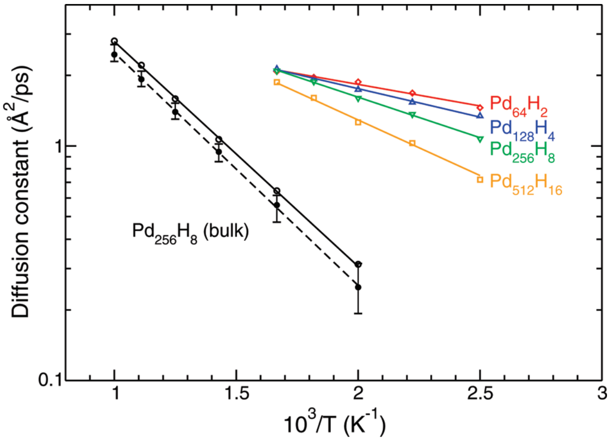


***Figure 4.*** Diffusion constant of hydrogen atoms in bulk and finite palladium systems, as obtained from classical (solid circles) or quantum (open symbols) simulations. The straight lines are Arrhenius fits.

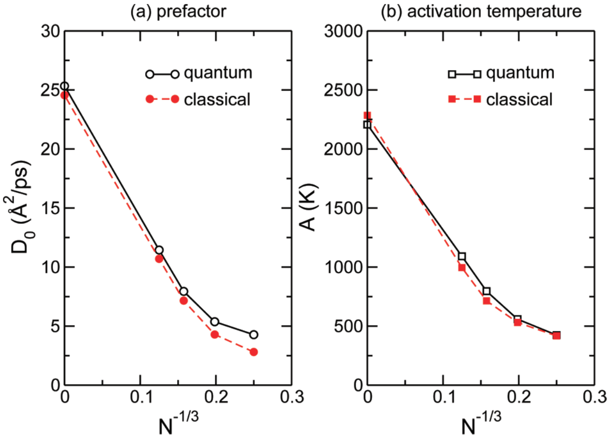


***Figure 5.*** Variations of Arrhenius parameters with inverse cluster size, as obtained from classical and RPMD simulations: (a) Arrhenius prefactor $D_0$ (in Å$^2$ /ps) and (b) activation temperature $A$ (in K).

heterogeneous dynamics, which is reflected on the variations of the mean square displacement with two distinct regimes.

The diffusion constants obtained from integrating the velocity autocorrelation functions of centroid or classical hydrogens are represented in Figure 4 as a function of inverse temperature. The results for both bulk and cluster systems are shown in the quantum case, classical results being displayed only for the bulk system in order to improve overall clarity. Error bars are also given for the bulk system, as estimated by the standard deviation between independent calculations from sets of 20 consecutive trajectories to which we added the 5% margin, corresponding to finite Trotter discretization.

The diffusion constants exhibit a linear behavior with inverse temperature when plotted in logarithmic scale, which is represented by the Arrhenius expression of eq 1 but with a size-dependent prefactor $D_0(N)$ and an activation energy $A(N)$. The surprisingly good fit of the simulation results onto the Arrhenius template of eq 1, with regression coefficients above 0.98, suggests that sufficient statistics were gathered in our simulations. These results indicate that diffusion is an activated process in both periodic and finite palladium hydrides. However, they are not as well represented by a $T^{-1/2}$ exp $(-A/k_BT)$ function, as would be obtained from the Flynn–Stoneham theory,[58] ruling out tunneling as a major contribution within the present quantum model.

The values found for the diffusion constant of bulk systems can be compared with other available results. Li and Wahnström[17] calculated $D(T = 800\text{ K}) \simeq 1.4-1.5$ Å$^2$/ ps for the periodic $Pd_{256}H_8$ system, depending on whether non adiabatic effects were neglected or included in the modeling, and $D(T = 1\,000\text{ K}) \simeq 2.8$ Å$^2$/ ps. The value reported by these authors at 630 K was found to agree well with available measurements.[59−61] Maeda and co-workers[18] measured $D(T \simeq 823\text{ K}) \simeq 1.3$ Å$^2$/ ps, but slightly lower values were experimentally determined by Goltsov et al.[62] and more recently by Powell and Kirkpatrick.[63] All aforementioned studies found the Arrhenius behavior to be well followed, though with different activation energies and prefactors between hydrogen and deuterium. The present RPMD simulations lead to $D(T = 800\text{ K}) \simeq 1.45$ Å$^2$/ ps, with the corresponding classical result being $D \simeq 1.30$ Å$^2$/ ps, in rather satisfactory agreement with published data.

Accounting for quantum delocalization generally enhances diffusion in bulk hydrides but by a small factor close to 10%. We repeated a limited number of simulations for deuterated samples, however, the diffusion appeared too slow for getting reliably converged results at temperatures below 800 K. At 1 000 K, our quantum calculations give a diffusion constant of $D \simeq 2.5$ Å$^2$ /ps, that is about 11% below the value for hydrogen. This effect is comparable in magnitude to the measured value as reported in the literature.[18,62,63]

Looking now more specifically at clusters, the diffusion constants are much higher than in the bulk system at the same hydrogen concentration, even at the lower common temperature of 500 K. Neglecting quantum vibrational effects, the Arrhenius plots are mainly shifted to slower diffusion by a few percent. More interestingly, the parameters $D_0$ and $A$ of the Arrhenius form are now found to decrease quite sensitively with cluster size. That hydrogen is more prone to diffusing in clusters than in the bulk has at least two causes. First, diffusion should be easier in the outer, less dense parts of a nanoparticle, which is precisely where the preferred absorption sites lie. Second, the mere presence of hydrogen atoms softens the cluster, and this premelting of the palladium host itself greatly enhances hydrogen motion. This Pd-assisted diffusion mechanism becomes less influential as the cluster size increases because the proportion of subsurface hydrogen atoms decreases concomitantly with an (related) increase in the melting point.

The finite size effects on the diffusion properties can be further quantified by representing the prefactor $D_0$ and the activation energy $A$ as a function of the inverse cluster radius $R^{-1} \propto N^{-1/3}$. As seen from Figure 5, both $D_0$ and $A$ display steady decreases with increasing $1/R$ and show some essentially linear variations for clusters containing 128 Pd atoms or more. Some deviations from this linear behavior are found in the smallest species as a manifestation of finite size effects beyond the surface contribution. More generally,

***Table II.*** Parameters of the Liquid Drop Model, eq 9[a]

| | classical | | quantum | |
|---|---|---|---|---|
| quantity $\chi$ | $D_0$, Å$^2$ /ps | $A$, $10^3$ K | $D_0$, Å$^2$ /ps | $A$, $10^3$ K |
| $\chi(\infty)$ | 25.33 | 2.285 | 26.14 | 2.207 |
| $\alpha_S$ | −93.22 | −11.449 | −96.18 | −7.073 |
| $\alpha_E$ | −363.47 | 0.603 | −346.80 | −30.437 |
| $\alpha_V$ | 1 510.50 | 61.413 | 1 537.21 | 120.927 |

[a] For the expansion of the Arrhenius prefactor $D_0$ and the activation temperature $A$ as a function of $N^{-1/3}$, where $N$ is the cluster size. The results are given for both classical and quantum dynamics.

the variations of $D_0$ and $A$ with the size $N$ can be approximately described by a liquid drop-like expansion:

$$\chi(N) = \chi(\infty) + \alpha_S N^{-1/3} + \alpha_E N^{-2/3} + \alpha_V N^{-1} + \mathcal{O}(N^{-1}) \tag{9}$$

where $\chi$ stands for either $D_0$ or $A$ and $\chi(\infty)$ is the value of $\chi$ at the bulk limit, taken here from the periodic sample. The constant coefficients $\alpha_S$, $\alpha_E$, $\alpha_V$, correspond to the contributions of the surface, edges, and vertices, respectively, relative to the volume. The values of these parameters for classical and quantum dynamics are collected in Table II.

These parameters show comparable values for the classical and quantum systems, except for the second- and third-order contributions to the activation temperature. This is in agreement with Figure 5, where $A$ is higher in the quantum case for the bulk system but lower for the $Pd_{512}H_{16}$ cluster.

The liquid drop extrapolation is especially useful for bridging the gap between the cluster and bulk regimes. The approximately linear rates at which $D_0$ and $A$ decrease with size will be mostly important for intermediate nanoparticle sizes of experimental relevance, in the $10^1$−$10^3$ nanometers range. At these sizes, our calculations predict that both the prefactor and activation temperature should be attenuated by about 20−30% with respect to the bulk limit.

The diffusion dynamics of hydrogen into Pd clusters can be further studied using the inherent structure approach, where the instantaneous atomic configuration is quenched into the nearest local minimum by standard local optimization. While the energy landscape of classical systems can be exactly partitioned into the basins of attractions of different inherent structures, for quantum systems, the vibrational wave function may extend over multiple basins. However, because both thermal and quantum delocalization are statistical in nature, repeating the quenches captures these combined effects but comparing classical and quantum results allows them to be eventually separated. In the quantum case, we have, thus, assimilated the current configuration as described by the centroids $\bar{\mathbf{R}}$ with weight 1 rather than assigning a weight $1/M$ to each configuration $\mathbf{R}_i$, $1 \leq i \leq M$. We have performed such a series of quenches for the $Pd_{256}H_8$ bulk or finite systems at 500 K. A number of 200 configurations periodically taken from classical or RPMD trajectories have been locally minimized, and for each resulting minimum, the local coordination of hydrogen atoms has been determined by enumerating the number $n_c$ of Pd atoms distant by less than 2.2 Å. The distributions of coordination numbers for the four situations studied are represented in Figure 6. In fcc palladium, the bimodal distributions indicate that the hydrogens occupy either tetrahedral ($n_c = 4$) or octahedral ($n_c = 6$) sites, the latter being highly favored by quantum effects. The octahedral occupancy of hydrogen for the quantum system is consistent with previous calculations[13,20] and with measurements on deuterated palladium[18,19] and is explained by the lower zero-point energy in these sites.

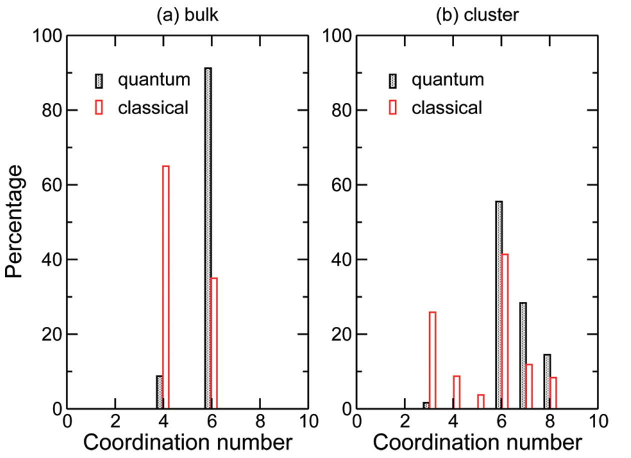


***Figure 6.*** Coordination probability of hydrogen atoms in bulk and finite $Pd_{256}H_8$ systems at 500 K, as obtained from quenching from the centroids or classical positions sampled in equilibrium trajectories.

Hydrogen atoms absorbed in the $Pd_{256}H_8$ cluster exhibit a broader variety of occupancies, covering the $n_c = 3$−$8$ range. Low-coordinated structures are actually a signature of surface hydrogen atoms, whereas highly coordinated configurations correspond to hydrogens lying near the most dense parts of the icosahedral clusters. Here again, there are marked differences between the inherent structures obtained from both classical and quantum trajectories. The centroids from the quantum dynamics tend to occupy higher coordinated sites but are essentially absent from the surface sites. This may partly result from our removal of trajectories that ended in desorption events, which were more frequent in the quantum case. The broader distribution of coordination numbers is a mere signature of not only the less ordered structure of icosahedral clusters, especially away from the magic numbers of 147 and 309, but also of the surface melted state at this temperature (vide supra). As in the bulk system, we interpret the higher average coordination of hydrogen atoms in the quantum case as another consequence of unfavorable zero-point motion in the low-coordinated sites. However, a direct comparison of the potential and zero-point energies of the inherent structures obtained from the quantum and classical trajectories would not be strictly relevant or even fair because clearly quantum effects drive the system to different parts of the landscape.

## IV. Summary and Conclusion

The diffusion of hydrogen in bulk nanoscale metals has found a renewed interest in the context of fuel cell technology. In the present work, we have theoretically studied hydrogen diffusion in palladium nanoparticles paying a particular attention to finite size scaling effects. Using the recently developed ring-polymer molecular dynamics method of

Manolopoulos and co-workers[36] and together with an explicit many-body potential fitted to reproduce energetic and structural properties of bulk palladium hydrides,[13,46,47] we have calculated the diffusion constant of hydrogen in periodic and finite palladium systems. At the low hydrogen concentration of 3%, our calculated diffusion constant for the bulk fcc palladium hydride agrees satisfactorily with previously published experimental results under similar conditions.[18,59−63] In particular, quantum delocalization effects are found to enhance diffusion by a few percents, but no clear signature of tunnel-assisted diffusion was found at the temperature variations of the diffusion constant, which follow some Arrhenius behavior in the range 500−1 000 K.

In Pd clusters containing between 64 and 512 atoms, diffusion is much faster than in the bulk sample. Hydrogen atoms tend to fill the clusters from the outside[9,13] and are found to destabilize or premelt the palladium host, in agreement with previous classical molecular dynamics simulations.[45] The faster diffusion of hydrogen in clusters was interpreted as being due to the less dense outer parts of nanoparticles, together with the mobile character of the corresponding premelted Pd atoms. Even though hydrogen diffuses faster in nanoparticles, the variations of the diffusion constant are still exponential with inverse temperature, with strongly size-dependent prefactor and activation energy. By plotting these Arrhenius parameters as a function of inverse cluster radius, a nearly linear behavior is found between the large clusters containing more than 128 Pd atoms and the bulk limit. Finite-size effects beyond the surface/volume contribution can be represented using a liquid drop expansion up to third order in inverse cluster radius.

While the diffusion constants extracted from quantum and classical trajectories do not overwhelmingly differ from each other, an inherent structure analysis conducted from the centroids positions reveals interesting differences between the two dynamics. When quantum effects are accounted for, hydrogens occupy preferentially octahedral sites in fcc palladium but tetrahedral sites in the classical case.[13,18−20] The much broader variety of occupancies in Pd nanoparticles conveys their less ordered, icosahedral character as well as their partially melted thermodynamical state. However, the same trends noted for the bulk system are found for clusters, namely quantum delocalization favors higher coordinated sites.

Some fundamental aspects touched in this work could be investigated further in the future. Quantifying hydrodynamic effects in the bulk system could be achieved either by estimating the viscosity or, more directly, by performing additional simulations for larger periodic boxes. It would then be interesting to determine whether the corresponding effect, which scales as $1/L \propto N^{-1/3}$ with the number of Pd atoms, is comparable in magnitude to the surface/volume ratio (parameter $\alpha_S$), which characterizes finite size effects in clusters. Quantum effects on heavier (deuterium) or lighter (muonium) particles could also be envisaged in a more systematic way, not only in the bulk but in clusters as well. Molecular simulation of finite systems would also allow looking into more detailed aspects of the diffusion near the free surfaces by quantifying the extent of diffusion parallel and perpendicular to the surface. It could then be useful to incorporate more robust ways of eliminating the risk of spontaneous desorption after adapting, for instance, the method of Li and co-workers[55] to the RPMD framework.

Finally, it would be useful to apply the ring-polymer molecular dynamics scheme to more complicated systems for which modeling has so far essentially relied on kinetic approximations. Amorphous metals,[28] membranes,[64] and nanoparticle arrays[16] are some examples of possible applications. Ruthenium hydride nanoparticles, for which recent experimental NMR evidence has shown a significant mobility of the hydrogens,[65] also offer promising candidates for testing the present methods in a related context.

**Acknowledgment.** The authors wish to thank the Pôle Scientifique de Modélisation Numérique (PSMN) for providing the generous computer resources needed.

## References

(1) Lewis, F. A. *The Palladium-Hydrogen System*; Academic Press: New York, 1967.

(2) *Hydrogen in Metals I and II*; Alefeld, G., Völkl, J., Eds.; Springer-Verlag: Berlin, Germany, 1978; vols 28 and 29;*Hydrogen in Metals III: Properties and Applications*; Wipf, H., Ed.; Springer-Verlag: Berlin, Germany, 1997.

(3) *Binary Alloy Phase Diagrams*; Massalski, T. B., Ed.; Americal Society for Metals: Metals Park, OH; 1986; vol II.

(4) Graham, T. *Proc. R. Soc.* **1869**, *212*, 17.

(5) Züttel, A. *Mater. Today* **2003**, *6*, 24.

(6) *The Chemical Physics of Solid Surfaces and Heterogeneous Catalysis*; King, D. A., Woodruff, D. P., Eds.; Elsevier: Amsterdam, The Netherlands, 1982.

(7) Huang, S.-Y.; Huang, C.-D.; Chang, B.-T.; Yeh, C.-T. *J. Phys. Chem. B* **2006**, *110*, 21783.

(8) Rather, S.; Zacharia, R.; Hwang, S. W.; Naik, M.; Nahm, K. S. *Chem. Phys. Lett.* **2007**, *438*, 78.

(9) Wolf, R. J.; Lee, M. W.; Ray, J. R. *Phys. Rev. Lett.* **1994**, *73*, 557.

(10) Pundt, A.; Dornheim, M.; Guerdane, M.; Teichler, H.; Ehrenberg, H.; Reetz, M. T.; Jisrawi, N. M. *Eur. Phys. J. D* **2002**, *19*, 333.

(11) Suleiman, M.; Jisrawi, N. M.; Dankert, O.; Reetz, M. T.; Bähtz, C.; Kirchheim, R.; Pundt, A. *J. Alloys Comp.* **2003**, *356−357*, 644.

(12) Pundt, A.; Suleiman, M.; Bähtz, C.; Reetz, M. T.; Kirchheim, R.; Jisrawi, N. M. *Mater. Sci. Eng., B* **2004**, *108*, 19.

(13) Calvo, F.; Carré, A. *Nanotechnology* **2006**, *17*, 1292.

(14) Ingham, B.; Toney, M. F.; Hendy, S. C.; Cox, T.; Fong, D. D.; Eastman, J. A.; Fuoss, P. H.; Stevens, K. J.; Lassesson, A.; Brown, S. A.; Ryan, M. P. *Phys. Rev. B: Condens. Matter* **2008**, *78*, 245408.

(15) Di Vece, M.; Grandjean, D.; Van Bael, M. J.; Romero, C. P.; Wang, X.; Decoster, S.; Vantomme, A.; Lievens, P. *Phys. Rev. Lett.* **2008**, *100*, 236105.

(16) van Lith, J.; Lassesson, A.; Brown, S. A.; Schulze, M.; Partridge, J. G.; Ayesh, A. *Appl. Phys. Lett.* **2007**, *91*, 181910.

(17) Li, Y.; Wahnström, G. *Phys. Rev. Lett.* **1992**, *68*, 3444. *Phys. Rev. B* **1992**, *46*, 14528.

(18) Maeda, T.; Naito, S.; Yamamoto, M.; Mabuchi, M.; Hashino, T. *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 899.

(19) Pitt, M. P.; Gray, E. MacA. *Europhys. Lett.* **2003**, *64*, 344.

(20) Caputo, R.; Alavi, A. *Mol. Phys.* **2003**, *101*, 181.

(21) Kang, B.-S.; Sohn, K.-S. *Physica B* **1995**, *205*, 163.

(22) Kamakoti, P.; Sholl, D. S. *J. Membr. Sci.* **2003**, *225*, 145.

(23) Jewell, L. J.; Davis, B. H. *Appl. Catal., A* **2006**, *310*, 1.

(24) Sundell, P. G.; Wahnström, G. *Phys. Rev. Lett.* **2004**, *92*, 155901.

(25) Dyer, M.; Zhang, C.; Alavi, A. *ChemPhysChem* **2005**, *6*, 1711.

(26) Olsen, R. A.; Kroes, G. J.; Løvvik, O. M.; Baerends, E. J. *J. Chem. Phys.* **1997**, *107*, 10652.

(27) Watson, G. W.; Wells, R. P. K.; Willcock, D. J.; Hitchings, G. J. *J. Phys. Chem. B* **2001**, *105*, 4889.

(28) Hao, S.; Sholl, D. S. *J. Chem. Phys.* **2009**, *130*, 244705.

(29) Gillian, M. J. *J. Phys. C* **1986**, *19*, 6169.

(30) Muranaka, T.; Uehara, K.; Takasu, M.; Hiwatari, Y. *Molec. Sim.* **1994**, *12*, 329.

(31) Gross, A.; Scheffler, M. *Phys. Rev. B: Condens. Matter* **1998**, *57*, 2493.

(32) Busnengo, H. F.; Pijper, E.; Kroes, G. J.; Salin, A. *J. Chem. Phys.* **2003**, *119*, 12553.

(33) Feynman, R. P.; Hibbs, A. R. *Quantum mechanics and path integrals*; McGraw-Hill: New York, 1965.

(34) Chen, B.; Gomez, M. A.; Sehl, M.; Doll, J. D.; Freeman, D. J. *J. Chem. Phys.* **1996**, *21*, 9686.

(35) Forsythe, K. M.; Makri, N. *J. Chem. Phys.* **1998**, *108*, 6819.

(36) Craig, I. R.; Manolopoulos, D. E. *J. Chem. Phys.* **2004**, *121*, 3368.

(37) Cao, J.; Voth, G. A. *J. Chem. Phys.* **1993**, *99*, 10070.

(38) Braams, B. J.; Manolopoulos, D. E. *J. Chem. Phys.* **2006**, *125*, 124105.

(39) Miller, T. F., III; Manolopoulos, D. E. *J. Chem. Phys.* **2005**, *122*, 184503.

(40) Miller, T. F., III; Manolopoulos, D. E. *J. Chem. Phys.* **2005**, *123*, 154504.

(41) Markland, T. E.; Habershon, S.; Manolopoulos, D. E. *J. Chem. Phys.* **2008**, *128*, 194506.

(42) Calvo, F. *J. Comput. Theor. Nanosci.* **2008**, *5*, 331.

(43) Jortner, J. *Z. Phys. D: At., Mol. Clusters* **1992**, *24*, 247.

(44) Wolf, R. J.; Lee, M. W.; Davis, R. C.; Fay, P. J.; Ray, J. R. *Phys. Rev. B: Condens. Matter* **1993**, *48*, 12415.

(45) Grönbeck, H.; Tománek, D.; Kim, S. G.; Rósen, A. *Chem. Phys. Lett.* **1997**, *264*, 39. *Z. Phys. D: At., Mol. Clusters* **1997**, *40*, 469.

(46) Zhong, W.; Li, Y. S.; Tománek, D. *Phys. Rev. B: Condens. Matter* **1991**, *44*, 13053.

(47) Rey, C.; Gallego, L. J.; García-Rodeja, J.; Alonso, J. A.; Iñiguez, M. P. *Phys. Rev. B: Condens. Matter* **1993**, *48*, 8253.

(48) Hone, T. D.; Rossky, P. J.; Voth, G. A. *J. Chem. Phys.* **2006**, *124*, 154103.

(49) Tuckerman, M. E.; Berne, B. J.; Martyna, G. J. *J. Chem. Phys.* **1992**, *97*, 1990.

(50) Pérez, A.; Tuckerman, M. E.; Müser, M. H. *J. Chem. Phys.* **2009**, *130*, 184105.

(51) Witt, A.; Ivanov, S. D.; Shiga, M.; Forbert, H.; Marx, D. *J. Chem. Phys.* **2009**, *130*, 194510.

(52) Dünweg, B.; Kremer, K. *J. Chem. Phys.* **1993**, *99*, 6983.

(53) Yeh, I.-C.; Hummer, G. *J. Phys. Chem. B: Condens. Matter* **2004**, *108*, 15873.

(54) Liu, P.; Harder, E.; Berne, B. J. *J. Phys. Chem. B* **2004**, *108*, 6595.

(55) Li, Y.; Krilov, G.; Berne, B. J. *J. Phys. Chem. B* **2005**, *109*, 463.

(56) Beck, T. L.; Marchioro, T. L., II *J. Chem. Phys.* **1990**, *93*, 1347.

(57) Values much higher than 24 could have been used for *M* if the Pd-H interactions had been additive, in which case the system could be partitioned into classical (Pd) and quantum (H) subsystems.

(58) Flynn, C. P.; Stoneham, A. M. *Phys. Rev. B: Condens. Matter* **1970**, *1*, 3966.

(59) Sköld, K.; Nelin, G. *J. Phys. Chem. Solids* **1967**, *28*, 2369.

(60) Rowe, J. M.; Rush, J. J.; de Graaf, L. A.; Ferguson, G. A. *Phys. Rev. Lett.* **1972**, *29*, 1250.

(61) Carlile, C. J.; Ross, D. K. *Solid State Commun.* **1974**, *15*, 1923.

(62) Goltsov, V. A.; Demin, V. B.; Vykhodets, V. B.; Kagan, G. Ye.; Geld, P. V. *Phys. Metals Metallog.* **1970**, *29*, 195.

(63) Powell, G. L.; Kirkpatrick, J. R. *Phys. Rev. B: Condens. Matter* **1991**, *43*, 6968.

(64) Semidey-Flecha, L.; Sholl, D. S. *J. Chem. Phys.* **2008**, *128*, 144701.

(65) Pery, T.; Pelzer, K.; Buntkowsky, G.; Philippot, K.; Limbach, H.-H.; Chaudret, B. *ChemPhysChem* **2005**, *6*, 605.

CT900554R

JCTC Journal of Chemical Theory and Computation

# Tautomerism in Reduced Pyrazinacenes

Roberto Scipioni,*[†,‡] Mauro Boero,[§] Gary J. Richards,[†,||] Jonathan P. Hill,*[†]
Takahisa Ohno,[†] Toshiyuki Mori,[||] and Katsuhiko Ariga[†]

*WPI Center for Materials Nanoarchitectonics (MANA), National Institute for Materials Science, Namiki 1-1, Tsukuba, Ibaraki 305-0044, Japan, Max Planck Institute for Polymer Research, Mainz, Ackermannweg 10, 55124, Germany, Institut de Physique et Chimie des Matériaux de Strasbourg (IPCMS), UMR 7504 CNRS-University of Strasbourg, 23 rue du Loess, F-67034 Strasbourg, France, Computational Materials Center, National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047 Japan, and Fuel Cell Materials Center, Nanoionics Group, National Institute for Materials Science (NIMS), 1-1 Namiki, Tsukuba, Ibaraki, 305-0044, Japan*

Received December 6, 2009

**Abstract:** Monoprotic and diprotic NH tautomerism in reduced oligoazaacenes, the pyrazinacenes, was studied by using first principles simulations. Stepwise reductions in the metadynamics-sampled free energy profile were observed during consecutive monoprotic tautomerizations, with energy barriers gradually reducing with increasing proton separation during monoprotic processes. This is accompanied by an increasing contribution from the quinoidal electronic structure, as evidenced by the computed highest occupied molecular orbital (HOMO) structure. An unusual odd−even effect in the free energy profiles is also observed upon changing the length of the pyrazinacene. Calculated HOMO structures reveal an increasing tendency for delocalization of pyrazine lone pairs with an increasing number of ring annelations. The influence of tautomerism on the pyrazine lone pair delocalization was also observed. Tautomers with protons situated centrally on the pyrazinacene backbone are predicted to be more stable due to a combination of (enamine) delocalization and a loss of Clar sextet resonance stabilization in tautomers with protons at terminal pyrazine rings. Experimental evidence suggesting the structure of pyrazinacene tautomers is included and discussed as a support to the calculation.

## Introduction

The acenes represent an important class of molecular materials which possess properties suitable for applications in organic semiconducting devices, including organic field effect transistors.[1] Pentacene (**1**),[2] the quintessential acene, has been extensively studied because of its high field effect mobility (up to 5 cm$^2$ V$^{-1}$ s$^{-1}$ for ultrapure samples[3]) but also in the development of the synthesis of more easily processable derivatives, since pentacene itself is a rather intractable and unstable substance. Thus, soluble, substituted pentacenes have been prepared, and improved stability against aerobic oxidation has also been obtained in several cases.[4] Because of their importance as organic electronic materials, the acenes have also attracted attention from a predictive point of view, and computational methods have been applied extensively in order to assess the benefits of preparing higher-order oligoacene structures.[5]

Our interest in oligoacenes stems from the synthesis of higher annulated oligopyrazines, which we term “pyrazinacenes”. Heteroacenes, including the pyrazinacenes, are important as *n*-type counterparts of the *p*-type semiconducting CH pentacenes.[6] During development of the synthetic methods, we were intrigued by the possibility of protic

* Corresponding author e-mail: scipioni@mpip-mainz.mpg.de (R.S.); Jonathan.Hill@nims.go.jp (J.P.H.).

[†] Center for Materials Nanoarchitectonics, National Institute for Materials Science.

[‡] Max Planck Institute.

[§] IPCMS.

[||] Fuel Cell Materials Center, National Institute for Materials Science.

10.1021/ct9006585
Published on Web 01/07/2010

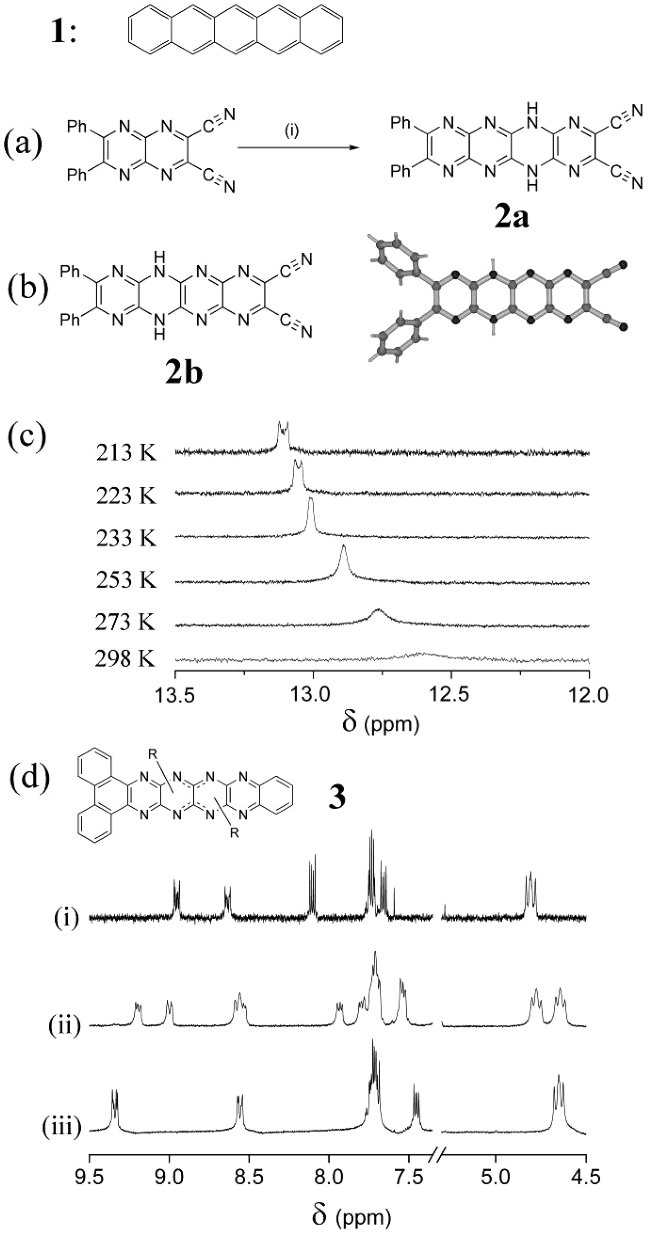


***Figure 1.*** Chemical structure of pentacene, **1**, and (a) synthesis of **2a**, (i) 2,3-dicyano-5,6-diaminopyrazine, dimethylsulfoxide, $Na_2CO_3$, 100 °C. (b) Chemical structure of **2b** and its molecular structure obtained by single-crystal X-ray crystallographic analysis. (c) VT-$^1$H NMR spectra of **2a·Na**$^+$ showing the splitting of the NH resonance at low temperatures. (d) Proposed chemical structure and $^1$H NMR spectra (i−iii) of compounds obtained by the *N*-alkylation of dihydro-5,6,7,8,13,14,15,16-octaaza-[*n*,*p*]-dibenzohexacene.

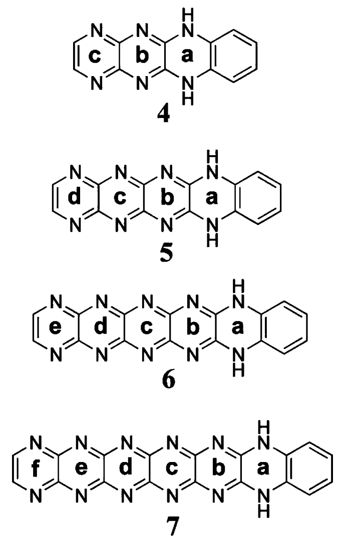


***Figure 2.*** Structures of the proposed compounds **4**−**7** studied for their tautomeric processes.

tautomerism in the reduced pyrazinacenes that does not have an analog in the CH pentacenes.[7] Thus, this feature might be a significant determinant of the physical properties of the pyrazinacenes. Initially, we studied dihydro-substituted compounds, the 2H-pyrazinacenes (Figure 1). The existence of tautomerism in the 2H-pyrazinacenes first became apparent during single-crystal X-ray crystallographic analysis of the compound that was expected to be 2,3-dicyano-5,12-dihydro-8,9-diphenyl-1,4,5,6,7,10,11,12-octaazatetracene, **2a**. Actually, its tautomeric analogue **2b** containing the 6−11-dihydro moiety was obtained (Figure 1b), having been formed during synthesis or at crystallization. A preliminary variable-temperature (VT) $^1$H NMR study on the monosodium salt of **2a** (**2a·Na**; used to avoid the appearance of too many tautomers; Figure 1c) revealed splitting of the resonance due to exchangeable amine protons at a depressed temperature consistent with the existence of two isomers of **2a**. Other, more circumstantial evidence for the tautomerism came from attempts to N-alkylate **2a/b** analog 2H-**3** using alkyl halides, which gave a mixture of several compounds of identical mass but with differing $^1$H NMR spectra, as illustrated in Figure 1d. Tautomerism should also influence the properties of the compounds. In particular, semiconductivity will be most likely modulated, while bulk proton mobility would make these compounds interesting materials for proton conduction applications (e.g., in fuel cells).

While investigations on the synthesis and structural analysis of the dihydropyrazinacenes continue in our laboratory, questions regarding their NH tautomerism and how this influences aromaticity and electronic properties of the molecules are addressable by using computational methods. We were especially curious about the energetics of potential tautomeric processes and how this might translate into proton delocalization. Furthermore, we were keen to determine what factors might delineate the relative yields of products of N-alkylation at the pyrazinacene nitrogen atoms. With this aim, we applied molecular dynamics simulations to study the deprotonation/reprotonation processes necessary for tautomerism in an isolated 2H-pyrazinacene molecule. For our computational study, we chose compounds **4**−**7** (containing a number of fused six-membered rings corresponding to their designated compound number), which possess one terminal benzene ring and, respectively, three, four, five, or six fused pyrazine rings (Figure 2). Pyrazine rings are labeled alphabetically, and monoprotic or diprotic tautomerism is denoted by "1H" and "2H", respectively. For example, if **6** undergoes a monoprotic shift from ring a to ring d, then the resulting compound is named **6d(1H)**, while a similar diprotic shift would result in **6d(2H)**. The fused benzo ring present in all of the compounds was placed as a fixed nontautomerizing group so that the effect on tautomerism of having a pyrazino

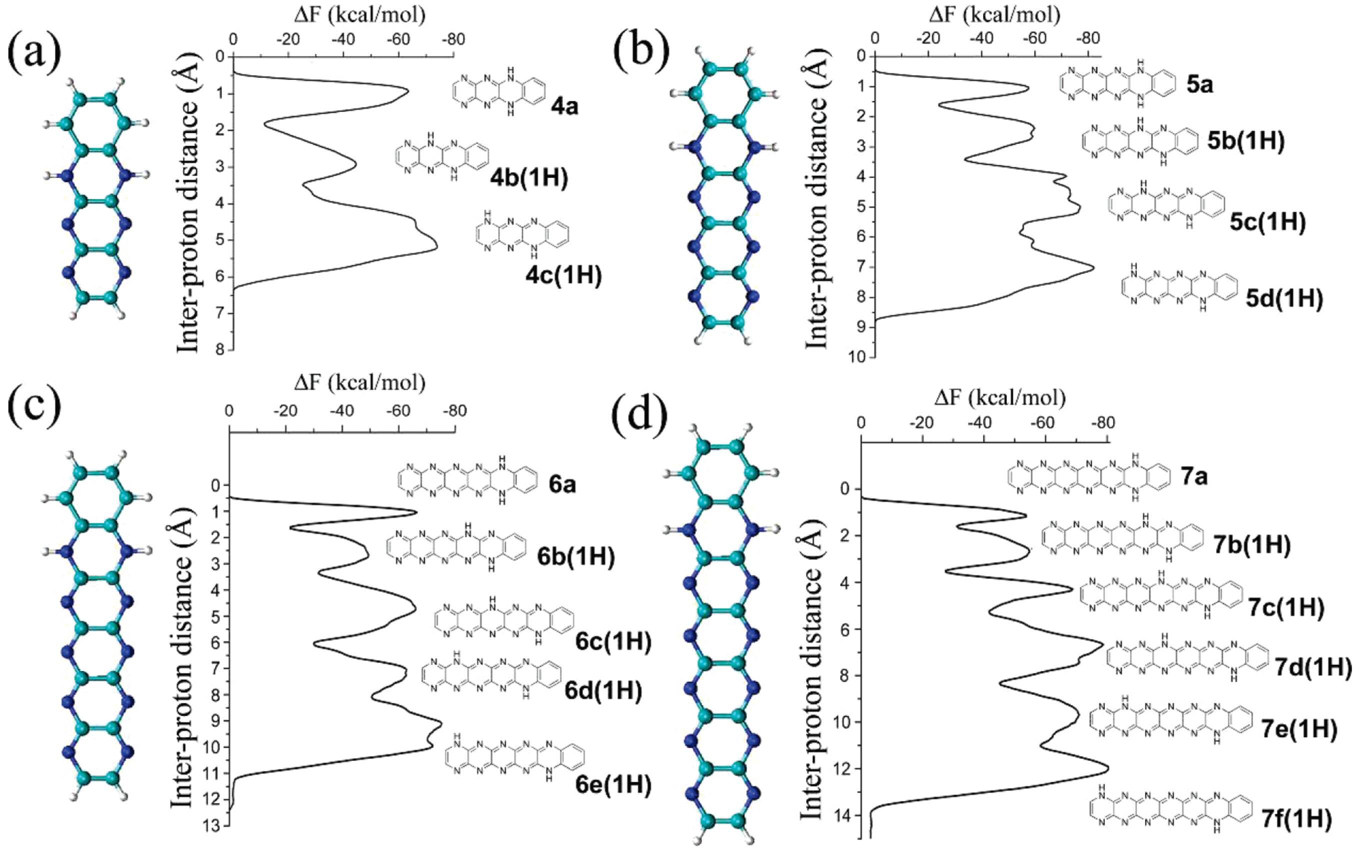


***Figure 3.*** Variation of free energy (Δ*F*) with distance between exchangeable protons.

group in a terminal position could be evaluated. This benzo group is also present since our real synthetic procedures often result in pyrazinacene molecules containing such a group.

## Computational and Experimental Details

**Computational.** All calculations were performed using the CPMD code.[8] The first principles molecular dynamics version adopted is the Car−Parrinello[9] approach; the density functional[10] used to describe the total energy included generalized gradient corrections to the exchange and correlation functionals after Becke[11] and Lee−Yang−Parr,[12] respectively. The core−valence interaction was described in terms of norm-conserving Troullier-Martins pseudopotentials,[13] and valence electron wave functions were represented in a plane-wave (PW) basis set with an energy cutoff of 70 Ry. All simulations were performed in an isolated cell according to the prescription of Barnett and Landman[14] for the release of periodic boundary conditions in PW approaches. The simulation of the displacement of a proton along the molecule and the related calculations of free energy barriers were done within the metadynamics approach,[15] which has already been extensively discussed in the literature and has been shown to be particularly suited to this class of problem.[16] The metadynamics collective variable (CV) used in the present set of simulations was the distance between one of the N atoms of the molecule and the proton which has to be displaced, CV = $[\mathbf{R}(\mathrm{H}) - \mathbf{R}(\mathrm{N})]$. This collective variable is included in the Lagrangean equations of motion with a fictitious mass $M_{\mathrm{CV}} = 25$ au and a force constant $k_{\mathrm{CV}} = 0.25$ au for the harmonic term. The penalty potential adopted was a superposition of small Gaussian functions, of amplitude $W(t)$, sampled uniformly between 0.02 and 0.18 kcal mol$^{-1}$, and each new Gaussian function was introduced after 100 steps of dynamics amounting to 10 fs. Further information on the computational procedure can be found in the Supporting Information. The total energies of the various molecules were obtained with a standard geometry optimization performed until the forces became lower than 0.001 eV/Bohr. The highest occupied molecular orbital (HOMO) and lowest unoccupied orbital (LUMO) structures were computed by exact diagonalization of the Kohn−Sham matrix on the final optimized geometries.

**Synthesis.** Compounds **2** and **3a**−**c** were synthesized as previously reported.[17] Details are provided in the Supporting Information.

## Results and Discussion

Free energy changes on the tautomerization of **4**−**7** are shown in Figure 3. The various minima in the plots of free energy versus the interproton distance are due to different tautomeric forms of the molecules, while the less negative peaks represent the barrier against the tautomerization process. We can observe one main effect in the free energies, namely, that the energy barrier for displacing the proton from its initial position systematically decreases by increasing the number of fused pyrazinacene rings in going from **4** to **7**. However, there is also a peculiar odd−even effect between molecules with even and odd numbers of rings. If we compare (see Figure 3) the free energy barriers for molecules **a** and **c**, we see that they decrease in passing from **a** to **c**.

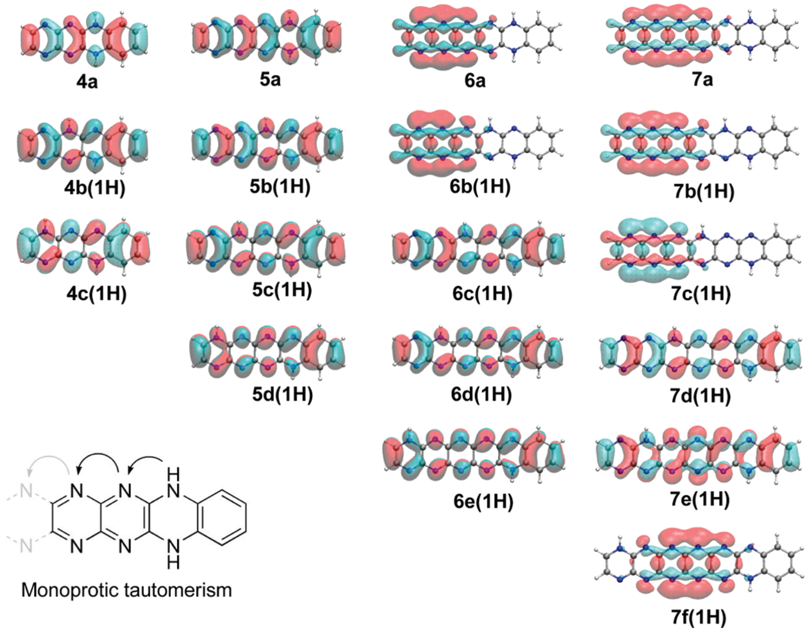


***Figure 4.*** Calculated structures of HOMOs of tautomers of **4**−**7** due to monoprotic migration ("1H" denotes one proton process).

However, they also decrease in going from even to odd, and in fact, molecules **b**, despite being composed of fewer rings, have a lower energy barrier for the first deprotonation than molecules **c**. There are therefore two effects: first, molecules with an odd number of total rings have energy barriers for first deprotonation, which are systematically lower (∼20 kcal/mol), and, second, increasing the number of rings leads to a reduction of the energy barriers. While the second effect can be explained by the fact that the longer molecules can more easily mesomerize in response to the proton shift, the peculiar odd−even effect does not have an immediate explanation. It may be a result of the combination of several different factors including symmetry, competition between aromaticity and antiaromaticity, and the role of entropic effects, including the fact that, as the molecule becomes longer, nonplanar geometries may be favored, although deviations from planarity were found to be small.

For the subsequent deprotonation steps, a combination of different effects occurs, and a different behavior is observed between odd and even ring molecules. For example, respective tautomerizations of **4a**−**7a** to **4b(1H)**−**7b(1H)** do not result in a reduction in the free energy activation barrier, and for **4a** and **6a**, an increase is observed, indicating that this is not a favored process. This is not entirely unexpected, since there are no mesomeric advantages apparent during this procedure. In fact, it is likely that formation of the **b** tautomers actually obstructs conjugation relative to the starting **a** tautomers (i.e., there are fewer fused aromatic six-membered rings). As the proton migrates toward the non-benzo-substituted end of the molecule, tautomers at each nitrogen "station" exhibit lower activation energies in comparison with the starting tautomer, which we attribute to the increasing conjugated quinoidal character of the fused pyrazine unit situated between singly reduced pyrazines, which can be seen in the calculated HOMO structures of compounds **4a**−**7a** to **4c(1H)**−**7f(1H)**, shown in Figure 4. However, there is the additional feature of pyrazino nitrogen atom lone pair delocalization[18] in the longer pyrazinacenes. In particular, where tautomers of **6** and **7** have at least three fused (and nonreduced) pyrazine rings, then delocalization occurs, as shown in **6a(1H)**, **6b(1H)**, **7a(1H)**, **7b(1H)**, **7c(1H)**, and **7f(1H)**.

In **4** and **5**, as monoprotic tautomerization occurs, an increasing quinoidal character appears, whereas in **6** and **7**, a similar tautomerization occurs at the expense of delocalization and is accompanied by increasing quinoidal character. Interestingly, in **7** the quinoidal character appears in **7d(1H)** and **7e(1H)** but is destroyed upon transfer of the single proton to the terminal pyrazine, giving **7f(1H)** in favor of a nitrogen-lone-pair delocalized structure. The structure of HOMO **7e(1H)** appears to be a chimera with an unusual structure between the delocalized and quinoidal forms. This structure occurs despite the presence of three fused pyrazine rings, which is thought to favor nitrogen atom lone pair delocalization. HOMO **7e(1H)** bears features of both quinoidal and delocalized orbitals, while, conversely, diprotic tautomerization does not yield such a structure, so that it is likely due to a mixing of quinoidal and delocalized nitrogen lone pair orbitals.

A very interesting phenomenon occurs in the case of diprotic tautomerism, graphically summarized in Figure 5, involving migration of the $\pi$-electron cloud in compounds **4**−**7**. We found that, when there are three fused pyrazine rings, a delocalization of the HOMO occurs along these rings with no (or just minor) contributions due to the remaining portion of the molecule. The shift of protons along the molecule in a certain direction causes a reduction of delocalization in the same direction or a transfer of electronic density in the opposite direction. This is a fundamental result that might have far-reaching consequences regarding the application of these molecules and their derivatives in molecular electronics and condensed matter devices. Thus, it can be argued that multiple tautomerizations or "conduction" of protons along a pyrazinacene backbone in one

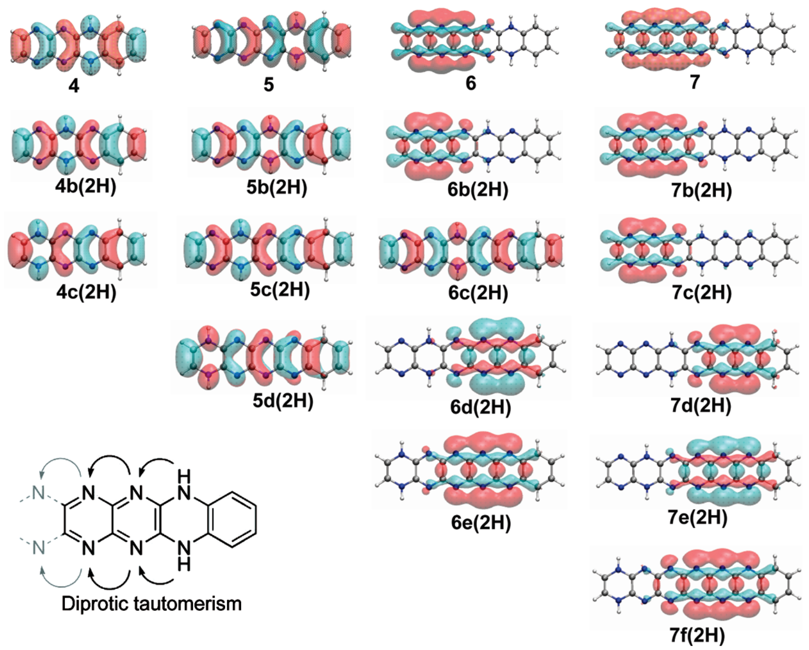


***Figure 5.*** Calculated structures of HOMOs of tautomers of **4**−**7** due to diprotic migration (“2H” denotes two proton process).

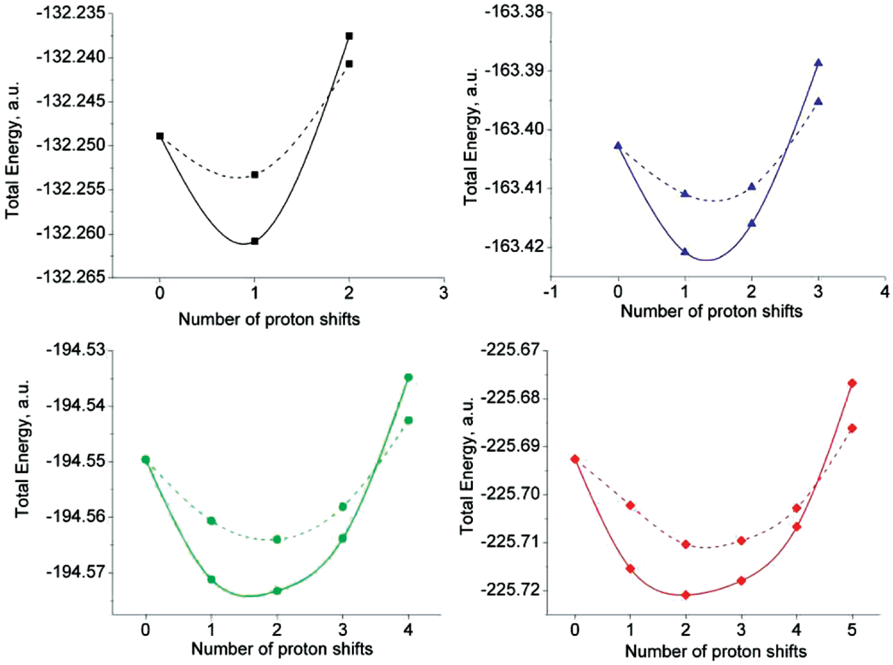


***Figure 6.*** Total energies of the tautomers from monoprotic/diprotic shifts. Dashed lines, monoprotic (1H); solid lines, diprotic (2H). Black line, **4**; blue line, **5**; green line, **6**; red line, **7**.

direction results in a net transfer of electron density in the opposite direction. If we consider a single extended pyrazinacene, then the result would be an accumulation of protons and electron density at opposing ends of the molecule. Finally, it should be noted that not all of the possible tautomers have been considered here. That is, tautomers due to mixed monoprotic/diprotic processes have been neglected for clarity. However, the tautomers on which we focused in this work exhibit the most important phenomena due to proton migration and HOMO structure so that the other tautomers, although certainly of interest, do not provide any further insight into this system.

**Total Energy Stability.** Next, we considered the stabilities, in terms of relative total energies, of the individual tautomers depending on mono- or diprotic tautomerization. The result of this analysis is summarized in Figure 6. Basically, shifting two protons (i.e., dihydropyrazine ring migration) yields a species of greater stability in all cases for **4**−**7**. A single proton migration results in a lower stability. There is an exception in that tautomers with two protons at the terminal pyrazine ring (i.e., **4c(2H)**, **5d(2H)**, **6e(2H)**, and **7f(2H)**) are all less stable than the corresponding monoprotic-shifted tautomers, although those compounds (i.e., **4c(1H)**, **5d(1H)**, **6e(1H)**, and **7f(1H)**) are still less stable than their respective parents (a symptom of the aforementioned fused benzo group). Tautomers with protons located toward the center of the molecule are more stable, and there is some precedent for this from our laboratory investigations and from the work of others.[5c,d] The molecules considered here are formally antiaromatic according to the Hückel rule, which classifies aromatic molecules as having $4n + 2$ $\pi$ electrons, while those with $4n$ $\pi$ electrons are antiaromatic. In this case, the resulting electronic structure is influenced by the fact that the molecule is composed of fused and partly reduced nitrogen heterocycles, so that delocalization may be also affected by polarity, electronegativity, and more crucially, tautomerism. The increased stability of the tautomers bearing protons at central rings of the molecules is an indication that these formally antiaromatic molecules can gain stabilization through structural rearrangements involving proton shifts. In longer molecules, $\pi$ electrons of the dihydropyrazine moieties can be more easily displaced to adjacent pyrazine rings, and this is one of the reasons why tautomers with protons at central positions are more stable. Furthermore, the shift of protons to a terminal pyrazine ring lowers the stabilities of the molecules because of a reduction in resonance stabilization as a result of the loss of a Clar six-membered benzenoid ring. This occurs in both monoprotic and diprotic processes, giving rise to the above tautomers characterized by highest energies. On the other hand, starting tautomers **4a**−**7a** are more stable than those with terminal mono- or dihydropyrazine groups, but they are less stable than those with centrally situated mono- or dihydropyrazines because of the presence of the stabilizing terminal benzenoid group (see Figure 6a). To determine which of the central rings is preferably reduced, there are two effects that must be considered. First, there should be a repulsive interaction between the electron-rich terminal benzo group and an adjacent electron-rich dihydropyrazine moiety. Second, dihydropyrazine groups are better accommodated at centrally positioned rings because of resonance stabilization effects.[5d] Thus, in **5** and **7**, where an odd number of fused rings is present, the most stable tautomer is the one with protons located at the central ring. In **6**, with two central rings, the one remote from the fused benzo group is favored. Energies of tautomers due to monoprotic processes reflect this observation, although the energetic benefits are less important, suggesting that diprotic tautomerism is preferred. Finally, if tautomerism of the pyrazinacenes can be modulated (or frozen, for instance, by *N*-alkylation), then they present an excellent opportunity for the study of how aromaticity (or antiaromaticity) varies depending on subtle variations in the structure of the molecules. Previous studies focused on this feature have given invaluable insights into the role of aromaticity on the stabilities of particular electronic structures.[5c,d]

Initially, we isolated compounds assigned the structures **2** and **3**, which we subjected to alkylation using simple alkyl halides at elevated temperatures (this was originally for purposes of derivatization to facilitate chemical analysis). *N*-alkylation of octaazatetracenes **2** and **3** gave simple mixtures, each of three compounds (with traces of other compounds of the same mass) which could be separated and identified as isomers of *N*,*N*′-dialkyl-octaazatetracenes. Proton NMR spectroscopy (see Figure 2) indicates which of the isomers is unsymmetrically substituted (i.e., *N*-alkyl groups on different pyrazine rings rather than on the same one). The evidence for existence of the di-*N*-alkylated pyrazinacenes is consistent with data from our calculations and also suggests that mono- and diprotic processes can both occur and that tautomers with protons located on central pyrazine rings are more stable. Hence, only three tautomers were isolated, presumably with alkyl groups on the central nitrogen atoms.[19] On the basis of the mixture of isomers obtained from the *N*-alkylation of octaazatetracene type compounds, **4**, we had expected that they might bear delocalized electronic (or delocalized protonic) systems, although this is not the case obtained from calculations (see Figure 4). Since the *N*-alkylation reaction is performed at elevated temperatures (100−140 °C), we believe that it is possible that formation of the delocalized state might be thermally activated, although it is probably not required for tautomerization to occur (since the *N*-alkylation reaction is performed under mildly basic conditions in a polar medium, e.g. dimethylsulfoxide, both of which are known to facilitate intra- and intermolecular protic reactions).

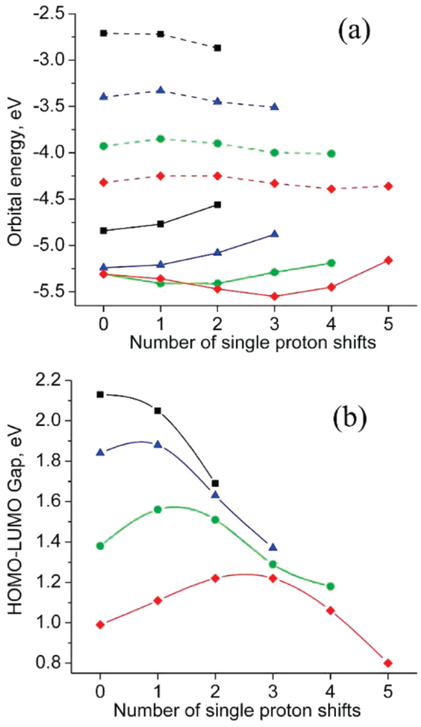


***Figure 7.*** (a) Proton location dependency of energies of highest occupied and lowest unoccupied molecular orbitals for tautomers of **4** (black), **5** (blue), **6** (green), and **7** (red) following consecutive single proton shifts. (b) Variation in HOMO−LUMO energy gap depending on the number of single proton shifts in **4** (black), **5** (blue), **6** (green), and **7** (red).

**HOMO−LUMO Levels and Gap.** Energy levels of HOMOs are somewhat influenced by the tautomerization processes, while those of the LUMOs are less affected (see Figures 7 and 8 for monoprotic and diprotic tautomerizations, respectively). During monoprotic shifts, HOMOs of tautomers of **4** and **5** increase gradually in energy, while **6** and

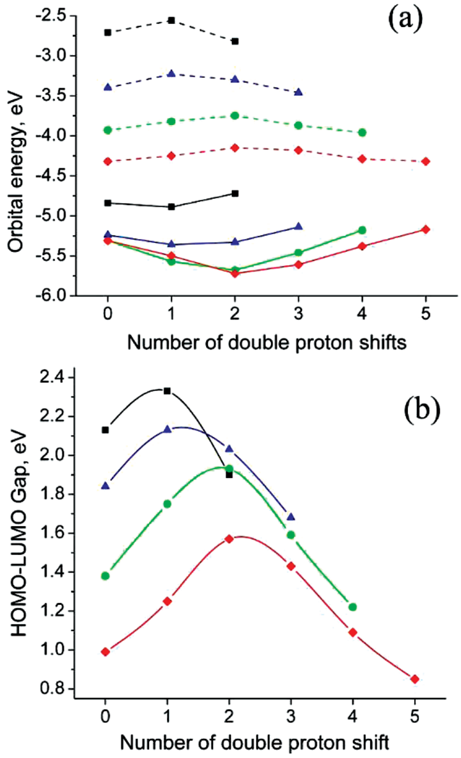


***Figure 8.*** (a) Proton location dependency of energies of highest occupied and lowest unoccupied molecular orbitals for tautomers of **4** (black), **5** (blue), **6** (green), and **7** (red) following consecutive double proton shifts. (b) Variation in HOMO−LUMO energy gap depending on the number of double proton shifts in **4** (black), **5** (blue), **6** (green), and **7** (red).

**7** go through a minimum after two and three proton shifts, respectively. For diprotic shifts, **4** and **5** tautomers have a slight minimum after a single shift, while **6** and **7** both reach their minimum orbital energy after two shifts. The curious uniting feature of these observations is that, for **6** and **7**, the degree of HOMO delocalization is at a minimum (according to the HOMO structures in Figures 4 and 5) in the HOMOs of lowest energy. However, this does not necessarily disagree with the statement of increased delocalization of the inner ring because, in the HOMO, we have contributions from all of the $\pi$ electrons of the molecule.

In 2H-pyrazinacenes **4**−**7**, the HOMO level is slightly stabilized with an increasing number of fused rings, while the LUMO level undergoes a much more significant stabilization. The latter has the effect of reducing the HOMO−LUMO gaps for **4**−**7**, which are also shown in Figures 7 and 8. The magnitude of the HOMO−LUMO gap gradually decreases with an increasing number of fused rings, as expected.[20] Tautomerism in the individual systems has the effect of increasing the HOMO−LUMO gap for molecules with centrally placed reduced pyrazine rings. This is due to an interruption in the standard conjugation and the occurrence of the Clar rule, as discussed before. These molecules possess properties appropriate for their incorporation into thin film FET devices or similar.[21]

Tautomerism in aza-acenes has been discussed as far back as the 1890s, when dihydro-5,7,12,14-tetraazapentacene **8** was erroneously ascribed a quinonoid structure, **8a**.[22] Proton

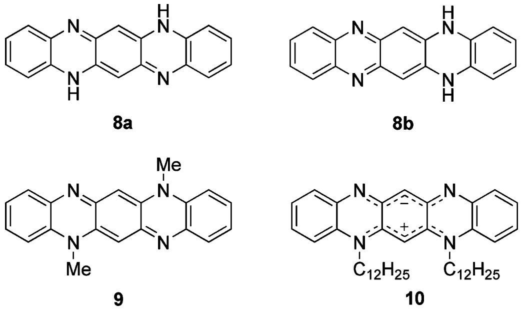


***Figure 9.*** Chemical structures of compounds **8a**,**b**, **9**, and **10**.

NMR measurements subsequently revealed the benzenoid form **8b**[7a] in solution, although other studies have found that the quinonoid form can be stabilized by **8**'s N-methylation, giving **9** (see Figure 9).[7b] Actually, direct or indirect introduction of N-alkyl groups into aza-acene compounds has been shown as a method for obtaining products with unusual zwitterionic electronic structures such as **10**.[23] In the case of the pyrazinacenes, the situation is complicated by the ability of the corresponding molecules to undergo tautomerization through proton transfer(s) to an adjacent pyrazine ring, and N-alkylation also results in unsymmetrically substituted products. The uniqueness of these compounds originates from the fusion of several pyrazine rings and the absence of interrupting carbon-only six-membered rings. Prior to the pyrazinacenes, only a few examples of fused pyrazines were available, and all contained bulky N-substituent side groups.[24] The potential importance of the pyrazinacenes and their relations has been emphasized by several recent computational investigations on their electronic structures.[5,13] Also, certain reduced derivatives, such as the dihydrodiazatetracene of Miao et al.,[5c] have provided insight into questions regarding the influence of reduced pyrazine rings on the aromaticity and stability of these compounds. Other studies have found that the introduction of a dihydropyrazine ring into oligoacenes can improve their properties (i.e., stability against oxidation) with regard to device preparation and operation.[22] This can be extended to the pyrazinacenes since we know that they have much greater stabilities and solubilities than the corresponding CH-only analogs.

NH tautomerism is a peculiarity of the present dihydropyrazinacene system that, at the same time as being scientifically important, has potential in some applications. One can imagine isolated lengthy reduced pyrazinacenes acting as discrete proton transporters in protonic devices or polymeric derivatives being used in proton-conducting membranes inside fuel cells. Also, the increased number of heteroatoms with smaller atomic radii in these systems allows for close intermolecular contacts when stacking, which should favor charge transport in their thin films and improve their potential as organic semiconducting materials.

## Conclusions

We have investigated protic tautomerism in a series of reduced nitrogen-rich oligoazaacenes, the pyrazinacenes, starting with experimental structures and using computational approaches to access information on proton shift and electronic structure not directly accessible to experimental probes. We found that the energy barrier for the displacement of a proton from an initial position systematically decreases by increasing the number of fused pyrazinacene rings. The energy barrier is also subject to a peculiar odd−even effect whose origin, although partly unknown, can be related to resonance effects depending on the ring multiplicity. Moreover, we found that both monoprotic and diprotic tautomerism in the pyrazinacenes influences strongly the structures of the highest occupied molecular orbital, especially in the acenes containing five or more fused rings. The latter is as a result of disruption of delocalization. However, in longer molecules, another effect appears: $\pi$ electrons of the intrinsically antiaromatic rings of the dihydropyrazine in the center of molecules can be more easily delocalized, partly accounting for the increased stability of the compounds with protons at central positions. Tautomers with terminal dihydropyrazine groups are destabilized by the loss of Clar resonance stabilization. At least three fused pyrazine rings are necessary for proper delocalization of the HOMO levels, and this feature is consistent with our experimental observation that the tautomerization of protons at reduced pyrazine rings of pyrazinacenes only occurs significantly in compounds containing at least four fused rings, such as *N*,*N*-dihydrooctaazatetracene (e.g., **2a**,**b**). Because of this feature, we suggest that protic tautomerization in pyrazinacenes containing a reduced ring is strongly associated with delocalization of the $\pi$ electrons of the remaining pyrazine groups. Thus, it might be inferred that multiple tautomerizations or "conduction" of protons along a pyrazinacene backbone in one direction results in a net transfer of electron density in the opposite direction. The connotations of the peculiarities of proton and electron transport for the properties of the pyrazinacenes remain to be seen, especially since intermolecular processes would be likely also involved in the operation of any condensed matter devices assembled using these compounds. However, the enhanced stability of these compounds over other acenes (and even other reduced heteroacenes) means that they are amenable to development as materials for thin film transistor electronics or as proton conductors. We expect to report other experimental observations of these compounds shortly.

**Acknowledgment.** This research was supported by a Grant-in-Aid for Scientific Research on Priority Area "Super-Hierarchical Structures" and by the World Premier International Research Center Initiative (WPI Initiative) on Materials Nanoarchitectonics (to R.S., J.P.H., and K.A.) from MEXT, Japan. The authors are also grateful to Luigi Delle Site for critical reading of the manuscript. Some of the simulations were performed on the Linux Cluster Shizuwo at the International Center for Young Scientists (ICYS), NIMS.

**Supporting Information Available:** Structures of lowest unoccupied molecular orbitals (LUMOs) of tautomers of compounds **4**, **5**, **6**, and **7**. Cartesian coordinates of calculated structures. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) (a) Horowitz, G. *Adv. Mater.* **1998**, *10*, 365–377. (b) Halik, M.; Klauk, H.; Zschieschang, U.; Kriem, T.; Schmid, G.; Radlik, W.; Wussow, K. *Appl. Phys. Lett.* **2002**, *81*, 289–291. (c) Zhang, Y.; Petta, J. R.; Ambily, S.; Shen, Y.; Ralph, D. C.; Malliaras, G. G. *Adv. Mater.* **2003**, *15*, 1632–1635. (d) Park, J. G.; Vasic, R.; Brooks, J. S.; Anthony, J. E. *J. Appl. Phys.* **2006**, *100*, 044511/1–044511/6. (e) Okamoto, T.; Senatore, M. L.; Ling, M.-M.; Mallik, A. B.; Tang, M. L.; Bao, Z. *Adv. Mater.* **2007**, *19*, 3381–3384. (f) Briseno, A. L.; Mannsfeld, S. C. B.; Ling, M. M.; Liu, S.; Tseng, R. J.; Reese, C.; Roberts, M. E.; Yang, Y.; Wudl, F.; Bao, Z. *Nature* **2006**, *444*, 913–917. (g) Newman, C. R.; Chesterfield, R. J.; Panzer, M. J.; Frisbie, C. D. *J. Appl. Phys.* **2005**, *98*, 084506/1–084506/6. (h) Kitamura, M.; Arakawa, Y. *J. Phys. (Paris)* **2008**, *20*, 184011/1–184011/16. (i) Meng, H.; Bendikov, M.; Mitchell, G.; Helgeson, R.; Wudl, F.; Bao, Z.; Siegrist, T.; Kloc, C.; Chen, C.-H. *Adv. Mater.* **2003**, *15*, 1090–1093.

(2) (a) Clar, E. *Polycyclic Hydrocarbons*; Academic Press: London, 1964; Vols. 1, 2. (b) Bjorseth, A. *Handbook of Polycyclic Aromatic Hydrocarbons*; Dekker: New York, 1983. (c) Harvey, R. G. *Polycyclic Aromatic Hydrocarbons*; Wiley-VCH: New York, 1997. (d) Bendikov, M.; Wudl, F.; Perepichka, D. F. *Chem. Rev.* **2004**, *104*, 4891–4945.

(3) (a) Kelley, T. W.; Muyres, D. V.; Baude, P. F.; Smith, T. P.; Jones, T. D. *Mater. Res. Soc. Symp. Proc.* **2003**, *771*, 169–179. (b) Eremtchenko, M.; Temirov, R.; Bauer, D.; Schaefer, J. A.; Tautz1, F. S. *Phys. Rev. B* **2005**, *72*, 115430. (c) Gundlach, D. J.; Lin, Y. Y.; Jackson, T. N.; Nelson, S. F.; Schlom, D. G. *IEEE Electron Device Lett.* **1997**, *18*, 87–89.

(4) (a) Anthony, J. E. *Chem. Rev.* **2006**, *106*, 5028–5048. (b) Anthony, J. E.; Gierschner, J.; Landis, C. A.; Parkin, S. R.; Sherman, J. B.; Bakus, R. C. *Chem. Commun.* **2007**, 4746–4748. (c) Lehnherr, D.; McDonald, R.; Ferguson, M. J.; Tykwinski, R. R. *Tetrahedron* **2008**, *64*, 11449–11461. (d) Sele, C. W.; Kjellander, B. K. C.; Niesen, B.; Thornton, M. J.; van der Putten, J. B. P. H.; Myny, K.; Wondergem, H. J.; Moser, A.; Resel, R.; van Breemen, A. J. J. M.; van Aerle, N.; Heremans, P.; Anthony, J. E.; Gelinck, G. H. *Adv. Mater.* **2009**, [Online] DOI: 10.1002/adma.200901548.

(5) (a) Bendikov, M.; Duong, H. M.; Starkey, K.; Houk, K. N.; Carter, E. A.; Wudl, F. *J. Am. Chem. Soc.* **2004**, *126* (24), 7416–7417. (b) Constantinides, C. P.; Koutentis, P. A.; Schatz, J. *J. Am. Chem. Soc.* **2004**, *12* (6), 16232–16241. (c) Miao, S.; Brombosz, S. M.; Schleyer, P. v. R.; Wu, J. I.; Barlow, S.; Marder, S. R.; Hardcastle, K. I.; Bunz, U. H. F. *J. Am. Chem. Soc.* **2008**, *130*, 7339–7344. (d) Wu, J. I.; Wannere, C. S.; Mo, Y.; Schleyer, P. v. R.; Bunz, U. H. F. *J. Org. Chem.* **2009**, *74*, 4343–4349. (e) Chen, H.-Y.; Chao, I. *Chem. Phys. Chem.* **2006**, *7*, 2003–2007.

(6) (a) Bunz, U. H. F. *Chem.—Eur. J.* **2009**, *15*, 6780–6789. (b) Appleton, A. L.; Miao, S.; Brombosz, S. M.; Berger, N. J.; Barlow, S.; Marder, S. R.; Lawrence, B. M.; Hardcastle, K. I.; Bunz, U. H. F. *Org. Lett.* **2009**, *11*, 5222–5225. (c) Miao, S.; Appleton, A. L.; Berger, N.; Barlow, S.; Marder, S. R.; Hardcastle, K. I.; Bunz, U. H. F. *Chem.—Eur. J.* **2009**, *15*, 4990–4993. (d) Miao, S.; Schleyer, P. v. R.; Wu, J. I.; Hardcastle, K. I.; Bunz, U. H. F. *Org. Lett.* **2007**, *9*, 1073–1076.

(7) Some oligoazaacenes do exhibit NH tautomerism. For example: (a) Sawtschenko, L.; Jobst, K.; Neudeck, A.; Dunsch, L. *Electrochim. Acta* **1996**, *41*, 123–131. (b) Tang, Q.; Liu, J.; Chan, H. S.; Miao, Q. *Chem.—Eur. J.* **2009**, *15*, 3965–3969.

(8) *CPMD*; IBM Corp.: Armonk, NY, 1990−2001; MPI für Festkörperforschung: Stuttgart, Germany, 1997−2004.

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

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

(11) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

(12) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

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

(14) Barnett, R. N.; Landman, U. *Phys. Rev. B* **1993**, *48*, 2081–2097.

(15) (a) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566. (b) Iannuzzi, M.; Laio, A.; Parrinello, M. *Phys. Rev. Lett.* **2003**, *90*, 238302.

(16) (a) Boero, M.; Ikeshoji, T.; Liew, C. C.; Terakura, K.; Parrinello, M. *J. Am. Chem. Soc.* **2004**, *126*, 6280–6286. (b) Iannuzzi, M. *J. Chem. Phys.* **2006**, *124*, 204710. (c) Cucinotta, C. S.; Ruini, A.; Catellani, A.; Stirling, A. *ChemPhysChem* **2006**, *7*, 1229–1234. (d) Boero, M.; Ikeda, T.; Ito, E.; Terakura, K. *J. Am. Chem. Soc.* **2006**, *128*, 16798–16807. (e) Glezakou, V. A.; Dupuis, M.; Mundy, C. J. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5752–5760. (f) Alfonso-Prieto, M.; Biarnes, X.; Vidossich, P.; Rovira, C. *J. Am. Chem. Soc.* **2009**, *131*, 11751–11761.

(17) (a) Richards, G. J.; Hill, J. P.; Subbaiyan, N. K.; D'Souza, F.; Karr, P. A.; Elsegood, M. R. J.; Teat, S. J.; Mori, T.; Ariga, K. J. *Org. Chem.* **2009**, [Online] DOI: 10.1021/jo901832n. (b) Richards, G. J.; Hill, J. P.; Okamoto, K.; Shundo, A.; Akada, M.; Elsegood, M. R. J.; Mori, T.; Ariga, K. *Langmuir* **2009**, *25*, 8408–8413.

(18) Winkler, M.; Houk, K. N. *J. Am. Chem. Soc.* **2007**, *129*, 1805–1815.

(19) An X-ray crystallographic study of the symmetrically and unsymmetrically N-substituted pyrazinacenes is currently underway.

(20) (a) Anthony, J. E. *Angew. Chem., Int. Ed.* **2008**, *47*, 452–483. (b) Zhang, X.; Côté, A. P.; Matzger, A. J. *J. Am. Chem. Soc.* **2005**, *127*, 10502–10503.

(21) (a) Miao, Q.; Nguyen, T.-Q.; Someya, T.; Blanchet, G. B.; Nuckolls, C. *J. Am. Chem. Soc.* **2003**, *125*, 10284–10287. (b) Tang, Q.; Zhang, D.; Wang, S.; Ke, N.; Xu, J.; Yu, J. C.; Miao, Q. *Chem. Mater.* **2009**, *21*, 1400–1405.

(22) (a) Fischer, O.; Hepp, E. *Chem. Ber.* **1890**, *23*, 2789–2793. (b) Fischer, O.; Hepp, E. *Chem. Ber.* **1900**, *33*, 1485–1498. (c) Hinsberg, O. *Liebigs Ann. Chem.* **1901**, *319*, 257–286. (d) Badger, G. M.; Pettit, R. *J. Chem. Soc.* **1951**, *73*, 3211–3215. (e) Nietszki, R. *Chem. Ber.* **1895**, *28*, 1357–1360. (f) Badger, G. M.; Pettit, R. *J. Chem. Soc.* **1951**, *4*, 3211–3215. (g) Beecken, H.; Musso, A. *Chem. Ber.* **1961**, *94*, 601–613.

(23) (a) Wudl, F.; Koutentis, P. A.; Weitz, A.; Ma, B.; Strassner, T.; Houk, K. N.; Khan, S. I. *Pure Appl. Chem.* **1999**, *71*, 295–302. (b) Riley, A. E.; Mitchell, G. W.; Koutentis, P. A.; Bendikov, M.; Kaszynki, P.; Wudl, F.; Tolbert, S. *Adv. Funct. Mater.* **2003**, *13*, 531–540.

(24) (a) Stöckner, F.; Beckert, R.; Gleich, D.; Birckner, E.; Günther, W.; Görls, H.; Vaughan, G. *Eur. J. Org. Chem.* **2007**, 1237–1243. (b) Stöckner, F.; Käpplinger, C.; Beckert, R.; Görls, H. *Synlett* **2005**, 643–645.

CT9006585

JCTC Journal of Chemical Theory and Computation

# Coarse-Grained Computer Simulations of Polymer/ Fullerene Bulk Heterojunctions for Organic Photovoltaic Applications

David M. Huang,* Roland Faller, Khanh Do, and Adam J. Moulé

*Chemical Engineering and Materials Science Department, University of California, Davis, California 95616*

Received September 17, 2009

**Abstract:** We develop coarse-grained (CG) computer simulation models of poly(3-hexylthiophene) (P3HT) and P3HT/fullerene-$C_{60}$ mixtures, in which collections of atoms from a physically accurate atomistic model are mapped onto a smaller number of "superatoms". These CG models allow much larger systems to be simulated for longer durations than is achievable atomistically, making it possible to study in molecular detail the morphology of polymer/fullerene bulk heterojunctions at length and time scales relevant to organic photovoltaic devices. We demonstrate that our CG models, parametrized at two state points, accurately capture the structure of atomistic systems at other points in the mixture phase diagram. Finally, we use our CG models to study the dynamic evolution of the microstructure of a P3HT/$C_{60}$ bulk heterojunction in a system approaching the device scale.

## 1. Introduction

Meeting the world's growing demand for energy with renewable, nonpolluting sources is one of the biggest challenges facing society.[1,2] Solar power is arguably the only source capable of supplying these needs into the next century.[3] But state-of-the-art crystalline silicon technology is currently not economically competitive with fossil fuels.[4] Organic photovoltaics (OPV), which includes polymer-based solar cells (PSCs), offers a cost-effective alternative to traditional crystalline silicon solar cells.[5] Advantages of OPV include solution processability, device flexibility, and the potential for high-volume reel-to-reel production, but device efficiencies must improve if OPV is to become commercially viable.[5]

Typical PSCs use a mixture of a light-absorbing semiconducting polymer as the electron donor and a fullerene derivative as the electron acceptor in the solar cell's photoactive layer. Because of the large discrepancy between the exciton diffusion length (~5−10 nm)[6] in the donor and the optimal active layer thickness for light absorption (>100 nm), the donor and acceptor phases are usually mixed together to form a bicontinuous network called a bulk heterojunction (BHJ),[7] in which generated excitons are (ideally) less than a diffusion length from a donor−acceptor interface. The delicate balance between maximizing interfacial area and maintaining percolating pathways for charge transport to the electrodes means that PSC device performance is sensitive to the BHJ morphology. Furthermore, the highly anisotropic nature of charge transport (e.g., intra- vs interchain or in the $\pi$-stacked vs side-chain direction) in semiconducting polymers[8] means that device efficiencies also depend on the molecular-scale organization of the donor and acceptor within the BHJ. Even with the same electron donor and acceptor materials, BHJ solar cell efficiencies can vary dramatically: for poly(3-hexylthiophene)/[6,6]-phenyl-$C_{61}$-butyric acid methyl ester (P3HT/PCBM) blends, one of the most widely used donor/acceptor combinations, varying the polymer chain length[9] and using processing steps such as heat tempering[10] and solvent soaking[11] can change device efficiencies from under 0.1% to the record efficiencies of ~5% that have recently been achieved.[11–13]

The fabrication of BHJ polymer solar cells is currently more of an art than a science: despite the crucial importance of the heterojunction morphology for device performance, it is still not known how to predict the microstructure of the photoactive layer based on the constituent materials or

* Corresponding author e-mail: dmhuang@ucdavis.edu.

10.1021/ct900496t
Published on Web 12/30/2009

processing conditions. Although a wealth of nanoscale morphological information on semiconducting polymer thin films and BHJs has become available from experimental techniques such as X-ray diffraction,[14–26] atomic force microscopy,[21,24–27] transmission electron microscopy[24,25] and tomography,[28,29] optical spectroscopy,[17,23,24,30] spectroscopic ellipsometry,[30] and near-edge X-ray absorption fine structure (NEXAFS) spectroscopy,[30] the molecular-scale structure is not easily resolved by any of these methods and must usually be inferred indirectly. An unambiguous assignment of the structure is thus often not possible. The lack of long-range order in semiconductor polymer films further complicates data interpretation. The phase behavior of polymer/fullerene mixtures and its relationship to charge transport and device performance has also been investigated experimentally, but even for the same P3HT/PCBM blend, there is some disagreement between the published phase diagrams;[31,32] difficulties in interpreting the measured data make resolving these discrepancies a challenge.

Accurate computer simulation models can play an important role in elucidating the phase behavior of polymer/fullerene mixtures, the morphology of BHJs, and the influence of the morphology on charge transport; they can also aid the interpretation of experimental data, because particle positions can be tracked exactly during the course of a simulation. A number of atomistic computer simulation studies[8,33] have examined the molecular structure of organic semiconductors[34,35] and its effect on charge transport.[36–38] However, the computational demands of atomistic simulations mean that systems of only a few nanometers can be readily studied. The study of domains the size of the exciton diffusion length (~5−10 nm), the length scale of interest for charge transport in polymer solar cells, is thus computationally prohibitive by these methods.

Coarse-grained (CG) simulations,[39–43] in which collections of atoms from an atomistic model are mapped onto a smaller number of "superatoms", allow multiple domains the size of the exciton diffusion length to be studied while retaining significant information about the molecular structure, thereby allowing the BHJ morphology and its effect on charge transport (if the simulation model is coupled to a charge transport model[8]) to be analyzed on length scales relevant to polymer photovoltaics.

In this article, we develop CG models of poly(3-hexylthiophene) (P3HT), one of the most widely used semiconducting polymers in OPV, and $C_{60}$, the simplest fullerene, and mixtures of the two materials. ($C_{60}$ is studied in this initial study because of its simplicity, but our future work will concentrate on PCBM, the fullerene most widely used in PSCs.) We use accurate atomistic force fields as the starting point for developing the CG models. We parametrize the CG interactions using simulations at 550 K and 1 atm to ensure that the simulated systems are in the fluid state and isotropic, because the nonbonded pair interactions are assumed to be isotropic in the CG models. We then verify that the CG models accurately capture the phase behavior of the atomistic models at various temperatures and mixture compositions. Finally we demonstrate that the CG models can be used to study the structure and dynamic evolution of the BHJ microstructure of polymer/fullerene mixtures for a system approaching the photovoltaic device scale and down to temperatures where phase separation is expected to occur.

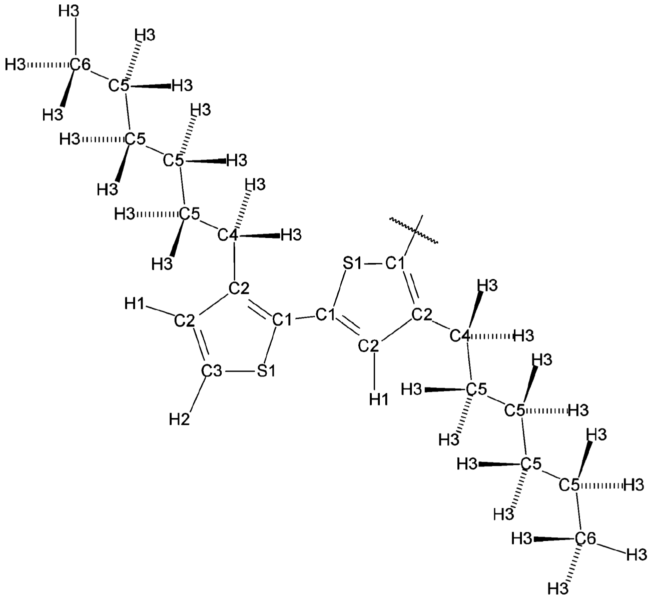


*Figure 1.* Chemical structure of last two monomers in a poly(3-hexylthiophene) (P3HT) chain. Atoms of different types in our atomistic model are labeled with different numerical suffixes.

## 2. Atomistic Simulation Models

The atomistic model of P3HT used in this work was adapted from the simulation model of tetrathiophene (T4) developed by Marcon and Raos[44] (which we will call the MR model from now on). The molecular geometry and atom types in our P3HT model are depicted in Figure 1. For the rest of this paper, we will focus on simulations of 100% regioregular P3HT (rr-P3HT), in which all monomers are joined head-to-tail; the models and methods presented in this paper are, however, equally applicable to P3HT with any degree of regioregularity.

Equilibrium distances and angles and partial charges on the atoms in the MR model were determined from ab initio density functional theory calculations.[44,45] In addition to Coulombic interactions between the point charge sites on the atoms, nonbonded atoms in the MR model also have van der Waals interactions described by the Lennard−Jones (LJ) potential, $V_{ij}^{LJ}(r) = 4\varepsilon_{ij}[(\sigma_{ij}/r)^{12} - (\sigma_{ij}/r)^{6}]$, truncated at a cutoff distance of 12 Å. The parameters for the LJ diameter $\sigma_{ij}$ and interaction strength $\varepsilon_{ij}$ in the MR model[44] were obtained from the OPLS-AA model,[46] in which heteronuclear interaction parameters are specified by the geometric mean of the homonuclear parameters, i.e., $\sigma_{ij} = (\sigma_{ii}\sigma_{jj})^{1/2}$ and $\varepsilon_{ij} = (\varepsilon_{ii}\varepsilon_{jj})^{1/2}$. As in the OPLS-AA model, atoms in the same molecule separated by more than three bonds have the same nonbonded interactions (LJ + Coulombic) with each other as atoms on different molecules, atoms separated by three bonds interact 1/2 as strongly, and atoms separated by one or two bonds have no nonbonded interactions with each other. The MR model has been found to give good agreement with experiment for the density, X-ray crystal structure, and heat of sublimation of T4[44] and therefore provides a good basis for our P3HT model.

The thiophene monomers in the MR model are slightly asymmetric, due to the significant chain-end effects in T4, whereas the thiophene units far from the ends in polythiophene (PT) should have reflection symmetry. For the PT backbone in our P3HT model, we have therefore used averaged equilibrium bond lengths and angles and bond stretching and bending force constants from the central monomers in the MR model and the charges from the half of the central monomers closest to the central intermonomer bond. The very slight asymmetry of the monomers in the MR model means that these modifications are small and should not have any noticeable quantitative impact on the results from our model. The simulation parameters for the alkyl side chain were taken directly from the OPLS-AA model,[46] except for the charge on the C4 carbon (see Figure 1), which was fixed by the requirement of monomer charge neutrality. The resulting partial charge of 0.0617e, where e is the elementary charge unit, is only slightly different from that of the equivalent site type ($CH_2$ bonded to aromatic C) of −0.005e in the OPLS-AA model. Marcon and Raos[45] used a similar strategy to determine the charge on the alkyl carbon directly bonded to the thiophene ring for their model of sexithiophene and tetrahexylsexithiophene, which were found to give reasonable agreement with experiment for the density and X-ray crystal structure.[45] The charges on the terminal carbon and hydrogen on the terminal monomers were adapted from those of the equivalent sites in the MR model,[44] with equal charges added to these two sites to ensure charge neutrality of the terminal monomers. The nonbonded, bond stretching, and bond bending parameters used in the atomistic simulation model of P3HT are given, respectively, in Tables S1−S3 in the Supporting Information.

All simulations, including the coarse-grained simulations described below, were carried out with the LAMMPS molecular dynamics simulation package.[47] Unless otherwise stated, electrostatic interactions were calculated using the particle−particle particle−mesh (PPPM) method.[48] C−H bond distances were constrained with the SHAKE algorithm.[49] All atomistic simulations were carried out at constant temperature and pressure (NPT ensemble), using a Nosé−Hoover thermostat[50] and Nosé−Hoover barostat.[51] A timestep of 1.5 to 2 fs was used, depending on the temperature.

High-level ab initio quantum calculations of bithiophene were used to determine the torsional potential of the intermonomer dihedral of T4 in the MR model,[52] in which constrained geometry optimizations and energy calculations were carried out at 30° intervals and the resulting points fit to a sixth-order cosine function. However, recent density functional theory calculations of thiophene and regioregular 3-hexylthiophene (3HT) oligomers[53] have shown substantial variations in the intermonomer torsional potential as a function of chain length, as shown in Figure 2 for 3HT dimers and 14-mers, with the potential only converging for chains 10 monomer units or longer. These variations were attributed to increasing electron delocalization with increasing chain length, making distortions of the chain from planarity less favorable for longer chains. Indeed, these recent calculations show that, although the torsional potential of thiophene dimers and 3HT dimers has a global minimum at a dihedral angle $\phi$ of around 150°[52,53] and 135°,[53] respectively, the global minimum for 8-mers and larger is at 180°,[53] corresponding to a planar chain with monomers in the anti conformation. The local minimum at small dihedral angles corresponding to the syn conformer also moves to smaller angles for longer chains.[53] A planar structure is most consistent with the available experimental data: according to X-ray diffraction measurements of crystals of 3,4′,4″-trimethyl-2−2′:5′,2″-terthiophene (the trimer of regioregular 3-methylthiophene, which would be less sterically hindered than 3HT), the equilibrium dihedral angle is ∼173°,[54] while similar measurements of rr-P3HT thin films also indicate a predominance of the planar conformation.[15]

We have therefore modeled the intermonomer torsional potential $\tilde{V}_{\rm dihed}(\phi)$ in our atomistic model for P3HT using the ab initio torsional potential for the longest rr-3HT oligomer in ref 53, 3HT14, rather than using the bithiophene torsional potential in the MR model. The torsional potential calculations in ref 53 were carried out for rotations around the central intermonomer bond using rigid monomers whose atoms were fixed at their global minimum configuration and using a smaller basis set than the bithiophene calculations used in the MR model,[52] because of the computational expense of calculations of this nature for such large molecules. However, it is likely that the errors associated with the large chain length dependence of the torsional potential far outweigh those associated with the approximations used in these calculations. In fact, the torsional potential for bithiophene in ref 53 is quantitatively quite similar to that used in the MR model,[52] indicating that errors due to the smaller basis set, different level of theory, and rigid rotor approximation may be small. The ab initio intermonomer torsional potential $\tilde{V}_{\rm dihed}(\phi)$ cannot be used directly in the simulation force field for our P3HT model because the model

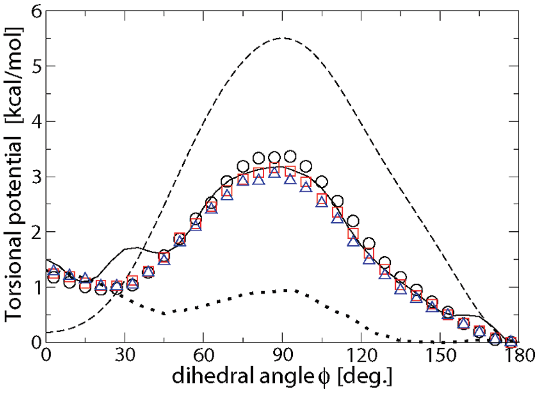


***Figure 2.*** Intermonomer torsional potential for a 3HT 14-mer (solid line) and dimer (dotted line) as a function of S1−C1−C1−S1 dihedral angle $\phi$ (see Figure 1 for definition of atom types) from ab initio calculations in ref 53 ($\tilde{V}_{\rm dihed}(\phi)$) and approximated as $-k_BT \ln P_{\rm dihed}(\phi)$ from constant NVT simulations at 300 (circles), 400 (squares), and 500 K (triangles) of a single 3HT hexamer in the gas phase with the atomistic simulation model used in this work, with intrinsic torsional potential $V_{\rm dihed}(\phi) = \sum_{i=0}^{8} c_i \cos^i(x)$ (dashed line), where $c_0 = 5.5121$, $c_1 = -0.0201$, $c_2 = -6.6011$, $c_3 = 1.1645$, $c_4 = 1.7991$, $c_5 = -5.1590$, $c_6 = 0.1496$, $c_7 = 4.1068$, and $c_8 = -0.7607$ kcal/mol.

includes nonbonded (LJ + Coulombic) interactions that are implicitly contained in $\tilde{V}_{\mathrm{dihed}}(\phi)$ in the ab initio calculations and which therefore must be subtracted from $\tilde{V}_{\mathrm{dihed}}(\phi)$ to obtain the "intrinsic" torsional potential $V_{\mathrm{dihed}}(\phi)$ that is used in the simulation force field. To account for the fact that the torsional potential of one dihedral angle depends on the orientation of more than just the adjacent monomers, we have determined an average torsional potential (in the "mean field" of monomers further along the chain) by carrying out constant NVT simulations of a single 3HT hexamer in the gas phase (longer chains gave indistinguishable results) at various temperatures and adjusting the intrinsic torsional potential $V_{\mathrm{dihed}}(\phi)$ of the S1−C1−C1−S1 dihedral so that $-k_BT \ln P_{\mathrm{dihed}}(\phi) \approx \tilde{V}_{\mathrm{dihed}}(\phi)$, where $P_{\mathrm{dihed}}(\phi)$ is the measured distribution of the central S1−C1−C1−S1 dihedral angle. (The torsional potentials of the C2−C1−C1−C2 and S1−C1−C1−C2 dihedrals were set to zero (see Figure 1 for the definitions of the atom types); other choices for the dihedral potentials, such as partitioning the total intrinsic torsional potential equally between the four dihedral angles, are not expected to affect results significantly.) Constant NVT simulations were carried out at 300, 400, and 500 K for 180, 36, and 18 ns, respectively. LJ and Coulombic interactions were truncated at a cutoff distance of 12 and 30 Å, respectively (the latter distance is larger than the molecule, so all Coulombic interactions were taken into account in this way). The intrinsic dihedral potential $V_{\mathrm{dihed}}(\phi)$ was specified by an eighth-order cosine series, $V_{\mathrm{dihed}}(\phi) = \sum_{i=0}^{8} c_i \cos^i(x)$. As shown in Figure 2, $-k_BT \ln P_{\mathrm{dihed}}(\phi)$ is relatively insensitive to temperature over the 200 K range used in the simulations, indicating that it is reasonable to approximate this torsional potential energy function by what is actually a free energy. We did not attempt to fit the smaller bumps in the torsional potential for 3HT14 from ref 53 shown in Figure 2, which the authors of ref 53 concede may be an artifact of the rigid-rotor approximation used in their calculations. For the purposes of developing a coarse-grained model of P3HT, this level of accuracy, namely getting the positions of the minima in the potential and the barrier height approximately correct, is sufficient. Parameters for the torsional potentials for the alkyl side chains of P3HT were taken from the OPLS-AA model.[46,55]

Comparison of the results of simulations with our atomistic P3HT model with experimental data indicate that the model accurately represents the structure of P3HT. The monomer density (0.931 ± 0.003 g/mL) from a 0.7-ns constant NPT simulation of 256 3HT monomers at 298 K and 1 atm agrees well with the measured density (0.936 g/mL)[56] at the same thermodynamic conditions. The simulated density (1.05 g/cm$^3$) from a constant NPT simulation of a crystal of 3HT 12-mers also agrees with the measured density (1.10 ± 0.05 g/cm $^3$)[57] of P3HT thin films. (This simulation involved annealing at 1 atm an fcc lattice of 72 aligned chains all in the anti conformation from 400 to 298 K over 1 ns and then equilibrating at 298 K for another 1 ns, during which the density was measured.) Although a direct comparison between oligomer and polymer data is not strictly correct, the 3HT 12-mers studied should be long enough that the structure of the crystal is similar to that of long chains. Unfortunately, experimental data that can be directly compared with our atomistic P3HT model is limited, because atomistic simulations of long P3HT chains, on which most experiments of P3HT morphology have focused, are not feasible; on the other hand, little experimental data exists for 3HT oligomers.

Turning to the atomistic model for $C_{60}$, the LJ parameters of the carbon atoms were taken from ref 58 and are given in Table S1 in the Supporting Information (the partial charges on the carbon atoms were all zero). The parameters in this model were obtained by computing the energy of an fcc crystal of $C_{60}$ (treating the $C_{60}$ molecule as a sphere with a surface of uniform density of carbon atoms) and comparing the results to experimental data for the heat of sublimation and lattice constant. The calculated compressibility from this model is close to the experimental value for an fcc crystal of $C_{60}$. The resulting parameters are also quite close to those in the OPLS-AA model for aromatic carbon atoms (see Table S1 in the Supporting Information),[46] indicating that combining our P3HT model (which uses OPLS-AA LJ parameters) with this $C_{60}$ model should give reasonable results. The equilibrium bond lengths of $C_{60}$ (see Table S2 in the Supporting Information) were taken from gas-phase electron diffraction measurements[59] and are consistent with the $C_{60}$ diameter in ref 58. Given these bond lengths and $C_{60}$'s icosahedral geometry, the equilibrium bond angles (see Table S3 in the Supporting Information) and improper dihedral angles in $C_{60}$ can readily be calculated. Harmonic improper dihedral potentials $V_{\mathrm{improp}}(\xi) = k_\xi(\xi - \xi_0)^2/2$ were defined so as to maintain the icosahedral geometry of $C_{60}$ and to maintain the planar geometry of the thiophene rings in P3HT. The force constant $k_\xi$ was 40 kcal/mol/rad$^2$ in all cases.

3HT 12-mers (3HT12) were used in the atomistic simulations, as oligomers of this length have been shown previously[60,61] to behave sufficiently like long-chain polymers to be used in the coarse-graining procedure. The initial system configurations used in the simulations comprised random polymer chains and randomly placed fullerenes. An initial energy minimization of the system with soft nonbonded interaction potentials given by a truncated cosine function was used to eliminate particle overlaps. Simulations were carried for a total time of at least $6\tau_2$ and at least $9\tau_2$ for the simulations used to calculate distribution functions used in the optimization of the CG models, where $\tau_2$ is a measure of the time scale for reorientation of the polymer chains and is obtained by fitting the autocorrelation function of the unit vector $\hat{\mathbf{u}}(t)$ between the polymer chain ends to the equation $\langle P_2[\hat{\mathbf{u}}(t)\cdot\hat{\mathbf{u}}(0)] \rangle \sim \exp(-t/\tau_2)$, where $P_2(x) \equiv (3x^2 - 1)/2$ is the second-order Legendre polynomial. Thus, total simulation times varied between around 5 and 35 ns. Mixtures with mole ratios that include those typically used in P3HT:PCBM solar cells were studied.[31] The thermodynamic conditions studied were chosen so as to include state points in the liquid phase and at or close to solid/liquid coexistence, based on the published experimental phase diagrams of P3HT:PCBM mixtures,[31,32] which are expected to exhibit solidification of the fullerene at a lower temperature than P3HT:$C_{60}$ mixtures, due to the disordering effect of the PCBM side chain. (Unfortunately, to the best of our

***Table 1.*** Temperatures and Mixture Ratios of Atomistic Systems Studied (pressure = 1 atm in all cases)[a]

| n(3HT12)/n($C_{60}$) | P3HT:$C_{60}$ (w/w) | P3HT:PCBM equiv[b] (w/w) | temperature (K) |
|---|---|---|---|
| 60/0 | 1.00:0 | 1.0:0 | 500, 550, 650 |
| 50/55 | 2.52:1 | 2.0:1 | 550 |
| 48/72 | 1.85:1 | 1.5:1 | 550, 650 |
| 42/92 | 1.27:1 | 1.0:1 | 550, 650 |

[a] Regioregular P3HT (rr-P3HT) was used in all cases. [b] P3HT:PCBM mixture with same mole ratio as P3HT:$C_{60}$ mixture.

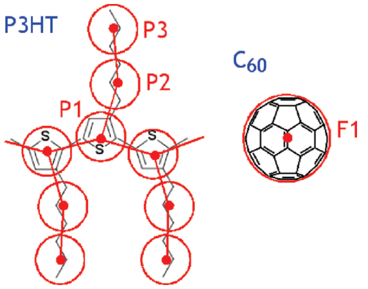


***Figure 3.*** Chemical structures of P3HT and $C_{60}$ with coarse-grained sites depicted and labeled.

knowledge, no experimental phase diagram of P3HT:$C_{60}$ mixtures exists for a more direct comparison.) As a point of reference, the polymer/fullerene = 1:1 w/w composition is the mostly commonly used mixture in P3HT/PCBM photovoltaic devices, although it has been suggested that a 2:1 w/w composition is optimal.[31] Although desirable, it is challenging to simulate well-equilibrated atomistic systems under thermodynamic conditions far into regions of the phase diagrams in which solids exist, because of the computational expense of the atomistic simulations, and so we have limited the atomistic simulations to temperature above 500 K. CG simulations at lower temperatures are, however, feasible. Table 1 summarizes the various P3HT:$C_{60}$ mixture ratios and temperatures studied.

## 3. Coarse-Grained Models and Methods

Our strategy in designing coarse-grained (CG) models of P3HT and $C_{60}$ was to use the simplest models that would accurately capture the structure of these molecules. To this end, we modeled the P3HT monomer using three sites: (1) the center-of-mass (COM) of the thiophene ring and the COM of the carbon atoms of the (2) first three and (3) last three side-chain methyl groups. A single site, the molecule's COM, was used for the CG model of $C_{60}$. Figure 3 illustrates the coarse-graining scheme used.

The interactions between CG sites were iteratively optimized to reproduce the atomistic system's structure (radial distribution functions (RDFs) of nonbonded sites and bond, angle, and dihedral distributions) using the iterative Boltzmann inversion (IBI) method, which has been described elsewhere in detail.[41,42] In our implementation of the IBI method, the potential energy $U_{i+1}(x)$ of a particular interaction type at the $(i + 1)$th iteration was calculated from the potential energy $U_i(x)$ at the $i$th iteration using

$$U_{i+1}(x) = U_i(x) + a_i k_B T \ln\left[\frac{P_i(x)}{P^{\text{target}}(x)}\right] \tag{1}$$

where $P_i(x)$ is the probability distribution of the variable $x$ calculated from the CG simulation during iteration $i$, $P^{\text{target}}(x)$ is the target distribution calculated from the atomistic simulation ($x$ can be the distance $r$, bond length $l$, bond angle $\theta$, proper dihedral angle $\phi$, or improper dihedral angle $\xi$ for nonbonded interactions, bond stretching, bond bending, or proper or improper torsions, respectively), and $0 \le a_i \le 1$ ($a_i$ was decreased or increased between iterations depending on whether the CG probability distribution diverged from or converged too slowly to the target distribution in the preceding iteration). For the initial potential energy function, we used

$$U_0(x) = -k_B T \ln[P^{\text{target}}(x)] \tag{2}$$

where $P_{\text{target}}(x) \propto g_{\text{target}}(r)$ (the RDF), $P^{\text{target}}_{\text{bond}}(l)/l^2$, $P^{\text{target}}_{\text{angle}}(\theta)/\sin\theta$, $P^{\text{target}}_{\text{dihed}}(\phi)$, and $P^{\text{target}}_{\text{improp}}(\xi)$, respectively, for nonbonded interactions, bond stretching, bond bending, and proper and improper torsions.

The CG simulations in which the CG interactions were optimized were carried out at constant temperature (NVT ensemble) with a Nosé–Hoover thermostat.[50] After optimization of the interactions using the above procedure, a linear correction,[41]

$$\Delta U_{jk}(r) = b_{jk}\left(1 - \frac{r}{r_c}\right), \quad r \le r_c \tag{3}$$

was added to all nonbonded interactions so that the pressure of the CG simulations matched those of the atomistic simulations (1 atm in all cases) and so that the RDFs were unchanged from the Boltzmann inversion step. Here $r_c$ is the cutoff distance in the nonbonded CG interactions ($r_c$ = 20, 25, and 27 Å, respectively, for the P3HT–P3HT, P3HT–$C_{60}$, and $C_{60}$–$C_{60}$ interactions) and $b_{jk}$ is a constant for each pair of site types $j$ and $k$ ($b_{jk}$ = −0.282, −0.158, and −2.063 kcal/mol, respectively, for the P3HT–P3HT, P3HT–$C_{60}$, and $C_{60}$–$C_{60}$ interactions).

The P3HT–P3HT CG interactions were optimized in simulations of pure P3HT (60 3HT 12-mers) at 550 K. Then, the P3HT–$C_{60}$ and $C_{60}$–$C_{60}$ CG interactions were optimized in simulations of 1.85:1 w/w P3HT:$C_{60}$ (48 3HT12 and 72 $C_{60}$) at 550 K with the P3HT–P3HT CG interactions fixed at their previously optimized values. Optimization of the CG interactions was carried out at 550 K to ensure that the systems were in the fluid state, because the use of isotropic interaction potentials between nonbonded pairs of sites implicitly assumes that the nonbonded site–site distributions in the atomistic systems from which the CG interactions are derived are isotropic as well. The use of isotropic nonbonded pair potentials in the CG models does not, however, preclude phase separation of the CG system as the temperature or mixture composition is varied, because phase separation arises from the collective interactions of many particles, nor does it mean that the CG models will not reasonably capture the behavior of the system as phase separation occurs. The CG nonbonded interaction potentials were defined numeri-

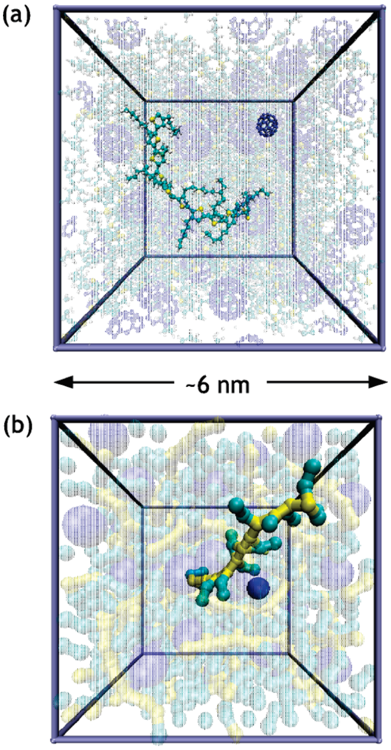


***Figure 4.*** Snapshots of configurations from (a) atomistic and (b) coarse-grained simulations of 48 3HT 12-mers and 72 $C_{60}$ molecules (P3HT:$C_{60}$ = 1.85:1 w/w) at 550 K and 1 atm. A single molecule of each type is highlighted.

cally at grid points, while the CG bonded interaction potentials were fit to polynomials in $l$, $\theta$, cos ($\phi$), and $\xi$, respectively, for bonds, angles, proper dihedrals, and improper dihedrals (see the Supporting Information for the model parameters). The end monomers of the oligomer chains were excluded from the calculation of distribution functions used in the optimization procedure to reduce end effects. A time-step of 5 fs was used in the CG simulations. Figure 4 shows typical snapshots from atomistic and CG simulations of mixtures of 3HT 12-mers and $C_{60}$.

## 4. Optimized Coarse-Grained Potentials

Approximately 10 Boltzmann inversion iterations each were required to optimize the CG interaction potentials for the pure 3HT12 system and for the 3HT12/$C_{60}$ mixture. Figures 5−9 depict, respectively, the radial distribution functions and bond length, bond angle, proper dihedral angle, and improper dihedral angle probability distributions for the pure 3HT12 system at 550 K from the atomistic simulation and from the CG simulation with the optimized CG interactions. The corresponding optimized CG interaction potentials are also shown. A few representative joint bond-length/bond-angle and bond-angle/dihedral-angle probability distributions from the atomistic and CG simulations are also plotted in the Supporting Information and show that the coarse-grained model accurately reproduces the cross-correlations between the bonded degrees of freedom in P3HT (the agreement between the joint probability distributions that are not shown is similarly good).

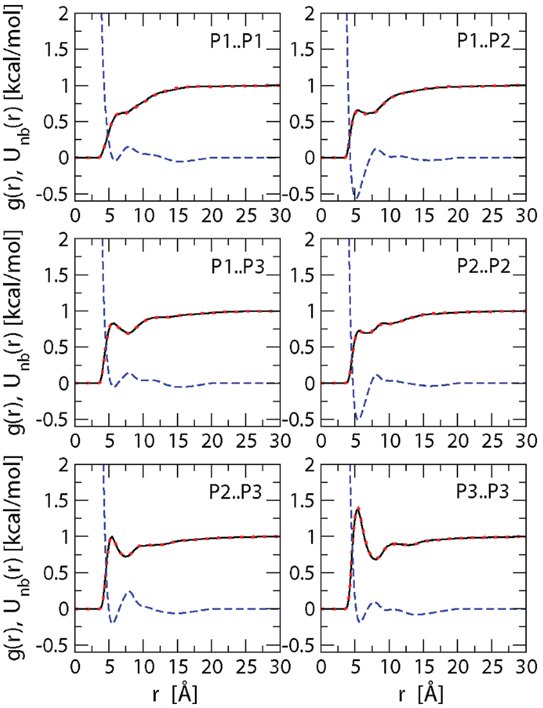


***Figure 5.*** Radial distribution functions $g(r)$ for nonbonded sites computed from constant NPT atomistic simulations of 60 P3HT 12-mers at 550 K and 1 atm (solid lines) and from constant NVT CG simulations at 550 K with optimized CG interaction potentials (dotted lines, which are almost indistinguishable from the solid lines). The optimized CG potentials $U_{nb}(r)$ are given by the dashed lines. (See Figure 3 for definitions of site types.)

Figure 6 and Figure 7 show that the bond length and angle distributions are unimodal and for the most part quite sharp, indicating that the bonds and angles in the CG model of P3HT are well-defined and physically meaningful. Note that because our rr-P3HT chains have directionality to them (much like a polypeptide has an *N* and a *C* terminus), distributions for the P1−P1−P2 and P2−P1−P1 angles and P1−P1−P2−P3 and P3−P2−P1−P1 proper dihedrals are different.

Figure 8 shows that the P2−P1−P1−P2 dihedral angle is roughly twice as likely to be 180° than to be 0°, indicating that on average every third monomer is in the syn conformation relative to one of its neighbors at 550 K. The higher probability of a P1−P1−P1−P1 dihedral angle of 0° compared with one of 180° shown in Figure 8 is consistent with configurations in which four monomers in a row are in the anti conformation being less common than those in which one pair is in the syn conformation. These findings are consistent with IR spectroscopic measurements on rr-P3HT thin films of the antisymmetric side-chain methylene stretch, which is sensitive to the chain conformation, which show a frequency more characteristic of a disordered chain than an all-anti conformation.[30] Other measurements also indicate that the conformation of P3HT chains, and rr-P3HT in particular, is not highly ordered: the vibrational frequency of the antisymmetric carbon−carbon stretch of the polymer backbone in rr-P3HT thin films indicate a conjugation length of five or six monomer units,[30,62] and although not strictly comparable with results on pure rr-P3HT films or melts, a

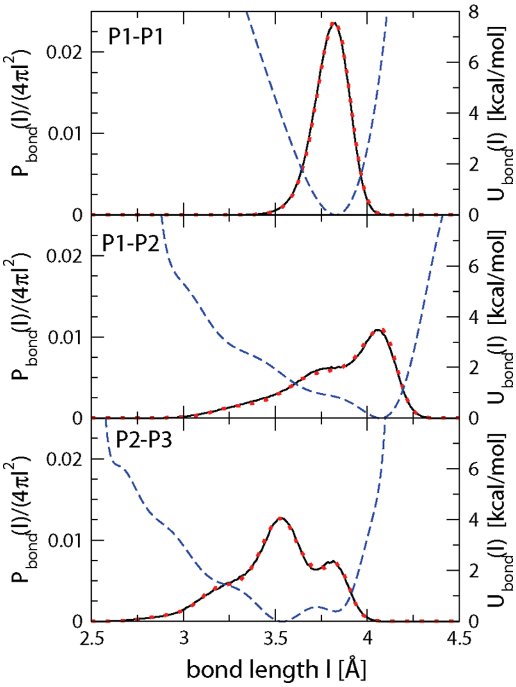


*Figure 6.* Bond length probability distributions $P_{bond}(l)$ computed from constant NPT atomistic simulations of 60 P3HT 12-mers at 550 K and 1 atm (solid lines) and from constant NVT CG simulations at 550 K (dotted lines) with optimized CG interaction potentials. The optimized CG bond ptentials $U_{bond}(l)$ are given by the dashed lines. (See Figure 3 for definitions of site types.)

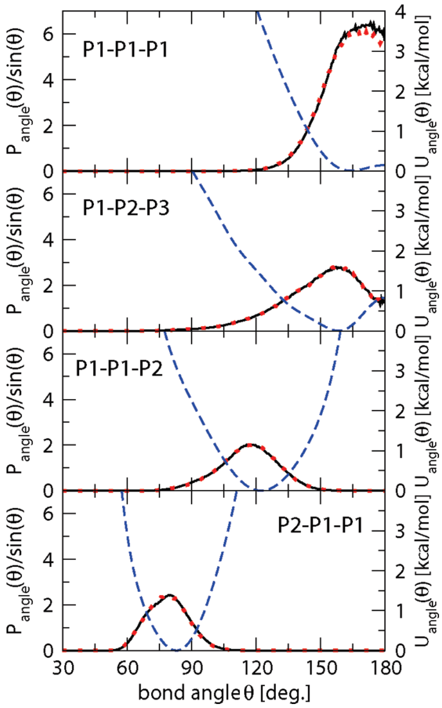


*Figure 7.* Bond angle probability distributions $P_{angle}(\theta)$ computed from constant NPT atomistic simulations of 60 P3HT 12-mers at 550 K and 1 atm (solid lines) and from constant NVT CG simulations at 550 K (dotted lines) with optimized CG interaction potentials. The optimized CG bond angle potentials $U_{angle}(\theta)$ are given by the dashed lines. (See Figure 3 for definitions of site types.)

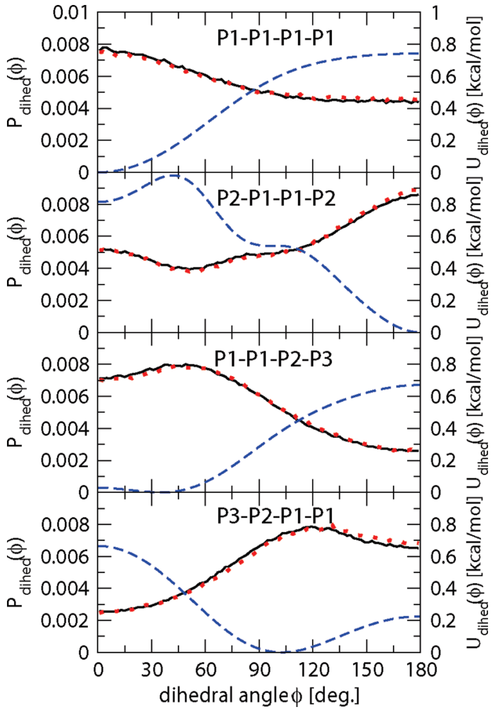


*Figure 8.* Dihedral angle probability distributions $P_{dihed}(\phi)$ computed from constant NPT atomistic simulations of 60 P3HT 12-mers at 550 K and 1 atm (solid lines) and from constant NVT CG simulations at 550 K (dotted lines) with optimized CG interaction potentials. The optimized CG dihedral potentials $U_{dihed}(\phi)$ are given by the dashed lines. (See Figure 3 for definitions of site types.)

persistence length of 2.4 ± 0.3 nm or roughly six monomer units has been measured for P3HT in THF at room temperature by light scattering.[63] These results indicate that P3HT does not exist in the completely ordered all-anti conformation, as is often depicted,[15,23,64] but is only ordered for roughly six monomer units on average. This value of five or six units (i.e., a 1:5 or 1:6 ratio of syn and anti conformers) measured experimentally at room temperature is consistent with a Boltzmann distribution of conformers with the same intermonomer torsion potential (the points in Figure 2) as that used in our simulation model, indicating that the torsion potential in our model provides an accurate description of the P3HT intermonomer dihedral. At the higher temperature of 550 K used in the simulations, this intermonomer torsion potential leads to the smaller 1:2 ratio of syn to anti conformers observed in Figure 8. The available X-ray diffraction data, while supporting the picture of the all-anti herringbone conformation of P3HT,[15,23,64] does not rule out the possibility of the occasional monomer pair existing in the syn conformation. This is a point worthy of further experimental study; our simulation model, which uses an intermonomer torsion potential from ab initio quantum calculations and which is consistent with several experimental results as discussed above, supports the picture of rr-P3HT existing mostly in the anti conformation but with a significant proportion of monomer pairs having a syn conformation.

Turning to the P3HT$-$C$_{60}$ and C$_{60}-$C$_{60}$ interactions, Figure 10 depicts the radial distribution functions from the 1.85:1

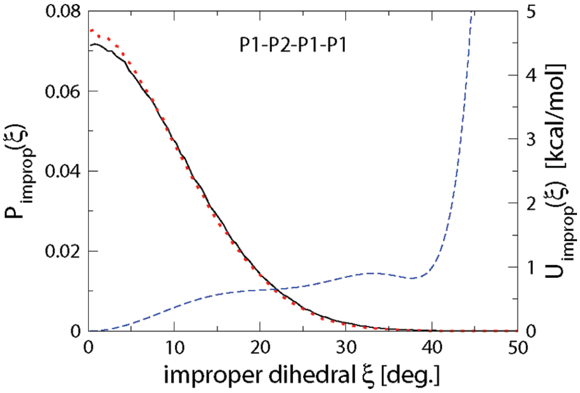


***Figure 9.*** P1−P2−P1−P1 improper dihedral angle probability distribution $P_{\mathrm{improp}}(\xi)$ (see Figure 3 for definitions of site types) computed from constant NPT atomistic simulations of 60 P3HT 12-mers at 550 K and 1 atm (solid line) and constant NVT CG simulations at 550 K (dotted line) with optimized CG interaction potentials. The optimized CG improper dihedral potential $U_{\mathrm{improp}}(\xi)$ is given by the dashed line.

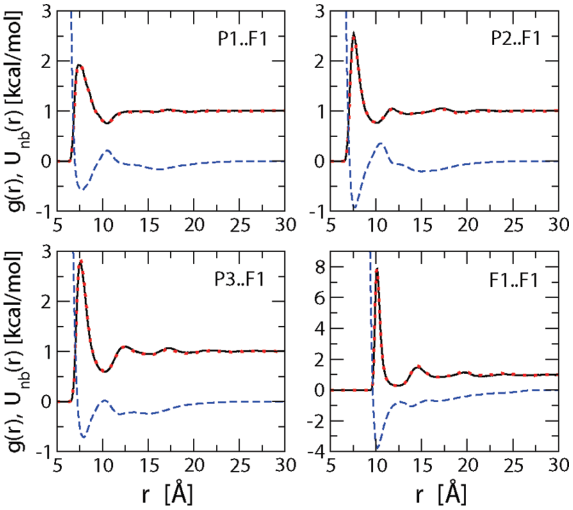


***Figure 10.*** Radial distribution functions $g(r)$ for nonbonded sites computed from constant NPT atomistic simulations of 48 P3HT 12-mers and 72 $C_{60}$ molecules at 550 K and 1 atm (solid lines) and constant NVT CG simulations with optimized CG interaction potentials at 550 K (dotted lines, which are almost indistinguishable from the solid lines). The optimized CG nonbonded interaction potentials $U_{\mathrm{nb}}(r)$ are given by the dashed lines. (See Figure 3 for definitions of site types.)

w/w mixture of P3HT 12-mers and $C_{60}$ at 550 K from the constant NPT atomistic simulation and from the constant NVT CG simulation with optimized CG interaction potentials. The corresponding optimized CG interaction potentials are also shown. Similarly good agreement between the atomistic and CG simulations to that shown in Figure 5 was found for the P3HT−P3HT distributions (not shown), even though the P3HT−P3HT interactions were not optimized in these simulations of P3HT/$C_{60}$ mixtures.

## 5. State-Point Dependence

In order to be generally useful, the CG models should accurately describe the behavior of the system at different thermodynamic states from those at which they were parametrized. We have therefore carried out constant NPT CG simulations at 1 atm of 3HT12/$C_{60}$ mixtures at temperatures other than 550 K and for P3HT:$C_{60}$ mixture ratios other than 1:0 and 1.85:1 w/w and have compared the resulting distributions with those of equivalent atomistic simulations. Figure 11 shows the RDFs for nonbonded sites from constant NPT atomistic and CG simulations of pure P3HT at 500 and 650 K at 1 atm. There is perfect agreement between the CG and atomistic simulations at both temperatures. The agreement between the atomistic and CG simulations for the bonded distributions (bond lengths, angles, and dihedrals), although not shown, is equally good.

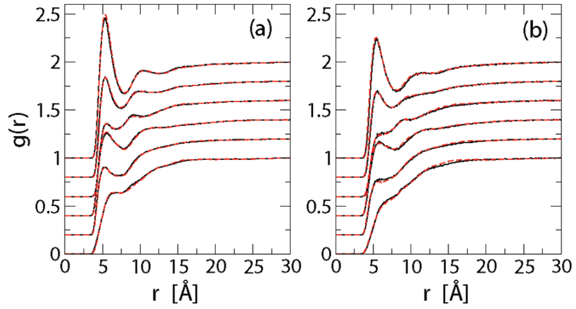


***Figure 11.*** Radial distribution functions (RDFs) for nonbonded sites from constant NPT atomistic (solid lines) and CG simulations (dashed lines, which are almost indistinguishable from the solid lines) of 60 P3HT 12-mers at (a) 500 and (b) 650 K. (RDFs have been shifted vertically for ease of viewing; site pairs (from bottom to top): P1..P1, P1..P2, P1..P3, P2..P2, P2..P3, and P3..P3 (see Figure 3 for definitions of site types).)

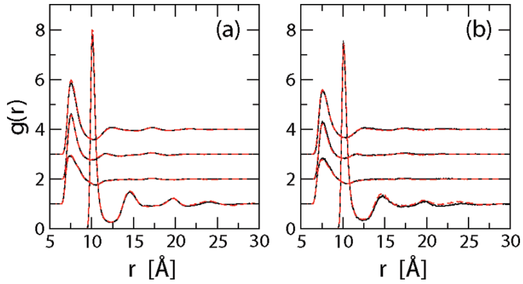


***Figure 12.*** Radial distribution functions (RDFs) for polymer−fullerene and fullerene−fullerene site pairs from constant NPT atomistic (solid lines) and CG simulations (dashed lines, which are almost indistinguishable from solid lines) for 3HT12:$C_{60}$ (w/w) of (a) 1.27:1 at 550 K and (b) 1.85:1 at 650 K. (RDFs have been shifted vertically for ease of viewing; site pairs (from bottom to top): P1..F1, P2..F1, P3..F1, and F1..F1 (see Figure 3 for definitions of site types).)

Figure 12 shows the RDFs for P3HT−$C_{60}$ and $C_{60}$−$C_{60}$ site pairs for 3HT12/$C_{60}$ mixtures at a couple of different temperatures and mixture compositions at 1 atm. For the systems presented in Figure 12 and also for all of the other thermodynamic states studied, all of the CG distributions agree with the atomistic ones, within the error bars on the points. (The P3HT−P3HT distributions are not shown for the P3HT/$C_{60}$ mixtures but display similarly good agreement.) The pure P3HT system at 500 K and the P3HT:$C_{60}$ = 1.27:1 w/w system at 550 K, in particular, are likely to be at or close to phase coexistence, given the experimental phase diagrams of P3HT:PCBM mixtures[31,32] and the highly

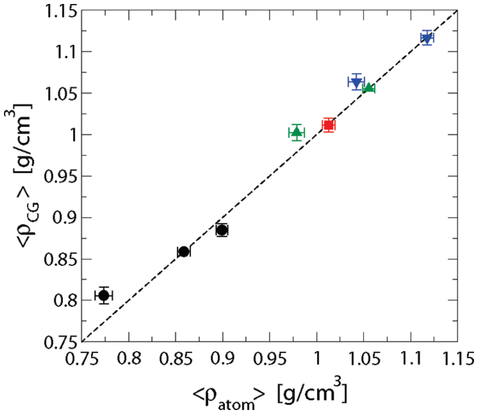


***Figure 13.*** Average total density $\langle\rho_{CG}\rangle$ from constant NPT CG simulations vs average total density $\langle\rho_{atom}\rangle$ from constant NPT atomistic simulations for the same mixture composition and temperature (3HT12:$C_{60}$ (w/w) = 1:0 (circles), 2.52:1 (squares), 1.85:1 (up triangles), and 1.27:1 (down triangles)).

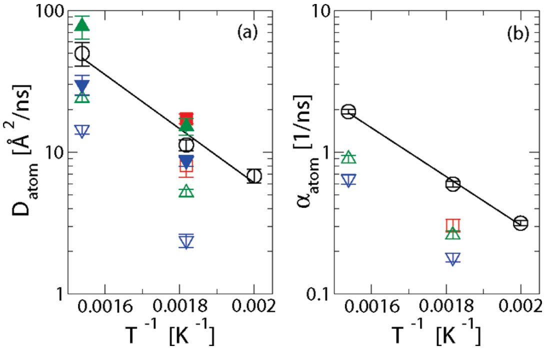


***Figure 14.*** (a) Translational diffusion coefficient $D_{atom}$ for 3HT12 monomers (empty symbols) and $C_{60}$ (filled symbols) and (b) rotational diffusion coefficient $\alpha_{atom}$ for 3HT12 chains from atomistic simulations as a function of $1/T$ for various 3HT12:$C_{60}$ weight ratios: 1:0 (circles), 2.52:1 (squares), 1.85:1 (triangles up), and 1.27:1 (triangles down). The lines are exponential (Arrhenius) fits to the circles.

structured $C_{60}-C_{60}$ RDF for the P3HT/$C_{60}$ mixture: the close agreement between the RDFs from the atomistic and CG simulations indicates that the CG models can reasonably describe the phase separation of these mixtures.

The results of all of our comparisons between the atomistic and CG simulations at different state points are summarized in Figure 13 in terms of the average total density $\langle\rho_{CG}\rangle$ measured in the CG simulations as a function of the average total density $\langle\rho_{atom}\rangle$ in the corresponding atomistic simulations. The line $\langle\rho_{CG}\rangle = \langle\rho_{atom}\rangle$ is also shown to indicate how closely the CG and atomistic densities agree with each other. It can be seen that the agreement is good, particularly at 500 and 550 K. The largest discrepancy between the densities of the CG and atomistic systems occurs for the pure P3HT system at 650 K, at which the CG system is 4 ± 2% more dense than the atomistic system. This difference is actually quite small for CG simulations models.[61,65,66] A 650 K temperature is also much higher than the temperatures that are typically used in processing polymer solar cells; the small discrepancy at this temperature shows that our CG models perform reasonably well even at thermodynamic conditions well outside those at which it is likely to be used.

## 6. Atomistic vs Coarse-Grained Time Scales

The CG interaction potentials that we have constructed for P3HT/$C_{60}$ mixtures have been optimized for the fluid structure (see section 3) and not for the dynamics. Therefore, it can be expected that the time scales for dynamics in the CG simulations will not be equivalent to those in the atomistic simulations.[39,67] No theory currently exists for predicting the time scales of a CG simulation relative to those of the atomistic simulation from which it was derived. Development of such a theory is beyond the scope of this work, but an estimate of the relative time scales of the atomistic and CG simulations is useful for comparing the CG simulation dynamics with experimental data.

We have estimated the relative time scales by comparing transport coefficients calculated from simulations of the atomistic and CG systems: translational time scales were determined by calculating the mean squared displacement (MSD) $\langle r^2(t)\rangle$ of the monomer center-of-mass in the P3HT chains (using only the central four monomers in each chain for the calculation) and of the $C_{60}$ center-of-mass; rotational time scales were obtained by calculating the orientational correlation function $\langle P_2[\hat{\mathbf{u}}(t)\cdot\hat{\mathbf{u}}(0)]\rangle \equiv \langle P_2[\cos\Theta(t)]\rangle$ of the unit vector $\hat{\mathbf{u}}$ connecting the polymer chain ends (defined by the centers-of-mass of the end thiophene rings). For diffusive translational motion,

$$\langle r^2(t)\rangle \sim 6Dt \tag{4}$$

and for diffusive rotational motion,

$$\langle P_2[\cos\Theta(t)]\rangle \sim \exp(-\alpha t) \tag{5}$$

where $D$ and $\alpha$ are the translational and rotational diffusion coefficients, respectively. For all of the simulations, we found, after an initial transient time, that $\langle r^2(t)\rangle$ and $\langle P_2[\cos\Theta(t)]\rangle$ were well fit by eqs 4 and 5, respectively, allowing well-defined values of $D$ and $\alpha$ to be extracted. The ratios $D_{CG}/D_{atom}$ and $\alpha_{CG}/\alpha_{atom}$, where the subscripts "atom" and "CG" denote quantities measured in the atomistic and CG simulations respectively, provide estimates of the relative time scales of the CG simulations compared with the atomistic simulations (the relative translational and rotational time scales defined by $D_{CG}/D_{atom}$ and $\alpha_{CG}/\alpha_{atom}$ are not necessarily the same but are expected to be comparable).

Before comparison of the atomistic and CG time scales, we first present in Figure 14 the translational and rotational diffusion coefficients measured in the atomistic simulations as functions of the temperature and the P3HT:$C_{60}$ mixture ratio. As expected, $D$ and $\alpha$ increase with $T$, with an approximately exponential dependence on $1/T$ that is indicative of activated diffusion. The activation energies for translational and rotational motion (given by the slope of the lines between the points of the same mixture composition) are comparable, although not equal. $D$ and $\alpha$ also decrease monotonically with increasing $C_{60}$ concentration. This is not surprising, given that the sublimation point of $C_{60}$[68] is substantially higher than the melting point of pure P3HT,[31] and so the addition of increasing amounts of $C_{60}$ appears to

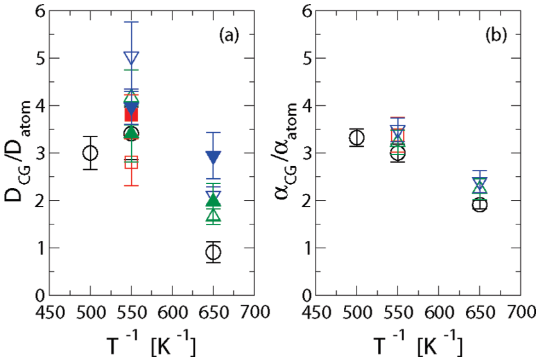


***Figure 15.*** (a) Ratio of translational diffusion coefficients, $D_{CG}/D_{atom}$, for 3HT12 monomers (empty symbols) and $C_{60}$ (filled symbols) and (b) ratio of rotational diffusion coefficients, $\alpha_{CG}/\alpha_{atom}$, from CG and atomistic simulations as a function of temperature for various 3HT12:$C_{60}$ weight ratios: 1:0 (circles), 2.52:1 (squares), 1.85:1 (triangles up), and 1.27:1 (triangles down).

move the systems closer to the freezing point of the mixture. We do not see evidence of a freezing point depression at intermediate fullerene concentrations that has been observed experimentally for P3HT:PCBM mixtures, which have a eutectic point for 65% P3HT (2:1 w/w P3HT:PCBM).[31] PCBM, however, mixes more readily with P3HT than does $C_{60}$ and is expected to have a lower melting/sublimation point due to the disordering effect of its side chain.

Turning to the relative time scales for the atomistic and CG simulations, $D_{CG}/D_{atom}$ and $\alpha_{CG}/\alpha_{atom}$ are plotted as a function of temperature in Figure 15 for the various P3HT:$C_{60}$ mixture ratios studied. It can be seen that, in almost all cases, the translational and rotational time scales are larger for the CG simulations than for the atomistic simulations The relative time scales are also fairly insensitive to mixture composition; the data suggests that they may increase with increasing $C_{60}$ content, but the large error bars on the points (particularly for the 3HT12 monomer translational diffusion coefficient) make it difficult to verify this hypothesis conclusively. The relative time scales also decrease with increasing temperature in almost all cases. The latter result is consistent with the expectation that the dynamics in the CG simulations occurs on a less rugged potential energy landscape with lower peaks and shallower valleys. The relative translational time scale for the pure P3HT system at 650 K does not follow the expected trend, possibly because of the higher average density in the CG simulation compared with the atomistic simulation (see section 5). The relative rotational time scale for pure P3HT at this temperature, however, displays the expected behavior: rotational motion of the polymer chains is not expected to be as sensitive to the density as translational motion, because in order to diffuse a polymer always needs to change place with its neighbors whereas reorientation can partially be achieved without displacement of other chains.

Taking into account the 2- to 5-fold difference between the CG and atomistic time scales and the 3-fold larger time step used in the CG simulations compared with the atomistic simulations, and the 10-fold speed-up in the simulation at each time-step for the systems studied, an overall speed-up of over 2 orders of magnitude is obtained in the CG simulations. This result underscores the huge advantage of the CG simulations over atomistic ones, particularly for large systems.

## 7. Device-Scale Molecular Simulations of Bulk Heterojunctions

Our CG model makes possible the study of the structure and dynamic evolution of the bulk heterojunction (BHJ) microstructure in systems approaching the device scale, because of the substantial speed-up over atomistic simulations discussed in section 6. As a proof-of-principle of the feasibility of studying device-scale bulk heterojunctions with our CG models, we have carried out a simulation of 768 P3HT 48-mers (MW ∼ 8 kDa) and 4608 $C_{60}$ molecules (1.85:1 w/w P3HT:$C_{60}$) in which the system, initially consisting of randomly placed polymer chains and fullerenes, was cooled over a period of 10 ns (CG time scale) from 550 to 490 K, after equilibrating at 550 K for 1 ns. The molecular weight of the polymer chains of 8 kDa is close to the ideal molecular weight for P3HT:fullerene solar cells of 13−34 kDa.[69] The average simulation box size length is roughly 25 nm, of the same order as the 50−100 nm[70,71] typically used for the thickness of the active layer in polymer solar cells. The simulation took approximately 24 h on 256 2.3-GHz AMD Opteron processors.

Figure 16 depicts snapshots of the system as a function of time as the system is cooled, in which P3HT and $C_{60}$ appear to begin to phase separate as time progresses. In this initial test of our CG models in a device-scale simulation, the cooling of the system has been carried out very rapidly (the total length of the simulation is shorter than the chain reorientational time scale $\tau_2$), and so the system is likely in a nonequilibrium state throughout the simulation. It is expected that phase separation would be even more evident if the cooling were carried out at a slower rate. Simulations at least 1 order of magnitude longer are feasible with our CG models for systems of this size and are in the process of being carried out.

It should be noted that the way in which the polymer/fullerene system was evolved in this simulation is not the same as the way in which polymer/fullerene solar cells are normally fabricated, in that the active layer in the latter case is deposited from a solvent. Rather, these simulations can describe the annealing step that is usually used in fabrication to improve device performance,[10,12] in which the solar cell is heated above its glass transition temperature to "improve" the BHJ morphology. CG simulation of polymer/fullerene mixtures in solution are, however, possible and represent a potential future step of this work.

## 8. Conclusions

In summary, we have developed coarse-grained (CG) computer simulation models of P3HT and P3HT/$C_{60}$ mixtures and verified that the models accurately describe the structure of these materials over a range of thermodynamic conditions other than those at which the CG models were parametrized. We have also demonstrated in a preliminary study of phase

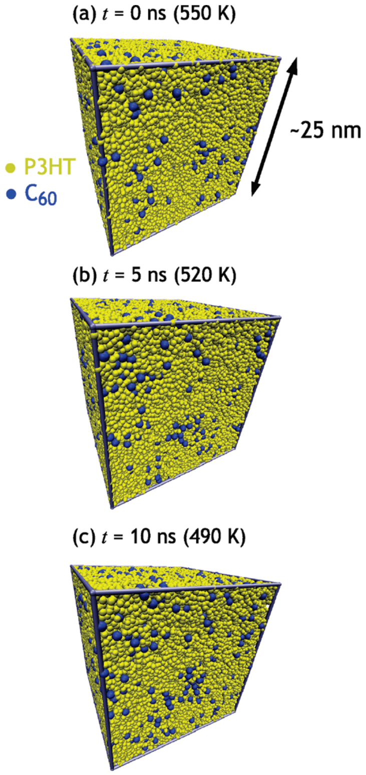


***Figure 16.*** Snapshots from a constant NPT CG simulation of 768 P3HT 48-mers (MW ~8 kDa) and 4608 $C_{60}$ molecules (1.85:1 w/w P3HT:$C_{60}$, 115 200 particles) in which the system is cooled from 550 to 490 K over a period of 10 ns (the system is periodically replicated in all three directions). The initial configuration consisted of randomly placed chains and fullerenes. P3HT and $C_{60}$ particles are in yellow and blue, respectively.

separation of a P3HT/$C_{60}$ mixture that the CG models can be used to study the structure and dynamic evolution of bulk heterojunctions at the molecular level for systems approaching the scale of organic photovoltaic devices.

In a subsequent publication, we will analyze in quantitative detail the structure and dynamic evolution of the BHJ microstructure (e.g., crystallinity, domain size, domain connectivity, chain persistence length, etc.) in these device-scale CG simulations as functions of the polymer:fullerene mole fraction and polymer chain length. We are also parametrizing a CG model of PCBM, the most widely used electron acceptor in polymer-based solar cells, having demonstrated the feasibility of CG simulations for $C_{60}$. In future work, we plan to study other polymer and fullerene types and to include solvent molecules in the CG simulations, thereby more closely mimicking the experimental processes by which BHJs are formed in organic solar cells.

**Acknowledgment.** A.J.M. acknowledges start-up funding from the University of California, Davis. This research also used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

**Supporting Information Available:** Atomistic and coarse-grained simulation model parameters for P3HT and $C_{60}$ and selected joint bond-length/bond-angle and bond-angle/dihedral-angle probability distributions from the atomistic and coarse-grained simulations. This information is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) U.S. Department of Energy, *Basic Research Needs for Solar Energy Utilization*, 2005.

(2) U.S. Department of Energy, *New Science for a Secure and Sustainable Energy Future*, 2008.

(3) Lewis, N. S. *MRS Bull.* **2007**, *32*, 808–820.

(4) Lewis, N. S. *Science* **2007**, *315*, 798–801.

(5) Brabec, C. J.; Hauch, J. A.; Schilinsky, P.; Waldauf, C. *MRS Bull.* **2005**, *30*, 50–52.

(6) Markov, D. E.; Amsterdam, E.; Blom, P. W. M.; Sieval, A. B.; Hummelen, J. C. *J. Phys. Chem. A* **2005**, *109*, 5266–5274.

(7) Mayer, A. C.; Scully, S. R.; Hardin, B. E.; Rowell, M. W.; McGehee, M. D. *Mater. Today* **2007**, *10*, 28–33.

(8) Coropceanu, V.; Cornil, J.; da Silva, D. A.; Olivier, Y.; Silbey, R.; Brédas, J. L. *Chem. Rev.* **2007**, *107*, 926–952.

(9) Schilinsky, P.; Asawapirom, U.; Scherf, U.; Biele, M.; Brabec, C. J. *Chem. Mater.* **2005**, *17*, 2175–2180.

(10) Padinger, F.; Rittberger, R. S.; Sariciftci, N. S. *Adv. Funct. Mater.* **2003**, *13*, 85–88.

(11) Li, G.; Shrotriya, V.; Huang, J.; Yao, Y.; Moriarty, T.; Emery, K.; Yang, Y. *Nat. Mater.* **2005**, *4*, 864–868.

(12) Ma, W.; Yang, C.; Gong, X.; Lee, K.; Heeger, A. J. *Adv. Funct. Mater.* **2005**, *15*, 1617–1622.

(13) Saunders, B. R.; Turner, M. L. *Adv. Colloid Interface Sci.* **2008**, *138*, 1–23.

(14) Chabinyc, M. L. *Polymer Rev.* **2008**, *48*, 463–492.

(15) Chen, T. A.; Wu, X. M.; Rieke, R. D. *J. Am. Chem. Soc.* **1995**, *117*, 233–244.

(16) Prosa, T. J.; Winokur, M. J.; Moulton, J.; Smith, P.; Heeger, A. J. *Macromolecules* **1992**, *25*, 4364–4372.

(17) Yang, C.; Orfino, F. P.; Holdcroft, S. *Macromolecules* **1996**, *29*, 6510–6517.

(18) Sirringhaus, H.; Brown, P. J.; Friend, R. H.; Nielsen, M. M.; Bechgaard, K.; Langeveld-Voss, B. M. W.; Spiering, A. J. H.; Janssen, R. A. J.; Meijer, E. W.; Herwig, P.; de Leeuw, D. M. *Nature* **1999**, *401*, 685–688.

(19) Aasmundtveit, K. E.; Samuelsen, E. J.; Guldstein, M.; Steinsland, C.; Flornes, O.; Fagermo, C.; Seeberg, T. M.; Pettersson, L. A. A.; Inganäs, O.; Feidenhans'l, R.; Ferrer, S. *Macromolecules* **2000**, *33*, 3120–3127.

(20) Hugger, S.; Thomann, R.; Heinzel, T.; Thurn-Albrecht, T. *Colloid Polym. Sci.* **2004**, *282*, 932–938.

(21) Kline, R. J.; McGehee, M. D.; Kadnikova, E. N.; Liu, J. S.; Fréchet, J. M. J.; Toney, M. F. *Macromolecules* **2005**, *38*, 3312–3319.

(22) Erb, T.; Zhokhavets, U.; Gobsch, G.; Raleva, S.; Stuhn, B.; Schilinsky, P.; Waldauf, C.; Brabec, C. J. *Adv. Funct. Mater.* **2005**, *15*, 1193–1196.

(23) Kim, Y.; Cook, S.; Tuladhar, S. M.; Choulis, S. A.; Nelson, J.; Durrant, J. R.; Bradley, D. D. C.; Giles, M.; McCulloch, I.; Ha, C. S.; Ree, M. *Nat. Mater.* **2006**, *5*, 197–203.

(24) Chang, J. F.; Clark, J.; Zhao, N.; Sirringhaus, H.; Breiby, D. W.; Andreasen, J. W.; Nielsen, M. M.; Giles, M.; Heeney, M.; McCulloch, I. *Phys. Rev. B* **2006**, *74*, 115318.

(25) Zen, A.; Saphiannikova, M.; Neher, D.; Grenzer, J.; Grigorian, S.; Pietsch, U.; Asawapirom, U.; Janietz, S.; Scherf, U.; Lieberwirth, I.; Wegner, G. *Macromolecules* **2006**, *39*, 2162–2171.

(26) Kline, R. J.; DeLongchamp, D. M.; Fischer, D. A.; Lin, E. K.; Richter, L. J.; Chabinyc, M. L.; Toney, M. F.; Heeney, M.; McCulloch, I. *Macromolecules* **2007**, *40*, 7960–7965.

(27) Zhang, R.; Li, B.; Iovu, M. C.; Jeffries-EL, M.; Sauve, G.; Cooper, J.; Jia, S. J.; Tristram-Nagle, S.; Smilgies, D. M.; Lambeth, D. N.; McCullough, R. D.; Kowalewski, T. *J. Am. Chem. Soc.* **2006**, *128*, 3480–3481.

(28) Andersson, B. V.; Herland, A.; Masich, S.; Inganäs, O. *Nano Lett.* **2009**, *9*, 853–855.

(29) van Bavel, S. S.; Sourty, E.; de With, G.; Loos, J. *Nano Lett.* **2009**, *9*, 507–513.

(30) Gurau, M. C.; Delongchamp, D. M.; Vogel, B. M.; Lin, E. K.; Fischer, D. A.; Sambasivan, S.; Richter, L. J. *Langmuir* **2007**, *23*, 834–842.

(31) Müller, C.; Ferenczi, T. A. M.; Campoy-Quiles, M.; Frost, J. M.; Bradley, D. D. C.; Smith, P.; Stingelin-Stutzmann, N.; Nelson, J. *Adv. Mater.* **2008**, *20*, 3510–3515.

(32) Kim, J. Y.; Frisbie, C. D. *J. Phys. Chem. C* **2008**, *112*, 17726–17736.

(33) Cheung, D. L.; Troisi, A. *Phys. Chem. Chem. Phys.* **2008**, *10*, 5941–5952.

(34) Corish, J.; Feeley, D. E.; Morton-Blake, D. A.; Béniere, F.; Marchetti, M. *J. Phys. Chem. B* **1997**, *101*, 10075–10085.

(35) Curco, D.; Aleman, C. *J. Comput. Chem.* **2007**, *28*, 1743–1749.

(36) Kirkpatrick, J.; Marcon, V.; Nelson, J.; Kremer, K.; Andrienko, D. *Phys. Rev. Lett.* **2007**, *98*, 227402.

(37) Troisi, A.; Cheung, D. L.; Andrienko, D. *Phys. Rev. Lett.* **2009**, *102*, 116602.

(38) Coropceanu, V.; Sanchez-Carrera, R. S.; Paramonov, P.; Day, G. M.; Brédas, J. L. *J. Phys. Chem. C* **2009**, *113*, 4679–4686.

(39) Tschöp, W.; Kremer, K.; Batoulis, J.; Bürger, T.; Hahn, O. *Acta Polym.* **1998**, *49*, 61–74.

(40) Baschnagel, J.; Binder, K.; Doruker, P.; Gusev, A. A.; Hahn, O.; Kremer, K.; Mattice, W. L.; Müller-Plathe, F.; Murat, M.; Paul, W.; Santos, S.; Suter, U. W.; Tries, V. *Adv. Polym. Sci.* **2000**, *152*, 41–156.

(41) Reith, D.; Pütz, M.; Müller-Plathe, F. *J. Comput. Chem.* **2003**, *24*, 1624–1636.

(42) Faller, R.; Reith, D. *Macromolecules* **2003**, *36*, 5406–5414.

(43) *Coarse-Graining of Condensed Phase and Biomolecular Systems*; Voth, G. A., Ed.; CRC Press: Boca Raton, FL, 2008.

(44) Marcon, V.; Raos, G. *J. Am. Chem. Soc.* **2006**, *128*, 1408–1409.

(45) Marcon, V.; Raos, G. *J. Phys. Chem. B* **2004**, *108*, 18053–18064.

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

(47) Plimpton, S. J. *J. Comput. Phys.* **1995**, *117*, 1–19; LAMMPS Molecular Dynamics Simulator: http://lammps.sandia.gov.

(48) Hockney, R. W.; Eastwood, J. W. *Computer Simulation Using Particles*; Institute of Physics Publishing, Bristol, 1988.

(49) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.

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

(51) Hoover, W. G. *Phys. Rev. A* **1986**, *34*, 2499–2500.

(52) Raos, G.; Famulari, A.; Marcon, V. *Chem. Phys. Lett.* **2003**, *379*, 364–372.

(53) Darling, S. B.; Sternberg, M. *J. Phys. Chem. B* **2009**, *113*, 6215–6218.

(54) Barbarella, G.; Zambianchi, M.; Bongini, A.; Antolini, L. *Adv. Mater.* **1994**, *6*, 561–564.

(55) Price, M. L. P.; Ostrovsky, D.; Jorgensen, W. L. *J. Comput. Chem.* **2001**, *22*, 1340–1352.

(56) Sigma-Aldrich Catalog: http://www.sigmaaldrich.com.

(57) Mardalen, J.; Samuelsen, E. J.; Gautun, O. R.; Carlsen, P. H. *Solid State Commun.* **1991**, *77*, 337–339.

(58) Girifalco, L. A. *J. Phys. Chem.* **1992**, *96*, 858–861.

(59) Hedberg, K.; Hedberg, L.; Bethune, D. S.; Brown, C. A.; Dorn, H. C.; Johnson, R. D.; De Vries, M. *Science* **1991**, *254*, 410–412.

(60) Reith, D.; Müller, B.; Müller-Plathe, F.; Wiegand, S. *J. Chem. Phys.* **2002**, *116*, 9100–9106.

(61) Ghosh, J.; Sun, Q.; Faller, R. Point Dependence and Transferability of Potentials in Systematic Structural Coarse-Graining. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*; Voth, G. A., Ed.; CRC Press: Boca Raton, FL, 2008; pp 69−82.

(62) Hernandez, V.; Casado, J.; Ramirez, F. J.; Zotti, G.; Hotta, S.; Navarrete, J. T. L. *Synth. Met.* **1996**, *76*, 277–280.

(63) Heffner, G. W.; Pearson, D. S. *Macromolecules* **1991**, *24*, 6295–6299.

(64) Mardalen, J.; Samuelsen, E. J.; Gautun, O. R.; Carlsen, P. H. *Solid State Commun.* **1991**, *80*, 687–689.

(65) Ghosh, J.; Faller, R. *Mol. Simul.* **2007**, *33*, 759–767.

(66) Carbone, P.; Varzaneh, H. A. K.; Chen, X.; Müller-Plathe, F. *J. Chem. Phys.* **2008**, *128*, 064904.

(67) Faller, R. *Polymer* **2004**, *45*, 3869–3876.

(68) Piacente, V.; Gigli, G.; Scardala, P.; Giustini, A.; Ferro, D. *J. Phys. Chem.* **1995**, *99*, 14052–14057.

(69) Ballantyne, A. M.; Chen, L.; Dane, J.; Hammant, T.; Braun, F. M.; Heeney, M.; Duffy, W.; McCulloch, I.; Bradley, D. D. C.; Nelson, J. *Adv. Funct. Mater.* **2008**, *18*, 2373–2380.

(70) Li, G.; Shrotriya, V.; Yao, Y.; Yang, Y. *J. Appl. Phys.* **2005**, *98*, 043704.

(71) Moulé, A. J.; Bonekamp, J. B.; Meerholz, K. *J. Appl. Phys.* **2006**, *100*, 094503.

CT900496T

JCTC Journal of Chemical Theory and Computation

# Interaction Analysis of the Native Structure of Prion Protein with Quantum Chemical Calculations

Takeshi Ishikawa[†,‡] and Kazuo Kuwata*[,†,‡]

*Division of Prion Research, Center for Emerging Infectious Disease, Gifu University, 1-1 Yanagido, Gifu 501-1194, Japan, and CREST Project, Japan Science and Technology Agency, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan*

Received August 28, 2009

**Abstract:** We examined the solvent interaction and intramolecular interaction of the native structure of prion protein (PrP) using quantum chemical calculations based on the fragment molecular orbital (FMO) method. The influence due to the geometrical fluctuation was taken into account by performing calculations on forty different conformations. Each FMO calculation was carried out at the MP2 level of theory with the cc-pVDZ in which the resolution of the identity approximation was employed to reduce the computational cost. The solvent interaction energies obtained from the calculations provided information about the hydrophilicity of the three α-helices. We examined the roles of the charged residues in retaining the native structure of PrP with the calculated intramolecular interaction energies. The analysis, focused on van der Waals interaction, showed that the hydrophobic residues were important for the stability of the native structure. Our results were also discussed in relation to the identified pathogenetic mutations of prion diseases. Additionally, we examined the distribution of the calculated values with 40 structures, in which we demonstrated the influence of geometrical fluctuations on quantum chemical calculations.

## 1. Introduction

The abnormal scrapie form of prion protein (PrP), which is a conformational isoform of the cellular form, causes transmissible spongiform encephalopathies,[1−3] for example, scrapie, bovine spongiform encephalopathy, and Creutzfeldt−Jakob disease. In these prion diseases, the conformational conversion from the cellular form to the scrapie form is a key event. Researchers have elucidated the structure of the cellular form at atomic resolution with experimental measurements.[4−8] However, despite many studies utilizing both experimental and theoretical approaches, the mechanism of the pathogenetic conversion as well as the structure of the scrapie form remain unclear. Here, we consider that detailed information about the interaction responsible for retaining the high-order structure of the cellular form will be helpful for the examination of the mechanisms of prion diseases.

With the growth of computer technology, a number of theoretical studies of large molecules including biomolecular systems with quantum chemical calculations have been reported. For example, the molecular-orbital derived polarization (MP) model,[9] in which an effective Hamiltonian of a whole system was introduced to reduce a computational effort, was reported by Gao. This model was developed for Monte Carlo simulations or molecular dynamics simulations of liquid−water systems with semiempirical quantum chemical methods. He performed statistical mechanical Monte Carlo simulations of a cubic box containing 267 water molecules, and averaging over millions of configurations was carried out.[10] If we focus our attention on ab initio quantum chemical methods, the fragment molecular orbital (FMO) method[11−13] is one of the most efficient approaches for calculations of large molecules. Some methodological concepts of the FMO scheme are analogous to those of the MP

* Corresponding author e-mail: kuwata@gifu-u.ac.jp.

† Gifu University.

‡ Japan Science and Technology Agency.

10.1021/ct900456v © 2010 American Chemical Society
Published on Web 12/30/2009

model, but detailed descriptions about the difference between two methods are beyond the scope of this Article.

The FMO method is known to be a powerful tool for analyzing the interaction of biomolecular systems because inter fragment interaction energy (IFIE) or pair interaction energy (PIE) is clearly defined.[14] The intermolecular interactions of many proteins with small compounds or DNA bases were examined using the IFIE, providing some useful information for fundamental research and drug discovery.[15−24] However, the intramolecular interaction of proteins (i.e., the interactions between two residues) has not been extensively examined with the FMO method despite the fact that such interactions are significant for understanding the higher-order structures of proteins. As one of a few examples, Kurisaki et al. developed a visualization method for the IFIEs, including the intramolecular interaction of a protein,[25] wherein the secondary structures of proteins were discussed with the matrix representation of the IFIEs.

The interaction of amino acid residues with the surrounding solvent molecules is also significant for retaining the native structures of proteins. In the previous studies with the FMO scheme, the solvent effect could be included using the two solvent models: the polarizable continuum model (PCM)[26] and the explicit solvent model.[27,28] We believe that the explicit solvent model is better than the PCM for the examination of the solvent interaction of the amino acid residues because not only electrostatic interactions but also charge transfer interactions can be included.

In most previous studies using the FMO method, researchers examined biomolecular systems using the results of a single structure (or a very few structures). However, a number of structures should be considered because proteins and solvent molecules have a significant geometrical fluctuation at body temperature. Very recently, Ishikawa et al. reported the FMO calculations using 20 different structures from the MD trajectory,[29] wherein they examined the influence of the geometrical fluctuation on the interaction energy between a protein and a small molecule. Their results indicated that the molecular interactions in biomolecular systems should be discussed using the averaged results of multiple structures.

In this work, we calculated the solvent interactions and intramolecular interactions of PrP with the FMO method. In these calculations, the solvent effect was included with the explicit solvent model, and the influence of the geometrical fluctuation was considered by performing multiple calculations with the different structures. Using our results, we will discuss the interactions retaining the native structure of PrP (i.e., the cellular form of PrP). In the following sections, we provide a brief description of the FMO method, and after presenting the computational details, we discuss the results of our calculations.

## 2. Method

**2.1. Brief Description of the FMO Method.** In the FMO method, a target molecule is divided into small fragments[11] by cutting C−C single bonds with projection operators.[12] The total energy is evaluated using the results of individual calculations of the fragments (referred to as monomers) and pairs of the fragments (referred to as dimers) with the following equation:[11]

$$E_{\mathrm{total}} = \sum_{I<J} E_{IJ} - (N_{\mathrm{f}} - 2)\sum_{I} E_I \qquad (1)$$

where $E_I$ and $E_{IJ}$ are energies obtained from the monomer and the dimer calculations, respectively, and $N_{\mathrm{f}}$ is the number of fragments. In such calculations, the electrostatic potential from the other fragments, which is generally referred to as the environmental electrostatic potential (ESP), is included.[11,14] At the HF level of theory, the total energy can be rewritten as:

$$E_{\mathrm{total}}^{\mathrm{HF}} = \sum_{I} E'^{\mathrm{HF}}_{I} + \sum_{I>J} \Delta E_{IJ}^{\mathrm{HF}} \qquad (2)$$

where $E'^{\mathrm{HF}}_{I}$ is the monomer energy without the ESP. Thus, one can consider that $\Delta E_{IJ}^{\mathrm{HF}}$ is the interaction energy between two fragments.[14] This value is the IFIE or PIE, the formulation of which can be found in a previous paper.[14]

As is generally known, electrostatic interactions and charge transfer interactions are included in HF calculations, but van der Waals interactions or dispersion interactions are not. Therefore, the MP2 calculation should additionally be performed to evaluate the van der Waals interactions. In such cases, the total energy is corrected using the MP2 results according to the following equations:

$$E_{\mathrm{total}}^{\mathrm{MP2}} = E_{\mathrm{total}}^{\mathrm{HF}} + \Big(\sum_{I} E_I^{\mathrm{corr}} + \sum_{I>J} \Delta E_{IJ}^{\mathrm{corr}}\Big) \qquad (3)$$

$$\Delta E_{IJ}^{\mathrm{corr}} = E_{IJ}^{\mathrm{corr}} - E_I^{\mathrm{corr}} - E_J^{\mathrm{corr}} \qquad (4)$$

where $E_I^{\mathrm{corr}}$, $E_{IJ}^{\mathrm{corr}}$, and $\Delta E_{IJ}^{\mathrm{corr}}$ are van der Waals contributions to the monomer energy, the dimer energy, and the IFIE, respectively.[30−32] Thus, the IFIE corrected with the MP2 method is

$$\Delta E_{IJ}^{\mathrm{MP2}} = \Delta E_{IJ}^{\mathrm{HF}} + \Delta E_{IJ}^{\mathrm{corr}} \qquad (5)$$

This equation clearly shows that one can obtain the interaction energies divided into two contributions, that is, $\Delta E_{IJ}^{\mathrm{HF}}$ (electrostatic interactions and charge transfer interactions) and $\Delta E_{IJ}^{\mathrm{corr}}$ (van der Waals interactions).

**2.2. Interaction Analysis.** In FMO calculations of typical biomolecular systems, amino acid residues and solvent molecules are basically treated as a single fragment. Thus, the interaction energy between a specific residue (assigned to fragment $I$) and all solvent molecules can be calculated as:

$$\Delta E_I^{\mathrm{solvent}} = \sum_{J\in \mathrm{solvent}} \Delta E_{IJ}^{\mathrm{MP2}} \qquad (6)$$

where the summation runs over the fragments assigned to the solvent molecules. The total interaction energy of the protein with the solvent molecules is obtained as:

$$\Delta E_{\mathrm{total}}^{\mathrm{solvent}} = \sum_{I\in \mathrm{protein}} \Delta E_I^{\mathrm{solvent}} \qquad (7)$$

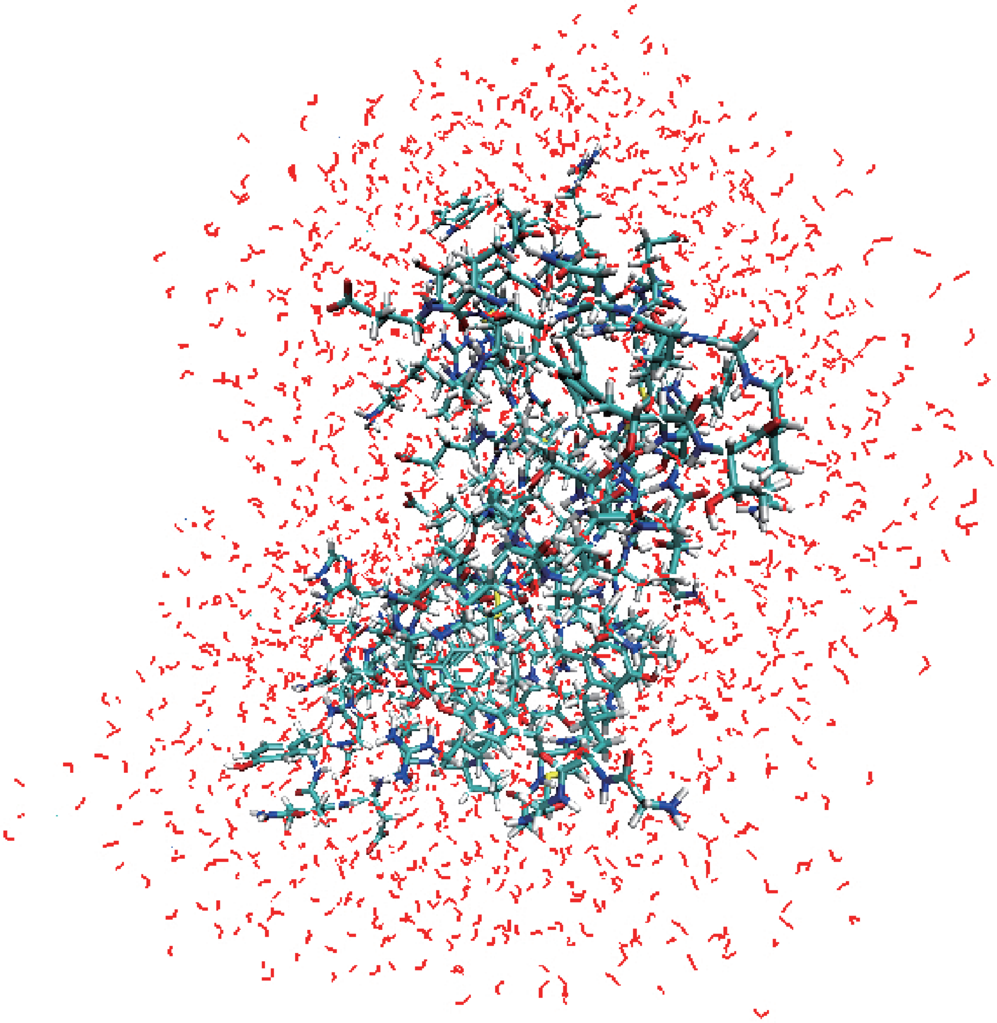


***Figure 1.*** Graphical representation of our system. In this work, the FMO calculations were performed with the 40 different structures (see text).

where the summation runs over the fragments assigned to the protein. In a similar way, the interaction energies between two residues are obtained from $\Delta E_{IJ}^{\mathrm{MP2}}$. However, the interaction energies between two fragments connected to each other are not calculated because of the theoretical requirements of the FMO method. Thus, we cannot obtain the interaction energies between two neighboring residues.

Because the $\Delta E_{IJ}^{\mathrm{MP2}}$ is calculated with eq 5, the above interaction energies can be divided into the two contributions ($\Delta E_{IJ}^{\mathrm{HF}}$ and $\Delta E_{IJ}^{\mathrm{corr}}$). If these values are calculated in the native structure, we can obtain detailed information about the interactions that retain this structure.

## 3. Computational Details

We believed that a single calculation with a specific structure, for example, the geometry optimized structure, is not sufficient for the examination of biomolecular systems that have a geometrical fluctuation at a physiological temperature. We expect that the effect of this fluctuation can be partially introduced into our analysis by taking the average over the results with a number of geometrical structures. Thus, 40 calculations with different structures were performed to obtain the averaged results in this study. These atomic coordinates were prepared according to the following method.

(1) We downloaded an initial structure of the globular domain of PrP containing the residues 124−226 from the Protein Data Bank[33] (PDB code: 1AG2[4]). We then generated the missing hydrogen atoms and the solvent molecules (water molecules, sodium ions, and chloride ions) around the PrP.

(2) After an energy minimization of this system, we performed a constant temperature and pressure (300 K and 1 atm) ensemble simulation for 2120 ps under the truncated octahedron boundary condition with FF03[34] and TIP3P[35] (AMBER 10 package[36]).

(3) Forty structures were randomly selected from the trajectory of the last 1000 ps.

(4) For each structure, we excluded all solvent molecules more than 8.0 Å from the PrP, resulting in approximately 1800 solvent molecules within our systems.

We affirmed the validity of the cutoff distance of the solvent molecules (8.0 Å) as shown in the following section. Figure 1 shows an example structure of our system. Because we extracted the atomic coordinates around the native conformation, our analysis yielded information about the interactions responsible for the stability of the native structure of the PrP.

In our FMO calculations, each amino acid residue was treated as a single fragment, except for C179 and C214, which were united into one fragment because of their S−S bond. The solvent molecules were essentially assigned as a single fragment, but ions and their hydration water molecules were collected into one fragment, in which the water molecules within 2.5 Å of the ions were treated as hydration waters.

We performed FMO calculations employing cc-pVDZ[37] at the MP2 level of theory together with the HF level of theory, wherein the total number of basis sets was about 60 000. In the case of the MP2 calculations, we utilized the resolution of the identity (RI) approximation[38] to reduce the computational efforts with auxiliary basis sets.[39] The RI-MP2 method was very recently introduced into the FMO scheme,[40] affording an advantageous increase in computational efficiency. As a result, timing of one FMO calculation of our system was about 70 h with the eight cores (Xeon E5420) and 2.0 GB memory per core. All calculations with the FMO scheme were performed using the PAICS program[29] developed in our laboratory.

As mentioned above, the statistical mechanical Monte Carlo simulations of liquid water systems were carried out with semiempirical quantum chemical methods.[10] However, to the best of our knowledge, this report is the first in which ab initio quantum chemical calculations of biomolecular systems including the explicit solvent molecules were carried out with 40 different structures.

## 4. Results and Discussion

**4.1. MD Simulation.** Prior to the discussions about the FMO calculations, we show the results of the MD simulation by which 40 structures used in our calculations were determined. This simulation was performed for 2120 ps with an energy minimized structure as an initial atomic coordinates. In this simulation, the atoms of PrP were restrained only for the first 20 ps, and, after that, this restraint was removed. The total energies of the system and the root-mean-square deviations (rmsd's) of the main-chains of PrP from the PDB structure are shown in Figure 2. Here, we selected the 40 structures from the trajectory of the last 1000 ps. In this range of the trajectory, the rmsd's were approximately 2.5−3.5 Å, indicating that a single structure calculation using the PDB structure is not sufficient for examinations of the native conformation of PrP.

**4.2. Thickness of Solvent Molecules.** As mentioned above, the explicit solvent molecules within 8.0 Å of the protein were included in our calculations to directly evaluate the interaction energies between residues and solvent molecules. Before starting the FMO calculations, we confirmed that the thickness of the solvent molecules was reasonable. The total solvent interaction energies ($\Delta E_{\mathrm{total}}^{\mathrm{solvent}}$) were calculated employing the various thicknesses. Figure 3 lists the results of calculations. In the case with a solvent molecule thickness of 12.0 Å, the interaction energy was −5204.9 kcal/mol. On the other hand, the interaction energy was −5091.4 kcal/mol for a 8.0 Å thickness, which was 97.8% of that with a 12.0 Å thickness. Judging from these calculations, we can safely say that our cutoff distance was reasonable for evaluating the interaction energies between residues and solvent molecules.

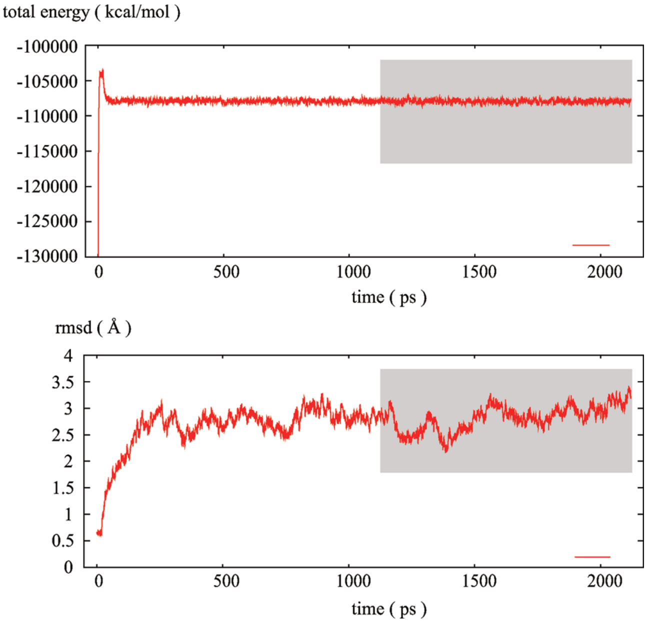


***Figure 2.*** The total energy of the system and the rmsd of main-chains of PrP from the PDB structure. The shadow squares present the range from 1120 to 2120 ps, from which the 40 structures used for the FMO calculations were picked up.

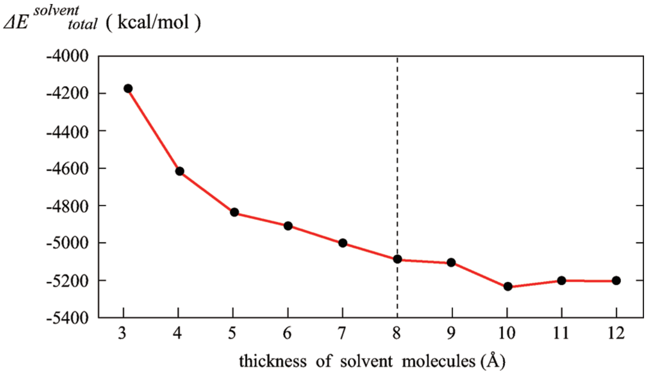


***Figure 3.*** The total interaction energies between PrP and solvent molecules ($\Delta E_{\mathrm{total}}^{\mathrm{solvent}}$) with different thicknesses of the explicit solvent molecules from 3.0 to 12.0 Å.

**4.3. Solvent Interactions of Residues.** In this subsection, we will discuss the interaction between amino acid residues and solvent molecules. Figure 4 shows the calculated interaction energies of each residue ($\Delta E_{I}^{\mathrm{solvent}}$), in which we added the result of C214 to that of C179 because these residues were treated as a single fragment. All of the interaction energies were obtained by taking the average over the results of the 40 structures.

First, we should note that several charged residues largely interacted with the solvent molecules, that is, E152, D167, K185, K194, E200, E211, K220, and E221. Two terminal residues (G124 and Y226) also largely interacted with the

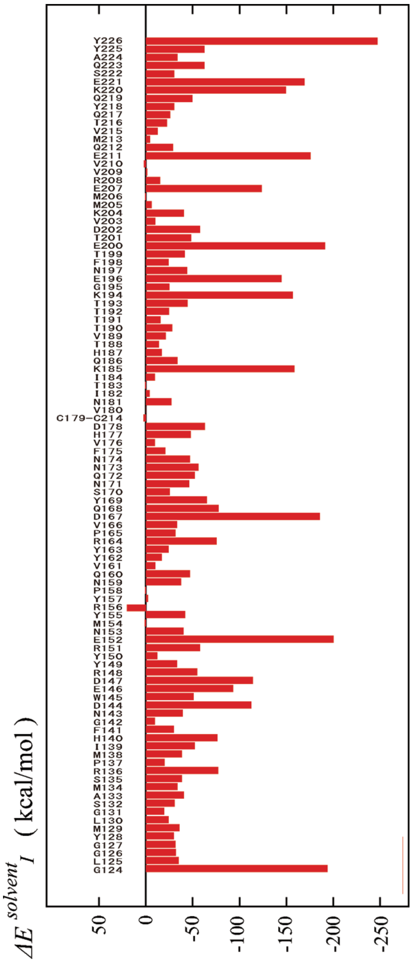


***Figure 4.*** The interaction energies of residues with the solvent molecules ($\Delta E_I^{\text{solvent}}$). Because C179 and C214 were treated as a single fragment, the interaction energy of C214 was collected into C179. All of the interaction energies were obtained by taking the average from the 40 selected structures (see text).

solvent due to the setting of our calculation; that is, the main-chains of the two terminal residues were set to $-\text{COO}^-$ or $-\text{NH}_3^+$. These results are consistent with the common picture of proteins: hydrophilic residues tend to locate on the surface of a protein and interact with the surrounding solvent molecules, ensuring that these residues play important roles in retaining the native structure of the protein. However, our results indicated the existence of charged residues with comparatively small solvent interaction energies; particularly, R156 had an unfavorable interaction energy. In the next subsection, we will discuss the difference between two types of charged residues, that is, those having large solvent interaction energies and those having small solvent interaction energies.

The helical structures of PrP are known to decrease in the pathogenetic conversion of prion diseases,[41] but the broken parts of the helices have not been specified. Thus, the nature of each α-helix is important for studying the mechanism of prion diseases. The solvent interaction energies of the three α-helices can be obtained by restricting the summation of eq 7 to the fragments belonging to each helix according to the following equation:

$$\Delta E_{\text{helix}}^{\text{solvent}} = \sum_{I \in \text{helix}} \Delta E_I^{\text{solvent}} \tag{8}$$

where $\Delta E_I^{\text{solvent}}$ is defined in eq 6. Table 1 summarizes the calculated results. The solvent interaction energy per residue averaged within HA (D144−N153) was −76.9 kcal/mol. This result indicated that HA was highly hydrophilic. On the other hand, the averaged solvent interaction energy of HB (Q172−K194) was −38.7 kcal/mol, indicating low hydrophilicity. In the case of HC (E200−A224), the averaged solvent interaction energy was −53.6 kcal/mol. The high hydrophilicity of HA was previously pointed out by Morrissey and Shakhnovich,[42] who used the two empirical scaling criteria[43,44] for estimating the hydrophilicity.

From a methodological point of view, our scheme for evaluating the hydrophilicity of the secondary structures has notable points. For example, we evaluated the hydrophilicity from the direct calculations of the interaction energies with the solvent molecules in accordance with the quantum chemical calculations, in which not only electrostatic interactions but also charge transfer interactions can be included. Additionally, because the solvent interaction energies were individually calculated for each residue under the conditions of the protein, the hydrophilicity was evaluated reflecting the side-chain exposure to the solvent.

**4.4. Ionic Interactions of Residues.** In this subsection, we will discuss the intramolecular interactions of PrP. As mentioned above, we calculated the interaction energies between two non-neighboring residues. Figure 5 shows the results of the pairs having the 25 lowest interaction energies. At first glance, one should note that seven pairs have significantly larger interaction energies. Table 2 provides detailed information about these pairs. The seven pairs possessed ionic interactions; that is, two residues had electrically opposite charges and were separated by sufficiently small distance. Such ionic interactions are generally referred to as salt bridges. Here, we may consider that these seven salt bridges were one of the important sources causing the stability of the native conformation of PrP. These salt bridges might be inferred only with the structural information. However, we think that salt bridges can be identified more definitely by additional use of the energetic information.

Two salt bridges, D144−R148 and D147−R151, were located in HA with interaction energies of −119.6 and −98.7 kcal/mol, respectively. Thus, the helical structure of HA was strongly retained. Another salt bridge, K204−E207, was located in HC, and its interaction energy was −127.3 kcal/mol. This salt bridge contributed to the conservation of the helical structure of HC. On the other hand, there was no salt bridge in HB, indicating that the helical structure of HB was weaker than those of the other two helices. The remaining four salt bridges, R164−D178, R156−D202, E146−R208, and R156 −E196, were constructed with the two residues whose sequence numbers were separated from each other. This fact indicated that these four salt bridges were helpful in retaining the tertiary structure of the PrP. The ionic interaction is a local interaction between two specific residues, similar to that of a disulfide bond. Thus, we can consider that the native structure of the PrP is

***Table 1.*** Solvent Interaction Energies of the Helices ($\Delta E^{\mathrm{solvent}}_{\mathrm{helix}}$) and the Whole Protein ($\Delta E^{\mathrm{solvent}}_{\mathrm{total}}$)[a]

| $\Delta E^{\mathrm{solvent}}_{\mathrm{helix}}$ | | | |
|---|---|---|---|
| HA | HB | HC | $\Delta E^{\mathrm{solvent}}_{\mathrm{total}}$ |
| −769.0<br>(−76.9) | −850.8<br>(−38.7) | −1286.6<br>(−53.6) | −5077.2<br>(−49.3) |

[a] The averaged interaction energies per residue are also shown in parentheses. These values are in kcal/mol. The interaction energies were obtained by taking the average from the 40 selected structures (see text).

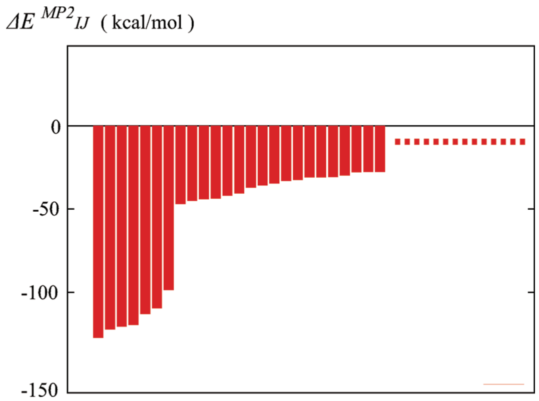


***Figure 5.*** The 25 largest interaction energies of all of the pairs. Seven pairs had extremely large interaction energies as compared to the other pairs. These values were averaged results over the selected 40 calculations (see text).

***Table 2.*** Interaction Energies of the Seven Pairs of Residues (kcal/mol)[a]

| types of residues | energy | distance |
|---|---|---|
| K204−E207 | −127.3 | 1.82 |
| R164−D178* | −122.4 | 1.99 |
| R156−D202* | −120.6 | 1.87 |
| D144−R148* | −119.6 | 2.05 |
| E146−R208* | −113.0 | 1.81 |
| R156−E196* | −109.7 | 1.87 |
| D147−R151 | −98.7 | 2.38 |

[a] In this table, the sequence number of residues and the distance (Å) are also shown. These interaction energies were obtained by taking the average from the 40 selected structures (see text). The asterisk indicates that the pathogenetic point mutation involved with the pair has been identified.[45]

stabilized by several local interactions, that is, the seven salt bridges and one disulfide bond between C179 and C214.

Currently, researchers have identified a number of pathogenetic mutations of prion diseases.[45] Our results associated with the salt bridges were consistent with some of them. From the list of such pathogenetic mutations,[45] five point mutations are involved with the salt bridges, that is, R148H, D178N, E196K, D202N, and R208H. The ionic interaction of the salt bridge is eliminated by these mutations, and the native structure of PrP becomes destabilized. As a result, the pathogenetic conversion of prion diseases progresses. This consistency of our results with the pathogenetic mutations shows the potential of quantum chemical calculations in the ongoing studies of prion diseases.

Next, we discuss the ionic interactions in cooperation with the solvent interactions of the residues. In the range of the sequence numbers from 124 to 226, there are 22 charged amino acid residues. According to Table 2, 13 charged residues were related to the salt bridges, and, consequently, nine charged residues were not. Table 3 shows the solvent interaction energy per residue averaged within the 13 charged residues as well as that within the other nine charged residues. As shown in this table, the solvent interaction energies of the charged residues forming the salt bridges were significantly smaller than those of the other charged residues. Particularly, the solvent interaction energy of R156, which was related to the two salt bridges, was unfavorable. This result can be interpreted as follows. The location of the charged residue not related to the salt bridge can be adjusted to the interaction with the solvent molecules; that is, the side-chain is exposed to the solvent. On the other hand, the charged residue forming the salt bridge is located at a suitable position for an ionic interaction with the other charged residue. As a result, the location of their side-chains cannot be adjusted to interact with the solvent molecules. Here, we can state that the 13 charged residues play an important role for the stability of the protein by forming the salt bridges; on the other hand, the nine charged residues also play an important role by interacting with the solvent molecules. As shown in this subsection, our analysis provided information about the roles of the residues contributing to the stability of the native conformation of PrP.

***Table 3.*** Averaged Solvent Interaction Energy of Charged Residues Related to the Salt Bridges and Not Related to the Salt Bridges (kcal/mol)[a]

| charged residues related to salt bridges | charged residues not related to salt bridges |
|---|---|
| −71.3 | −162.1 |

[a] These interaction energies were obtained by taking the average from the 40 selected structures (see text).

**4.5. van der Waals Interaction of Residues.** In this subsection, we discuss the van der Waals interactions, which is the main source of the interactions between the hydrophobic residues. Here, to discuss the interactions retaining the high-order structure of PrP, we focus our attention on the interaction between two residues with sequence numbers separated from each other. Thus, we calculated the following values:

$$\Delta E^{\mathrm{vdw}}_{I} = \sum_{J\in \mathrm{protein},|I-J|>5} \Delta E^{\mathrm{corr}}_{IJ} \qquad (9)$$

where $\Delta E^{\mathrm{corr}}_{IJ}$ is van der Waals contribution to the interaction energy between the two fragments (see eq 5). The summation of $J$ runs over the fragments whose index are different from $I$ by more than five; that is to say, we accumulated the interaction energies between the residues separated from each other by more than five sequence numbers.

Figure 6 illustrates the calculated results. Phenylalanines and tyrosines, which have a benzene ring, showed large interaction energies associated with the van der Waals interaction. Additionally, several hydrophobic residues, leucine, methionine, valine, and isoleucine, had relatively large interaction energies, thus indicating that these residues contribute to the conservation of the native structure of PrP. These results are consistent with the common picture of

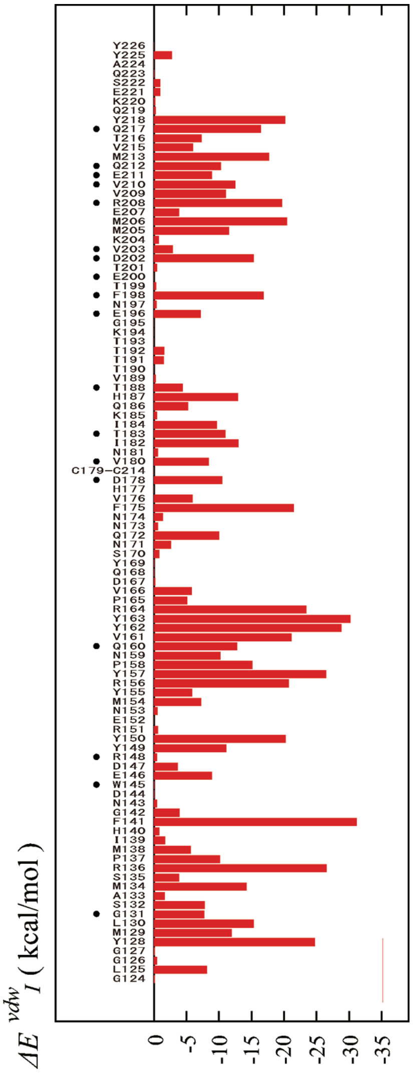


***Figure 6.*** van der Waals interaction energies of each residue with the other residues separated by more than six sequence numbers ($\Delta E_I^{\mathrm{vdw}}$). Because C179 and C214 were treated as a single fragment, the interaction energy of C214 was collected into C179. These values were averaged results over the selected 40 calculations (see text). The dots over the residue name indicate that the pathogenetic point mutation of the residue has been identified.[45]

proteins: the side-chains of hydrophobic residues tend to be directed toward the inside of the protein and interact with the other hydrophobic residues, by which these residues play an important role in retaining the native conformation of the protein. Roughly speaking, in the case of PrP, several regions of the residues, Y128−F141, M154−V166, and M205−Y218, largely contribute to the retention of the native structure via van der Waals interactions. From the results in Figure 6, we can qualitatively say that van der Waals interaction is delocalized to many residues unlike the ionic interactions of the salt bridges. Several of the amino acid residues related to pathogenetic mutations[45] were labeled with black dots, some of which have large van der Waals interactions. Such results indicated a possibility that van der Waals interactions between residues may be important in the pathogenetic mechanism of prion diseases.

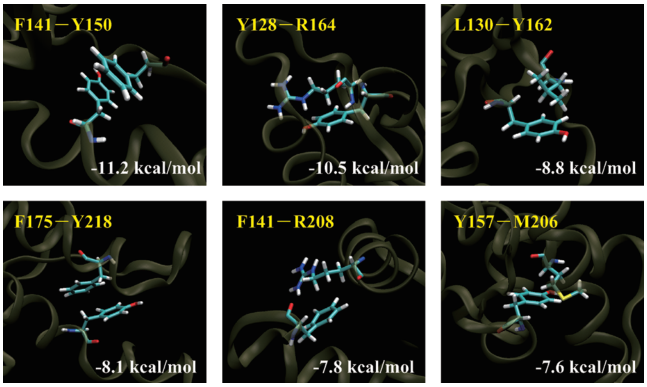


***Figure 7.*** The six pairs of amino acid residues having the largest interaction energies associated with the van der Waals interaction ($\Delta E_{IJ}^{\mathrm{corr}}$). Two pairs had $\pi/\pi$ type interactions, and four pairs had CH/$\pi$ type interactions.

van der Waals interaction in biomolecular systems can be categorized as $\pi/\pi$ interactions and CH/$\pi$ interactions,[46] which are constructed from aromatic rings and C−H bonds. In recent years, such interactions have been considered to be important because many aromatic rings and C−H bonds exist in proteins. In Figure 7, we show the six pairs of amino acid residues with the largest van der Waals interaction energies in the PrP. Two pairs had $\pi/\pi$ interactions (F141−Y150 and F175−Y218), and the other four pairs had CH/$\pi$ interactions (Y128−R164, L130−Y162, F141−R208, and Y157−M206). Although only six pairs were illustrated in this Article, there are many pairs having such types of interactions. Therefore, $\pi/\pi$ and CH/$\pi$ interactions are considered to be important in retaining the high-order structure of the protein despite interaction energies smaller than those of the ionic interactions of the salt bridges.

**4.6. Influence of Geometrical Fluctuation.** Of theoretical interest, we discuss the influence of the geometrical fluctuation on the two quantities, that is, the solvent interaction energy of a protein ($\Delta E_{\mathrm{total}}^{\mathrm{solvent}}$) and the total energy of a protein ($E_{\mathrm{total}}^{\mathrm{protein}}$). The solvent interaction energy is defined in eq 7, and the total energy of protein can be calculated by restricting the summations of eq 3 within the fragments belonging to the protein.

Figure 8 illustrates the histograms of the 40 values obtained from calculations with the different structures. Although the rigorous discussion of the fluctuation is not possible here, we can roughly discuss the fluctuation of our results. In the case of the solvent interaction energy, the standard deviation and the difference between the maximum and minimum values were 142.5 and 507.0 kcal/mol, respectively. On the other hand, the standard deviation was 62.0 kcal/mol, and the difference between the maximum and minimum values was 233.1 kcal/mol in the total energy of the protein. The larger fluctuation of the solvent interaction energy reflected the high mobility of the solvent molecules. These results indicated that values obtained from the quantum chemical calculations for biomolecular systems might strongly depend on the selection of the atomic coordinates. Thus, we should always discuss the nature of proteins using averaged results with the various structure instead of one result with a single structure. Additionally, we intend to use our results

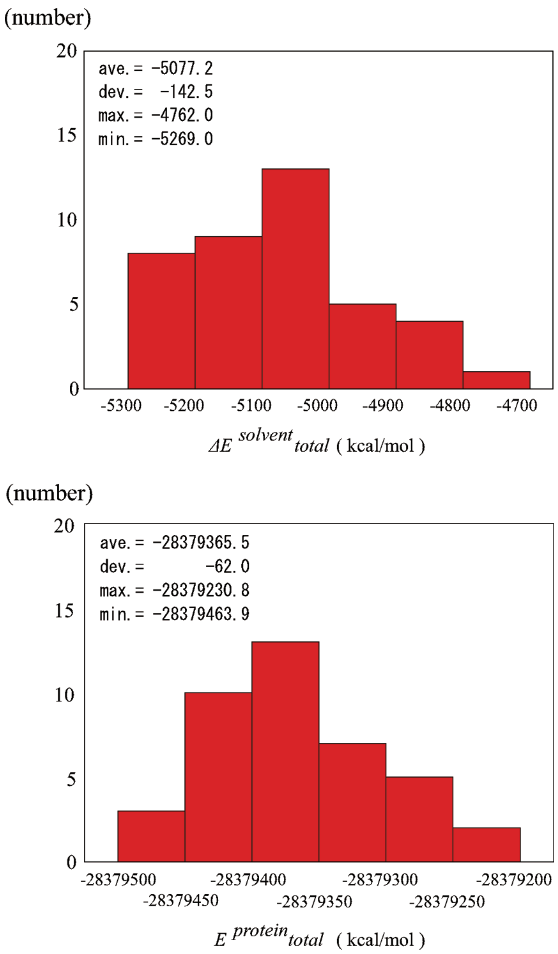


***Figure 8.*** The two histograms of the energy distribution of the solvent interaction energy of the protein ($\Delta E^{\text{solvent}}_{\text{total}}$) and the total energy of the protein ($E^{\text{protein}}_{\text{total}}$). These data were obtained from the results using the selected 40 structures (see text). The average values, standard deviations, maximum values, and minimum values are also given.

in this subsection as fundamental data concerning the influence of the geometrical fluctuation of biomolecular systems in quantum chemical studies.

## 5. Summary

In this work, we examined the interactions retaining the native structures of PrP with quantum chemical calculations based on the FMO method. The solvent interactions could be included with the explicit solvent model, and the influence of the geometrical fluctuation was taken into account using the 40 results with different structures.

The direct calculation of the interaction energies between residues and solvent molecules showed that two types of charged residues exist, that is, those having large solvent interaction energies and those having small solvent interaction energies. The difference between them could be explained in connection with salt bridges. The calculated solvent interaction energy also revealed that HA had a high hydrophilicity while HB had a low hydrophilicity. Next, the intramolecular interaction energies provided information about the seven salt bridges of PrP. Two salt bridges, D144−R148 and E147−R151, contributed to retaining the helical structure of HA, and one salt bridge, K204−E207, stabilized the helical structure of HC. The remaining four salt bridges, R164−D178, R156−D202, E146−R208, and R156−E196, were helpful for conservation of the tertiary structure. These results about the salt bridges were consistent with some of the pathogenetic mutations of the prion diseases. Finally, we carried out the analysis of van der Waals interactions, which showed that several hydrophobic residues contributed to the stability of the native conformation. Our analysis indicated that several regions of the residues, Y128−F141, M154−V166, and M205−Y218, had large interaction energies associated with van der Waals interactions. It was also found that several residues related to the pathogenetic mutations had large van der Waals interactions, indicating an importance of van der Waals interactions in the pathogenetic mechanism. We expect that our results will be utilized in future studies to elucidate the mechanism of prion diseases.

From a theoretical point of view, we examined the influence of the geometrical fluctuation on the results of quantum chemical calculations. The standard deviation of the total energy and the solvent interaction energy were 62.0 and 142.5 kcal/mol, respectively. The larger fluctuation of the solvent interaction energy was considered to be caused by the high mobility of the solvent molecules. These results will be utilized as fundamental data concerning the influence of the geometrical fluctuation in ab initio quantum chemical studies for biomolecular systems.

**Acknowledgment.** This study was supported by a Grant-In-Aid for Young Scientist (Start-up) No. 20850020 from the Japan Society for the Promotion of Science. This study was also supported by the Program for Promotion of Fundamental Studies in Health Science of the National Institute of Biomedical Innovation. Calculations necessary in this study were performed using the computational resources at Gifu University.

## References

(1) Prusiner, S. B. Novel proteinaceous infectious particles cause scrapie. *Science* **1982**, *216*, 136–144.

(2) Prusiner, S. B. Molecular biology and transgenetics of prion diseases. *Crit. Rev. Biochem. Mol. Biol.* **1991**, *26*, 397–438.

(3) Prusiner, S. B. Prions. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 13363–13383.

(4) Riek, R.; Hornemann, S.; Wider, G.; Billeter, M.; Glockshuber, R.; Wuthrich, K. NMR structure of the mouse prion protein domain PrP(121−231). *Nature* **1996**, *382*, 180–182.

(5) Donne, D. J.; Viles, J. H.; Groth, D.; Mehlhorn, I.; James, T. L.; Cohen, F. E.; Prusiner, S. B.; Wright, P. E.; Dyson, H. J. Structure of the recombinant full-length hamster prion protein PrP(29−231): The N terminus is highly flexible. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 13452–13457.

(6) James, T. L.; Liu, H.; Ulyanov, N. B.; Farr-Jones, S.; Zhang, H.; Donne, D. G.; Kaneko, K.; Groth, D.; Mehlhorn, I.; Prusiner, S. B.; Cohen, F. E. Solution structure of a 142-residue recombinant prion protein corresponding to the infectious fragment of the scrapie isoform. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10086–10091.

(7) Gossert, A. D.; Bonjour, S.; Lysek, D. A.; Fiorito, F.; Wuthrich, K. Prion protein NMR structures of elk and of mouse/elk hybrids. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 646–650.

(8) Lysek, D. A.; Schorn, C.; Nivon, L. G.; Esteve-Moya, V.; Christen, B.; Calzolai, L.; von Schroetter, C.; Fiorito, F.; Herrmann, T.; Guntert, P.; Wuthrich, K. Prion protein NMR structures of cats, dogs, pigs, and sheep. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 640–645.

(9) Gao, J. Toward a molecular orbital derived empirical potential for liquid simulations. *J. Phys. Chem. B* **1997**, *101*, 657–663.

(10) Gao, J. A molecular-orbital derived polarization potential for liquid water. *J. Chem. Phys.* **1998**, *109*, 2346–2354.

(11) Kitaura, K.; Sawai, T.; Asada, T.; Nakano, T.; Uebayasi, M. Pair interaction molecular orbital method: an approximate computational method for molecular interactions. *Chem. Phys. Lett.* **1999**, *312*, 319–324.

(12) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. Fragment molecular orbital method: an approximate computational method for large molecules. *Chem. Phys. Lett.* **1999**, *313*, 701–706.

(13) Fedorov, D. G.; Kitaura, K. Extending the power of quantum chemistry to large systems with the fragment molecular orbital method. *J. Phys. Chem. A* **2007**, *111*, 6904–6914.

(14) Nakano, T.; Kaminuma, T.; Sato, T.; Fukuzawa, K.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. Fragment molecular orbital method: use of approximate electrostatic potential. *Chem. Phys. Lett.* **2002**, *351*, 475–480.

(15) Fukuzawa, K.; Mochizuki, Y.; Tanaka, S.; Kitaura, K.; Nakano, T. Molecular interactions between estrogen receptor and its ligand studied by the ab initio fragment molecular orbital method. *J. Phys. Chem. B* **2006**, *110*, 16102–16110.

(16) Fukuzawa, K.; Komeiji, Y.; Mochizuki, Y.; Kato, A.; Nakano, T.; Tanaka, S. Intra- and intermolecular interactions between cyclic-AMP receptor protein and DNA: Ab initio fragment molecular orbital study. *J. Comput. Chem.* **2006**, *27*, 948–960.

(17) Ito, M.; Fukuzawa, K.; Mochizuki, Y.; Nakano, T.; Tanaka, S. Ab initio fragment molecular orbital study of molecular interactions between liganded retinoid X receptor and its coactivator: Roles of helix 12 in the coactivator binding mechanism. *J. Phys. Chem. B* **2007**, *111*, 3525–3533.

(18) Ito, M.; Fukuzawa, K.; Mochizuki, Y.; Nakano, T.; Tanaka, S. Ab initio fragment molecular orbital study of molecular interactions between liganded retinoid X receptor and its coactivator; Part II: Influence of mutations in transcriptional activation function 2 activating domain core on the molecular interactions. *J. Phys. Chem. A* **2008**, *112*, 1986–1998.

(19) Ito, M.; Fukuzawa, K.; Ishikawa, T.; Mochizuki, Y.; Nakano, T.; Tanaka, S. Ab initio fragment molecular orbital study of molecular interactions in liganded retinoid X receptor: Specification of residues associated with ligand inducible Iinformation transmission. *J. Phys. Chem. B* **2008**, *112*, 12081–12094.

(20) Amari, S.; Aizawa, M.; Zhang, J.; Fukuzawa, K.; Mochizuki, Y.; Iwasawa, Y.; Nakata, K.; Chuman, T.; Nakano, T. VISCANA: Visualized cluster analysis of protein−ligand interaction based on the ab initio fragment molecular orbital method for virtual ligand screening. *J. Chem. Inf. Model.* **2006**, *46*, 221–230.

(21) Nakanishi, Y.; Fedorov, D. G.; Kitaura, K. Molecular recognition mechanism of FK506 binding protein: An all-electron fragment molecular orbital study. *Proteins: Struct., Funct., Bioinf.* **2007**, *68*, 145–158.

(22) Iwata, T.; Fukuzawa, K.; Nakajima, K.; Aida, S. H.; Mochizuki, Y.; Watanabe, H.; Tanaka, S. Theoretical analysis of binding specificity of influenza viral hemagglutinin to avian and human receptors based on the fragment molecular orbital method. *Comput. Biol. Chem.* **2008**, *32*, 198–211.

(23) Ishikawa, T.; Mochizuki, Y.; Amari, S.; Nakano, T.; Tokiwa, H.; Tanaka, S.; Tanaka, K. Fragment interaction analysis based on local MP2. *Theor. Chem. Acc.* **2007**, *118*, 937–945.

(24) Ishikawa, T.; Mochizuki, Y.; Amari, S.; Nakano, T.; Tanaka, S.; Tanaka, K. An application of fragment interaction analysis based on local MP2. *Chem. Phys. Lett.* **2008**, *463*, 189–194.

(25) Kurisaki, I.; Fukuzawa, K.; Komeiji, Y.; Mochizuki, Y.; Nakano, T.; Imada, J.; Chmielewski, A.; Rothstein, S. M.; Watanabe, H.; Tanaka, S. Visualization analysis of interfragment interaction energies of CRP-cAMP-DNA complex based on the fragment molecular orbital method. *Biophys. Chem.* **2007**, *130*, 1–9.

(26) Fedorov, D. G.; Kitaura, K.; Li, H.; Jensen, J. H.; Gordon, M. S. The polarizable continuum model (PCM) interfaced with the fragment molecular orbital method (FMO). *J. Comput. Chem.* **2006**, *27*, 976–985.

(27) Komeiji, Y.; Ishida, T.; Fedorov, D. G.; Kitaura, K. Change in a protein's electronic structure induced by an explicit solvent: An ab initio fragment molecular orbital study of ubiquitin. *J. Comput. Chem.* **2007**, *28*, 1750–1762.

(28) Ishikawa, T.; Mochizuki, Y.; Nakano, T.; Amari, S.; Mori, H.; Honda, H.; Fujita, T.; Tokiwa, H.; Tanaka, S.; Komeiji, Y.; Fukuzawa, K.; Tanaka, K.; Miyoshi, E. Fragment molecular orbital calculations on large scale systems containing heavy metal atom. *Chem. Phys. Lett.* **2006**, *427*, 159–165.

(29) Ishikawa, T.; Ishikura, T.; Kuwata, K. Theoretical study of the prion protein based on the fragment molecular orbital method. *J. Comput. Chem.* **2009**, *30*, 2594–2601.

(30) Mochizuki, Y.; Nakano, T.; Koikegami, S.; Tanimori, S.; Abe, Y.; Nagashima, U.; Kitaura, K. A parallelized integral-direct second-order Møller-Plesset perturbation theory method with a fragment molecular orbital scheme. *Theor. Chem. Acc.* **2004**, *112*, 442–452.

(31) Mochizuki, Y.; Koikegami, S.; Nakano, T.; Amari, S.; Kitaura, K. Large scale MP2 calculations with fragment molecular orbital scheme. *Chem. Phys. Lett.* **2004**, *396*, 473–479.

(32) Fedorov, D. G.; Kitaura, K. Second order Møller-Plesset perturbation theory based upon the fragment molecular orbital method. *J. Chem. Phys.* **2004**, *121*, 2483–2490.

(33) Protein Data Bank (PDB); http://www.pdb.org/.

(34) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(35) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(36) Case, D. A.; Darden, T. A.; Cheatham, T. E., III.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvàry, I.; Wong, K. F.; Paesani, F.;

Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*; University of California: San Francisco, CA, 2008.

(37) Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(38) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. Use of approximate integrals in ab initio theory. An application in MP2 energy calculations. *Chem. Phys. Lett.* **1993**, *208*, 359–363.

(39) Weigend, F.; Köhn, A.; Hätting, C. Efficient use of the correlation consistent basis sets in resolution of the identity MP2 calculations. *J. Chem. Phys.* **2002**, *116*, 3175–3183.

(40) Ishikawa, T.; Kuwata, K. Fragment molecular orbital calculation using the RI-MP2 method. *Chem. Phys. Lett.* **2009**, *474*, 195–198.

(41) Nguyen, J.; Baldwin, M. A.; Cohen, F. E.; Prusiner, S. B. Prion protein peptides induce α-Hhelix to *β*-sheet conformational transitions. *Biochemistry* **1995**, *34*, 4186–4192.

(42) Morrissey, M. P.; Shakhnovich, E. I. Evidence for the role of PrPC helix 1 in the hydrophilic seeding of prion aggregates. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11293–11298.

(43) Radzicka, A.; Wolfenden, R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* **1988**, *27*, 1664–1670.

(44) Kuhn, L. A.; Swanson, C. A.; Pique, M. E.; Tainer, J. A.; Getzoff, E. D. Atomic and residue hydrophilicity in the context of folded protein structures. *Proteins: Struct., Funct., Bioinf.* **1995**, *23*, 536–547.

(45) Kong, Q.; Surewicz, W. K.; Petersen, R. B.; Zou, W.; Chen, S. G.; Gambetti, P.; Parchi, P.; Capellari, S.; Goldfarb, L.; Montagna, P.; Lugaresi, E.; Piccardo, P.; Ghetti, B. In *Inherited Prion Diseases. Prion Biology and Diseases*, 2nd ed.; Prusiner, S. B., Ed.; Cold Spring Harbor Laboratory Press: New York, 2004; Chapter 14, pp 673−775.

(46) Nishio, M.; Hirota, M.; Umezawa, Y. *Specific Interactions in Protein Structures. The CH/π Interaction*; Wiley−VCH: New York, 1998; Chapter 11, pp 175−202.

CT900456V

JCTC Journal of Chemical Theory and Computation

# Utility of the Hard/Soft Acid−Base Principle via the Fukui Function in Biological Systems

John Faver and Kenneth M. Merz, Jr.*

*Quantum Theory Project, Department of Chemistry, University of Florida, 2328 New Physics Building, P.O. Box 118435, Gainesville, Florida 32611-8435*

Received September 25, 2009

**Abstract:** The hard/soft acid−base (HSAB) principle has long been known to be an excellent predictor of chemical reactivity. The Fukui function, a reactivity descriptor from conceptual density functional theory, has been shown to be related to the local softness of a system. The usefulness of the Fukui function is explored and demonstrated herein for three common biological problems: ligand docking, active site detection, and protein folding. In each type of study, a scoring function is developed on the basis of the local HSAB principle using atomic Fukui indices. Even with necessary approximations for its use in large systems, the Fukui function remains a useful descriptor for predicting chemical reactivity and understanding chemical systems.

## 1. Introduction

Computational biochemistry is an expanding field that relies heavily on the increasing efficiency of computers and clever algorithms to approach very large and complex problems. While methods for high-level quantum mechanical (QM) calculations have been developed and proven to be very successful in calculating energies, equilibrium structures, vibrational frequencies, and more properties of small- to medium-sized molecules, the computational resources required for very large systems (e.g., a protein of many hundreds of atoms) are usually unattainable.[1–3] For these very large systems, more approximate modeling tools are often used such as molecular mechanics.[4,5] These more approximate methods greatly accelerate the speed at which energy calculations are performed, but they do not explicitly account for electronic structure. This can be a disadvantage, because there is a significant amount of information encoded in the electronic structure of a system.

Conceptual density functional theory (CDFT) defines many reactivity descriptors for a system based on its electron density and provides a large set of tools for use in the prediction and understanding of chemical reactivity. An extensive review of CDFT and the myriad of possible descriptors has been compiled by Geerlings, De Proft, and Langenaeker.[6] These descriptors have been used in the past for a diverse set of chemical systems.[7–9] More recently, they have been used with some success in biochemically relevant systems including the detection of metabolic sites in known drug molecules, the understanding of metal binding to porphyrin, and enzymatic catalysis.[10–12] A beneficial characteristic of these descriptors is that the majority of them depend on quantities such as electron density that can be obtained from any QM method, including semiempirical QM Hamiltonians.[13,14]

In the past two decades, advances in algorithms have allowed computational chemists to perform QM calculations on large systems such as proteins.[15,16] One such method is the divide and conquer method.[17–21] By dividing a molecule into smaller subsystems and performing separate calculations followed by the formation of a global density matrix, the method greatly accelerates calculations for large systems. An important result of this development is that electron density and descriptors based on electron density can now be calculated for large molecules as well as small molecules. Khandogin and York recently described a few such useful descriptors for divide and conquer semiempirical calculations.[22]

Pearson's hard/soft acid−base (HSAB) principle states that chemical species can be described as being either hard or soft acids or bases.[23] Soft species tend to be easily polarizable and large in volume, have low charge, and have small HOMO−LUMO gaps. Hard species tend to have the opposite characteristics: they are not easily polarized, are small in

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

10.1021/ct9005085 © 2010 American Chemical Society
Published on Web 01/07/2010

volume, are highly charged, and have large HOMO−LUMO gaps. The HSAB concept can be summarized as one simple rule: hard species favor interacting with hard species, and soft species tend to favor interacting with soft species. The HSAB concept has been successful in predicting reactivity preferences in many systems since its inception.[24–32]

Researchers have devised various methods of quantifying hardness and softness. Although empirical approximations have been used in the past, in this paper we will describe the use of one related reactivity descriptor from CDFT called the Fukui function which has been shown to carry information about chemical softness.[33–35] This work then explores its applicability to biological problems, specifically ligand docking, active site detection, and protein folding.

## 2. Background

According to density functional theory, changes in the electronic energy $\mathrm{d}E[\rho(r)]$ are related to changes in the number of electrons $N$ and changes in the external potential $v(r)$ felt by the electron distribution (which usually refers to the nuclear positions in chemical systems):

$$\mathrm{d}E[\rho(r)] = \mu\,\mathrm{d}N + \int \rho(r)\,\mathrm{d}v(r)\,\mathrm{d}r \tag{1}$$

For simplicity, consider a molecule at a given geometry in its ground state so that $\mathrm{d}v(r)$ is zero. Thus, the partial derivative of energy with respect to the number of electrons $N$ at constant geometry is the electronic chemical potential $\mu$:

$$\mu = \left(\frac{\partial E}{\partial N}\right)_{v(r)} = -\chi \tag{2}$$

This quantity has been related conceptually to the electronegativity $\chi$ of a system.[36] This definition agrees with chemical intuition, as more energetically favorable changes in electron number yield higher values of electronegativity. Consider now the second partial derivative of the energy with respect to the electron number

$$\eta = \left(\frac{\partial^2 E}{\partial N^2}\right)_{v(r)} = \left(\frac{\partial \mu}{\partial N}\right)_{v(r)} \tag{3}$$

which has been defined as $\eta$, or chemical hardness as described by Pearson.[33] This definition can be understood by the analogy of a spring constant in classical physics. The spring constant is the second derivative of energy with respect to displacement and measures the difficulty of displacing a spring from its equilibrium position. Equation 3 can be thought of as measuring the difficulty of changing a system's number of electrons, which is conceptually similar to nonpolarizability, or hardness. Since softness is the opposite of hardness, it has been defined as the inverse of hardness:

$$S = \frac{1}{\eta} = \left(\frac{\partial N}{\partial \mu}\right)_{v(r)} \tag{4}$$

Parr and Yang have also defined a distance-dependent version of softness, called the local softness, as

$$s(r) = \left(\frac{\partial \rho(r)}{\partial \mu}\right)_{v(r)} = \left(\frac{\partial \rho(r)}{\partial N}\right)_{v(r)}\left(\frac{\partial N}{\partial \mu}\right)_{v(r)} = f(r)\,S \tag{5}$$

The local softness function identifies the softest regions of a molecule. A system has a total softness $S$ that is distributed throughout the molecule by a function $f(r)$ called the Fukui function:

$$f(r) = \left(\frac{\partial \rho(r)}{\partial N}\right)_{v(r)} = \frac{s(r)}{S} \tag{6}$$

The Fukui function is normalized to unity so that the local softness integrates over all space to yield the total softness. Furthermore, the Fukui function can be viewed as containing the same information as the local softness, since the two are proportional to each other by a constant, $S$. Although there exist several descriptors for local hardness, the problem of defining it has not been resolved.[37,38] In this work, low values of local softness are assumed to be locally hard. From the equations of DFT, we now have the Fukui function, a descriptor that identifies the softest (and hardest) regions of a molecule. With this knowledge in hand, one can begin to make predictions about chemical reactivity.

One issue that arises when calculating the Fukui function is that it is a derivative of the electron number, which is by nature an integer. Although recent studies have examined ways to circumvent this apparent discontinuity, these methods are impractical at this time for the large systems considered here.[39,40] Limiting the calculations to changes with integer electrons, it is necessary to use finite difference derivatives. With the finite difference formulas, there is the option of taking the derivative from the left, right, or center:

$$f(r)^- \cong \frac{1}{\Delta N}[\rho(r,N) - \rho(r,N-\Delta N)] \tag{7}$$

$$f(r)^+ \cong \frac{1}{\Delta N}[\rho(r,N+\Delta N) - \rho(r,N)] \tag{8}$$

$$f(r)^0 \cong \frac{1}{2}[f(r)^+ + f(r)^-] \tag{9}$$

The Fukui function taken from the left is the difference in electron density between the reference system and the system with an electron removed, e.g., a ground state and its cation (eq 7). Because maxima in this function represent areas where electron density is most favorably decreased, they are interpreted as areas in a molecule most favorable for electrophilic attack. The Fukui function taken from the right has maxima that are interpreted as areas most favorable for nucleophilic attack, since it detects areas where electron density increases most favorably under addition of electrons (eq 8). The centered derivative is simply the average of the two other derivatives and has often been interpreted as showing areas most favorable for attack by a radical (eq 9). A recent study has explored the validity of this interpretation and found that it may not be quite as easy to interpret as the other derivatives.[41] While the left and right derivatives are clearly understood in terms of two classical reaction mechanisms, the middle derivative can for now be viewed as the best approximation of the derivative at the reference state.

In addition to finite difference derivatives, a second common approximation of the Fukui function is the condensed Fukui function, which is composed of atomic Fukui indices.[42] Within this approximation, atomic partial charges are used to replace the electron density in the expression for the Fukui function. Though this may be a crude approximation of the full electron density and Fukui function, several studies have been successful with its use.[7,8,10,12,43–45] In general, one must choose a density partitioning scheme which unfortunately can depend heavily on the QM method or basis set and thus introduce error. Because of this, Fukui indices are sometimes negative, which seems unphysical. A negative Fukui index implies that addition of electrons to a system decreases density in locations in the system or vice versa. Though some example molecules have been shown to have this interesting property, it should not be as common as the use of Fukui indices suggests.[46,47] Keeping this in mind, Fukui indices were used in this work rather than full Fukui functions, simply because full Fukui functions are considerably more expensive to calculate.

A third common approximation is the use of the frozen orbital approximation, in which a single calculation is done to obtain the eigenstates of the system which are assumed to be "frozen" in place as electrons are added or removed. Clearly, changing the electron number in a system will alter the forces felt by the remainder of the electrons, and the eigenstates will be altered in a phenomenon called orbital relaxation. Examples have been shown in which Fukui indices based on the frozen orbital approximation fail to predict correct reactivity in small organic molecules.[47] Orbital relaxation effects were taken into account in this work by performing separate calculations for the ground-state system, the system with added electrons, and the system with electrons removed, rather than using the frozen orbital approximation.

Khandogin et al. have described the calculation and interpretation of several QM-based reactivity descriptors for biomolecules, including the Fukui function and local softness.[22] The present work concentrates on one of them, the Fukui function, and attempts to determine in what kinds of applications it can be used and for what kinds of interactions it can account and then determine the extent of its reliability. With these approximations (divide and conquer, AM1, finite difference, Mulliken atomic charges), five specific types of problems were addressed: finding correct ligand poses in an active site, detecting active binders from a set including decoy ligands, ranking binding affinities of ligands, finding reactive sites in a protein, and detecting native from decoy protein structures. In each of these systems, it was hypothesized that molecular interactions are favorable when hard areas are near hard areas and soft areas are near soft areas. In each of the five problems, a scoring function based on this hypothesis is developed and used to predict the preferred molecular interactions.

## 3. General Computational Details

All calculations were performed with the semiempirical AM1 Hamiltonian in the DivCon program utilizing the divide and conquer strategy for proteins.[17–19,48] Standard unrestricted AM1 calculations were done in DivCon for all ligand molecules. Unrestricted divide and conquer calculations were done for the proteins in the 1F40 and 2FOM docking studies, and restricted divide and conquer calculations were done for the 1EFY docking study and the 1ORC and 1I6C protein folding study. Atomic Fukui indices were calculated from centered finite difference derivatives of Mulliken charges. The derivatives were calculated by varying the electron number by 1 for ligand molecules. The electron number was varied by 5 in the case of 1F40 and 2FOM, by 8 for 1EFY, and by 4 for 1ORC and 1I6C. These values were chosen to roughly correspond to the size of the proteins. Fukui indices and molecular surfaces were visualized with the program PYMOL.[49]

**3.1. Docking.** Docking studies were performed on three protein/ligand systems: FKBP12 (PDB ID 1F40), dengue virus type 2 NS3 protease (PDB ID 2FOM), and poly(ADP-ribose) polymerase (PDB ID 1EFY). 1F40 is an NMR structure bound to a synthetic ligand, GPI-1046.[50] 2FOM is an X-ray crystal structure (1.50 Å resolution) without a bound ligand.[51] Several known active binders with experimental $IC_{50}$ values for the dengue protease were taken from a previous docking study.[52,53] 1EFY is an X-ray crystal structure (2.20 Å resolution) with bound inhibitor. The structure was taken from the DUD (directory of useful decoys) data set along with 32 active inhibitors with experimental $K_i$ values taken from Tikhe et al.[54,55] Schrödinger's Glide program was used for all docking studies with the XP scoring function except for where it is mentioned otherwise in the 1F40 study (where AutoDock was used).[56,57] Hydrogens were added to the crystal structures, and the structures were relaxed with the OPLS 2001 force field within the Maestro program prior to grid generation and docking. These final structures from Maestro (receptors and ligand poses) were used in the AM1 single-point calculations to obtain atomic Fukui indices.

*3.1.1. Ranking Ligand Poses in a Receptor.* The first docking test was to determine the correct pose of a ligand in the active site of a receptor. The ligand from the FKBP12 system was docked to FKBP12 with the Autodock program.[58] Ten of the best poses from the docking results were taken to evaluate the hardness and softness matching between atoms in the docked conformations. A score was developed to measure the complementarity of a given ligand and its receptor, hereafter called the FRMSD, or the root-mean-square difference in the Fukui index. For each atom in the docked ligand ($L_i$), the nearest atom of the protein ($R_i$) was matched to it to form a closest match atomic pair. The difference in Fukui indices for each atomic pair is squared and then averaged over all ligand atoms (eq 10). A lower value of FRMSD represents a better ligand pose with respect to the match between the hardness and softness of the atoms in the two molecules.

$$\mathrm{FRMSD} = \sqrt{\frac{\sum_i (f_{\mathrm{L}_i} - f_{\mathrm{R}_i})^2}{N}} \tag{10}$$

A score was calculated for the 10 best poses generated by Autodock, and the results are shown in Figure 1. The two

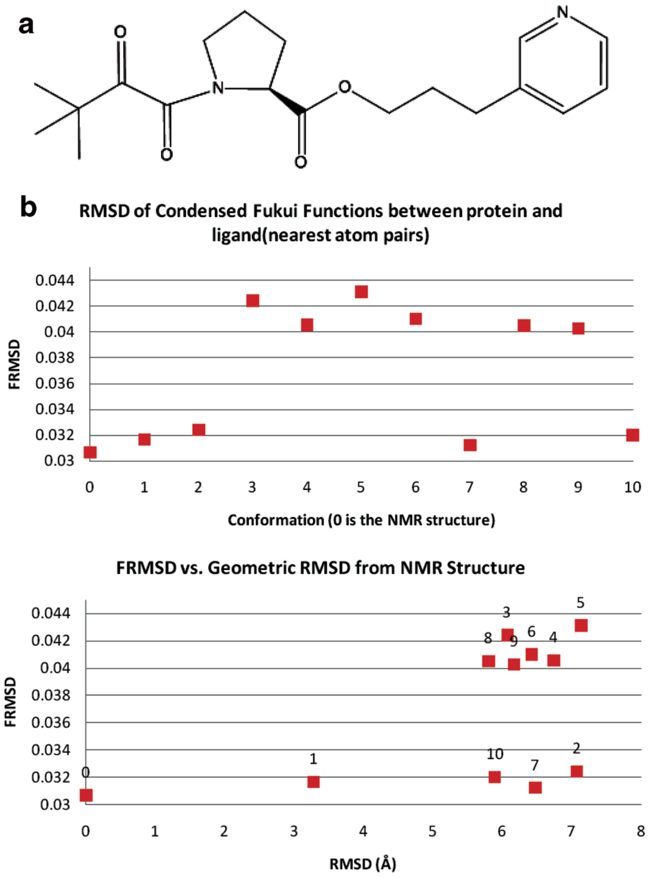


***Figure 1.*** (a) GPI molecule used in the docking procedure for FKB12. (b, top) Docking of a known binder (GPI-1046) to FKBP12 (PDB ID 1F40). Each point on the horizontal axis represents a different pose taken from the top 10 poses generated by the program Autodock, arranged by decreasing Autodock score. Zero on the horizontal axis represents the observed NMR structure. Low FRMSD values indicate a better hard/soft match between the ligand and its receptor. (b, bottom) FRMSD vs geometric rmsd from the NMR docked structure.

best scoring poses from the Autodock run (poses 1 and 2) score well with the FRMSD score, and worse poses from the Autodock score generally score worse with the FRMSD score. Perhaps the most interesting finding is that the observed pose from the NMR structure (pose 0) has the best FRMSD score, meaning the observed pose is among the docked poses with the best soft/hard matching between closest atom pairs. To show that the NMR pose is actually an acceptable reference pose, an energy minimization was carried out in AMBER for the ligand in the restrained active site. The relaxed ligand structure had an rmsd of 0.255 Å with respect to the NMR ligand structure, which, in our opinion, is a negligible difference.

Figure 1b (bottom) plots the FRMSD of each pose vs the geometric rmsd with respect to the pose from the NMR structure. It was observed that the ligands are divided almost evenly into those with good FRMSD scores and those with poorer FRMSD scores. Upon visualization of the good poses, it was seen that poses 0 and 1 are actually very similar, with the major differences being a rotation of the pyridine ring and a rotation of the *tert*-butyl group. Pose 7 had the same placement of the central pyrrolidine ring but had the positions of the *tert*-butyl and pyridyl groups swapped (i.e., a molecular rotation by 180°). This pose was also observed in a docking study by Wang et al. in which it was shown to match NMR chemical shift data fairly well.[59]

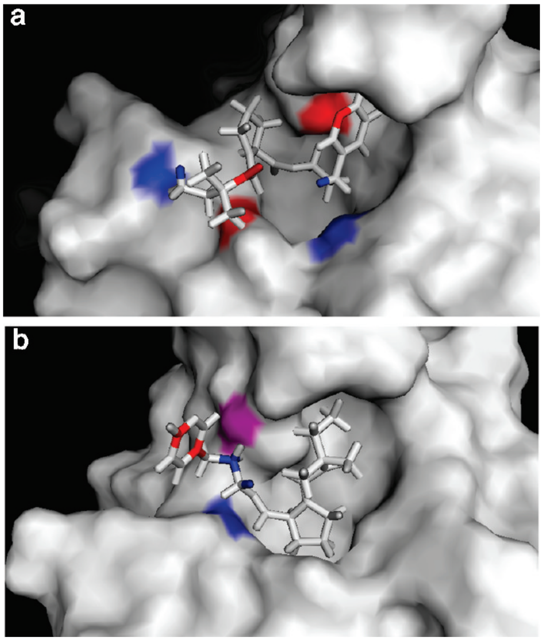


***Figure 2.*** Two docked poses for the FKBP/GPI complex. The active site is shown as a white surface, and the ligand is shown as white sticks. Good hard/soft matching atom pairs are shown in blue and poor hard/soft matches are shown in red on both the ligand and the protein surface. (a) Pose number 5 from the docking procedure. (b) Pose 0, the NMR pose.

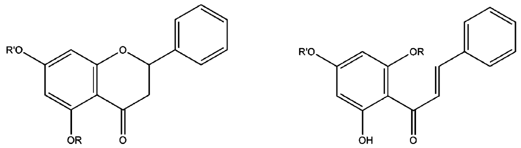


***Figure 3.*** Active binders to dengue type 2 protease taken from a previous docking study by Othman et al.:[53] (a, left) pinostrobin (R = H; R′ = Me), pinocembrin (R = H; R′ = H), alpinetin (R = Me; R′ = H), (b, right) pinostrobin chalcone (R = H; R′ = Me), pinocembrin chalcone (R = H; R′ = H), cardamonin (R = Me; R′ = H). Reprinted from ref 53. Copyright 2008 American Chemical Society.

A benefit of this closest atom pair scoring is that the resulting data can be qualitatively analyzed by simply searching for the best and worst matched pairs. A simple script can analyze the data and produce input for visualization programs such as PyMOL, as demonstrated in Figure 2. Such visual and qualitative measurement of hard/soft matching could be useful in the drug design process, as the human eye can easily detect the best and worst hard/soft matches. In addition, it provides a method of verifying the FRMSD results. In Figure 2, good contacts are marked by shades of blue and poor contacts are marked by shades of red. Of course one could show as many contacts as desired, but here only the two best matches and the two worst matches are shown.

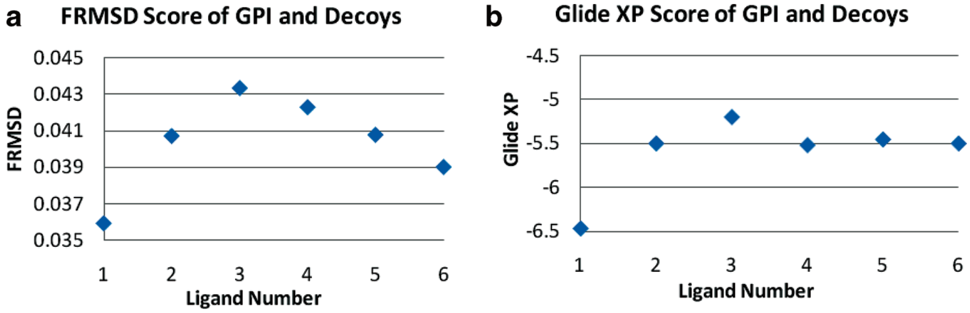


***Figure 4.*** (a) FRMSD scores of the best scoring poses of the GPI molecule and decoy ligands to FKBP12. (b) Glide XP scores. The ligands are numbered as (1) GPI-1046, (2) alpinetin, (3) pinocembrin, (4) pinocembrine chalcone, (5) pinostrobin, and (6) pinostrobin chalcone. Both scores correctly discriminate the known binder from the decoy ligands.

Pose number 5 from the docking run had the worst FRMSD score out of all 10 poses, and its best and worst pairs are shown in Figure 2a. This figure highlights an important interaction not contained in the Fukui function—hydrogen bonding. One of the worst hard/soft mismatches is between the pyridyl nitrogen and the hydroxyl hydrogen from a tyrosine residue inside the binding pocket, which are at a distance of 2.26 Å from each other. This should be a favorable interaction, but is considered a poor interaction from a hard/soft perspective. This example suggests that the Fukui function alone would not be able to account for all types of molecular interactions and would need contributions from additional terms in a scoring function (such as an electrostatic or hydrogen-bonding term) to be universally applicable or to correctly predict binding affinity. In the meantime it is assumed that ligands of similar construction with similar types of interactions can be analyzed by hardness and softness alone.

The NMR pose is shown in Figure 2b. The native pose has its best contacts just outside the binding pocket, and the worst contacts are with the terminal pyridyl group, which faces outward from the binding site. A tyrosine residue near the pocket is shaded purple because it makes both good contacts with the carbon chain of the ligand and poor contacts with the pyridyl group. The color-coding helps in qualitatively understanding why the native pose is a good docking pose. The pyridyl group may have poor hard/soft matching with the receptor, but it is directed outward from the binding site, making the interaction longer ranged and possibly less unfavorable. This would suggest that if distance were taken into account in the scoring function, the native pose would be even more preferred by FRMSD. From visualizing this pose it seems that another way to improve the FRMSD score would be to include a distance dependence, which is introduced in section 3.1.3.

*3.1.2. Selection of Active Ligands from Decoys.* The second docking test involved the same receptor (FKBP12) with its active binder, GPI-1046, along with a set of decoy ligands from the data set for the dengue virus protease shown in Figure 3. Though these are known binders for the dengue virus type 2 protease, they are assumed to be nonbinding decoys for the FKBP12 system. The top 10 docked poses of each ligand from Glide XP were retained and scored with the closest atom pair FRMSD score. Figure 4a shows only the FRMSD scores for the best scoring pose of each ligand. Figure 4b shows the Glide XP scores of the best scoring poses for comparison.

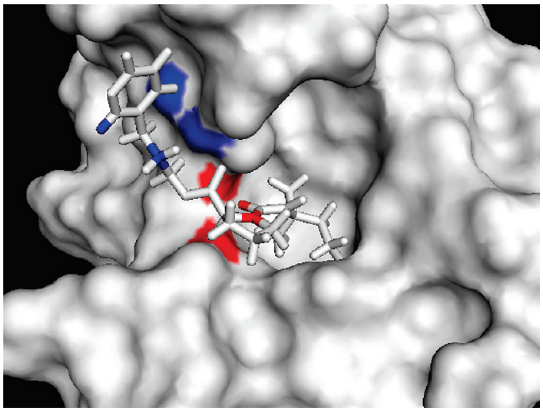
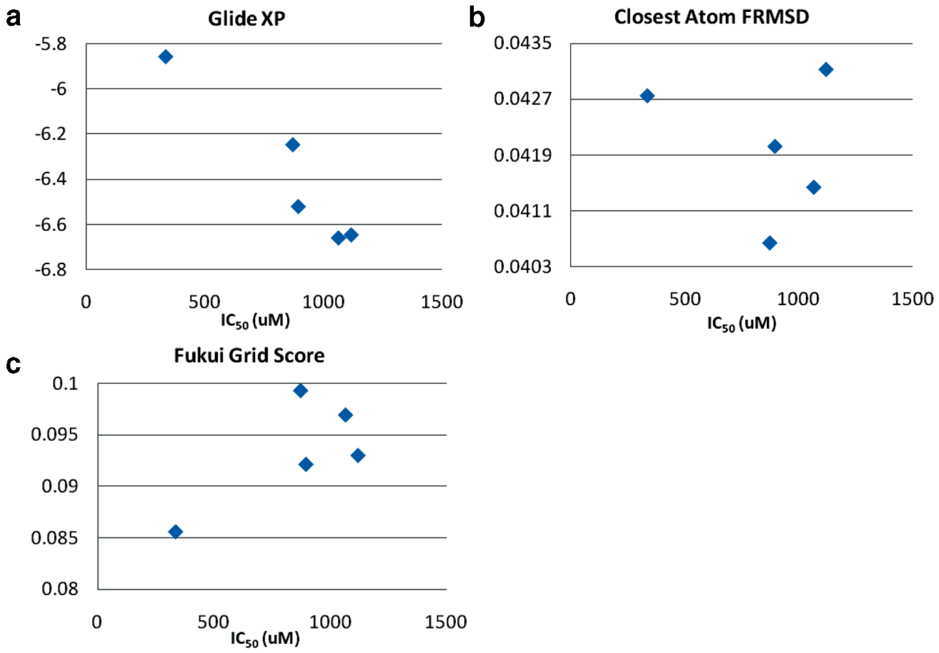


***Figure 5.*** Worst of 10 poses of the correct ligand in the binding site of FKB12 as determined by the FRMSD score. The poorest hard/soft matches (shown in red) are closer contacts than the good matches (shown in blue), showing that this pose is poorly docked according to hard/soft matching.

Both FRMSD and Glide XP were able to score the correct ligand, GPI-1046, as the best binder. In fact, almost all 10 of the poses generated by Glide scored better than all of the decoy poses. Upon visualization of the worst FRMSD scoring pose of GPI-1046 (Figure 5), the poorest matching pairs are a carbonyl oxygen in the ligand with a $\gamma$-methyl hydrogen from an isoleucine residue (at 2.74 Å) and the nitrogen from the ligand's pyrrolidine ring with the $\alpha$-hydrogen of a valine residue (at 3.66 Å). Both of these pairs should represent somewhat favorable electrostatic interactions, which are not captured by the FRMSD score. The best pairs are between the ligand and a tyrosine group, but both of these pairs are at a distance greater than 3.0 Å, leading to a match that is probably overaccounted for in the distance-independent FRMSD score. This pose provides more evidence that distance should be accounted for in a score based on Fukui indices.

*3.1.3. Ranking of Different Ligands by Binding Affinity.* The third type of docking experiment for hardness/softness-based scoring was to rank ligand molecules by binding affinity using the Fukui indices for the ligands and receptor. From visualization of the previous docking results it is apparent that a distant-dependent score is necessary. Here a second score is introduced, hereafter referred to as the Fukui grid score, in which distance dependence to the Fukui indices



d

| Ligand Number | Ligand Name | $IC_{50}$ Exp. (μM) | Glide XP | FRMSD | Fukui-Grid |
|---|---|---|---|---|---|
| 1 | pinostrobin | 334.76 | -5.8571 | 0.042745 | 0.08561 |
| 2 | cardamonin | 872.65 | -6.24693 | 0.04064 | 0.0993 |
| 3 | alpinetin | 895.37 | -6.52093 | 0.04202 | 0.09215 |
| 4 | pinocembrin chalcone | 1065.8 | -6.66056 | 0.041436 | 0.0969 |
| 5 | pinocembrin | 1119.6 | -6.64687 | 0.043122 | 0.0930 |
| 6 | pinostrobin chalcone | - | -7.32997 | 0.041459 | 0.0950 |

***Figure 6.*** Docking and scoring results from known binders to dengue virus type 2 protease. The Glide XP (a), closest atom pair FRMSD (b), and Fukui grid (c) scores are presented along with the experimental $IC_{50}$ values (d). The plotted scores represent the pose that yielded the best score for the particular scoring function.

is included. As mentioned before, atomic Fukui indices approximate the full Fukui function as a collection of points centered at atomic positions, which introduces error by ignoring a substantial amount of information about the topology of the Fukui function. In addition, the closest atom pair approach would not properly account for hard/soft matches or mismatches between functional groups. A distance-dependent score could reduce errors caused by both of these factors by allowing many atomic indices to contribute to a value for a given point in space.

The score is calculated by placing a grid of points over each conformation of each ligand. At each grid point, all atomic Fukui indices are first scaled according to their distance to the grid point and then summed. A grid of values is calculated for the receptor's active site as well as for each ligand pose. The grid of each ligand pose is then compared to the receptor grid to generate a score based on eq 11. The grids are superimposed, and each overlapping grid point is used to produce an rmsd between grids. A lower score implies a better match between the hard and soft areas of the receptor and ligand grids. The grids used here were cubic with each side 10 Å in length. Grid points were spaced by 1.0 Å. Since the distance dependence was unknown, the indices were divided by distance raised to the α power. The parameter α was varied to find the best discrimination between ligands, and for this case a value of α = 0.5 was found to be appropriate.

$$\text{grid score} = \sqrt{\frac{\sum_{k}^{N_{\mathrm{GP}}}[\mathrm{GP}_{\mathrm{ligand}} - \mathrm{GP}_{\mathrm{receptor}}]^2}{\text{number of grid points}}}$$
$$\mathrm{GP}_i = \sum_{j}^{N_i}\left(\frac{f_j}{r_{jk}^{\alpha}}\right) \qquad (11)$$

The first data set tested was the dengue virus type 2 protease (PDB ID 2FOM) and a collection of known binders that has been previously reported along with $IC_{50}$ values (Figure 3).[52,53] The ligands are known to be allosteric binders, as discussed by Othman et al.[53] The same binding site was used in this study as was used previously, but whereas Othman used the standard Glide score, here the Glide XP score is used. The top 10 poses of each ligand

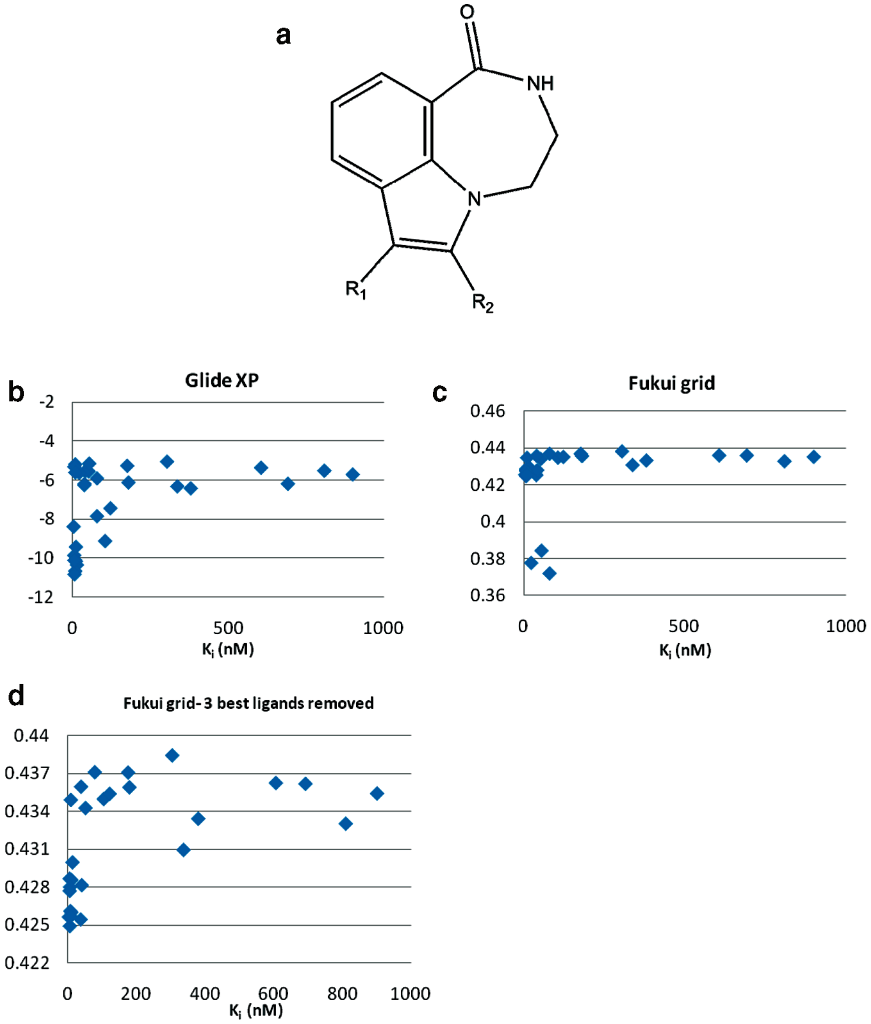


*Figure 7.* (a) Shared molecular scaffold of the 32 ligands from the DUD data set for the poly(ADP-ribose) polymerase receptor. (b) Results from Glide XP. (c) Fukui grid score. (d). Fukui grid score with the three best ligands removed to display the smaller differences between the remaining ligands.

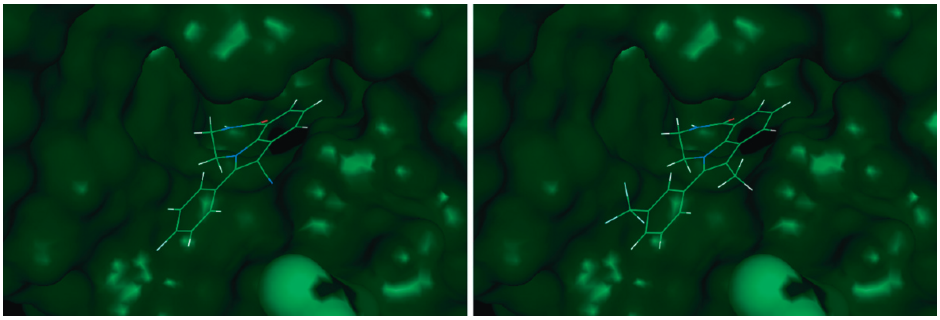


*Figure 8.* Trifluoromethyl-containing ligand (a) and cyano-containing ligand (b) in their docked poses. The ligand atoms are colored by atom name and are shown as sticks. The protein receptor is shown as a surface that is colored by its atomic Fukui indices. Dark green corresponds to lower values of the Fukui function, and bright colored areas are local maxima in the Fukui function. The relatively hard trifluoromethyl group in (a) points away from the binding site, and the relatively soft cyano group of (b) points toward a soft area just outside the binding pocket.

were saved and scored by their Glide XP score, the closest atom pair FRMSD, and the Fukui grid score. For each scoring function, the best scoring pose from the 10 poses was used to rank the ligands. The results are shown in Figure 6.

The Glide XP score predicts ligand 6, pinostrobin chalcone, to be the best binder and ligand 1, pinostrobin (the best binder experimentally), to be the worst. The score fails to predict the correct trend in binding affinity and actually predicts the reverse trend. The FRMSD score also fails to rank the ligands correctly or show any kind of trend in binding affinity. In contrast, the distance-dependent Fukui grid score captures the correct trend in binding affinity. It correctly predicts pinostrobin to be the best binder, and the

***Table 1.*** Results from the Active Site Search Experiment[a]

| receptor (PDB ID) | active site residue no. | percentile rank |
|---|---|---|
| Ampc (1XGJ) | 317 | 93.7 |
| Ar (1XQ2) | 873 | 93.0 |
| Fgfr1 (1AGW) | 512 | 97.5 |
| | 545 | 92.4 |
| | 563 | 91.4 |
| | 567 | 92.1 |
| Fxa (1F0R) | 98 | 92.3 |
| | 220 | 98.7 |

[a] Four of the twelve studied enzymes had residues within 7 Å of the active site with average atomic Fukui indices ranking higher than those of 90% of the total number of residues in the enzyme.

others are separated from it by hard/soft compatibility. Experimentally this is the case: pinostrobin is by far the strongest binder, while the other ligands are grouped together at lower activity. In this case, hardness/softness matching is able to pick out the best binder from a set of ligands for dengue 2 NS3 protease.

A second test of the Fukui grid score was to rank a set of known binders for poly(ADP-ribose) polymerase (PDB ID 1EFY) taken from the DUD database.[54,55] The ligands are all similar in structure, with substitutions made at both the $R_1$ and $R_2$ positions of the molecular scaffold shown in Figure 7a. The grid score was used as it was in the case for 2FOM, except here a cutoff radius is introduced for the protein Fukui indices due to the larger size of 1EFY. The cutoff used was 10 Å. Increasing this cutoff radius changed individual scores for ligands but did not alter the relative rankings of the ligands. Results from the docking of the 32 DUD ligands are shown in Figure 7.

Here the Glide XP score does fairly well at showing a correct trend in experimental binding affinity (Figure 7b). The Fukui grid score shows only small differences in affinity for the different ligands except for three which clearly display better hard/soft contacts with the receptor than the rest of the set (Figure 7c). If those three are removed from the plot, upon closer examination (Figure 7d) the Fukui grid score still shows better hard/soft matches for many of the stronger binders ($K_i$ < 50 nM) than the weaker binders ($K_i$ >50 nM). There were three ligands with $K_i$ less than 60 nM that had relatively poor Fukui grid scores. Two of these ligands had oxime groups participating in hydrogen bonding, and the third had two hydroxyl groups both participating in hydrogen bonding. As stated previously, the Fukui function based score does not include this stabilizing interaction and would need an additional hydrogen-bonding term to be used as a more general scoring function.

Upon examination of two of the three that stood out as the best hard/soft matching poses, it was found that the best pose evaluated by the Fukui grid score was the only ligand in the DUD set with a trifluoromethyl group (more specifically, 3-(trifluoromethyl)phenyl) which was placed in the $R_2$ position. Fluorine is usually a very hard atom, so it can be hypothesized that the active site favors a chemically hard group in the area where the trifluoromethyl group is docked. Another of the three best ligands was the only one containing a cyano group in the $R_1$ position. Cyano groups are usually chemically soft, so the receptor may favor soft species in this position. After visualization of the docked poses of these two ligands, it was discovered that the best poses according to the Fukui grid score had nearly identical binding conformations, with the trifluoromethyl and cyano groups of the two ligands pointing in opposite directions (Figure 8). It would be interesting to test the binding affinity of a ligand containing both of these groups in the two positions to see whether the measured affinity increases due to HSAB preferences. Such a ligand was built and placed in the bound conformation of Figure 8 and then scored by the Fukui grid score. The hypothetical ligand received a score of 0.379, which places it among the best three known ligand molecules in terms of hard/soft compatibility.

**3.2. Active Site Detection.** The Fukui function is generally used to detect favorable sites of interaction between molecules. Maxima in the function are interpreted as areas in a molecule most favorable for changes in electron density. One can hypothesize that, by mapping a Fukui function, one could easily pick out reactive areas in a molecule. In the realm of proteins, these are called active sites. Fukushima et al. have previously studied the link between the locations of active site residues and localized frontier orbitals.[60] Among their results was that, for the 112 enzymes under their study, about 20% of the active site residues had molecular orbitals localized on them that lay within a spread of 10 molecular orbitals around the HOMO−LUMO gap.

To use the Fukui function to find active sites of proteins under the finite difference approximation, it is useful to take

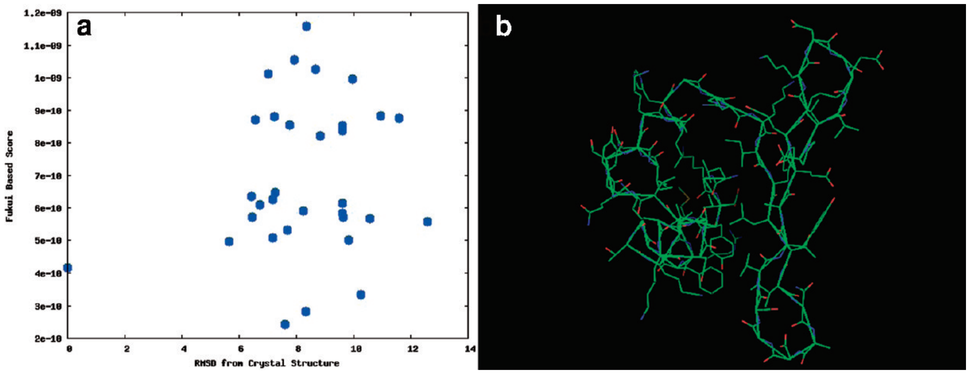


***Figure 9.*** (a) Fukui index based folding score (eq 13) plotted against the rmsd from the native structure for the cro repressor mutant (PDB ID 1ORC) and its decoy folds. The decoy structures have a wide range of scores, and the native structure is among the best scoring folds. (b) Crystal structure of 1ORC.

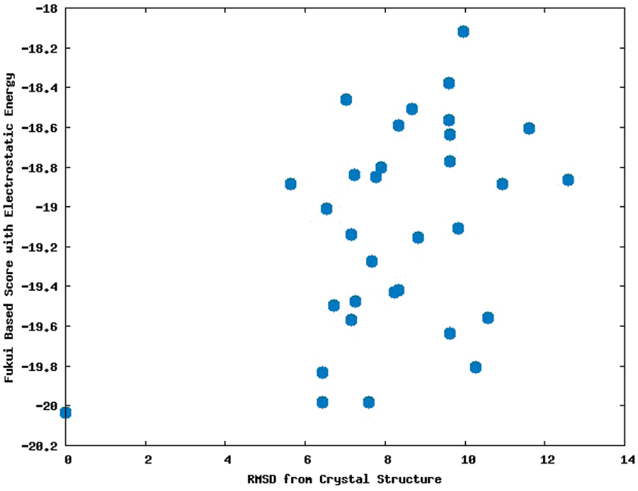


***Figure 10.*** Hybrid score (eq 14) composed of the Fukui function based score (eq 13) and the electrostatic energy score (eq 15) vs the rmsd with respect to the native structure. The native fold scored the best in terms of hard and soft matching, but is not clearly distinguished from the decoy set.

the finite difference derivative by varying the number of electrons by more than 1. Here the electron number was varied by 8. Increasing this number is analogous to increasing the span of MOs searched in Fukushima's study. A calculation was performed on the ground-state system with 8 electrons added and the ground-state system with 8 electrons removed. The centered finite difference Fukui function is then

$$f(r)^0 = \frac{\rho(r,N+8) - \rho(r,N-8)}{2 \times 8} \qquad (12)$$

Twelve receptors taken from the DUD database were used in this study.[54] The experimentally observed bound structures were examined, and all residues within 7 Å of the bound ligand were considered to be part of the active site. The atomic Fukui indices were averaged on each residue to yield one characteristic value of the Fukui function for each amino acid. Each amino acid was then sorted by average Fukui index, and the active site residues were examined by a percentile rank (e.g., 90% means a Fukui score higher than that of 90% of the total number of residues in the enzyme). Among the 12 receptors, 4 of them had active site residues with percentile rankings higher than 90% (Table 1). One of them, fibroblast growth factor receptor 1 (PDB ID 1AGW), had four active site residues ranking higher than 90% of the total number of residues in the protein. This would suggest that polarization of electron density is important for binding to this receptor, which of course is a characteristic that the Fukui function is designed to detect.

If we were to assume these 12 receptors were a randomly chosen collection, then it would make sense that this Fukui function based approach had a success rate similar to that of the Fukushima study. The Fukui function is by definition quite similar to the frontier molecular orbital type analysis used in their earlier work, which is analogous to using the frozen orbital approximation in a Fukui function based approach.

**3.3. Protein Folding.** The third application explored was the detection of native protein folds from a collection of decoy folds. It was hypothesized that better protein folds should have more favorable hard and soft interactions between residues than poorly folded proteins. To test this hypothesis, a distance-dependent score was introduced, in which Fukui indices of atoms in different amino acids are compared. The score can be written as

$$\sum_i^{N_a} \left(\frac{1}{N_n}\right) \times \begin{cases} \sum_j^{N_n} \left(\frac{f_i - f_j}{r_{ij}^{\alpha}}\right) & j \notin i_{\text{res}}, r_{ij} < r_{\max} \\ 0 & \text{otherwise} \end{cases} \qquad (13)$$

where $N_a$ is the total number of atoms and $N_n$ is the number of neighboring atoms from other residues within a cutoff distance $r_{\max}$. This score was used to rank the X-ray structure and decoy structures of a mutated cro repressor (PDB ID 1ORC). The decoy folds were the same set used by He et al. generated with Rosetta.[61,62] The values of $r_{\max}$ and α were optimized and found to make the best predictions at $r_{\max}$ = 10 Å. Varying α did not seem to have a significant impact on the rankings between folds. Figure 9 plots the scores of the X-ray structure and decoys of 1ORC against their rmsd from the X-ray structure. Here $r_{\max}$ = 10 Å and α = 0.2. Only three of the decoys score better than the native fold, and several of the decoys are far separated from the native structure by the Fukui based score.

As shown in Figure 9, although the Fukui function based score gave the native fold one of the best scores, it could not discriminate the native structure from the set of decoys. As discussed in the docking studies, the hardness and softness interactions do not seem to include electrostatic interactions, which are relevant in looking at protein decoys. Therefore, an electrostatic energy term was added to the Fukui function based score to create a hybrid score given by

$$\text{hybrid score} = x(\text{FF score}) + E_{\text{el}}\ \text{score} \qquad (14)$$

where $x$ is a parameter introduced to scale the Fukui function based scores (FF score) to the range of the electrostatic energy scores ($E_{\text{el}}$ score). The electrostatic energy score used was simply

$$E_{\text{el}} = \sum_{i=1}^{N_a} \sum_{j=i+1}^{N_a} \frac{q_i q_j}{r_{ij}} \qquad (15)$$

where $q_i$ is the Mulliken charge of atom $i$ from the AM1 calculation, $N_a$ the number of atoms, and $r_{ij}$ the distance between atoms. An appropriate value for $x$ was found to be around 1.2 × $10^9$. While the hybrid score was able to rank the native structure the best of all folds, it did not offer clearly superior discrimination ability (Figure 10).

The Fukui function based folding score was also tested on NMR structures of the Pin 1 WW domain (PDB ID 1I6C). This system was chosen because the effect of electron correlation (and attractive van der Waals energy) has been shown to be vital in determining its native fold from a set of decoys.[61] Of the possible nonbonded interactions of molecules, chemical softness seems to be the most relevant to these types of interactions. Using the same parameters as

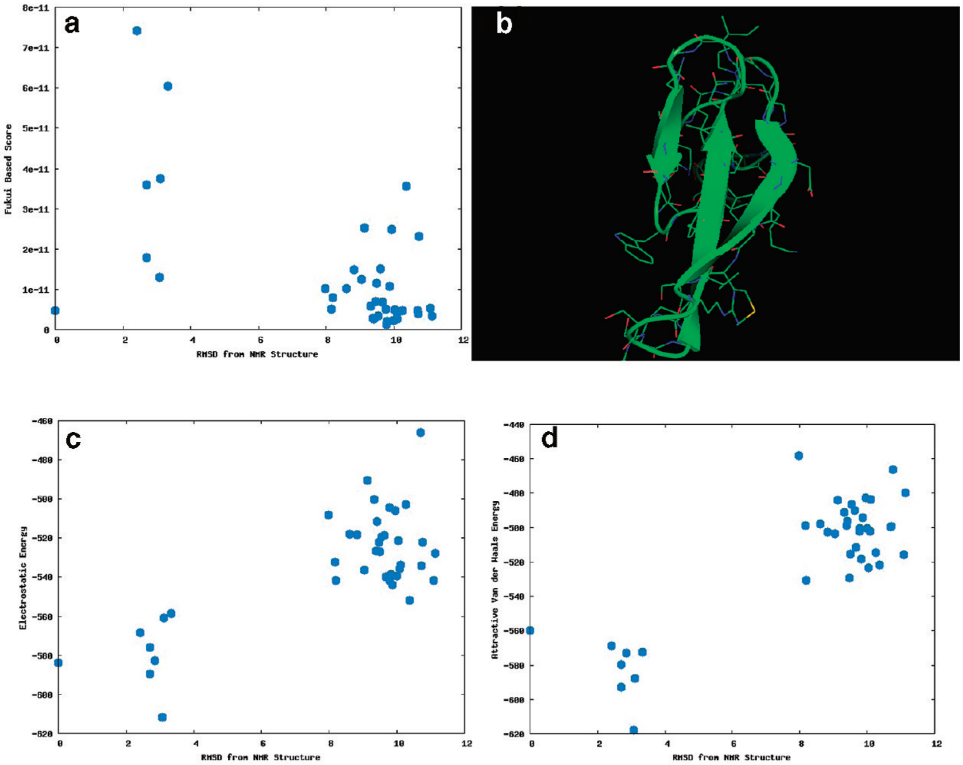


***Figure 11.*** (a) Fukui based folding score for the Pin 1 WW domain (PDB ID 1I6C) vs the rmsd of an NMR structure. The score fails to clearly distinguish the NMR models from the decoys, presumably because the $\beta$-sheet conformation keeps the side chains apart (b). (c) Electronic energy from Amber vs the rmsd from the NMR structure. (d) Attractive van der Waals energy from Amber vs rmsd.

with the 1ORC system, the results of the Fukui based score did not distinguish the NMR structures from the decoys (Figure 11a). Using a hybrid score such as eq 14 would be meaningless in this situation since electronic energy (Figure 11c) or van der Waals energy (Figure 11d) alone do very well in discriminating native from decoy folds. The optimum weighting factor for the Fukui based score would be very close to zero.

One possible source of failure for the Fukui based score would be that the Pin 1 WW domain consists mostly of a $\beta$-sheet formation, which is held together by a chain of hydrogen bonds and points side chains outward away from one another. In this kind of conformation, there is little chance for side chains to interact in a hard/soft type of interaction. In contrast, 1ORC (Figure 9b) has side chains pointed inward near each other. This suggests that any kind of hard/soft scoring for protein folding would most likely be only useful for specific types of protein folds with well-defined cores.

## 4. Conclusions

The utility of the Fukui function was explored in three significant problems in computational biochemistry: docking, active site detection, and protein folding. We hypothesized that hard/soft acid−base matching concepts would allow us to gain new insights into these problems. To make use of the Fukui function for large protein systems, several approximations were used including the AM1 Hamiltonian, the divide and conquer algorithm, Mulliken charges, and the finite difference derivative. Even with these approximations Fukui based scoring functions correctly determined the binding conformation of a ligand in an active site, distinguished between active binders and nonbinders for a receptor, determined the best binders from a set of known binders, detected possible active sites, and ranked an observed protein fold among the best of a set of native and decoy folds.

It was observed that not all types of molecular interactions are captured by Fukui based scoring functions and that additional terms (such as an electrostatic term) are necessary to make them more broadly applicable. It was also observed that strictly using atomic indices is not always as effective as approximating the full Fukui function by adding distance dependence, especially in the case of docking several different ligands to a binding site. A clear advantage of these types of analyses is that molecular surfaces can easily be visualized and colored by hardness or softness, aiding the chemist in deciding which parts of two molecules interact favorably or unfavorably from a hardness/softness perspective. The concepts presented herein offer new descriptors that can be used in QSAR studies and present alternative ways to examine biological problems such as protein−ligand interactions.

**Acknowledgment.** We thank the NIH for generously supporting this research via Grant GM044974.

## References

(1) Szabo, A.; Ostlund, N. S. The Hartree Fock Approximation. *Modern Quantum Chemistry: Introduction to Advanced*

*Electronic Structure Theory*, 1st ed.; Dover Publications: Mineola, NY, 1996; Vol. 1, pp 108−151.

(2) Levine, I. N. Ab Initio and Density Functional Treatments of Molecules. *Quantum Chemistry*, 6th ed.; Prentice Hall: Upper Saddle River, NJ, 2008; Vol. 1, pp 480−625.

(3) Jensen, F. Electronic Structure Methods: Independent-Particle Models. *Introduction to Computational Chemistry*, 2nd ed.; Wiley: Chichester, England, 2006; Vol. 1, pp 80−132.

(4) Cramer, C. J. Molecular Mechanics. *Essentials of Computational Chemistry: Theories and Models*, 2nd ed.; Wiley: Chichester, England, 2004; Vol. 1, pp 17−68.

(5) Leach, A. Emperical Force Field Models: Molecular Mechanics. *Molecular Modelling: Principles and Applications*, 2nd ed.; Prentice Hall: Harlow, England, 2001; Vol. 1, pp 165−252.

(6) Geerlings, P.; De Proft, F.; Langenaeker, W. *Chem. Rev.* **2003**, *103*, 1793.

(7) Sengupta, D.; Chandra, A. K.; Nguyen, M. T. *J. Org. Chem.* **1997**, *62*, 6404.

(8) Roy, R. K.; Krishnamurti, S.; Geerlings, P.; Pal, S. *J. Phys. Chem. A* **1998**, *102*, 3746.

(9) Roy, R. K.; Tajima, N.; Hirao, K. *J. Phys. Chem. A* **2001**, *105*, 2117.

(10) Feng, X.-T.; Yu, J.-G.; Lei, M.; Fang, W.-H.; Liu, S. *J. Phys. Chem. B* **2009**, *113*, 13381.

(11) Beck, M. E. *J. Chem. Inf. Model.* **2005**, *45*, 273.

(12) Roos, G.; Geerlings, P.; Messens, J. *J. Phys. Chem. B* **2009**, *113*, 13465.

(13) Giessner, C.; Pullman, A. *Theor. Chim. Acta* **1972**, *25*, 83.

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

(15) Goedecker, S. *Rev. Mod. Phys.* **1999**, *71*, 1085.

(16) Galli, G. *Phys. Status Solidi B* **2000**, *217*, 231.

(17) Dixon, S. L.; Merz, K. M. *J. Chem. Phys.* **1996**, *104*, 6643.

(18) Dixon, S. L.; Merz, K. M. *J. Chem. Phys.* **1997**, *107*, 879.

(19) Van der Vaart, A.; Gogonea, V.; Dixon, S. L.; Merz, K. M. *J. Comput. Chem.* **2000**, *21*, 1494.

(20) Yang, W. T.; Lee, T. S. *J. Chem. Phys.* **1995**, *103*, 5674.

(21) Lee, T. S.; Lewis, J. P.; Yang, W. T. *Comput. Mater. Sci.* **1998**, *12*, 259.

(22) Khandogin, J.; York, D. M. *Proteins* **2004**, *56*, 724.

(23) Pearson, R. G. *J. Am. Chem. Soc.* **1963**, *85*, 3533.

(24) Datta, D.; Singh, S. N. *J. Chem. Soc., Dalton Trans.* **1991**, 1541.

(25) Datta, D. *J. Chem. Soc., Dalton Trans.* **1992**, 1855.

(26) Benedetti, L.; Gavioli, G. B.; Fontanesi, C. *J. Chem. Soc., Faraday Trans.* **1992**, *88*, 843.

(27) Langenaeker, W.; Coussement, N.; Deproft, F.; Geerlings, P. *J. Phys. Chem.* **1994**, *98*, 3010.

(28) Deproft, F.; Amira, S.; Choho, K.; Geerlings, P. *J. Phys. Chem.* **1994**, *98*, 5227.

(29) Deka, R. C.; Vetrivel, R.; Pal, S. *J. Phys. Chem. A* **1999**, *103*, 5978.

(30) Mondal, P.; Hazarika, K. K.; Deka, R. C. *PhysChemComm* **2003**, 24.

(31) Flores-Sandoval, C. A.; Zaragoza, I. P.; Maranon-Ruiz, V. F.; Correa-Basurto, J.; Trujillo-Ferrara, J. *J. Mol. Struct.: THEOCHEM* **2005**, *713*, 127.

(32) Wisniewski, M.; Gauden, P. A. *Appl. Surf. Sci.* **2009**, *255*, 4782.

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

(34) Parr, R. G.; Yang, W. T. *J. Am. Chem. Soc.* **1984**, *106*, 4049.

(35) Yang, W. T.; Parr, R. G. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 6723.

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

(37) Ayers, P. W.; Parr, R. G. *J. Chem. Phys.* **2008**, *128*, 184108.

(38) Chattaraj, P. K.; Roy, D. R.; Geerlings, P.; Torrent-Sucarrat, M. *Theor. Chem. Acc.* **2007**, *118*, 923.

(39) Ayers, P. W.; De Proft, F.; Borgoo, A.; Geerlings, P. *J. Chem. Phys.* **2007**, *126*, 224107.

(40) Fievez, T.; Sablon, N.; De Proft, F.; Ayers, P. W.; Geerlings, P. *J. Chem. Theory Comput.* **2008**, *4*, 1065.

(41) Chandra, A. K.; Nguyen, M. T. *J. Chem. Soc., Faraday Discuss.* **2007**, *135*, 191.

(42) Balawender, R.; Komorowski, L. *J. Chem. Phys.* **1998**, *109*, 5203.

(43) Mineva, T.; Russo, N.; Sicilia, E.; Toscano, M. *Theor. Chem. Acc.* **1999**, *101*, 388.

(44) Mineva, T.; Parvanov, V.; Petrov, I.; Neshev, N.; Russo, N. *J. Phys. Chem. A* **2001**, *105*, 1959.

(45) Madjarova, G.; Tadjer, A.; Cholakova, T. P.; Dobrev, A. A.; Mineva, T. *J. Phys. Chem. A* **2005**, *109*, 387.

(46) Ayers, P. W. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3387.

(47) Melin, J.; Ayers, P. W.; Ortiz, J. V. *J. Phys. Chem. A* **2007**, *111*, 10017.

(48) Vincent, J. J.; Dixon, S. L.; Merz, K. M. *Theor. Chem. Acc.* **1998**, *99*, 220.

(49) DeLano, W. L.; Lam, J. W. *Abstr. Pap.—Am. Chem. Soc.* **2005**, *230*, U1371.

(50) Sich, C.; Improta, S.; Cowley, D. J.; Guenet, C.; Merly, J. P.; Teufel, M.; Saudek, V. *Eur. J. Biochem.* **2000**, *267*, 5342.

(51) Erbel, P.; Schiering, N.; D'Arcy, A.; Renatus, M.; Kroemer, M.; Lim, S. P.; Yin, Z.; Keller, T. H.; Vasudevan, S. G.; Hommel, U. *Nat. Struct. Mol. Biol.* **2006**, *13*, 372.

(52) Kiat, T. S.; Pippen, R.; Yusof, R.; Ibrahim, H.; Norzulaani, K.; Rahman, N. A. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 3337.

(53) Othman, R.; Kiat, T. S.; Khalid, N.; Yusof, R.; Newhouse, E. I.; Newhouse, J. S.; Alam, M.; Rahman, N. A. *J. Chem. Inf. Model.* **2008**, *48*, 1582.

(54) Huang, N.; Shoichet, B. K.; Irwin, J. J. *J. Med. Chem.* **2006**, *49*, 6789.

(55) Tikhe, J. G.; Webber, S. E.; Hostomsky, Z.; Maegley, K. A.; Ekkers, A.; Li, J. K.; Yu, X. H.; Almassy, R. J.; Kumpf, R. A.; Boritzki, T. J.; Zhang, C.; Calabrese, C. R.; Curtin, N. J.; Kyle, S.; Thomas, H. D.; Weng, L. Z.; Calvert, A. H.; Golding, B. T.; Griffin, R. J.; Newell, D. R. *J. Med. Chem.* **2004**, *47*, 5467.

(56) Halgren, T. A.; Murphy, R. B.; Banks, J.; Mainz, D.; Klicic, J.; Perty, J. K.; Friesner, R. A. *Abstr. Pap.—Am. Chem. Soc.* **2002**, *224*, U345.

(57) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. *J. Med. Chem.* **2006**, *49*, 6177.

(58) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639.

(59) Wang, B.; Westerhoff, L. M.; Merz, K. M. *J. Med. Chem.* **2007**, *50*, 5128.

(60) Fukushima, K.; Wada, M.; Sakurai, M. *Proteins* **2008**, *71*, 1940.

(61) He, X.; Fusti-Molnar, L.; Cui, G. L.; Merz, K. M. *J. Phys. Chem. B* **2009**, *113*, 5290.

(62) Bonneau, R.; Strauss, C. E. M.; Rohl, C. A.; Chivian, D.; Bradley, P.; Malmstrom, L.; Robertson, T.; Baker, D. *J. Mol. Biol.* **2002**, *322*, 65.

CT9005085

JCTC Journal of Chemical Theory and Computation

# Coupling Constant pH Molecular Dynamics with Accelerated Molecular Dynamics

Sarah L. Williams,*,† César Augusto F. de Oliveira,†,§ and J. Andrew McCammon†,‡,§,||

*Department of Chemistry & Biochemistry, University of California San Diego, La Jolla, California 92093-0365, Center for Theoretical Biological Physics, University of California San Diego, La Jolla, California 92093, Howard Hughes Medical Institute, University of California San Diego, La Jolla, California 92093-0365, Department of Pharmacology, University of California San Diego, La Jolla, California 92093-0365*

Received October 6, 2009

**Abstract:** An extension of the constant pH method originally implemented by Mongan et al. (*J. Comput. Chem.* **2004**, *25*, 2038−2048) is proposed in this study. This adapted version of the method couples the constant pH methodology with the enhanced sampling technique of accelerated molecular dynamics, in an attempt to overcome the sampling issues encountered with current standard constant pH molecular dynamics methods. Although good results were reported by Mongan et al. on application of the standard method to the hen egg-white lysozyme (HEWL) system, residues which possess strong interactions with neighboring groups tend to converge slowly, resulting in the reported inconsistencies for predicted p$K_a$ values, as highlighted by the authors. The application of the coupled method described in this study to the HEWL system displays improvements over the standard version of the method, with the improved sampling leading to faster convergence and producing p$K_a$ values in closer agreement to those obtained experimentally for the more slowly converging residues.

## Introduction

It is well-known that the structure and function of a protein are highly dependent on the pH of its surrounding aqueous environment due to pH-mediated changes in the protonation state of titratable residues. The protonation state of a titratable residue in a protein is determined by its p$K_a$ and the solution pH, the former being a measure of the relative acidity of the residue, which is influenced by interactions with neighboring residues, including titratable residues. These changes in protonation equilibrium, which are essentially of electrostatic nature, are closely linked with the conformation and are fundamental in the definition of the often-narrow pH range for the functioning protein, beyond which unfavorable conformational change and denaturation of the protein structure may occur.

The important pairing of protonation state and protein conformation is not accounted for in standard molecular dynamics (MD) simulations. Currently, the majority of standard simulations of biological systems use fixed, predetermined protonation states for titratable residues, which are generally based on the p$K_a$ values of the isolated residue in solution. In addition, protonation states are usually assigned during the preparation of the system and are not changed throughout the standard MD simulation. This method of protonation state assignment is a severe approximation, as the p$K_a$ values of titratable residues are frequently shifted from that of the model residue in solution, making the assignment a nontrivial task. Furthermore, protonation states are not single constant values; they are subject to the changing electrostatic environment surrounding the titratable group. Therefore, incorporating pH as an input variable in MD simulations is highly desirable, as it would allow a more

* Corresponding author phone: 858-822-0168; fax: 858-534-4974; e-mail: swilliam@mccammon.ucsd.edu.

† Department of Chemistry & Biochemistry.

§ Howard Hughes Medical Institute.

‡ Center for Theoretical Biological Physics.

|| Department of Pharmacology.

10.1021/ct9005294 © 2010 American Chemical Society
Published on Web 01/14/2010

accurate study of pH-coupled conformational phenomena, such as reaction mechanisms, ligand binding, and the determination of the structure and function of proteins as a function of pH.

Over approximately the past 15 years, several methods have been proposed which enable MD to be carried out at a constant pH with changing protonation states. These constant-pH MD (CpHMD) methods can be largely classified into two categories, discrete[1-4] and continuous.[5-7] Several reviews have been published which compare and contrast the different methods.[8-10] In the following paragraphs, a brief description of some of these methods is given.

Continuous protonation state models, such as that of Börjesson and Hünenberger[6,7] and Baptista et al.,[5] consider protonation state as a continuous titration parameter, which advances simultaneously with the atomic coordinates of the system. However, these methods use a mean-field approximation, whereby they do not take into account any interaction with other nearby titratable residues that may occur, and the titratable groups are represented by fractional, nonphysical protonation states, intermediate between the protonated and unprotonated forms.[11,12] These factors cause the models to perform poorly for tightly coupled residues[7] and result in inadequate estimation of physical observables. The more recent work of Lee et al.[13] overcomes the issues with unphysical fractional protonation states with the use of $\lambda$ dynamics with the addition of an artificial titration barrier along the continuous titration coordinate between the fully protonated and deprotonated end points. This has the effect of forcibly lengthening simulation time in the fully protonated or deprotonated values. The authors report good correlation between the predicted and experimental p$K_a$ values for the hen egg-white lysozyme (HEWL), turkey ovomucoid, and bovine pancreatic trypsin inhibitor, although convergence issues were encountered for these systems, and even for the simpler aspartate model. Khandogin and Brooks[14] developed an extension to this method, a two-dimensional $\lambda$-dynamics method using GBSW[15] solvation. The two dimensions are the two reaction coordinates: the deprotonation process and the interconversion between proton tautomers, to account for proton tautomerism in simulations involving histidine and carboxyl residues. The authors observe significant quantitative improvement over the previous work of Lee et al.[13] and note that the method could be further improved with enhanced sampling and an improved solvent model. In other work by the same group,[16] the continuous titration method is coupled with replica exchange (coupled method known as REX-CPHMD), used with an improved GB solvent model[17] and a salt-screening function, to achieve more accurate predictions of p$K_a$ shifts when applied to 10 test protein systems, all possessing residues with significant p$K_a$ shifts.

The majority of the more recent studies have involved the use of discrete protonation state models, which avoid the nonphysical intermediate charge states. These methods use MD simulations for conformational sampling, with periodic sampling of discrete protonation states through trial Monte Carlo (MC) moves. The main differences between these methods lie in their choice of solvent model and the protocol for updating the protonation states.[1-4] The methods employing explicit solvent are computationally expensive, and MC trial moves are attempted relatively infrequently, causing long convergence times for systems with multiple titration sites. Both Bürgi et al.[18] and Baptista[1] et al. developed methods using explicit solvent. Baptista et al. used Poisson−Boltzmann (PB) electrostatics for the calculation of protonation state energies to be used for the MC test. However, the PB calculations are time-consuming and introduce a solvent potential different from that used for the explicit-solvent dynamics. Bürgi et al. avoid the discrepancy in the potentials with the use of thermodynamic integration (TI) under the same explicit solvent conditions as used for the dynamics, to determine the transition energies for the MC test. However, this has the effect of perturbing the trajectory, even when the MC trial is rejected, since the final trajectory is formed from the concatenation of the TI segments. In addition, the length of time over which the TI calculations are performed (~20 ps) makes their significance uncertain, and the expense of explicit solvent is a probable contributor to the apparent poor convergence of the simulations.[8] Stern introduces a method whereby the issues associated with instantaneous protonation state change when using explicit solvent are circumvented, with the use of a hybrid Monte Carlo procedure.[19] In this method, the trial moves comprise relatively short MD trajectories, which employ a time-dependent potential energy that interpolates between the old and new protonation states. The method has been successfully applied to acetic acid in water but has not yet been applied to protein systems.

Methods employing implicit solvent for both the dynamics and MC steps include the work of Dlugosz and Antosiewicz,[2,3] who use PB calculations in the calculation of transition energies, and the analytical continuum electrostatics (ACE/GB) model of CHARMM for dynamics. Again, this method has the problems associated with the expense of PB calculations and the mismatch in potentials used, although the method reports fair agreement with experiment when applied to a heptapeptide derived from the ovomucoid third domain and succinic acid. Mongan et al. use GB solvation for both the MC steps and the dynamics,[4] therefore avoiding the discrepancy in the potentials used, with p$K_a$ predictions agreeing well with experimental results on application to the HEWL system, although convergence issues are noted for some residues of the system.

In this work, we propose an extension to the constant pH model of Mongan et al., whereby the methodology is coupled with the enhanced sampling technique of accelerated molecular dynamics (aMD) to increase the sampling and alleviate the reported convergence issues. This version of the method has been implemented in AMBER8 and has been applied to the popular test case, the HEWL system. The results show improvement in the sampling compared with the standard Mongan et al. method, with p$K_a$ results close to those obtained experimentally for the more problematic, more slowly converging residues.

## Background

The standard constant pH (Mongan et al.) and aMD methods (de Oliveria et al.) coupled in this study are described in detail elsewhere, and so only outlines of the techniques are given here. The CpHMD method described here differs from the standard method in that the sections of conventional MD are replaced with the enhanced sampling technique, aMD. The combined method is denoted as CpHaMD in this work.

The method employs GB-solvated aMD with periodic MC sampling of protonation states, also using the same GB electrostatics. At each MC step, a titratable residue and a new protonation state are chosen at random, with the total transition energy, $\Delta G$, being used as the Metropolis criterion for the decision of protonation state. The calculation of $\Delta G$ is shown in eq 1, where $k_B$ is the Boltzmann constant, $T$ is the temperature, pH is the specified solvent pH, $pK_{a,ref}$ is the $pK_a$ of the reference compound, $\Delta G_{elec}$ is the electrostatic energy component of the titratable residue, and $\Delta G_{elec,ref}$ is the electrostatic component of the transition energy for the reference compound. If the MC move is accepted, the protonation state of the residue will change to the new state and MD is continued; if not, the simulation will continue with the residue remaining in the unchanged protonation state.

$$\Delta G = k_B T(\mathrm{pH} - \mathrm{pK}_{a,\mathrm{ref}})\ln 10 + \Delta G_{\mathrm{elec}} - \Delta G_{\mathrm{elec,ref}} \tag{1}$$

In the previous implementation of the method, conventional MD was employed between the MC steps. In the version of the method reported in this work, standard MD is replaced with aMD. As mentioned previously, a limitation of constant pH methods is often convergence,[8,16] therefore implying the performance of the method may be improved by the use of enhanced sampling techniques, as shown by Khandogin and Brooks with their REX-CPHMD method.[16] Here, a recently modified version of the dual-boost aMD method (referred to as aMDtT[b] in the literature) by de Oliveira et al. is used (a modification of the Hamelberg et al. aMD methodology[20]), which has been found to be useful in improving the accuracy and convergence of TI simulations. This approach increases conformational sampling through the modification of the energy landscape by lowering energy barriers while leaving the potential surface in the vicinity of the minima unchanged. The energy barriers are reduced through the application of a boost potential, $\Delta V(r)$, to the true potential surface, $V(r)$, in cases where the true potential exceeds a predefined energy level, $E$. The boost potential is implemented in the method according to eq 2, where $\alpha$ modulates the shape of the modified potential (lower values of $\alpha$ result in a flatter modified potential, and higher values approach the unmodified potential).

$$\Delta V(r) = \begin{cases} \dfrac{(V(r) - E)^2}{\alpha + (V(r) - E)}, & V(r) \geq E \\ 0, & V(r) < E \end{cases} \tag{2}$$

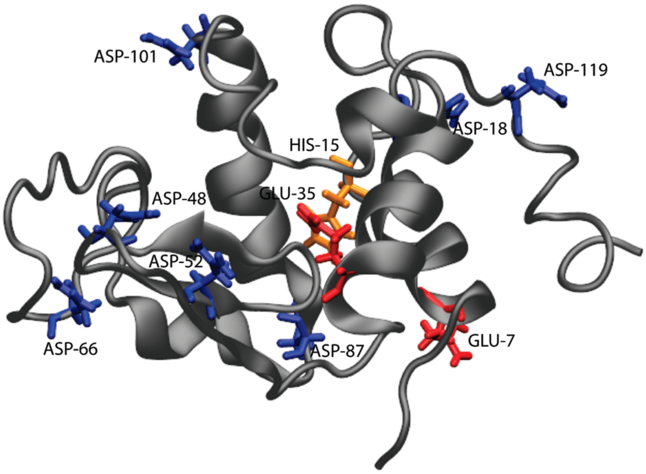


*Figure 1.* The HEWL enzyme (PDB ID: 1AKI) with titratable groups highlighted in liquorice style (aspartates, blue; glutamates, red; and histidine, orange).

In cases where the true potential exceeds the boost energy level $E$, the boost potential is subtracted from the true potential, and the simulation is performed on this modified potential surface $V^*(r) = V(r) - \Delta V(r)$. At times where the true potential lies below the boost energy level, $E$, the simulation is performed on the true potential, $V^*(r) = V(r)$.

In this work, the dual-boost approach[21] has been used in order to increase the sampling of both the torsional degrees of freedom and the atomic packing arrangements. The first boost is applied to only the torsional terms, $V_t(r)$, and the second boost is added to the total potential energy, $V_T(r) = V_0(r) + V_t(r)$ (eq 3).

$$V^*(r) = \{V_0(r) + [V_t(r) - \Delta V_t(r)]\} - \Delta V_T(r) \tag{3}$$

The correct canonical averages of an observable, in this case $pK_a$, are calculated from configurations sampled on the modified potential energy surface and are fully recoverable by reweighting each point in configuration space by eq 4.

$$\exp\{-\beta[\Delta V_t(r) + \Delta V_T(r)]\} \tag{4}$$

## Test Case: Hen Egg White Lysozyme

HEWL is a 129-residue monomeric single-domain enzyme which catalyzes the hydrolysis of polysaccharides found in many bacterial cell walls (see Figure 1). The enzyme is known to possess several residues with $pK_a$ values significantly shifted from the model isolated residue values.[22,23] Additionally, it is a well-known example of an enzyme which employs a proton donor and a catalytic nucleophile (Asp 52 and Glu 35)[24] within the active site, located in a cleft between an all-$\alpha$ and a $\beta$-rich region. Owing to extensive experimental study of this system, and the challenging nature of the $pK_a$ shifts of some of the residues, HEWL has been a popular test system for many of the $pK_a$ calculation methods. In this study, focus is placed on the acidic residues of this enzyme, which have been experimentally determined to possess the most significant $pK_a$ shifts of the system.

## Molecular Dynamics Simulations

The standard CpHMD and coupled CpHaMD methods have been implemented in the AMBER 8 molecular dynamics

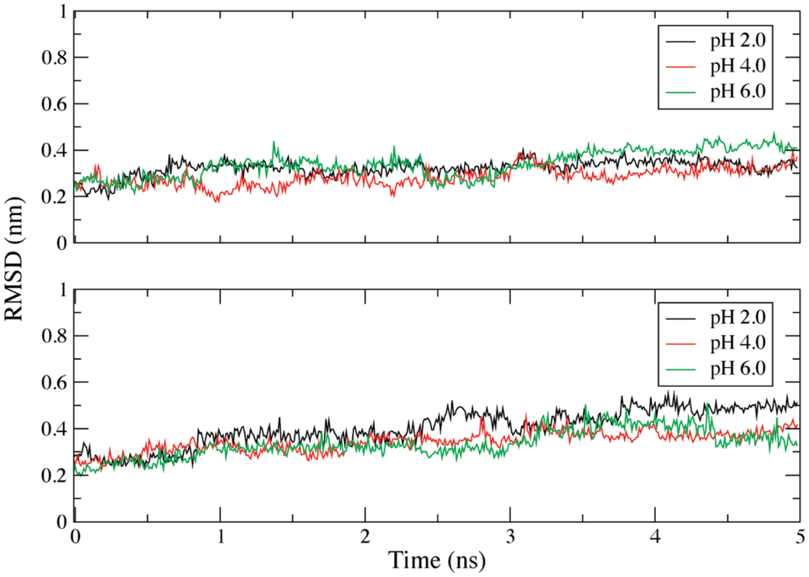


***Figure 2.*** The RMSD of CA atoms with respect to the crystal structure, over the duration of 5 ns simulations carried out at pH 2.0, 4.0, and 6.0 using CPH-aMD (lower plot) and CpHMD (upper plot).

program. This study follows from the work of Mongan et al., and the parameters used match those used in their work. All simulations described employed the AMBER99 forcefield[25] and the GB solvent model[26−28] (igb=2). Salt concentrations were set at 0.1 M, and a 30 Å cutoff value for nonbonded interactions and effective Bond radii calculations was used. The SHAKE algorithm was used to constrain all bonds involving hydrogen, allowing a time step of 2 fs to be used. The temperature was maintained at 300 K using the Berendsen temperature coupling method with a time constant of 2 ps. A period of 10 fs of MD or aMD separated the MC trials.

For the HEWL system, values for the boost energy, $E$, applied to the torsional degrees of freedom and the total potential energy were estimated on the basis of the average torsional and total potential energies and the root mean square (RMS) deviation in these energies over CpHMD simulations carried out on the unmodified potential at the pH of interest. The parameter, $E$, was calculated from subtracting the sum of twice the RMS deviation from the average potential and torsional energies. The value of the $\alpha$ parameter for the total potential energy was estimated to be ~5 kcal/atom, and for the torsional potential, a value of ~30% of the average dihedral potential energy, obtained from the simulation carried out on the unmodified potential, was found to be efficient.

All simulations were started from the minimized 1AKI (PDB ID) crystal structure, as prepared by Mongan et al. (details given in ref 4). Simulations of 5 ns in length were carried out in the pH range 2−6.5 at 0.5 pH unit intervals, using both CpHMD and CpHaMD methods. GLU and ASP residues were set to titrate from pH 2.0 to pH 6.5, with the addition of HIS residues from pH 4.5 to pH 6.5. HIS residues were not set to titrate for the most acidic simulations, as they are likely to remain in the protonated state in this pH range, as indicated by its model p$K_a$ value of 6−7.[29] Models for the terminal residues have not been developed yet for this system, so these residues were set to their neutral protonation states. This approximation has been deemed to be sufficient for these simulations, as explained in the prior work on this system, by Mongan et al. All nontitrating residues were set to their expected protonation states.

**Extended Simulations.** To further investigate the effects of CpHaMD, simulations at pH 3 and pH 6.5 were extended to 40 ns in triplicate, the further two simulations initialized from different random seeds, and generated from re-equilibration of the minimized structure. The pH values were chosen as they represented two different regions of the acidic pH range, pH 3 being close to the experimental p$K_a$ values for the majority of the residues and pH 6.5 being more challenging in obtaining convergence and accurate p$K_a$ evaluations due to the many residues of interest being in the deprotonated state.

**Principal Component Analysis (PCA).** Details of PCA can be found elsewhere in the literature.[30,31] The GROMACS analysis program,[32] g_covar, was used for the calculation and diagonalization of the covariance matrix, with the analysis of the resultant eigenvectors performed using g_anaeig. The covariance matrix of positional fluctuation was calculated for atoms of the residue of interest and atoms in the vicinity, within a distance 7.5 Å, from the 12 concatenated trajectories of 40 ns (CpHMD: pH 3.0, three simulations; pH 6.5, three simulations; CpHaMD: pH 3.0, three simulations; pH 6.5, three simulations). The two-dimensional plots were generated from the projection of the trajectories onto the first two eigenvectors.

## Results

**Simulation Stability.** Initially, simulations of 5 ns in length were performed (as described in the Molecular Dynamics Simulations section). Figure 2 monitors the root-mean squared deviation (RMSD) of the Cα atoms, with respect to the crystal structure, over the duration of the 5 ns simulations at pH 2.0, 4.0, and 6.0. Simulations employing the standard CpHMD and the adapted CpHaMD method are

***Table 1.*** pKa Predictions of Titratable Residues of the HEWL Enzyme over the Acidic pH Range[a]

| | simulation pH | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| residue | pH 2.0 | pH 2.5 | pH 3.0 | pH 3.5 | pH 4.0 | pH 4.5 | pH 5.0 | pH 5.5 | pH 6.0 | pH 6.5 | av. | exptl. value |
| ASP-18 | 2.62 | 2.92 | 2.47 | 2.24 | 2.41 | 2.34 | 2.58 | 2.38 | – | – | 2.50 | 2.66 |
| | (2.50) | (2.99) | (2.53) | (1.69) | (2.07) | (2.27) | (2.15) | (2.78) | (2.26) | (–) | (2.36) | |
| ASP-48 | 2.70 | **0.35** | 2.83 | 2.38 | 3.34 | 1.98 | 1.96 | 3.06 | **1.32** | 2.01 | 2.19 | 2.50 |
| | (2.39) | (2.23) | (2.70) | (2.82) | (2.86) | (3.24) | (**0.30**) | (3.42) | (2.66) | (1.80) | (2.44) | |
| ASP-52 | **1.99** | **1.17** | **2.36** | **2.17** | **2.14** | **−0.1** | **1.34** | 3.27 | 3.37 | **4.71** | **2.24** | 3.68 |
| | (**2.18**) | (**2.45**) | (**2.52**) | (**1.33**) | (**1.72**) | (**2.45**) | (**2.25**) | (2.87) | (–) | (**2.57**) | (**2.26**) | |
| ASP-66 | 2.71 | **3.28** | 2.47 | 2.15 | **3.12** | 2.14 | 2.62 | 2.31 | 2.52 | 2.93 | 2.63 | 2.00 |
| | (2.89) | (2.79) | (2.87) | (2.59) | (**3.44**) | (**3.43**) | (**1.79**) | (–) | (2.14) | (2.50) | (2.72) | |
| ASP-87 | 2.51 | 2.25 | 2.38 | 1.76 | 2.25 | 2.69 | 2.43 | 2.85 | **3.29** | 3.21 | 2.56 | 2.07 |
| | (2.42) | (2.41) | (2.91) | (2.80) | (2.70) | (1.84) | (3.00) | (2.21) | (3.23) | (**3.18**) | (2.67) | |
| ASP-101 | 3.46 | 3.42 | **2.82** | 3.17 | 3.93 | **2.88** | 3.45 | 3.73 | – | 3.63 | 3.50 | 4.09 |
| | (3.19) | (3.29) | (**2.39**) | (3.12) | (**2.59**) | (3.24) | (3.48) | (3.96) | (3.33) | (3.91) | (3.25) | |
| ASP-119 | **2.07** | 3.55 | 2.25 | 2.52 | 2.50 | 2.35 | 2.64 | **1.92** | 2.36 | **1.21** | 2.34 | 3.20 |
| | (2.73) | (**2.08**) | (2.21) | (2.90) | (2.36) | (2.45) | (2.17) | (3.06) | (**2.00**) | (2.40) | (2.44) | |
| GLU-7 | 3.62 | 3.77 | 3.61 | 3.66 | 3.70 | 3.67 | 3.77 | 3.81 | 3.85 | 3.56 | 3.70 | 2.85 |
| | (3.64) | (3.53) | (3.73) | (3.63) | (3.70) | (3.60) | (3.80) | (3.68) | (4.11) | (4.12) | (3.75) | |
| GLU-35 | 5.51 | 5.76 | 6.06 | 5.61 | **5.02** | 6.22 | **4.92** | 5.13 | **4.72** | **4.54** | 5.35 | 6.20 |
| | (**4.65**) | (**4.75**) | (**4.76**) | (5.79) | (**4.17**) | (6.33) | (5.51) | (**3.05**) | (5.91) | (5.46) | (**5.04**) | |
| HIS-15 | NM | NM | NM | NM | NM | **3.94** | 5.20 | 5.48 | 6.52 | **7.25** | 5.68 | 5.71 |
| | | | | | | (**4.09**) | (5.47) | (4.85) | (**7.28**) | (**7.45**) | (5.83) | |

[a] Results generated using the standard constant-pH methodology (lower values) and using the aMD-modified approach (upper values). Average values (av.) were calculated for each residue from 5 ns simulations performed at the indicated pH values. The pKa of HIS-15 was not measured (NM) at lower pH values. Where a value is missing (−), the pKa of that residue was unable to be measured owing to zero transitions in protonation state occurring over the duration of the simulation. Values highlighted in bold are >1 pKa unit from the experimentally reported range.[34]

shown to be reasonably stable, with no major domain motion over the 5 ns. This is in agreement with experimental evidence; the HEWL enzyme has been experimentally reported to be stable over a wide range of pH values, including a pH stability screen carried out in the range of pH 3−8 which revealed HEWL to be very stable at pH 4, 5, and 8.[33]

**p$K_a$ Predictions Calculated from 5 ns Simulations.** A summary of p$K_a$ predictions for residues titratable over the acidic pH range, calculated from the set of 5 ns simulations performed in this study, is shown in Table 1. At each pH value, the predicted p$K_a$ was calculated according to the Henderson−Hasselbach equation, with the ratio of time that a titratable group spends in the protonated and deprotonated states used as a ratio of concentrations. For CpHaMD simulations, each state is reweighted by eq 4, before the ratio of concentrations is calculated. For comparison with experimental p$K_a$ values, a composite p$K_a$ value for each residue was obtained from the combination and averaging of the individual p$K_a$ values generated at the various pH conditions by the CpHaMD and CpHMD simulations.

Over the pH range studied, the simulations predict p$K_a$ values of titratable residues which correlate well with experimental values (good correlation is deemed as <1 p$K_a$ unit deviation from the experimental range given in ref 34). However, both methodologies predict p$K_a$ values which deviate more than 1 p$K_a$ unit from the experimental results for several residues (highlighted in bold in Table 1), and significant variation is observed for some residues between predictions made for the same residue at different pH values (for example, the calculated p$K_a$ for ASP-52 at pH 4.5 is −0.1, and that at pH 5.5 is 3.27), indicating a lack of convergence for these residues.

Mongan et al.[4] employed the constant-pH method for a set of 1 ns simulations of the HEWL system and reported p$K_a$ predictions for a range of pH values. As observed in this study, they also obtain good overall calculated p$K_a$ values with inconsistency in predictions obtained from different simulations for some of the more strongly interacting titratable groups. They suggest one limitation of the method to be its inability to sufficiently sample conformational space, as, due to the dependence of instantaneous p$K_a$ on conformation, limited conformational sampling would restrict the accuracy of p$K_a$ prediction, especially for the more slowly converging residues of the system (e.g., the more buried Asp-52 and Glu-35 residues). In this study, measurement of the calculated p$K_a$ over the duration of the simulations reported here (Figure 3) shows that 5 ns is still an insufficient simulation time to observe convergence for all residues, even for simulations using the enhanced sampling methodology.

The most problematic case across the pH range is shown to be ASP-52, one of the catalytically crucial residues, residing within the active site (see Figure 1). Residues located within the protein experience a very different electrostatic environment from the isolated model residue, resulting in significant shifts in p$K_a$. It is common for such residues to form strong interactions with residues in the vicinity, which often causes sampling problems with the use of conventional MD, owing to the slower convergence of these residues. In addition to convergence issues, deficiencies in the GB solvent model used have been highlighted in the literature.[16] For buried residues, the GB model underestimates the desolvation energy and buried charge−charge interactions owing to neglect of the solvent excluded volume. Although an improved GB model would certainly improve results, this study is only focused on the sampling issues associated with the constant pH method and improvement of results with the use of enhanced sampling. In the case of ASP-52, an interaction with ASN-46 is shown to persist for the duration of several simulations, causing the calculated p$K_a$ to be

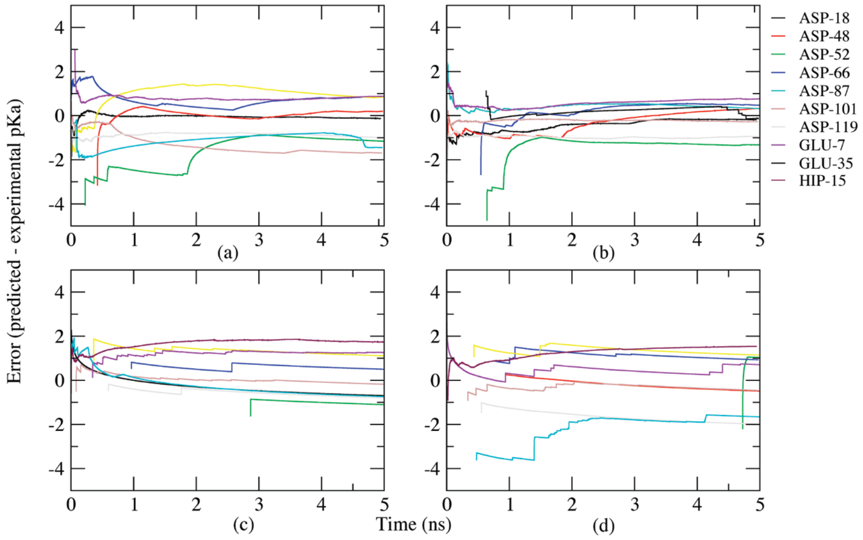


***Figure 3.*** p$K_a$ values of titratable residues over the duration of 5 ns simulations employing CpHMD (top two plots) and CpHaMD (lower two plots) at pH 3.0 (plots a and c) and pH 6.5 (plots b and d).

***Table 2.*** RMS Error of pKa Values Calculated from 5 ns Simulations over the pH Range of 2−6.5, with Respect to the Mid-Point Experimental pKa Value

| | RMS error | | |
|---|---|---|---|
| | CpHaMD | CpHMD | null model |
| all residues | 0.73 | 0.80 | 1.39 |
| aspartates | 0.75 | 1.46 | 1.34 |
| glutamates | 0.85 | 1.04 | 1.68 |
| histidine | 0.03 | 0.12 | 0.69 |

notably lower than the experimentally reported p$K_a$ range of 3.6−3.76. As also observed in the constant-pH study of Mongan et al., persistent interactions with neighboring residues also exist with GLU-35, but to a lesser degree when compared with ASP-52.

**RMS Error Analysis.** The RMS error was calculated for groups of residue types, with respect to the midpoint of the experimental value range[34] (Table 2). The RMS error was also calculated for the null model, against which both methods are shown to perform considerably better.

As stated earlier, although the simulations do not appear to have reached convergence for all residues, overall, the CpHaMD method is indicated to predict p$K_a$ values which are closer to experimental results, confirmed by the lower RMS error values reported in Table 2. The RMS error for the single histidine residue included in the calculations has the lowest error and is reported experimentally to only have a small shift in p$K_a$ from the model compound. This histidine residue resides on the surface, away from the active site, and possesses interactions with neighboring residues, including Thr-89, which are overestimated in a few of the simulations, resulting in the higher predicted p$K_a$ values. The groups of aspartates and glutamates both contain residues exhibiting larger p$K_a$ shifts and reside in more buried locations of HEWL, resulting in the relatively higher RMS errors.

In addition, the results of both the CpHMD and CpHaMD methods are shown to perform well on comparison with those achieved using other CpHMD methodologies (Table 3), where the leading results have RMS errors in the range of 0.65−0.8 for the acidic residues. The constant-pH methods achieving good results when applied to the HEWL system include the more recent method of Khandogin and Brooks,[16] who combine the constant-pH method with replica exchange and an improved GB implicit solvent model, attaining p$K_a$ predictions with a RMS error of 0.65 and 1.19 (RMS error calculated from simulations including and excluding salt effects, respectively). Machuqueiro and Baptista[35] achieve a RMS error value of 0.84 with the inclusion of proton tautomerism in their CpHMD method, and incorporating the generalized reaction field (GRF) for the treatment of long-range electrostatics. The RMS error increases to 0.79−0.93 when changing to the particle-mesh Ewald (PME) algorithm. Bürgi et al.[18] use constant-pH with TI and MC for the protonation state determination and achieve only qualitative p$K_a$ results, denoted by a RMS error of 2.97, which has been attributed to inadequate simulation time for convergence.

***Table 3.*** RMS Error Values Calculated for Acidic Residues of HEWL Listed in Table 1[a]

| methodology | RMS error |
|---|---|
| null model | 1.39 |
| Bürgi et al. (2002)[18] | 2.97 |
| Lee et al. (2004)[13] | 2.28 |
| Mongan et al. (2004)[4] | 0.82 |
| Khandogin and Brooks (2006)[16] | 0.65,[b] 1.19[c] |
| Machuqueiro and Baptista (2008)[35] | PME: 0.79−0.93;[d] GRF: 0.84 |
| this work | CpHMD: 0.8; CpHaMD: 0.73[e] |

[a] Values calculated with respect to the mid-point of the experimental range given in ref.[34] [b] Calculated from simulations carried out in the implicit solvent model in the presence of salt effects. [c] Carried out in the implicit solvent model in the absence of salt effects. [d] Range of values for the dielectric constant used. [e] Calculated from 5 ns CpHaMD simulations.

Good results have also been achieved for HEWL using Poisson−Boltzmann-based p$K_a$ calculations.[36−42] However, no good method has yet been developed which accounts for significant conformational change, and generally, the current methods are likely to be insufficient in cases where conformational change has a large influence on residue p$K_a$.[11] The

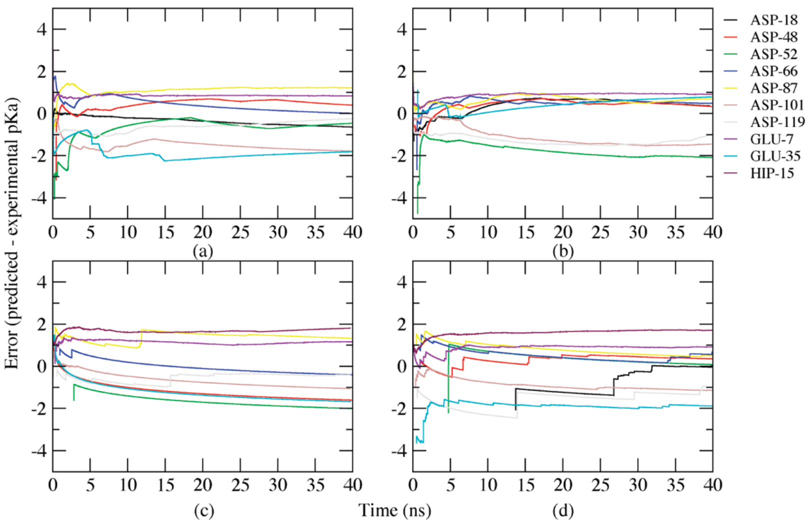


***Figure 4.*** p$K_a$ values of titratable residues over the duration of 40 ns CPMD (top two plots) and CpHaMD (lower two plots) simulations at pH 3.0 (plots a, c) and pH 6.5 (plots b, d).

***Table 4.*** Average pKa Predictions Calculated from Three 40 ns Simulations Using CpHMD and CpHaMD Methods[a]

| | CpHMD | | CpHaMD | | exptl. |
|---|---|---|---|---|---|
| residue | pH 3.0 | pH 6.5 | pH 3.0 | pH 6.5 | value |
| ASP-52 | **2.47** (1.19) | **1.67** (−) | 3.73 (0.67) | 3.62 (0.78) | 3.68 |

[a] The standard deviation is noted in brackets. (−) indicates pKa prediction only possible in one of three simulations.

CpHMD methods, such as those described here, are attractive since they incorporate flexibility and offer the ability to study the dynamics of pH-dependent phenonoma.

**Extended Simulations.** Simulations were extended to 40 ns and performed in triplicate for two pH values in different regions of the acidic pH range (pH 3.0 and pH 6.5), increasing the opportunity for conformational change and, thus, to test whether a further increase in conformational sampling would improve the accuracy of the p$K_a$ prediction for the more challenging residues.

As Figure 4 demonstrates, the extended simulations carried out using the CpHaMD method appear to converge faster and progress closer toward the experimental values for a larger number of residues when compared with simulations carried out using the standard CpHMD method. This is especially pronounced for simulations carried out at pH 3.0, which is expected, since the majority of the experimental p$K_a$ values lie closer to this pH. At pH 6.5, the majority of the measured residues are in the deprotonated state, and the calculation of the p$K_a$ becomes very sensitive as a result of the relative magnitude of unprotonated to protonated states becoming very small, due to the considerably smaller number of transitions between protonation states.

The p$K_a$ predictions for the previously mentioned problematic residue, ASP-52, are significantly improved using the CpHaMD method, with all six simulations (three simulations at pH 3 and three simulations at pH 6.5) generating values within 1 p$K_a$ unit of the experimental range (see Table 4). A greater variation, indicated by the larger standard deviations, is observed between the simulations employing the standard CpHMD method, with no value at all calculated for two of the three simulations carried out at pH 6.5. For these simulations, the residue has a strong tendency to become stuck in the deprotonated state, owing to a continued hydrogen bond with residue ASN-46. This interaction is present in the crystal structure and was noted to persist in the study using the standard version of this constant pH method.[4] ASN-46 resides in a loop region of HEWL, which does not display significant mobility in the CpHMD simulations. However, in simulations employing CpHaMD, this region undergoes increased conformational motion, as highlighted by the larger sampling area of the CpHaMD simulations, shown in the two-dimensional plots, generated from PCA analysis (Figure 5). This increased loop motion facilitates the dissociation of the interaction between the two residues, as when the loop moves away, the interaction is lost and ASP-52 is able to interchange to the protonated state (Figure 6). Within CpHaMD simulations, the aforementioned interaction is observed to repeatedly dissociate and reform, depending on local conformational change, illustrated by the number of transitions occurring between the protonated and deprotonated forms of the residue throughout the simulation (example shown in Figure 7). Over the three CpHaMD simulations at pH 6.5, an average of 181 transitions were recorded, whereas for the one CpHMD simulation at pH 6.5, for which a p$K_a$ could be calculated, only two transitions took place throughout the 40 ns of CpHMD simulation. For less problematic residues, the number of transitions is far higher during CpHMD simulations at 6.5, with >10 000 transitions recorded for some residues.

Overall, the initial application of this newly coupled aMD enhanced sampling technique to the standard constant pH methodology of Mongan et al. signifies the CpHaMD technique to be promising in improving the convergence of constant-pH simulations, providing more accurate p$K_a$ pre-

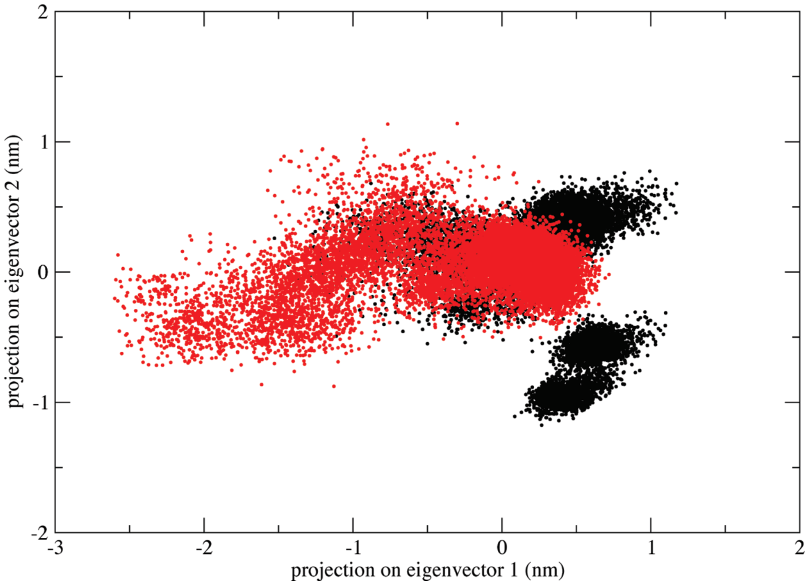


***Figure 5.*** Conformational sampling of residues (502 atoms) within 7.5 Å of ASP-52 demonstrated by PCA analysis. Eigenvectors generated from the concatenation of trajectories of simulations carried out at pH 6.5. Red: sampling from simulation carried out using CpHaMD. Black: sampling from simulation carried out using CpHMD.

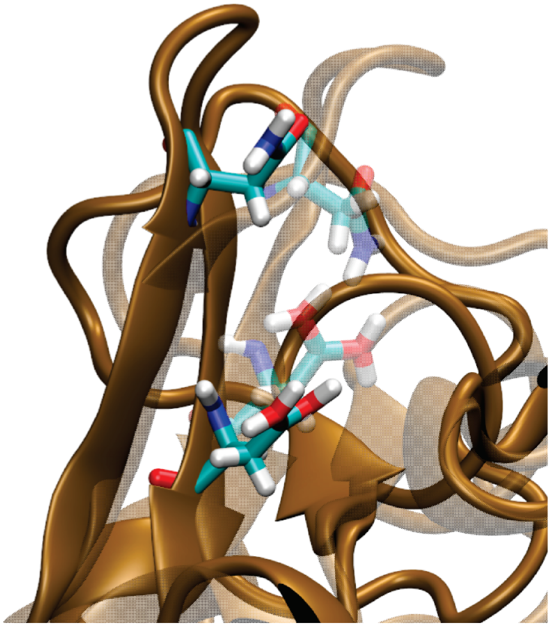


***Figure 6.*** Motion of loop allowing the dissociation of the ASP-52–ASN-46 interaction in simulations employing CpHaMD methodology.

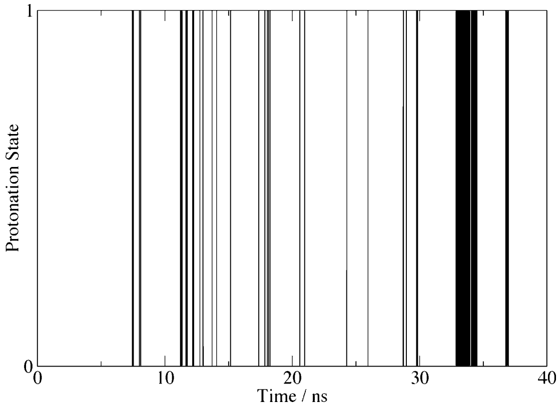


***Figure 7.*** Transitions between protonated (1) and deprotonated (0) states of ASP-52 over a 40 ns CpHaMD simulation at pH 6.5.

dictions and dynamics of titratable residues at a range of pH conditions.

## Conclusions

This study has introduced a new technique whereby the CpHMD method of Mongan et al.[4] has been coupled with the aMD enhanced sampling method of de Oliveira et al. and Hamelberg et al.[20,43] This coupled technique substitutes the conventional MD employed in the standard CpHMD method with aMD, a method previously demonstrated to enhance sampling by lowering the energy barriers of the energy landscape, while leaving the minima unchanged, with the capability of fully recovering the correct canonical averages of observables, in this case, p$K_a$. CpHaMD utilizes the same GB implicit solvation, with Monte Carlo sampling based on GB-derived energies as used in the standard method.

The initial results generated in this study show the CpHaMD method to more efficiently sample conformational space compared with the standard CpHMD method, resulting in faster convergence of constant pH simulations and improved agreement of calculated p$K_a$ values with those obtained experimentally. In addition, the calculated RMS error between the predicted and experimental p$K_a$ values of the acidic residues of HEWL demonstrate the CpHaMD methodology to generate results close to the leading results reported in the literature for other CpHMD methods. Owing to the improved conformational sampling, this method has proved to be advantageous over the CpHMD method in obtaining more accurate and consistent p$K_a$ predictions for the more buried residues of the system, which are typically more problematic to obtain owing to their slow convergence. This has been highlighted by the considerably improved results of the most problematic residue of HEWL, the catalytically important ASP-52, where the enhanced conformational motion observed in the vicinity of this residue in simulations utilizing CpHaMD clearly demonstrates the link between protonation state and conformation.

From this initial study, the RMS error measured between the calculated and experimental results are close to the leading results reported in the literature for HEWL. It is hoped that this method would assist in the study of biomolecular systems, in gaining more accurate thermodynamics and capturing important pH-coupled conformational events in a more time-efficient manner.

**Acknowledgment.** We thank Mikolai Fajer for his helpful suggestions and discussions in this work. This work was supported in part by the NSF, NIH, HHMI, CTBP, and NBCR.

## References

(1) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. *J. Chem. Phys.* **2002**, *117*, 4184–4200.

(2) Dlugosz, M.; Antosiewicz, J. M.; Robertson, A. D. *Phys. Rev. E* **2004**, *69*, 021915.1021915.10.

(3) Dlugosz, M.; Antosiewicz, J. M. *Chem. Phys.* **2004**, *302*, 161–170.

(4) Mongan, J.; Case, D. A.; McCammon, J. A. *J. Comput. Chem.* **2004**, *25*, 2038–2048.

(5) Baptista, A. M.; Martel, P. J.; Petersen, S. B. *Proteins: Struct., Funct., Genet.* **1997**, *27*, 523–544.

(6) Borjesson, U.; Hunenberger, P. H. *J. Chem. Phys.* **2001**, *114*, 9706–9719.

(7) Borjesson, U.; Hunenberger, P. H. *J. Phys. Chem. B* **2004**, *108*, 13551–13559.

(8) Mongan, J.; Case, D. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 157–163.

(9) Chen, J. H.; Brooks, C. L.; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140–148.

(10) Baker, N. A.; Bashford, D.; Case, D. A. In *New Algorithms for Macromolecular Simulation*; Leimkuhler, B., Chipot, C., Elber, R., Laaksonen, A., Mark, A., Schlick, T., Schütte, C., Skeel, R., Eds.; Springer: New York, 2006; Vol. 49, pp 263−295.

(11) Bashford, D. *Front. Biosci.* **2004**, *9*, 1082–1099.

(12) Tanford, C.; Roxby, R. *Biochemistry* **1972**, *11*, 2192–2198.

(13) Lee, M. S.; Salsbury, F. R.; Brooks, C. L. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 738–752.

(14) Khandogin, J.; Brooks, C. L. *Biophys. J.* **2005**, *89*, 141–157.

(15) Im, W. P.; Lee, M. S.; Brooks, C. L. *J. Comput. Chem.* **2003**, *24*, 1691–1702.

(16) Khandogin, J.; Brooks, C. L. *Biochemistry* **2006**, *45*, 9363–9373.

(17) Chen, J. H.; Im, W. P.; Brooks, C. L. *J. Am. Chem. Soc.* **2006**, *128*, 3728–3736.

(18) Burgi, R.; Kollman, P. A.; van Gunsteren, W. F. *Proteins: Struct., Funct., Bioinf.* **2002**, *47*, 469–480.

(19) Stern, H. A. *J. Chem. Phys.* **2007**, *126*, 164112.

(20) de Oliveira, C. A. F.; Hamelberg, D.; McCammon, J. A. *J. Chem. Phys.* **2007**, *127*, 175105–175112.

(21) Hamelberg, D.; de Oliveira, C. A. F.; McCammon, J. A. *J. Chem. Phys.* **2007**, 127.

(22) Bartik, K.; Redfield, C.; Dobson, C. M. *Biophys. J.* **1994**, *66*, 1180–1184.

(23) Takahashi, T.; Nakamura, H.; Wada, A. *Biopolymers* **1992**, *32*, 897–909.

(24) Vocadlo, D. J.; Davies, G. J.; Laine, R.; Withers, S. G. *Nature* **2001**, *412*, 835–838.

(25) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(26) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297–1304.

(27) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

(28) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 383–394.

(29) Kyte, J. In *Structure in Protein Chemistry*; Garland Publishing, Inc: New York, 1995; p 64.

(30) Balsera, M. A.; Wriggers, W.; Oono, Y.; Schulten, K. *J. Phys. Chem.* **1996**, *100*, 2567–2572.

(31) Mu, Y. G.; Nguyen, P. H.; Stock, G. *Proteins: Struct., Funct., Bioinf.* **2005**, *58*, 45–52.

(32) Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(33) Yeh, A. P.; McMillan, A.; Stowell, M. H. B. *Acta Crystallogr., Sect. D* **2006**, *62*, 451–457.

(34) Demchuk, E.; Wade, R. C. *J. Phys. Chem.* **1996**, *100*, 17373–17387.

(35) Machuqueiro, M.; Baptista, A. M. *Proteins: Stuct., Funct., Bioinf.* **2008**, *72*, 289–298.

(36) Beroza, P.; Case, D. A. *J. Phys. Chem.* **1996**, *100*, 20156–20163.

(37) Yang, A. S.; Honig, B. *J. Mol. Biol.* **1993**, *231*, 459–474.

(38) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. *J. Mol. Biol.* **1994**, *238*, 415–436.

(39) Gibas, C. J.; Subramaniam, S. *Biophys. J.* **1996**, *71*, 138–147.

(40) van Vlijmen, H. W. T.; Schaefer, M.; Karplus, M. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 145–158.

(41) Baptista, A. M.; Soares, C. M. *J. Phys. Chem. B* **2001**, *105*, 293–309.

(42) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. *Biophys. J.* **2002**, *83*, 1731–1748.

(43) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919–11929.

CT9005294

JCTC Journal of Chemical Theory and Computation

# Magnetostructural Dynamics from Hubbard-*U* Corrected Spin-Projection: [2Fe−2S] Complex in Ferredoxin

Nisanth N. Nair,*,† Jordi Ribas-Arino, Volker Staemmler, and Dominik Marx

*Lehrstuhl für Theoretische Chemie, Ruhr-Universität Bochum, 44780 Bochum, Germany*

Received October 16, 2009

**Abstract:** A Hubbard-corrected spin-projected two-determinant approach, EBS+$U_{\mathrm{scf}}$, is introduced to treat low-spin ground states of antiferromagnetically coupled transition metal complexes. In addition to providing access to total energies, forces, and ab initio simulations, it allows one to readily compute Heisenberg's exchange coupling $J(t)$ on the fly. By studying the binuclear [2Fe−2S] cofactor in a metalloprotein, *Anabaena* Fd, within this consistent nonempirical procedure in combination with a QM/MM framework, it is illustrated that spin-projection, self-interaction corrections, thermal fluctuations, and protein matrix shifts are crucial in obtaining $\langle J \rangle$ close to the experiment.

## 1. Introduction

The interplay of magnetic interactions and geometric structure[1] is at the heart of many important phenomena ranging from strongly correlated materials,[2] via transition metal coordination chemistry[3] to the redox biophysics of iron−sulfur proteins.[4] However, the computation of the magnetic exchange coupling constant $J$ for even rather small transition metal complexes, which is crucial for describing magnetic clusters,[5–7] molecular magnets,[8] and metalloproteins,[9–11] to name but a few, is still a challenge to electronic structure calculations.[12] This applies to wave function-based methods because the computational complexity rapidly explodes beyond feasibility, although it is well-known what should be done in the systematic world of configuration interaction (CI) calculations. For instance, even the computation of $J$ for the small antiferromagnetic $[Fe_2S_2(SH)_4]^{2-}$ complex in vacuo, being the most stripped-down model for the [2Fe−2S] cofactor in Ferredoxins (Fd) as a major class of mobile electron carrier in contemporary biology, is beyond current capabilities when it comes to convergence of multireference character, electron correlation, and basis set in concert.

The roots of the problem are quite different in the realm of density-based methods where conceptual difficulties dominate. It is well-known that LDA or GGA functionals suffer severely from spurious self-interactions,[13] producing artificially delocalized spin densities. This, in turn, induces stronger bonding, contraction of structures, and a dramatic overestimation of $|J|$.[14] This problem can be tackled in various ways such as using approximate self-interaction corrections[15–17] or hybrid functionals[14,18,19] including the idea to tune the admixture of Fock exchange appropriately.[20] In addition, constrained density functional theory[21] has been transferred successfully from solids to molecular systems[22–26] by controlling spin-density in real space.

A distinctly different route is to invoke a Hubbard-*U* correction.[27,28] In practice, this can be carried out by adjusting *U* (semi)empirically or by determining it self-consistently,[29] $U_{\mathrm{scf}}$, using linear-response theory.[30] The GGA+$U_{\mathrm{scf}}$ method yields favorable results without any fit to experimental data at computational costs that are basically the same as those of the underlying GGA calculation in addition to allowing readily for ab initio molecular dynamics.[29,31–33]

All aforementioned density-based approaches draw on the idea to improve a single-determinantal representation of the antiferromagnetic ground state of such spin-coupled systems. A conceptually different approach to compute $J$ is based on spin-projection schemes and relies on evaluating more than one total energy[34–36] and thus using more than one Kohn−Sham (KS) determinant. Along such lines, the "extended broken symmetry" (EBS) approach has been introduced,[37,38] which provides a general and efficient two-determinantal formulation of the antiferromagnetic low-spin

* Corresponding author e-mail: nnair@iitk.ac.in.

† Present address: Department of Chemistry, Indian Institute of Technology Kanpur, Kanpur 208016, India.

10.1021/ct900547w © 2010 American Chemical Society
Published on Web 01/15/2010

state of transition metal dimers. The EBS approach allows one to compute both geometrical structures and $J$ values consistently by using identical spin-projection techniques. In addition, it can be easily used in ab initio molecular dynamics, thus opening the doorway to compute spectral densities $J(\omega)$ from the time evolution of the exchange coupling constant, $J(t)$, and thus to "magnetostructural dynamics".[37,38]

Here, we propose a technique, EBS+$U_{\mathrm{scf}}$, that builds on the strengths of a systematic spin-projection scheme, EBS, combined with a linear-response GGA+$U_{\mathrm{scf}}$ treatment of the underlying open-shell KS determinants. Most importantly, this method is very practical and has the accuracy of state-of-the-art multireference CI calculations. In addition, being readily amenable to ab initio molecular dynamics,[39] EBS+$U_{\mathrm{scf}}$ provides access to the dynamics of magnetic properties. Here, this will be demonstrated by investigating the [2Fe−2S] cofactor in *Anabaena* Ferredoxin (Fd) within a QM/MM framework.[39]

## 2. Methods

**2.1. Evaluation of Magnetic Exchange Coupling Constants.** The magnetic properties of transition metal dimers may be represented by a Heisenberg Hamiltonian:

$$\hat{H} = -2J\hat{S}_{\mathrm{A}}\hat{S}_{\mathrm{B}} \tag{1}$$

where $\hat{S}_{\mathrm{A}}$ and $\hat{S}_{\mathrm{B}}$ are effective local spin angular momentum operators at the two sites A and B, respectively; $J < 0$ ($J > 0$) implies antiferromagnetic (ferromagnetic) coupling. $J$ can be expressed[37] as

$$J = \frac{E^{\mathrm{BS}} - E^{\mathrm{HS}}}{S_{\max}^2 - S_{\min}^2 - \Theta^{\mathrm{BS}} + \Theta^{\mathrm{HS}}} \tag{2}$$

upon invoking generalized spin-projection ideas in conjunction with Löwdin's exact formulation of the expectation value of the total spin operator $\hat{S} = \hat{S}_{\mathrm{A}} + \hat{S}_{\mathrm{B}}$. Here, $S_{\min} = |S_{\mathrm{A}} - S_{\mathrm{B}}|$ and $S_{\max} = S_{\mathrm{A}} + S_{\mathrm{B}}$ are the minimum and maximum total spin quantum number corresponding to the exact high-spin (HS) and low-spin (LS) eigenstates of eq 1, respectively. $E^{\mathrm{HS}}$ and $E^{\mathrm{BS}}$ are the total energies of the HS and broken symmetry (BS) states obtained using spin-polarized KS determinants with appropriate integer occupation numbers. Because these determinants are not spin eigenfunctions, they contain spin contaminations that are generally different for the different spin states. However, the expression for $J$ in eq 2 contains systematic corrections for these spin contaminations via

$$\Theta^X = N_{\mathrm{nmag}}^{\beta,X} + 2\int \Gamma^X(\mathbf{r}_1\alpha, \mathbf{r}_2\beta|\mathbf{r}_1\beta, \mathbf{r}_2\alpha)\, \mathrm{d}\mathbf{r}_1\, \mathrm{d}\mathbf{r}_2 \tag{3}$$

$\Theta^X$ is exact[40] if $\Gamma^X(\mathbf{r}_1\alpha, \mathbf{r}_2\beta|\mathbf{r}_1\beta, \mathbf{r}_2\alpha)$ is the spin-off-diagonal $(\alpha\beta|\beta\alpha)$ element of the exact two-particle density matrix;[38] $N_{\mathrm{nmag}}^{\beta,X}$ is the number of paired $\beta$ electrons, and $N^{\alpha,X} \geq N^{\beta,X}$.

**2.2. The EBS + $U_{\mathrm{scf}}$ Approach.** The problem with the economical GGA functionals is that they suffer severely from self-interaction and thus from an unacceptable overestimation of $|J|$. Keeping an eye on the computational efficiency and acknowledging the excellent performance of the self-consistent linear-response-based GGA+$U_{\mathrm{scf}}$ method[29,30] for challenging molecular systems[29,31,33] prompted us to combine this approach with the extended spin-projection scheme. Self-interaction results in an unphysical curvature of the GGA energy curve as a function of "local" electron occupation for noninteger (or fractional) values of local occupation as discussed in detail in ref 30 in the context of Hubbard corrections. Most importantly, by knowing the unphysical curvature of the GGA energy curve, we can repair the self-interaction artifacts of the pure-GGA functional, thus attaining a linear behavior in the GGA energy with respect to the local occupation. The curvature of this function, which is in fact the $U$ parameter, can be obtained conveniently using the linear response approach.[30] The curvature of the GGA energy, or the $U$ parameter, then depends on the system, the definition of "local" occupation, and the density functional that we have chosen. Such a GGA+$U$ functional in turn provides new energy and a new set of orbitals to represent the ground-state density, which can in many cases be qualitatively different from the non-$U$ case. Thus, a self-consistent approach[29] is required to obtain the true numerical $U$ value for the ground-state wave function.

In this spirit, the Hubbard functional[30]

$$E_U^X = \frac{1}{2}\sum_{I,\sigma} U_{\mathrm{scf}}^{I,X} \mathrm{Tr}[\mathbf{n}^{I,\sigma}(\mathbf{1} - \mathbf{n}^{I,\sigma})] \tag{4}$$

is added to the GGA functional used within EBS to describe the HS and BS states; $I$ runs over all selected atoms where the Hubbard correction is applied. The occupation matrix $\mathbf{n}^{I,\sigma}$ is

$$n_{j,k}^{I,\sigma} = \sum_i f_i^\sigma \langle\psi_i^\sigma|\chi_j^I\rangle\langle\chi_k^I|\psi_i^\sigma\rangle \tag{5}$$

where the sum runs over all spin orbitals $\psi_i^\sigma$ with occupation $f_i^\sigma$. In the present case, $\chi_j^I$ is the set of five orthonormal pseudo atomic d-orbitals of iron atoms.

In addition to providing direct access to $J$, this two-determinant EBS+$U_{\mathrm{scf}}$ approach allows for a convenient and general spin-projected representation of the total energy of the antiferromagnetic LS ground state via

$$E^{\mathrm{LS}} = (1 + c)E^{\mathrm{BS}} - cE^{\mathrm{HS}} = \mathscr{P}E^{\mathrm{BS,HS}} \tag{6}$$

$$c = \frac{S_{\max} - S_{\min} + \Theta^{\mathrm{BS}}}{S_{\max}^2 - S_{\min}^2 - \Theta^{\mathrm{BS}} + \Theta^{\mathrm{HS}}} \tag{7}$$

where $\mathscr{P}$ projects the energy of the LS state from the energies of the two single-determinant BS and HS states.[37,38] Having access to the total energy and its derivatives enables spin-projected multideterminant ab initio molecular dynamics via $M_I\ddot{\mathbf{R}}_I = -\nabla_I\mathscr{P}E^{\mathrm{BS,\,HS}}$. We stress that this allows one to compute both $J(t)$ and the geometrical structure of the complexes consistently using the identical spin-projection and electronic structure methods at variance with the standard approach to obtain only $J$ from spin-projection. For convenience, we have used the strong localization approximation of magnetic orbitals for $\Theta^X$ as assessed thoroughly in ref 38.

***Table 1.*** Selected Structural Properties (Distances in Å and Angles in deg) of the $Fe_2S_2$ Core for the Fully Optimized BS, EBS, BS+$U_{scf}$, and EBS+$U_{scf}$ Structures of the $[Fe_2S_2(SH)_4]^{2-}$ Complex in Vacuo Together with the Exchange Coupling Constant *J* (Reported in cm $^{-1}$) of These Structures Computed Using Various Methods as Indicated; See "Statics"[a]

| | statics | | | | dynamics | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | BS | BS+*U* | EBS | EBS+*U* | EBS | EBS+*U* | EBS @ Fd | EBS+*U* @ Fd | X-ray @ Fd |
| *r*(Fe1−Fe2) | 2.68 | 2.78 | 2.62 | 2.72 | 2.64 | 2.75 | 2.62 | 2.74 | 2.75 |
| *r*(Fe1−S1) | 2.20 | 2.24 | 2.17 | 2.22 | 2.18 | 2.24 | 2.22 | 2.27 | 2.29 |
| *r*(Fe1−S2) | 2.19 | 2.24 | 2.17 | 2.22 | 2.17 | 2.24 | 2.16 | 2.22 | 2.23 |
| *r*(Fe2−S1) | 2.20 | 2.24 | 2.18 | 2.22 | 2.17 | 2.24 | 2.22 | 2.29 | 2.23 |
| *r*(Fe2−S2) | 2.20 | 2.25 | 2.18 | 2.23 | 2.18 | 2.24 | 2.17 | 2.24 | 2.18 |
| $\theta$(Fe1−S1−Fe2) | 74.9 | 76.6 | 74.2 | 75.4 | 74.8 | 76.0 | 72.2 | 74.0 | 75.1 |
| $\theta$(Fe1−S2−Fe2) | 75.1 | 76.6 | 74.2 | 75.3 | 74.7 | 76.0 | 74.3 | 76.0 | 76.8 |
| $\theta$(Fe−S−Fe-S) | 0.4 | 0.2 | 0.1 | 1.0 | 0.1 | 0.0 | 8.4 | 9.7 | 8.7 |
| $J_{PBE}$ | −390 | −192 | −435 | −223 | −402 (±50) | −202 (±40) | −386 (±47) | −175 (±35) | −182 (±20) |
| $J_{CAS\text{-}CI}^{[10,10]}$ | −43 | −30 | −51 | −36 | | | | | |
| $J_{MCAS\text{-}CI}^{[10,10]}$ | −127 | −92 | −155 | −112 | | | | | |
| $J_{CASSCF}^{[22,16]}$ | −186 | −137 | −216 | −159 | | | | | |

[a] Thermal averages of the same properties for the same systems as obtained from molecular dynamics simulations are reported for the EBS and EBS+$U_{scf}$ methods again in vacuo and in the protein (denoted EBS @ Fd and EBS+*U* @ Fd); see "Dynamics". Structural data of Fd and average *J* values obtained from these simulations are also compared to the X-ray structure[57] (PDB code: 1qt9, see X-ray @ Fd) and to the experimental *J* value;[54] for the latter, the rmsd in case of simulations and the experimental errors are reported in parentheses, ($\sigma_J$).

**2.3. Density Functional Calculations.** This EBS+$U_{scf}$ method has been implemented in CPMD.[41] The simulations of the $[Fe_2S_2(SH)_4]^{2-}$ complex in vacuo were performed within spin-unrestricted Kohn−Sham DFT using the CPMD code.[41,42] For the EBS and EBS+*U* calculations, we used the plane wave/pseudopotential formulation[39] of Kohn−Sham DFT together with the PBE functional.[43,44] The core electrons were represented by ultrasoft pseudopotentials[45] with a plane-wave cutoff of 30 Ry. Additional d-projectors were considered for the sulfur atoms, and scalar relativistic corrections and semicore states were taken for Fe. A cubic box of 40 au with finite cluster boundary conditions[46] was used to decouple the negatively charged periodic images.[39]

In the case of the Hubbard corrected EBS+$U_{scf}$ approach, the self-consistent[29] linear-response procedure[30] is used to compute the $U_{scf}$ parameters acting on the two iron atoms separately for the HS and BS states using the optimized EBS equilibrium structure of $[Fe_2S_2(SH)_4]^{2-}$ in vacuo. This yields $U_{scf,0}^{HS} = 3.45$ eV and $U_{scf,0}^{BS} = 3.50$ eV for the two required Kohn−Sham determinants. It is important to note that within this recently introduced framework the Hubbard-*U* correction is not adjusted such that *J* fits any experimental data. Rather it is a property that is extracted self-consistently from the underlying electronic structure theory, that is, the PBE density functional together with a plane wave/pseudopotential representation of the orbitals. Furthermore, the "strongly localized" approximation underlies the calculation of *J* using the EBS and EBS+$U_{scf}$ approaches, which is consistent with the assumptions underlying the Heisenberg Hamiltonian.[47,38]

**2.4. Configuration Interaction Calculations.** The two wave function-based methods that we used were a conventional CAS-CI (complete active space configuration interaction) and a modified CAS-CI (MCAS-CI) approach. The Bochum suite of open-shell ab initio programs was used for these calculations.[48−51] The first step is a restricted open-shell Hartree−Fock (ROHF) calculation for the HS state of the complex. This yields a set of orthogonal occupied molecular orbitals, with the 10 3d orbitals at the two Fe atoms singly occupied and all other orbitals at the Fe cores, the bridging $S^{2-}$ anions, and the ligands doubly occupied. These orbitals are then used in a subsequent configuration interaction calculation (CAS-CI), in which all possible configurations with 10 electrons in the 10 3d orbitals are included in the active space, that is, CAS-CI(10,10). The basis set used for $[Fe_2S_2(SH)_4]^{2-}$ in vacuo had approximately aVTZ quality (augmented valence triple-$\zeta$ with two sets of polarization functions) and consisted of 500 basis functions.

It is well-known[52,53] that *J* values determined by CAS-CI are by a factor 2−5 too small mainly because the orbitals determined for the covalent (cov) states yield only a poor description for the charge-transfer (ct) configurations. This leads to too large energy denominators in the perturbation estimate:

$$J_{AF} = -\sum_{ct} \frac{\langle \Psi_{cov} | \hat{H} | \Psi_{ct} \rangle^2}{E_{ct} - E_{cov}} \tag{8}$$

for the antiferromagnetic coupling constant $J_{AF}$, thus yielding too small values of $J_{AF}$.

Two different schemes have been employed for improving this description. First, we optimized the wave functions for the ct-configurations, which requires nonorthogonal CI. Even this demanding procedure yields only about 70% of $J_{AF}$. We have therefore used a simple but efficient modification of CAS-CI by introducing a correction *R* into the energy denominator in eq 8:

$$\Delta\tilde{E} = (E_{ct} - R) - E_{cov} \approx \tilde{U} \tag{9}$$

to account for the relaxation of the ct wave functions, where *R* is computed by separate CASSCF calculations.[53] This MCAS-CI approach yields $J_{AF}$ in good agreement with elaborate multireference-CI at the cost of economical CAS-CI. Including the bridging S 3p-orbitals in the active space of multireference CASSCF[22, 16] yields *J* values better than CAS-CI, but they still deviate significantly from MCAS-CI. Thus, the MCAS-CI results in Table 1 are the most reliable wave function-based values for *J* to date.

Note that we have used the letter "$U$", describing an on-site Coulomb repulsion, in slightly different but closely related contexts. In the Hubbard functional eq 4, $U_{scf}$ prevents the unpaired d-electrons from being too strongly delocalized, thus reducing spurious self-interactions. In MCAS-CI, $\tilde{U}$ is the cov → ct excitation energy. For $[Fe_2S_2(SH)_4]^{2-}$, the unmodified energy denominators are ∼23.0 eV, whereas $R \approx 15.0$ eV, such that the "true" excitation energies amount to $\tilde{U} \approx 8.0$ eV.

**2.5. Molecular Dynamics Simulations.** The Car−Parrinello molecular dynamics scheme[39,55] was used for performing molecular dynamics simulations in the approximate low-spin (LS) ground state of [2Fe−2S] systems within the EBS, EBS+$U_{scf}$, and EBS+$U_{scf}$ QM/MM schemes. We employed a recently introduced spin-projected Car−Parrinello Lagrangian for the LS dynamics, which is described in detail in refs 37, 38. Hydrogen masses were substituted by deuterium masses for technical reasons,[39] and a time step of 4 au corresponding to ∼0.12 fs together with a fictitious orbital mass parameter of 500 au was used. The HS and BS Kohn−Sham wave functions were thermostatted separately using Nosé−Hoover chain thermostats[39,56] to keep the orbitals close to their instantaneous ground state. Similarly, the nuclei and atoms in the QM and MM systems, respectively, were connected to separate Nosé−Hoover chain thermostats at 300 K. After equilibration of ∼2−5 ps, trajectories of ∼18 and 8 ps were collected for the in vacuo complex (using both the EBS and the EBS+$U_{scf}$ approaches) and in protein (using EBS+$U_{scf}$ QM/MM), respectively.

**2.6. Protein Model and System Setup.** The protein model is based on the oxidized *Anabaena* PCC7119 Fd[57] (PDB code: 1qt9, chain B) where standard protonation states of all amino acids were assumed.[37] In addition to the water molecules resolved in the crystal structure, the system was solvated using 13 265 TIP3P water molecules. In addition, 23 $Na^+$ and 5 $Cl^-$ ions were added to establish charge neutrality of the whole system. The protein was described using the AMBER94 force field,[58] and the partial charges for the [2Fe−2S] core are based on Bader analysis[59] of the electron density of the $[Fe_2S_2(SH)_4\ ]^{2-}$ cluster in vacuo. The nonbonding interactions for Fe were obtained from ref 60, and cysteinyl parameters were used for S.

For our hybrid QM/MM simulations,[37,61,62] we use the efficient CPMD/Gromos interface[63] within the CPMD program package[41,42] both extended by the EBS and EBS+$U_{scf}$ techniques. In addition to including the spin-projection to describe the LS ground state of the antiferromagnetically coupled [2Fe−2S] core, this approach allows for the consistent computation of spin-projected $J(t)$ values "on the fly" along the trajectories at no extra cost.[38] The QM subsystem contained the [2Fe−2S] cluster as well as the $S_\gamma$, $C_\beta$, and $H_\beta$ atoms of the four cysteinyl ligands. The dangling bonds at $C_\beta$ were saturated using capping H atoms constrained to the $C_\alpha$−$C_\beta$ connecting line. Thus, the QM part of the protein is $[Fe_2S_2(S-CH_3)_4]^{2-}]$ carrying a total charge of −2.

As a part of the cysteins connected to the [2Fe−2S] core are treated by QM and the other part by MM, the charges of the MM atoms need to be reparameterized to obtain a total charge of zero separately for the MM part of the cysteines. This is achieved by distributing the residual charge (of +0.09870) on the $C_\alpha$ and $H_\alpha$ atoms because they are mostly screened by the rest of the residues and are not involved in direct interactions with other residues; the charges of $C_\alpha$ and $H_\alpha$ atoms are increased from −0.0351 to 0.0000 and from 0.0508 to 0.1144, respectively.

For incorporating the QM-MM electrostatic interaction, we have chosen the well-established methods provided in ref 63; see also ref 39. In particular, all of the MM atoms lying within 17 Å of the QM region, which includes all of the protein MM atoms and several solvating water molecules, interact directly with the full QM electronic density using screened electrostatics[63] to counteract electron spill-out.[39] In the far field, the partial charges of all of those MM atoms lying beyond 17 Å with respect to the QM region interact with the multipole moments generated from the electronic charge density distribution of the QM system.[63]

## 3. Results and Discussion

First, the performance of EBS+$U_{scf}$ to compute $J$ is compared to various other approaches in Table 1 for the $[Fe_2S_2(SH)_4]^{2-}$ complex in vacuo. One first notes that the Hubbard-$U_{scf}$ correction leads to an elongation of the Fe−Fe and Fe−S distances by 0.10 and 0.05 Å, respectively, while the Fe−S−Fe angles remain unaffected, ∼(75.0 ± 1)°. A small amount of out of plane bending (about 1°) is observed in the EBS+$U_{scf}$ case, while the rest of the optimized $[Fe_2S_2(SH)_4]^{2-}$ structures feature a nearly planar core. However, a very symmetric planar average structure is obtained from the MD simulations using the EBS+$U_{scf}$ method. As a side remark, we mention that the minimum energy Fe−Fe distance of 2.62 Å determined by Li and Noodleman[64] for the $[Fe_2S_2(SCH_3)_4]^{2-}]$ complex using the spin-projected ground-state energy is exactly matching our EBS result obtained from direct energy minimization of the EBS density functional. This is not a surprising result because the EBS functional is based on the same spin projection as employed for the calculations in ref 64.

This structural expansion is connected with a reduction of $|J|$ by only ∼50 cm $^{-1}$ as judged from MCAS-CI; see Table 1. The main effect of the Hubbard correction, however, is an electronic-structure-based reduction of $|J|$ by almost a factor of 2! In particular, the EBS+$U_{scf}$ result for $J$ yields better agreement with the MCAS-CI than the EBS. Here, it is important to stress that using the less compact BS optimized structure, as mostly done in standard $J$ calculations, is intrinsically inconsistent with the spin-projection used to obtain $J$ itself!

As a next step, temperature and thus fluctuation effects on $J$ are assessed via ab initio molecular dynamics at 300 K by computing the probability distribution function, $P(J)$, from $J(t)$ using the EBS and EBS+$U_{scf}$ methods. First, a significant improvement is achieved with EBS+$U_{scf}$: the average $\langle J \rangle$ is appropriately shifted from about −400 to −200 cm$^{-1}$ when adding the Hubbard correction to the spin-projected PBE functional (see Figure 1). Second, it is crucial to note that the fluctuations $\sigma_J$ are dramatic: during its dynamics, $J(t)$

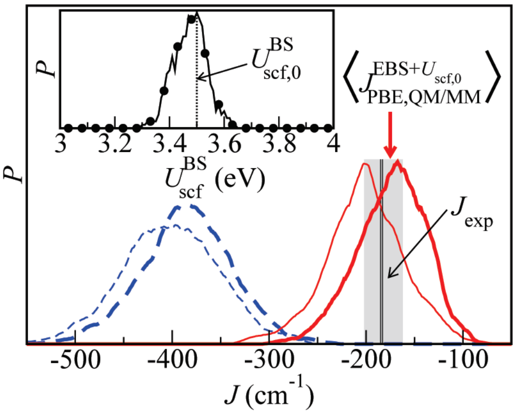


***Figure 1.*** Distribution functions $P(J)$ of $[Fe_2S_2(SH)_4]^{2-}$ in vacuo (thin lines) and of [2Fe−2S] cofactor in *Anabaena* Fd (thick lines) using EBS (dashed blue lines) and EBS+$U_{scf}$ (solid red lines) at 300 K. The value $J_{exp}$ measured in various Fd proteins[66,67,54] is marked by vertical lines and experimental errors[54] by a shaded region, to be compared to the computed average $J$ value in the protein ($J^{EBS+U_{scf,0}}_{PBE,QM/MM}$), marked by the red arrow. The inset shows the distribution function of the linear-response Hubbard-$U$ parameter computed for 100 configurations sampled from 5 ps trajectory of the in vacuo EBS+$U_{scf}$ simulation; $U^{BS}_{scf,0}$ at the EBS equilibrium structure is marked by a vertical line.

spans the range from about −300 to −100 cm$^{-1}$, thus yielding rmsd values as large as ±40 cm$^{-1}$; see Table 1.

As a technical note, we extract from the inset of Figure 1 that the value of $U^{BS}_{scf}$ is not strongly dependent on fluctuations at 300 K that drive the complex away from its equilibrium structure and thus from $U^{BS}_{scf,0}$; a similar picture holds for $U^{HS}_{scf,0}$ (not shown). This supports the established procedure[31,33] to compute $U_{scf,0}$ for some reference structure and to keep it fixed during structural relaxation or ab initio molecular dynamics.

In the crucial step of full EBS+$U_{scf}$ QM/MM simulations of the *Anabaena* Fd, $|J|$ is again significantly shifted toward smaller values; see Figure 1. Together with the finite-temperature shift from −223 to −202 cm$^{-1}$ in vacuo and the vacuum-to-protein shift of another ~30 cm$^{-1}$ at 300 K, the resulting $\langle J \rangle$ value of −175 ± 35 cm$^{-1}$ compares favorably to the experimental values of −183,[66] −185,[67] and −182 ± 20 cm$^{-1}$ [54] measured in vitro for several Fd species that are closely related to the one considered here fully in silico. Thus, a sound electronic structure treatment in conjunction with the protein environment fluctuating at finite temperatures is necessary for a reliable nonempirical computation of average $J$ values in proteins.

Similar to what has been observed for calculations in vacuo, an expansion in the structure of the core has been observed (Table 1). It is interesting to note that structural parameters predicted by the EBS+$U_{scf}$ QM/MM method are closer to the crystal structure[57] than those obtained from the simpler EBS QM/MM scheme. Like in the EBS QM/MM calculations,[37] an asymmetry in the hydrogen-bonding pattern near the iron−sulfur core is also observed in the case of the EBS+$U_{scf}$ QM/MM calculations. The Fe1−S1 (Fe1−S2) and Fe2−S1 (Fe2−S2) bond lengths are nearly the same, while Fe1−S1 (Fe2−S1) and Fe1−S2 (Fe1−S1) are quite different, which is ascribed to the hydrogen bonds to the [2Fe−2S] core within the solvated Fd. Different Fe2−S1 and Fe2−S2 bond lengths between EBS+$U_{scf}$ and the crystal structure can be traced back to a distinct hydrogen-bonding topology near the [2Fe−2S] cofactor as compared to the crystal structure as already amply discussed earlier.[37] This different topology might originate from a combination of solvation and finite temperature effects, which can easily change hydrogen-bonding patterns in view of the small energies involved. In fact, the structural data obtained for synthetic analogues of Fd[65] not only deviate from our calculations, but they also differ from the crystal structure of oxidized *Anabaena* Fd.[57] This is not surprising as the structure of the [2Fe−2S] core varies with its local environment, noting that the synthetic analogues have very different ligands as compared to wild-type Fd; no hydrogen bonds exist with the core, and counterions are present in the cyrstal structure.

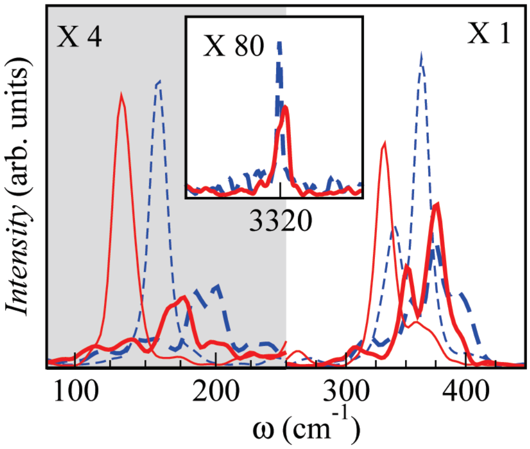


***Figure 2.*** Spectral densities $J(\omega)$ at 300 K of $[Fe_2S_2(SH)_4]^{2-}$ in vacuo (thin lines) and of [2Fe−2S] cofactor in *Anabaena* Fd (thick lines); see Figure 1 for labeling.

Last but not least, having direct access to the dynamics of the exchange coupling, $J(t)$, allows one to evaluate its power spectrum $J(\omega)$ via Fourier transforming its autocorrelation function; see Figure 2. One striking feature, both in vacuo and in protein, is the red-shift or softening of $J(\omega)$ due to the Hubbard correction. A detailed normal-mode analysis[37,38] shows that the two major peaks around 130 and 320 cm$^{-1}$ in vacuo can be solely assigned to $A_{g,A}$ and $A_{g,D}$ vibrational modes involving mainly angles and distances, respectively, of the $Fe_2S_2$ core. Furthermore, a weak peak at ~260 cm$^{-1}$ and the shoulder at 350−370 cm$^{-1}$ is due to vibrational coupling with the four SH-ligands. In the protein, however, $J(\omega)$ is systematically blue-shifted with respect to in vacuo as a result of structural constraints on the entire [2Fe−2S] cofactor imposed by the protein. In addition to this trend, the power spectrum $J(\omega)$ is much richer in the protein, which can be traced back to the mutual coupling of [2Fe−2S] motion to various skeleton vibrations. Most strikingly, very high-frequency modulations of $J(\omega)$ at about 3320 cm$^{-1}$ can be related to hydrogen bonding to the core.

## 4. Conclusion

We introduced a multideterminant QM/MM dynamics approach that combines systematically spin-projection with a linear-response Hubbard-$U_{scf}$ correction to compute exchange couplings including their time-evolution, $J(t)$, for antiferromagnetic transition metal dimers in complex molecular environments. Studying the [2Fe−2S] cofactor in Ferredoxin as a first example, it is shown that $J$ depends crucially on the subtle interplay of the quality of spin-projection, reduction of self-interaction artifacts, thermal fluctuations, protein matrix shifts, and a consistent treatment of geometrical structure and magnetic coupling. Taking into account all of these effects, consistently EBS+$U_{scf}$ QM/MM simulations are shown to yield excellent agreement with experiment. Transcending the specific case and implementation, the established framework can be generalized to other systems containing antiferromagnetically coupled centers, including polynuclear transition metal complexes, organic radicals, and molecular magnets. Most interestingly, this method can also be applied to magnetic states other than the ground state, which so far are not accessible to molecular dynamics techniques.

**Acknowledgment.** J.R.-A. is grateful to AvH and Catalan Government for Humboldt and Beatriu de Pinós Fellowships. D.M. acknowledges financial support from DFG (MA 1547/7) and FCI. Calculations were done at NIC (Jülich), Bovilab@RUB (Bochum), and Rechnerverbund-NRW.

## References

(1) Willett, R. D.; Gatteschi, D.; Kahn, O. Magneto-Structural Correlations in Exchange Coupled Systems; Reidel: Dordrecht, 1985; pp 1−632.

(2) Pickett, W. E. *Rev. Mod. Phys.* **1989**, *61*, 433–512.

(3) Ribas Gispert, J. *Coordination Chemistry*; Wiley-VCH: Weinheim, 2008; pp 295−338.

(4) Beinert, H.; Holm, R. H.; Münck, E. *Science* **1997**, *277*, 653–659.

(5) Gatteschi, D.; Caneschi, A.; Pardi, L.; Sessoli, R. *Science* **1994**, *265*, 1054–1058.

(6) Wernsdorfer, W.; Sessoli, R. *Science* **1999**, *284*, 133–135.

(7) Mandal, S.; Green, M. A.; Pati, S. K.; Natarajan, S. *J. Mater. Chem.* **2007**, *17*, 980–985.

(8) Schnelzer, L.; Waldmann, O.; Horvatic, M.; Ochsenbein, S. T.; Kraemer, S.; Berthier, C.; Guedel, H. U.; Pilawa, B. *Phys. Rev. Lett.* **2007**, *99*, 087201.

(9) Rees, D. C.; Howard, J. B. *Science* **2003**, *300*, 929–931.

(10) Howard, J. B.; Rees, D. C. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17088–17093.

(11) Pati, S. K.; Mallajosyula, S. S. *Angew. Chem., Int. Ed.* **2009**, *48*, 4977–4981.

(12) Pati, S. K.; Ramasesha, S.; Sen, D. Exact and Approximate Theoretical Techniques for Quantum Magnetism in Low Dimensions. In *Magnetism: Molecules to Materials IV*; Miller, J. S., Drillon, M., Eds.; Wiley-VCH Verlag GmbH & Co.: KGaA Weinheim, Germany, 2002; pp 119−171.

(13) Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981**, *23*, 5048–5079.

(14) Martin, R. L.; Illas, F. *Phys. Rev. Lett.* **1997**, *79*, 1539–1542.

(15) Svane, A.; Gunnarsson, O. *Phys. Rev. Lett.* **1990**, *65*, 1148–1151.

(16) Filippetti, A.; Fiorentini, V. *Phys. Rev. Lett.* **2005**, *95*, 086405.

(17) Akande, A.; Sanvito, S. *J. Chem. Phys.* **2007**, *127*, 034112.

(18) Rivero, P.; de P. R. Moreira, I.; Illas, F.; Scuseria, G. E. *J. Chem. Phys.* **2008**, *129*, 184110.

(19) Kudin, K. N.; Scuseria, G. E.; Martin, R. L. *Phys. Rev. Lett.* **2002**, *89*, 266402.

(20) Herrmann, C.; Yu, L.; Reiher, M. *J. Comput. Chem.* **2006**, *27*, 1223–1239.

(21) Dederichs, P. H.; Blügel, S.; Zeller, R.; Akai, H. *Phys. Rev. Lett.* **1984**, *53*, 2512–2515.

(22) Wu, Q.; Van Voorhis, T. *Phys. Rev. A* **2005**, *72*, 024502.

(23) Behler, J.; Delley, B.; Lorenz, S.; Reuter, K.; Scheffler, M. *Phys. Rev. Lett.* **2005**, *94*, 036104.

(24) Rudra, I.; Wu, Q.; Van Voorhis, T. *J. Chem. Phys.* **2006**, *124*, 024103−1−9.

(25) Sit, P. H.-L.; Cococcioni, M.; Marzari, N. *Phys. Rev. Lett.* **2006**, *97*, 028303.

(26) Schmidt, J. R.; Shenvi, N.; Tully, J. C. *J. Chem. Phys.* **2008**, *129*, 114110.

(27) Anisimov, V. I.; Zaanen, J.; Andersen, O. K. *Phys. Rev. B* **1991**, *44*, 943–954.

(28) Liechtenstein, A. I.; Anisimov, V. I.; Zaanen, J. *Phys. Rev. B* **1995**, *52*, R5467–R5470.

(29) Kulik, H. J.; Cococcioni, M.; Scherlis, D. A.; Marzari, N. *Phys. Rev. Lett.* **2006**, *97*, 103001.

(30) Cococcioni, M.; de Gironcoli, S. *Phys. Rev. B* **2005**, *71*, 035105.

(31) Sit, P. H.-L.; Cococcioni, M.; Marzari, N. *J. Electroanal. Chem.* **2007**, *607*, 107–112.

(32) Scherlis, D. A.; Cococcioni, M.; Sit, P. H.-L.; Marzari, N. *J. Phys. Chem. B* **2007**, *111*, 7384–7391.

(33) Kulik, H. J.; Marzari, N. *J. Chem. Phys.* **2008**, *129*, 134314.

(34) Noodleman, L. *J. Chem. Phys.* **1981**, *74*, 5737–5743.

(35) Noodleman, L.; Peng, C. Y.; Case, D. A.; Mouesca, J.-M. *Coord. Chem. Rev.* **1995**, *144*, 199–244.

(36) Noodleman, L.; Lovell, T.; Liu, T.; Himo, F.; Torres, R. A. *Curr. Opin. Chem. Biol.* **2002**, *6*, 259–273.

(37) Schreiner, E.; Nair, N. N.; Pollet, R.; Staemmler, V.; Marx, D. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 20725–20730.

(38) Nair, N. N.; Schreiner, E.; Pollet, R.; Staemmler, V.; Marx, D. *J. Chem. Theory Comput.* **2008**, *4*, 1174–1188.

(39) Marx, D.; Hutter, J. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*; Cambridge University Press: Cambridge, UK, 2009; pp 1−285.

(40) Wang, J.; Becke, A. D.; Smith, V. H., Jr. *J. Chem. Phys.* **1995**, *102*, 3477–3480.

(41) CPMD, Version 3.11; Copyright IBM Corp. 1990−2009, MPI für Festkörperforschung Stuttgart, 1997−2001.

(42) Marx, D.; Hutter, J. Ab Initio Molecular Dynamics: Theory and Implementation. In *Modern Methods and Algorithms of Quantum Chemistry*; Grotendorst, J., Ed.; John von Neumann Institute for Computing (NIC), Forschungszentrum Jülich: Germany, 2000; Vol. 3, pp 301−449.

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

(44) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396–1399.

(45) Vanderbilt, D. *Phys. Rev. B* **1990**, *41*, 7892–7895.

(46) Martyna, G. J.; Tuckerman, M. E. *J. Chem. Phys.* **1999**, *110*, 2810–2821.

(47) Adamo, C.; Barone, V.; Bencini, A.; Broer, R.; Filatov, M.; Harrison, N. M.; Illas, F.; Malrieu, J. P.; de P. R. Moreira, I. *J. Chem. Phys.* **2006**, *124*, 107101.

(48) Staemmler, V. *Theor. Chim. Acta* **1977**, *45*, 89–94.

(49) Wasilewski, J. *Int. J. Quantum Chem.* **1989**, *36*, 503–524.

(50) Meier, U.; Staemmler, V. *Theor. Chim. Acta* **1989**, *76*, 95–111.

(51) Fink, R.; Staemmler, V. *Theor. Chim. Acta* **1993**, *87*, 129–145.

(52) Illas, F.; Casanovas, J.; García-Bach, M. A.; Caballol, R.; Castell, O. *Phys. Rev. Lett.* **1993**, *71*, 3549–3552.

(53) Fink, K. *Chem. Phys.* **2006**, *326*, 297–307.

(54) Petersson, L.; Cammack, R.; Rao, K. K. *Biochim. Biophys. Acta* **1980**, *622*, 18–24.

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

(56) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635–2643.

(57) Morales, R.; Chron, M. H.; Hudry-Clergeon, G.; Petíllot, Y.; Norager, S.; Medina, M.; Frey, M. *Biochemistry* **1999**, *38*, 15764–15773.

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

(59) Henkelman, G.; Arnaldsson, A.; Jónsson, H. *Comput. Mater. Sci.* **2006**, *36*, 354–360.

(60) Giammona, D. A. An Examination of Conformational Flexibility in Porphyrins and Bulky-Ligand Binding in Myoglobin. Ph.D. Thesis, University of California, Davis, CA, 1994.

(61) Rousseau, R.; Kleinschmidt, V.; Schmitt, U. W.; Marx, D. *Angew. Chem., Int. Ed.* **2004**, *43*, 4804–4807.

(62) Mathias, G.; Marx, D. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 6980–6985.

(63) Laio, A.; VandeVonde, J.; Rothlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941–6947.

(64) Li, J.; Noodleman, L. Electronic Structure Calculations: Density Functional Methods for Spin Polarization, Charge Transfer, and Solvent Effects in Transition Metal Complexes. In *Spectroscopic Methods in Bioinorganic Chemistry, ACS Symposium Series*; Solomon, E. I., Hodgson, K. O., Eds.; American Chemical Society: Washington, DC, 1998; Vol. 692, pp 179−196.

(65) Mayerle, J. J.; Denmark, S. E.; DePamphilis, B. V.; Ibers, J. A.; Holm, R. H. *J. Am. Chem. Soc.* **1975**, *97*, 1032–1045.

(66) Palmer, G.; Dunham, W. R.; Fee, J. A.; Sands, R. H.; Iizuka, T.; Yonetani, T. *Biochim. Biophys. Acta* **1971**, *245*, 201–207.

(67) Anderson, R. E.; Dunham, W. R.; Sands, R. H.; Bearden, A. J.; Crespi, H. L. *Biochim. Biophys. Acta* **1975**, *408*, 306–318.

CT900547W

JCTC Journal of Chemical Theory and Computation

# Performance of CASPT2 and DFT for Relative Spin-State Energetics of Heme Models

Steven Vancoillie,† Hailiang Zhao,† Mariusz Radoń,‡ and Kristine Pierloot*,†

*Department of Chemistry, University of Leuven, Celestijnenlaan 200F,*
*B-3001 Heverlee-Leuven, Belgium and Faculty of Chemistry, Jagiellonian University,*
*ul. Ingardena 3, 30-060 Kraków, Poland*

Received October 26, 2009

**Abstract:** The accuracy of the relative spin-state energetics of three small $Fe^{II}$ or $Fe^{III}$ heme models from multiconfigurational perturbation theory (CASPT2) and density functional theory with selected functionals (including the recently developed M06 and M06-L functionals) was assessed by comparing with recently available coupled cluster results. While the CASPT2 calculations of spin-state energetics were found to be very accurate for the studied $Fe^{III}$ complexes (including FeP(SH), a model of the active site of cytochrome P450 in its resting state), there is a strong indication of a systematic error (around 5 kcal/mol) in favor of the high-spin state for the studied $Fe^{II}$ complexes (including FeP(Im), a model of the active site of myoglobin). A larger overstabilization of the high-spin states was observed for the M06 and M06-L functionals, up to 22 and 11 kcal/mol, respectively. None of the tested density functionals consistently provides a better accuracy than CASPT2 for all model complexes.

## 1. Introduction

Because of their important role in biological systems as the active centers or prosthetic groups of heme proteins[1] iron porphyrins have over the years received a lot of attention. The elucidation of both the geometric and the electronic structures of these compounds is of paramount importance for the detailed understanding of the complex mechanisms of biological systems.[2] An important aspect of iron porphyrins is that during the catalytic processes the spin state of the central iron changes. Both $3d^6$ Fe(II) and $3d^5$ Fe(III) porphyrins can access low-spin (LS; singlet or doublet), intermediate-spin (IS; triplet or quartet), and high-spin (HS; quintet or sextet) states. Consequently, a good description of the relative spin-state energetics is required for any method aimed at achieving a good accuracy for describing heme−ligand bond formation.

For the quantum chemical treatment of iron porphyrins, the methods that are most readily employed are either density functional theory (DFT) or multiconfigurational second-order perturbation theory (CASPT2). The former is dominant in the area of bioinorganic molecules because it accounts for electron correlation at a low computational cost, allowing for treatment of large molecules. Unfortunately, the results are significantly dependent on the functional, especially for predicting relative spin-state energetics in transition-metal complexes.[3−25] The CASPT2 method on the other hand is the only feasible ab initio alternative for DFT in cases of relatively large transition-metal compounds. This method was shown to outperform DFT[20,22] with several traditional (GGA or hybrid) functionals in two comparative studies of the HS-LS splittings of a number of six-coordinated ferrous compounds.[20,22] However, there are also indications that CASPT2 may in fact significantly overstabilize higher with respect to lower spin states at least in some (critical) cases. A typical example is ferrous porphin FeP (P = porphin), for which CASPT2 is unable to predict the correct $^3A_{2g}$ ground state. Instead, HS $^5A_{1g}$ is the calculated ground state, 5 kcal/mol below the $^3A_{2g}$ state.[23,26] In this respect, it should be noted that the CASPT2 method, in its present implementation in the MOLCAS 7.x software,[27] already includes in its zeroth-order Hamiltonian $\hat{H}^{(0)}$ an ionization potential−electronic affinity (IPEA) shift technique to properly discriminate the HS and LS states. Without this shift, the error on the

* Corresponding author fax: +32 16 32 79 92; e-mail: Kristin.Pierloot@chem.kuleuven.be.

† University of Leuven.

‡ Jagiellonian University.

10.1021/ct900567c © 2010 American Chemical Society
Published on Web 01/06/2010

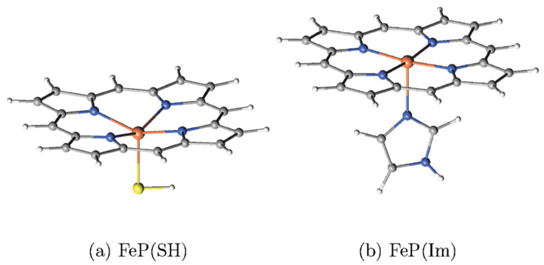

(a) FeP(SH) (b) FeP(Im)

*Figure 1.* Molecular structures of heme models.

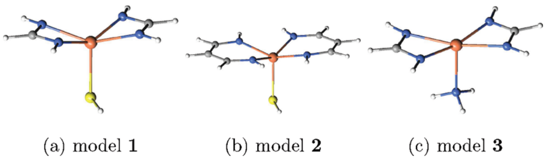

(a) model **1** (b) model **2** (c) model **3**

*Figure 2.* Molecular structures of the small models.

$^5A_{1g}$−$^3A_{2g}$ gap in FeP is significantly larger, up to 10 kcal/mol.[26] In the original implementation of CASPT2 in MOLCAS, the IPEA shift was absent. However, already in one of the first systematic test studies of the method it was shown that the original $\hat{H}^{(0)}$ would systematically favor HS over LS states.[28,29] A first remedy was formulated in the so-called $g1$, $g2$, and $g3$ modifications of $\hat{H}^{(0)}$,[30] but later the IPEA-modified $\hat{H}^{(0)}$ was introduced[31] and became the standard zeroth-order Hamiltonian in MOLCAS 6.4. The standard IPEA shift was set to 0.25 au, based on systematic tests of dissociation, ionization, and excitation energies in atoms and simple molecules. However, in a recent CASPT2 study on spin-cross-over complexes with Fe(II)N$_6$ architecture it was suggested[32] that a shift of 0.25 au is in fact too small to properly describe the adiabatic HS−LS gap in these systems, and a shift of 0.5−0.7 au was proposed instead.

In a recent study by Oláh and Harvey, the performance of several popular DFT functionals for treating NO bonding to heme groups with ferric or ferrous iron was investigated.[25] To this end, DFT calculations were performed on FeP(Im) and FeP(SH) (Figure 1). The latter models the cystein-ligated Fe(III) heme group which is commonly found in cytochrome P450s and other hemoproteins, while the former serves as a model for histidine-ligated Fe(II) porphyrins found in the active site of many enzymes, e.g., in myoglobin. In order to assess the accuracy of the DFT functionals for the spin-state energetics a series of benchmark calculations was performed by means of the CCSD(T) method and different basis sets. Because this was not possible for the FeP(SH) and FeP(Im) molecules, three smaller model systems were used using two chelating amidine ligands instead of the full porphyrin ring (Figure 2).

The main purpose of the present study is to test the accuracy/error of CASPT2 for describing the spin-state energetics in ferrous and ferric porphyrins. To this end, CASSCF/CASPT2 calculations with different basis sets were first performed on the small heme models introduced by Oláh and Harvey so as to compare to their CCSD(T) benchmark results. The role played by the IPEA shift in $\hat{H}^{(0)}$ was investigated by performing test calculations where this shift was increased to 0.5 au. As a second point of interest, in a continued search for improved exchange-correlation functionals for the property at hand we decided to include in this work also some DFT test calculations, in particular with the M06 and M06-L exchange-correlation functionals. Both are part of the recently introduced Minnesota 2006 suite of exchange-correlation functionals.[33−35] M06-L is a local functional, while M06 is a hybrid functional which was parametrized including both transition metals and nonmetals. Both Minnesota functionals include also the density of electron kinetics energy, as characteristic of so-called meta-GGA functionals.[36] In an extensive series of test calculations, these two functionals were shown to perform well for organometallic and inorganometallic thermochemistry.[35] In this work, their performance for spin-state energetics was tested, both for the three small complexes and for the more realistic models FeP(SH) and FeP(Im). More traditional functionals were also employed for comparison, including some common hybrid (B3LYP and B3LYP*),[5,37] pure (BP86, OLYP),[38−40] meta-GGA (TPSS),[36] and hybrid-meta-GGA (TPSSh)[36] functionals.

## 2. Computational Details

All CASSCF[41]/CASPT2[42,43] and some DFT (Minnesota 2006 class of functionals[33−35]) calculations were performed with the MOLCAS 7.4 package[27] using a Cholesky decomposition technique[44] for approximating the two-electron integrals, with the convergence threshold set to $10^{-6}$ au. DFT calculations for the other functionals were performed with Gaussian 03[45] (B3LYP, B3LYP*, OLYP, BP86, TPSS) or Gaussian 09[46] (TPSSh). In all calculations scalar relativistic effects were included via the second-order Douglas−Kroll−Hess transformation.[47] All DFT calculations were done using the spin-unrestricted formalism. In all CASPT2 calculations, an imaginary level shift of 0.1 au was used to improve convergence and avoid intruder states. These calculations were performed with either the default IPEA shift of 0.25 au or an increased shift of 0.5 au. Core electrons were kept frozen during the CASPT2 step. For the small heme models, the Fe(3s,3p) electrons were also not included in the correlation, in order to compare to the CASPT2 results to the CCSD(T) results from Oláh and Harvey,[25] which did not include these electrons either. For the larger models the Fe(3s,3p) electrons were included, so as to to be able to compare the results obtained here to our previous CASPT2 results for FeP(Im).[23]

For all five model complexes (Figure 1) single-point CASPT2 and DFT calculations were performed on the B3PW91-optimized geometries from ref 25 (where they were used for the CCSD(T) calculations). Two different types of basis sets were used. The first, correlation-consistent type basis sets, is the same as those used previously for the CCSD(T) calculations. They consist of the Douglas−Kroll recontraction[48] of the cc-pVDZ, cc-pVTZ, and cc-pVQZ basis sets by Dunning et al.[49] for H, C, N, O, and S atoms, combined with cc-pVTZ and cc-pVQZ basis sets for iron, developed by Balabanov and Peterson, also in forms adapted for use with Douglas−Kroll one-electron integrals.[50] Three combinations were used for iron and the ligand atoms, labeled A = cc-pVTZ/cc-pVDZ, B = cc-pVTZ/cc-pVTZ,

***Table 1.*** Number of Contracted Functions Included in Each of the Basis Sets Used in This Work

| basis | Fe | N | C | H | S |
|---|---|---|---|---|---|
| A = cc-pVTZ/cc-pVDZ | 7s6p4d2f1g | 3s2p1d | 3s2p1d | 2s1p | 4s3p1d |
| B = cc-pVTZ/cc-pVTZ | 7s6p4d2f1g | 4s3p2d1f | 4s3p2d1f | 3s2p1d | 5s4p2d1f |
| C = cc-pVQZ/cc-pVDZ | 8s7p5d3f2g1h | 3s2p1d | 3s2p1d | 2s1p | 4s3p1d |
| I = ANO-rcc/ANO-s | 7s6p5d2f1g | 4s3p1d | 3s2p1d | 2s | 5s4p2d |
| II = ANO-rcc/ANO-rcc | 7s6p5d3f2g1h | 4s3p2d1f | 4s3p1d | 3s1p | 5s4p3d2f |
| III = ANO-rcc/ANO-rcc | 10s9p8d6f4g2h | 5s4p3d2f1g | 4s3p2d1f | 3s2p1d | 6s5p4d3f2g |

***Table 2.*** Relative Energy (kcal/mol) of the Low- and Intermediate-Spin States with Respect to the High-Spin State of the Small Heme Models **1** ($^6$A′), **2** ($^6$A′), and **3** ($^5$A″) from CCSD(T) and CASPT2 Calculations

| | | CCSD(T)[a] | | | | CASPT2 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A | B | C | ∞ | A | B | C | I | II | III | III[b] |
| **1** | $^2$A′ | 37.8 | 38.5 | 33.2 | 29.8 | 35.3 | 35.8 | 30.4 | 33.0 | 30.6 | 30.1 | 25.7 |
| | $^4$A″ | 18.8 | 20.4 | 16.0 | 13.9 | 19.2 | 20.4 | 16.0 | 18.5 | 17.2 | 17.1 | 15.4 |
| **2** | $^2$A″ | 0.6 | | −4.1 | −7.4 | 0.6 | 1.6 | −4.7 | −1.2 | −3.1 | −4.7 | −8.5 |
| | $^4$A″ | 7.4 | | 4.3 | 2.2 | 6.9 | 7.3 | 3.7 | 5.5 | 4.1 | 3.3 | 1.9 |
| **3** | $^1$A′ | 37.2 | | 32.9 | 30.4 | 40.2 | 40.3 | 36.1 | 38.6 | 37.3 | 35.9 | 33.8 |
| | $^3$A′ | 21.8 | | 19.0 | 17.3 | 23.9 | 24.5 | 21.1 | 23.8 | 22.2 | 21.5 | 20.7 |

[a] Taken from ref 25. [b] IPEA shift of 0.5 au.

and C = cc-pVQZ/cc-pVDZ. Calculations were also performed with a second type of atomic natural orbital (ANO) basis sets. In these sets, labeled I−III, ANO-rcc basis sets on iron were combined with either ANO-s (basis I) or ANO-rcc (basis II, III) on the other atoms.[51,52] All basis set contractions are given in Table 1. All six basis set combinations were used for the calculations on the small models **1**−**3**. Calculations on the two larger models, FeP(SH) and FeP(Im), were only performed with basis C and II.

The active space used in the CASSCF and CASPT2 calculations was constructed by starting from a distribution of 6 or 5 electrons in the five Fe 3d orbitals and adding a second 3d′ shell to describe the double-shell effect.[53] To account for nondynamic correlation effects associated with covalent Fe−ligand interactions, a doubly occupied bonding $\sigma$ (Fe−N$_{\text{ring}}$) orbital is added, with an additional bonding $\sigma$ (Fe−NH$_3$) orbital for model **3** and FeP(Im), leading to a total of 10 electrons in 12 orbitals, and two additional bonding $\sigma$ (Fe−SH) and $\pi$ (Fe−SH) orbitals for models **1**, **2**, and FeP(SH), leading to a total of 11 electrons in 13 orbitals. In cases where orbitals from the 3d′ shell correlating empty 3d orbitals were found to rotate out of the active space, they were removed. The empty 3d orbitals themselves were kept active and also maintained their character during the orbital optimization process.

## 3. Results and Discussion

**3.1. Small Heme Models 1, 2, and 3.** All three models consist of an iron atom surrounded by two bidentate amidine ligands, each bonded to iron by 2 nitrogen atoms, and a third axial ligand, either the sulfur-bonded SH (Fe$^{\text{III}}$ complexes **1** and **2**) or the nitrogen-bonded NH$_3$ (Fe$^{\text{II}}$ complex **3**) ligand. Interaction with these ligands gives rise to a splitting of the Fe 3d orbitals, forming pairs of bonding−antibonding molecular orbitals: $\sigma$, $\sigma$* (Fe d$_{xy}$−N; Fe d$_{z^2}$−NH$_3$/SH) and $\pi$, $\pi$* (Fe d$_{yz}$−SH). The lowest states for each of the different spin multiplicities are $^6$A′, $^4$A″, and $^2$A′ for model **1**, $^6$A′, $^4$A″, and $^2$A″ for model **2**, and $^5$A″, $^3$A′, and $^1$A′ for model **3**. The principal CASSCF configurations (occupations of Fe 3d orbitals) for these states are given in the Supporting Information.

The CCSD(T), CASPT2, and DFT relative energies for the different spin states of the three small heme models are collected in Table 2 for each of the basis sets. Since the DFT results for different basis sets were very similar, only the values of basis set C (cc-pVQZ/cc-pVDZ on iron/ligand) are included in the discussion.

As can be seen from Table 2, all relative energies obtained from either CCSD(T) and CASPT2 using the same basis sets (A−C) are in close agreement, with absolute differences below 3 kcal/mol. This is particularly the case for the Fe(III) model **2**, with differences less than 1 kcal/mol for all states/basis sets. The same is also true for the $^4$A″−$^6$A′ energy difference in the smaller Fe(III) model **1**. On the other hand, the low-spin state $^2$A″ in this case seems to be overstabilized by about 2 kcal/mol by CASPT2 (as compared to CCSD(T)). This is opposed to the results obtained for the Fe(II) model **3**, for which CASPT2 systematically favors the high-spin $^5$A′ ground state, giving rise to energy differences which are larger by 2−3 kcal/mol than the corresponding CCSD(T) results.

Table 2 also includes an extrapolation of the CCSD(T) results to the infinite basis set limit, based on the results obtained with basis A and C.[25] The choice of only these two sets was based on the observation that the quality of the basis set on iron (cc-pVQZ in basis C versus cc-pVTZ in basis A, combined with the same basis sets on the amidines) influences the spin-state energies of model **1** to a much larger extent than the size of the basis set on the amidine ligands (cc-pVTZ in B versus cc-pVDZ in A, combined with the same basis set on iron). The same is also observed for the CASPT2 results for all three models. Between basis A and C, the relative energies of the high- and intermediate-spin states systematically improve by 3−5 kcal/mol. On the other hand, going from basis A to basis B has a much smaller, and opposite, effect.

In a previous study we made use of ANO-type basis sets, contracted as in basis I and II, to study the bonding of CO,

***Table 3.*** Relative Energy (kcal/mol) of the Low- and Intermediate-Spin States with Respect to the High-Spin State of the Small Heme Models **1** ($^6$A′), **2** ($^6$A′), and **3** ($^5$A″) from DFT Calculations (basis C)

| | | B3LYP | B3LYP* | OLYP | BP86 | TPSS | TPSSh | M06 | M06-L | CCSD(T)(∞) |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | $^2$A′ | 21.4 | 17.5 | 22.7 | 6.5 | 4.9 | 12.3 | 43.8 | 35.0 | 29.8 |
| | $^4$A″ | 6.4 | 3.5 | 6.5 | −4.8 | −4.1 | 1.3 | 19.0 | 16.8 | 13.9 |
| **2** | $^2$A″ | −5.8 | −12.2 | −12.4 | −31.1 | −29.4 | −17.8 | 14.9 | 3.7 | −7.4 |
| | $^4$A″ | −4.2 | −7.4 | −3.7 | −16.2 | −14.8 | −9.2 | 8.8 | 7.3 | 2.2 |
| **3** | $^1$A′ | 26.3 | 22.3 | 27.2 | 12.6 | 12.1 | 20.0 | 35.9 | 27.1 | 30.4 |
| | $^3$A′ | 10.1 | 7.0 | 9.1 | −1.8 | −2.1 | 4.1 | 21.0 | 15.9 | 17.3 |

NO, and $O_2$ to Fe(II) heme systems.[23] So as to be able to compare the results from that study with the present results, the CASPT2 calculations on models **1**−**3** were also performed with the same basis sets. Furthermore, the smaller size of the present models also allows us to extend the contraction of the ANO-rcc basis sets even further, thus giving basis III (see Table 1). The size of the contracted basis set I is comparable to basis sets A and B (actually the ANO contraction on iron contains one more *d* function than the cc-pVTZ set). Still, as Table 2 indicates, the relative energies obtained with basis I are superior to these correlation-consistent sets, predicting, for example, the correct $^2$A″ ground state for model **2**. The ANO-rcc basis sets II and III contain many more functions on the ligands, yet for the metal they are comparable (either slightly smaller or larger) to cc-pVQZ. The fact that these three basis sets give similar relative energies is another confirmation that the ligand basis set size is of minor importance for the property at hand. The results obtained with basis III should be close to the basis set limit for this ANO-rcc basis set on iron. Still, with respect to the CCSD(T) infinite basis set limit, the CASPT2 results in Table 2 are invariably too high, indicating that in all cases the high-spin state is overstabilized with respect to the low- and intermediate-spin states. This is primarily a basis set effect, pointing to the need of extremely large basis sets on the metal, both primitive and contracted, for an accurate description of the relative spin-state energetics in transition-metal complexes.

The difference between the best CASPT2 results and the CCSD(T)(∞) results is largest for the Fe(II) heme model **3**: 4−6 kcal/mol. Here, it should partly be traced back to an inherent tendency of CASPT2 to overstabilize higher with respect to lower spin states in ferrous complexes. Let us note that a comparable error (at least 5 kcal/mol, not accounting for ZPVE) was found in our previous study of the $^5A_{1g}$−$^3A_{2g}$ splitting in the four-coordinate FeP complex, calculated there with basis II[23] and in our recent study of the doublet−quartet transition in some $\{FeNO\}^7$ complexes.[54] Similar errors were also found in a recent CASPT2 study of the adiabatic quintet−singlet splitting in a number of ferrous pseudo-octahedral $FeN_6$ complexes.[32] On the basis of the results of their study the authors proposed replacing the standard IPEA shift of 0.25 au in the zeroth-order Hamiltonian by a larger value, 0.50−0.70 au, for these specific adiabatic gap calculations. In order to investigate whether their proposal can be made more general also for the ferrous and ferric complexes considered here, we decided to repeat the CASPT2 calculations with basis III using an IPEA shift of 0.5 au. The results are given in the rightmost column of the CASPT2 data in Table 2. As one can see, the success of the shift operation is not unequivocal. For the Fe(II) model **3**, the relative energies are indeed shifted toward the CCSD(T) results. However, the effect of the IPEA shift increase is too limited, only 0.8 kcal/mol for the $^3$A′ and 2.1 kcal/mol for the $^1$A′ state. The latter value is only about one-half of what was found for the series of seven $FeN_6$ complexes studied in ref 32, showing a systematic increase of their adiabatic HS−LS splitting with 3.6−4.3 kcal/mol with an increase of the IPEA shift to 0.5 au, and further with 3.3−3.8 kcal/mol with a further increase to 0.75 au On the other hand, for the ferric models **1** and **2**, where the original CASPT2 results proved to be excellent, the effect of increasing IPEA is significantly larger than for model **3**, up to 4.4 kcal/mol for the $^2$A′ state, but it is obviously deteriorating. This set of results, although limited, seems to indicate that changing the IPEA shift in the CASPT2 $\hat{H}^{(0)}$ should be done with care, if at all, as the lack of systematics in the approach may easily turn CASPT2 into a semiempirical method.

Turning next to the DFT results (given in Table 3) we first note that, since for this method basis set convergence should be much faster than for traditional correlated ab initio methods, basis C should be large enough to provide results close to the basis set limit. The DFT results in Table 3 should therefore rather be confronted with the infinite basis set CCSD(T) results rather than with the results obtained with basis C. The CCSD(T)(∞) results are therefore included as a reference in the rightmost column of Table 3. A first look at this table already shows that all DFT results significantly differ from the CCSD(T)(∞) results. Notably, none of the tested functionals yields an agreement better than ±5 kcal/mol for all three models simultaneously, not rarely the errors exceeding 10 kcal/mol. The DFT energetics are substantially dependent on the functional, in a qualitative agreement with a trend already recognized in the literature: pure functionals systematically overstabilizing low-spin relative to high-spin states, and the hybrid functionals favoring high spins more as the contribution of Hartree−Fock exchange is increased.[5,16,17,19,20,22,23] Herein, this trend is most clearly evidenced by comparing the TPSS (pure) with the TPSSh (hybrid) results or the B3LYP* (hybrid, 15% of exact exchange) with the B3LYP results (hybrid, 20%). The (meta-GGA) TPSS functional in fact behaves in a very similar way as the BP86 functional. In contrast, the OLYP functional yields significantly different spin-state energetics than the traditional pure functionals (here epitomized by BP86), in agreement with previous observations.[23,55] In fact, the present OLYP results are similar to B3LYP* or B3LYP ones. Going to the recently introduced Minnesota functionals (M06 and M06-L) one can note that they both predict a much higher energy of the IS state (with respect to the HS state) than

***Table 4.*** Relative Energy (kcal/mol) of the Low- and Intermediate-Spin States with Respect to the High-Spin State of the Large Heme Models FeP(SH) ($^6A'$) and FeP(Im) ($^5A'$) from CASPT2 Calculations

| | | CASPT2 | | | | | |
|---|---|---|---|---|---|---|---|
| basis | | C | II | C[a] | II[a] | II[b] | II[c] |
| FeP(SH) | $^2A''$ | 7.0 | 6.8 | 4.1 | 3.6 | | |
| | $^4A''$ | 10.2 | 9.2 | 8.7 | 7.5 | | |
| FeP(Im) | $^1A'$ | 11.2 | 10.6 | 14.4 | 13.7 | 14.0 | 13.0 |
| | $^3A''$ | 10.3 | 9.5 | 10.3 | 9.5 | 8.6 | 8.5 |

[a] With Fe(3s,3p) core electron correlation included. [b] From ref 23 using PBE0 structures (and omitting the ZPVE contribution). [c] From ref 23 using BP86 structures (and omitting the ZPVE contribution).

any other of the tested functionals; the same holds true for the energy of the LS state (with respect to the HS state), except for model 3 for M06-L. Two more observations can be made for the Minnesota functionals. First, the local functional M06-L in general performs better than the M06 functional. The latter functional systematically overstabilizes the high-spin with respect to the intermediate-spin and even more with respect to the low-spin state, with errors amounting up to 22 kcal/mol. A second observation is that the ferric complexes, with model **2** in particular, are described considerably worse than the ferrous complex **3**. For the latter model, quite reasonable results are in fact obtained with both functionals, M06 overestimating the relative energies by 2−4 kcal/mol and M06-L underestimating them with 1−3 kcal/mol. In contrast, for model **2**, both functionals severely overestimate the stability of the high-spin $^6A'$ state, incorrectly predicting this state to be the ground state. With M06, the ordering of the two other states $^2A''$ and $^4A''$ is also not correctly reproduced.

**3.2. Large Heme Models FeP(SH) and FeP(Im).** Interaction of the iron atom with the surrounding porphyrin and axial ligand gives rise to a splitting of the d orbitals similar to that of the small model complexes, thus forming pairs of bonding−antibonding molecular orbitals: $\sigma$, $\sigma$* (Fe $d_{xy}$−P; Fe $d_{z^2}$−Im/SH) and $\pi$, $\pi$* (Fe $d_{xz}$−SH). The lowest states for each spin are $^6A'$, $^4A''$, and $^2A''$ for FeP(SH) and $^5A'$, $^3A''$, and $^1A'$ for FeP(Im). The relative energies of the different spin states are presented in Table 4.

Compared to the small models **1** and **3**, the CASPT2 relative energies of the LS and IS states of FeP(SH) and FeP(Im) are much lower, though the HS state remains the ground state for both of the large heme models. The $^2A''$ state of FeP(SH) is lower in energy than the $^4A''$ state, but unlike in model **2** it remains above the $^6A'$ state. The relative energies of the $^1A'$ and $^3A''$ states of FeP(Im) are similar, about 10 kcal/mol above the $^5A'$ ground state. Going from basis C to the ANO basis set II, an increased stabilization of the low- and intermediate-spin states is observed for both complexes. The same basis set was used by Radoń and Pierloot[23] to calculate the relative energies of the different spin states of FeP and FeP(Im), using the same active space but different geometries, that is PBE0 and BP86 optimized and Fe(3s,3p) core correlation. The results for FeP(Im) were added to Table 4. In order to compare with these results, extra CASPT2 calculations with Fe(3s,3p) core correlation were performed for basis C and II. We can see that the energies are very similar to the values obtained here using the B3PW91-optimized structures from ref 25. Both singlet and triplet excited-state energies are within a range of 0.5 kcal/mol around 13.5 and 9.0 kcal/mol, respectively. We also note that the effect of Fe(3s,3p) correlation is rather significant (to about 3 kcal/mol) and opposite for both complexes: in FeP(Im) the LS and IS states are stabilized with respect to the HS state, while in FeP(SH) the LS state is destabilized and the IS state is unaffected. In view of these irregularities we believe that the 3s,3p electrons should be preferably correlated in ab initio calculations of spin-state energetics in first-row transition-metal complexes.

Going to the DFT results (Table 5), we first note that they compare well to the previous DFT calculations on the heme complexes (FeP(Im) and FeP(SH)) available in the literature.[23,25,56−59] As could be expected, a similar behavior of the different DFT functionals is found for the large heme models as for the small models **1**−**3**. In this respect we note again that the energies of the LS and IS states (with respect to the HS state) are lower with the pure (BP86, TPSS) than with the hybrid functionals (B3LYP, B3LYP *,TPSSh), with OLYP giving results close to the hybrid functionals. A more concrete discussion of the accuracy of the DFT results for FeP(Im) and FeP(SH) requires a comparison to some reliable reference results. A reasonable estimate of the spin splitting in these complexes may be obtained from the available CASPT2 results (Table 4), assuming that this method has similar errors for the large models as were noted for models **1**−**3**. Our “best” estimate of the splittings is given in the rightmost column of Table 5. As one can see, the results of the hybrid functionals (B3LYP, B3LYP*) and the pure OLYP functional are reasonably close to this estimate, with errors typically ≤6 kcal/mol. The results obtained from the pure TPSS and BP86 functionals are worse and again (see also Table 3) very similar, overstabilizing the IS and LS states by 12−16 kcal/mol with respect to the HS state for FeP(SH) and by 16−20 kcal/mol for FeP(Im). The error is reduced in TPSSh; however, somewhat suprisingly, this method now overshoots the relative energy of the $^4A''$ state in FeP(SH).

A more important question is whether M06 or M06-L can outperform the traditional functionals. This is obviously not the case. Similar to the small models, we find that both functionals tend to overstabilize the HS with respect to the IS and even more with respect to the LS state. The M06-L functional again clearly outperforms M06. As was also found for models **1**−**3** both Minnesota functionals describe the ferrous complex FeP(Im) much better than the ferric complex FeP(SH). In fact, for the former (ferrous) complex, the M06-L functional yields quite accurate spin-state energetics. This success should however be put into perspective, given the much larger error of more than 10 kcal/mol obtained with this functional for the $^2A''$ state in FeP(SH).

## 4. Conclusion

In this investigation we attempted to benchmark the accuracy of CASPT2 and selected DFT methods for spin-state energetics of selected heme complexes of Fe(II) and Fe(III),

***Table 5.*** Relative Energy (kcal/mol) of the Low- and Intermediate-Spin States with Respect to the High-Spin State of the Large Heme Models FeP(SH) ($^6A'$) and FeP(Im) ($^5A'$) from DFT Calculations with Basis C

| | | B3LYP | B3LYP* | OLYP | BP86 | TPSS | TPSSh | M06 | M06-L | "best" estimate[a] |
|---|---|---|---|---|---|---|---|---|---|---|
| FeP(SH) | $^2A''$ | 3.8 | −1.2 | 1.9 | −16.3 | −15.8 | −6.1 | 24.1 | 14.5 | −0.7−3.5 |
| | $^4A''$ | 2.0 | −1.0 | 2.6 | −9.5 | −8.2 | 10.3 | 14.9 | 12.0 | 4.2−7.2 |
| FeP(Im) | $^1A'$ | 6.5 | 1.1 | 4.2 | −14.4 | −13.1 | −2.5 | 17.7 | 5.6 | 6.8−8.7 |
| | $^3A''$ | −0.6 | −3.6 | −1.6 | −12.4 | −12.1 | −6.0 | 10.7 | 4.8 | 4.6−6.5 |

[a] CASPT2 results from Table 4, corrected for the errors found for the small models **1**−**3**. The "best" estimates were obtained by adding to the CASPT2 results (including (3s,3p) correlation) the difference between the CASPT2 and CCSD(T)(∞) calculations for the corresponding small models. When using either the results from basis C and II and either models **1** or **2** for FeP(SH) the range of values given in the rightmost column is obtained.

including the models of the active site of cytochrome P450 (in its resting state) and myoglobin. While there are no benchmark results for the large heme complexes (FeP(Im) and FeP(SH)), the CCSD(T) calculations for their smaller mimics (complexes **1**−**3**) were recently published.[25] It must be mentioned here that although we believe in a high accuracy of the reference CCSD(T) data, one should not forget that they also might be subject to errors related to the absence of higher order terms in the CC expansion, the multiconfigurational character of the wave function, or the basis set extrapolation procedure.

The performance of CASPT2 is excellent for the ferric complexes **1** and **2** (an error within chemical accuracy) and worse for the ferrous complex **3**, thereby confirming previous suspicions that this method overstabilizes the high-spin state in some Fe(II) complexes.[23,32,54] Let us note that all these problematic cases concern the ligand-field transitions from a nonbonding ($d_{x^2-y^2}$) to an antibonding ($d_{xy}$) orbital of Fe. The error can be estimated as slightly above 5 kcal/mol but definitely less than 10 kcal/mol. It should be stressed that CASPT2 errors of this size are rather exceptional in transition-metal chemistry, even for $d_{x^2-y^2} \rightarrow d_{xy}$ transitions. This is illustrated by the excellent performance of this method for complexes **1** and **2** as well as by previous numerous applications. Unfortunately, it seems that for the presently studied complexes the error cannot be easily reduced by changing the zeroth-order Hamiltonian of CASPT2 (increasing the IPEA shift), as was successful in the previous study of some FeN$_6$ complexes. This indicates that playing with the zeroth-order Hamiltonian of CASPT2 should be done with care (and preferably avoided); otherwise, one may easily turn this ab initio method into a de facto semiempirical approach.

The present investigation also explored the accuracy of several DFT methods. With respect to the extrapolated CCSD(T) reference results, all tested functionals lead to errors above 5 kcal/mol at least for one complex. This is also true for the recently introduced M06 and M06-L functionals from the Minnesota 2006 set, among which the second one (M06-L) performs much better. However, its overall performance for the complexes studied here is not any better than of some more traditional functionals, like B3LYP or OLYP.

In summary, the present investigation confirms an overall high accuracy of CASPT2, although a systematic error of CASPT2 for the ferrous complexes (possibly one of the most difficult cases for CASPT2 calculations) is definitely pinpointed. We believe that CASPT2 calculations on transition-metal systems will become more and more common and useful in the field of bioinorganic chemistry.

**Acknowledgment.** This investigation was supported by grants from the Flemish Science Foundation (FWO) and the Concerted Research Action of the Flemish Government (GOA) and by the Polish State Ministry of Science and Higher Education (MNiSW) from the funds for scientific research (2009−2010). The computational grant from the Academic Computer Center CYFRONET AGH is also acknowledged.

**Supporting Information Available:** Plots of the active orbitals for the small models **1** and **3** and the large models FeP(Im) and FeP(SH); occupation numbers of the Fe 3d orbitals in the lowest state of each spin and symmetry in the different models; graphical representation of the differences obtained with different methods for the (IS-HS) and (LS-HS) splittings with respect to CCSD(T)(∞)). This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) In *Iron Porphyrins*; Lever, A. B. P., Gray, H. B., Eds.; Addison-Wesley: Reading, MA, 1983.

(2) In *The Porphyrin Handbook*; Kadish, K., Smith, K., Guillard, R., Eds.; Academic Press: New York, 2000.

(3) Kozlowski, P. M.; Spiro, T. G.; Bérces, A.; Zgierski, M. Z. *J. Phys. Chem. B* **1998**, *102*, 2603–2608.

(4) Paulsen, H.; Duelund, L.; Winkler, H.; Toftlund, H.; Trautwein, A. X. *Inorg. Chem.* **2001**, *40*, 2201–2203.

(5) Reiher, M.; Salomon, O.; Hess, B. A. *Theor. Chem. Acc.* **2001**, *107*, 48–55.

(6) Reiher, M. *Inorg. Chem.* **2002**, *41*, 6928–6935.

(7) Salomon, O.; Reiher, M.; Hess, B. A. *J. Chem. Phys.* **2002**, *117*, 4729–4737.

(8) Baranović, G. *Chem. Phys. Lett.* **2003**, *369*, 668–672.

(9) Ghosh, A.; Vangberg, T.; Gonzalez, E.; Taylor, P. R. *J. Porphyrins Phthalocyanins* **2001**, *5*, 345–356.

(10) Ghosh, A.; Persson, B. J.; Taylor, P. R. *J. Biol. Inorg. Chem.* **2003**, *8*, 507–511.

(11) Ghosh, A.; Tangen, E.; Ryeng, H.; Taylor, P. R. *Eur. J. Inorg. Chem.* **2004**, 4555–4560.

(12) Paulsen, H.; Trautwein, A. X. *Top. Curr. Chem.* **2004**, *235*, 197–219.

(13) Harvey, J. N. *Struct. Bonding (Berlin)* **2004**, *112*, 151–183.

(14) Deeth, R. J.; Fey, N. *J. Comput. Chem.* **2004**, *25*, 1840–1848.

(15) Swart, M.; Groenhof, A. R.; Ehlers, A. W.; Lammertsma, K. *J. Phys. Chem. A* **2004**, *108*, 5479–5483.

(16) Fouqueau, A.; Casida, S. M. M. E.; Daku, L. M. L.; Hauser, A.; Neese, F. *J. Chem. Phys.* **2004**, *120*, 9473–9486.

(17) Fouqueau, A.; Casida, M. E.; Daku, L. M. L.; Hauser, A.; Neese, F. *J. Chem. Phys.* **2005**, *122*, 044110.

(18) Daku, L. M. L.; Vargas, A.; Hauser, A.; Fouqueau, A.; Cassida, M. E. *Chem. Phys. Chem.* **2005**, *6*, 1393–1410.

(19) Ganzenmüller, G.; Berkane, N.; Fouqueau, A.; Casida, M. E. *J. Chem. Phys.* **2005**, *122*, 234321.

(20) Pierloot, K.; Vancoillie, S. *J. Chem. Phys.* **2006**, *125*, 124303.

(21) Strickland, N.; Harvey, J. N. *J. Phys. Chem. B* **2007**, *111*, 841–852.

(22) Pierloot, K.; Vancoillie, S. *J. Chem. Phys.* **2008**, *128*, 034104.

(23) Radoń, M.; Pierloot, K. *J. Phys. Chem. A* **2008**, *112*, 11824–11832.

(24) Khvostichenko, D.; Choi, A.; Boulatov, R. *J. Phys. Chem. A* **2008**, *112*, 3700–3711.

(25) Oláh, J.; Harvey, J. N. *J. Phys. Chem. A* **2009**, *113*, 7338–7345.

(26) Pierloot, K. *Mol. Phys.* **2003**, *101*, 2083–2094.

(27) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222–239.

(28) Andersson, K.; Roos, B. O. *Int. J. Quantum Chem.* **1993**, *45*, 591–607.

(29) Roos, B. O.; Andersson, K.; Fülscher, M. P.; Malmqvist, P.-Å.; Serrano-Andrés, L.; Pierloot, K.; Merchán, M. Multiconfigurational Perturbation Theory: Applications in Electronic Spectroscopy. In *Advances in Chemical Physics: New Methods in Computational Quantum Mechanics*; Prigogine, I., Rice, S. A., Eds.; John Wiley & Sons: New York, 1996;, Vol. XCIII, pp 219−332.

(30) Andersson, K. *Theor. Chim. Acta* **1995**, *91*, 31–46.

(31) Ghigo, G.; Roos, B. O.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **2004**, *396*, 142–149.

(32) Kepenekian, M.; Robert, V.; Le Guennic, B. *J. Chem. Phys.* **2009**, *131*, 114702.

(33) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.

(34) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126–13130.

(35) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(36) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401–146404.

(37) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

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

(39) Perdew, J. P. *Phys. Rev. B* **1986**, *33*, 8822–8824.

(40) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403–412.

(41) Malmqvist, P.-Å.; Rendell, A.; Roos, B. O. *J. Phys. Chem.* **1990**, *94*, 5477–5482.

(42) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. *J. Chem. Phys.* **1990**, *94*, 5483.

(43) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218.

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

(45) Frisch, M. J. *Gaussian 03*, Revision C.02; Gaussian Inc., Wallingford, CT, 2004.

(46) Frisch, M. J. *Gaussian 09*, Revision A.02; Gaussian Inc.: Wallingford, CT, 2009.

(47) Reiher, M.; Wolf, A. *J. Chem. Phys.* **2004**, *121*, 10945–10956.

(48) de Jong, W. A.; Harrison, R. J.; Dixon, D. A. *J. Chem. Phys.* **2001**, *114*, 48–53.

(49) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(50) Balabanov, N. B.; Peterson, K. A. *J. Chem. Phys.* **2005**, *123*, 064107.

(51) Roos, B. O.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, V.; Widmark, P.-O. *J. Phys. Chem. A* **2005**, *109*, 6575–6579.

(52) Pierloot, K.; Dumez, B.; Widmark, P.-O.; Roos, B. O. *Theor. Chem. Acc.* **1995**, *90*, 87.

(53) Andersson, K.; Roos, B. O. *Chem. Phys. Lett.* **1992**, *191*, 507–514.

(54) Radoń, M.; Broclawik, E.; Pierloot, K. *J. Phys. Chem. B* **2010**, in press.

(55) Conradie, J.; Ghosh, A. *J. Phys. Chem. B* **2007**, *111*, 12621–12624.

(56) Groenhof, A. R.; Swart, M.; Ehlers, A. W.; Lammertsma, K. *J. Phys. Chem.* **2005**, *109*, 3411–3417.

(57) Shaik, S.; Kumar, D.; de Visser, S. P.; Altun, A.; Thiel, W. *Chem. Rev.* **2005**, *105*, 2279–2328.

(58) Rydberg, P.; Sigfridsson, E.; Ryde, U. *J. Biol. Inorg. Chem.* **2004**, *9*, 203–223.

(59) Strickland, N.; Harvey, J. N. *J. Phys. Chem. B* **2007**, *111*, 841–852.

CT900567C

JCTC Journal of Chemical Theory and Computation

# Relation between Free Energy Landscapes of Proteins and Dynamics

Gia G. Maisuradze, Adam Liwo, and Harold A. Scheraga*

*Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301*

Received October 29, 2009

**Abstract:** By using principal component analysis (PCA) to examine the molecular dynamics (MD) of protein folding trajectories, generated with the coarse-grained UNRES force field, for the B-domain of staphylococcal protein A and the triple $\beta$-strand WW domain from the formin binding protein 28 (FBP), we demonstrate how different free energy landscapes (FELs) and folding pathways of trajectories can be, even though they appear to be very similar by visual inspection of the time dependence of the root-mean-square deviation (rmsd). Approaches to determine the minimal dimensionality of FELs for a correct description of protein folding dynamics are discussed. The correlation between the amplitude of the fluctuations of proteins and the dimensionality of the FELs is shown. The advantage of internal-coordinate PCA over Cartesian PCA for small proteins is also illustrated.

## 1. Introduction

Protein folding is a rapid and complex process that is difficult to characterize because folding does not refer to the progressive pathway of a single conformation. Instead, it pertains to interconversions among ensembles of conformations in a back-and-forth progression from the non-native to the native state. In addition, the non-native and native states themselves can consist of large ensembles of conformations, interconverting at a rapid rate, that are characterized by basins with many minima in each state. A folding pathway is not always defined in terms of a two-state model consisting of the non-native and the native state separated by an energetically unfavorable transition state. Proteins can fold through intermediate states[1,2] or undergo one-state downhill folding.[1,3] Therefore, finding the coordinates along which the intrinsic folding pathways of biological molecules (containing thousands of degrees of freedom) can be identified still remains a challenge.

A study of free energy landscapes (FELs) provides an understanding of how proteins fold and function.[4−6] It should be noted that the FELs determined from canonical molecular dynamics (MD) simulations at temperatures significantly lower than the folding transition temperature are usually nonequilibrium landscapes because canonical simulations take very long to equilibrate. Generalized-ensemble algorithms,[7] in which walks in temperature or energy space are performed, converge much faster than canonical sampling and should be used to obtain equilibrium FELs. On the other hand, the nonequilibrium FELs resulting from canonical simulations are also valuable, because they provide condensed information about the frequency of visiting particular regions of conformational space during the simulated folding. It must be borne in mind, however, that these FELs are dependent on simulation setup parameters, such as the trajectory length, the number of trajectories run at a given temperature, and even the starting conformation(s). In this article, we discuss the FELs calculated from canonical trajectories, which, as remarked above, are generally not equilibrated. However, because we ran our calculations close to the folding transition temperatures for both proteins studied, which lowers the free energy barriers between conformational states, the FELs should be close to equilibrium FELs. Molecular dynamics (MD) simulations based on atomic[8] and coarse-grained[9] models provide the atomic- and coarse-grained-level pictures, respectively, of protein motion and the connection to the underlying FEL. The commonly used reaction coordinates [radius of gyration ($R_g$), root-mean-square deviation (rmsd) with respect to the native state, and so on] are arbitrary and do not necessarily capture the features of protein energy landscapes. To overcome these problems,

* Corresponding author e-mail: has5@cornell.edu.

10.1021/ct9005745 © 2010 American Chemical Society
Published on Web 01/13/2010

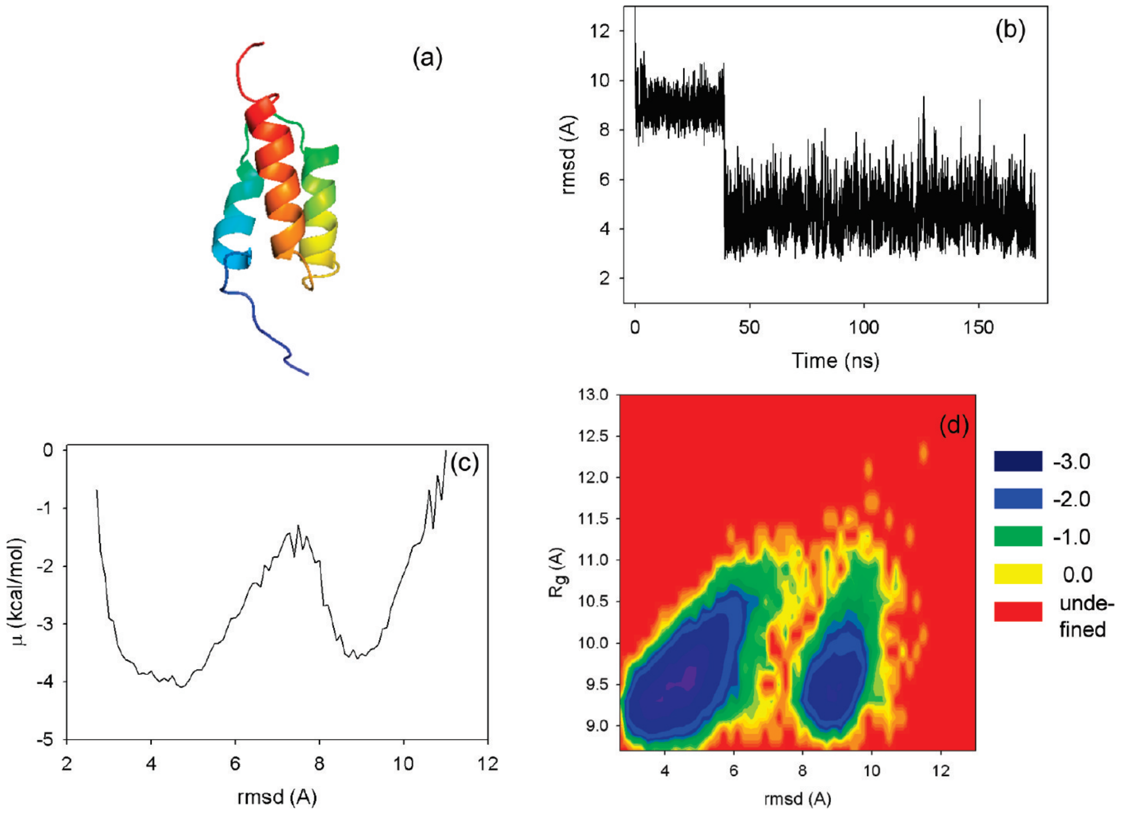


***Figure 1.*** (a) Experimental NMR structure of B-domain of staphylococcal protein A, (b) rmsd from the native structure as a function of time, (c) free energy profile (FEP) (in kcal/mol) plotted as a function of rmsd, and (d) FEL (in kcal/mol) plotted as a function of rmsd and radius of gyration for 1BDD.

many different methods have been developed over the past two decades, for example, the approaches based on transition networks,[10,11] an unprojected representation of FEL. Another frequently used method for defining reaction coordinates is a covariance-matrix-based mathematical technique, called principal component analysis (PCA),[12] that typically captures most of the total displacement from the average protein structure during a simulation with the first few principal components (PCs).

Although PCA reduces the dimensionality of a complex system dramatically, the low-dimensional [one-dimensional (1-D) or two-dimensional (2-D)] representation of an FEL does not always provide a correct picture and can lead to serious artifacts.[13,14] How complete are 1-D and 2-D FELs? How correct are the protein-folding kinetics and diffusive behavior described by 1-D and 2-D FELs? These questions were addressed in a preliminary way in our recent study.[15] An analysis of the different-dimensional FELs for a folding/unfolding trajectory of the B-domain of staphylococcal protein A (1BDD), a 46-residue three-α-helical protein,[16] showed that the low-dimensional FELs are not always sufficient for the description of folding/unfolding processes.[15]

In the present work, we continue our study of the relation between FELs and a correct description of folding dynamics. For this purpose, we ran 110 trajectories of canonical MD simulations with the coarse-grained united-residue (UNRES) force field[17−22] at different temperatures for both 1BDD and the 37-residue triple-$\beta$-stranded WW domain from the formin binding protein 28 (FBP) (1E0L),[23] and we investigated one folding trajectory in detail for each protein. Based on the rmsd's as functions of time, the behaviors of the two proteins are simple and similar to each other (panels labeled b in Figures 1 and 2). In particular, both proteins fold directly from the unfolded state to the nativelike conformation and remain there for the rest of the simulations.

In our recent preliminary study,[15] we investigated a more complex trajectory of 1BDD in which frequent transitions between the native and unfolded structures occurred; consequently, the question arises as to whether the complexity of the pathway could be the reason that a one- or two-dimensional FEL sometimes fails to describe the behavior of the system. We demonstrate how to determine the lowest-dimensional FEL for each trajectory that can describe the folding dynamics correctly and show the correlation between the percentage of the fluctuations captured by the PCs and the dimensionality of the FEL necessary for a correct description of folding/unfolding processes. We also demonstrate that the FELs of coarse-grained folding trajectories obtained from internal-coordinate PCA[24−27] are more rugged than those constructed by traditional Cartesian PCA.

It should be noted that both 1BDD and 1E0L proteins have been the subject of extensive theoretical[8,9,15,27−41] and experimental[2,42−46] studies because of their small size, fast-folding kinetics, and biological importance. As a related phenomenon, the formation of intermolecular $\beta$-sheets is thought to be a crucial event in the initiation and propagation of amyloid diseases such as Alzheimer's disease[47] and spongiform encephalopathy.[48]

This article is organized as follows: The UNRES force field and PCA method are reviewed in section 2. The results are discussed in section 3. A summary and conclusions are presented in section 4.

## 2. Methods

**2.1. UNRES Model and Simulation Details.** The UNRES model of polypeptide chains[18,19,22,49,50] is illustrated in Figure 3. A polypeptide chain is represented as a sequence of α-carbon ($C^\alpha$) atoms linked by virtual $C^\alpha \cdots C^\alpha$ bonds with united peptide groups halfway between the neighboring $C^\alpha$'s and united side chains, whose sizes depend on the nature of the amino acid residues, attached to the respective $C^\alpha$'s by virtual $C^\alpha \cdots$SC bonds. The effective energy is expressed by the equation[22]

$$\begin{aligned} U = {} & w_{\text{SC}} \sum_{i<j} U_{\text{SC}_i\text{SC}_j} + w_{\text{SCp}} \sum_{i \neq j} U_{\text{SC}_i\text{p}_j} + \\ & w_{\text{pp}} f_2(T) \sum_{i<j-1} U_{\text{p}_i\text{p}_j} + w_{\text{tor}} f_2(T) \sum_i U_{\text{tor}}(\gamma_i) + \\ & w_{\text{tord}} f_3(T) \sum_i U_{\text{tord}}(\gamma_i, \gamma_{i+1}) + w_{\text{b}} \sum_i U_{\text{b}}(\theta_i) + \\ & w_{\text{rot}} \sum_i U_{\text{rot}}(\alpha_{\text{SC}_i}, \beta_{\text{SC}_i}, \theta_i) + w_{\text{bond}} \sum_i U_{\text{bond}}(d_i) + \\ & \sum_{m=3}^{6} w_{\text{corr}}^{(m)} f_m(T) U_{\text{corr}}^{(m)} + w_{\text{SS}} \sum_i U_{\text{SS};i} \end{aligned} \quad (1)$$

with[22]

$$f_m(T) = \frac{\ln(\text{e} + \text{e}^{-1})}{\ln\left\{\exp\left[\left(\frac{T}{T_0}\right)^{m-1}\right] + \exp\left[-\left(\frac{T}{T_0}\right)^{m-1}\right]\right\}}, \quad T_0 = 300 \text{ K} \quad (2)$$

where the successive terms represent side chain−side chain, side chain−peptide, and peptide−peptide interaction energies; torsional, double-torsional, bond-angle bending, and side-chain local (dependent on the angles α and $\beta$ of Figure 3) energies; distortion energies of virtual bonds; multibody (correlation) interaction energies; and energy of formation of disulfide bonds, respectively. $w$ represents the relative weights of each term. The correlation terms arise from a cumulant expansion[50,51] of the restricted free energy function of the simplified chain obtained from the all-atom energy surface by integrating out the secondary degrees of freedom. The temperature-dependent factors of eq 2, introduced in our recent work[22] and discussed further in ref 52, reflect the fact that the UNRES effective energy is an approximate cumulant expansion of the restricted free energy. The virtual-bond vectors are the variables used in molecular dynamics.

For 1BDD, we ran canonical UNRES molecular dynamics trajectories[38] at 11 temperatures at 5 K intervals between 290 and 340 K, with 10 trajectories at each temperature (for a total of 110 trajectories). The force field parametrized on 1GAB[22] was used. For 1E0L, we carried out canonical MD runs at the 11 temperatures 280, 290, 300, 310, 320, 330, 335, 340, 345, 350, and 360 K, with 10 trajectories at each temperature (for a total of 110 trajectories), using the force field parametrized on 1E0L and 1ENH.[53] The Berendsen

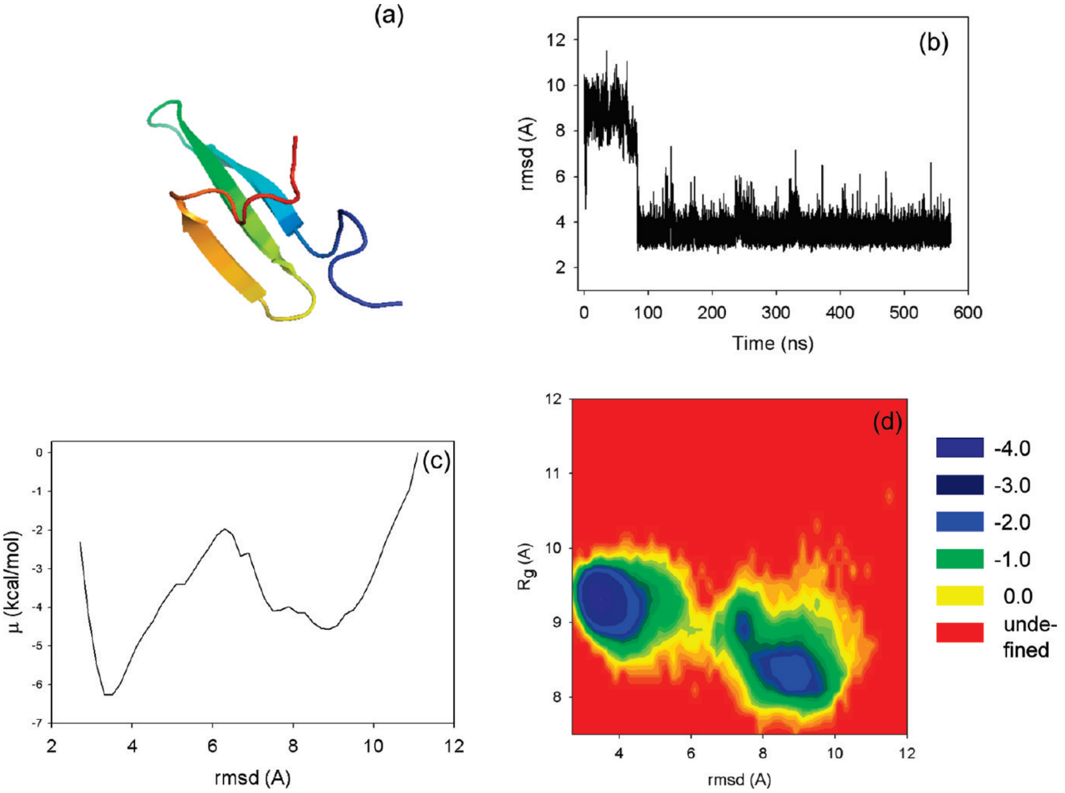


***Figure 2.*** (a) Experimental NMR structure of the WW domain of formin binding protein 28, (b) rmsd from the native structure as a function of time, (c) free energy profile (FEP) (in kcal/mol) plotted as a function of rmsd, and (d) FEL (in kcal/mol) plotted as a function of rmsd and radius of gyration for 1E0L.

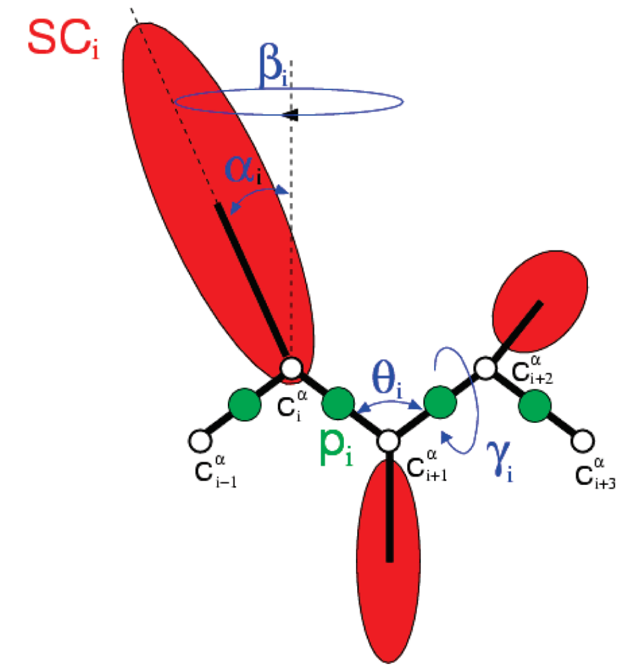


***Figure 3.*** UNRES model of polypeptide chains. The interaction sites are red side-chain centroids (SC) of different sizes, and the peptide-bond centers (p) are indicated by green circles; the α-carbon atoms (small empty circles) are introduced only to assist in defining the geometry. The virtual C$^\alpha$···C$^\alpha$ bonds have a fixed length of 3.8 Å, corresponding to a trans peptide group; the virtual-bond ($\theta$) and virtual-dihedral ($\gamma$) angles are variable. Each side chain is attached to the corresponding α-carbon with a fixed "bond length", $b_{SC_i}$; a variable "bond angle", $\alpha_i$, formed by SC$_i$ and the bisector of the angle defined by $C^\alpha_{i-1}$, $C^\alpha_i$, and $C^\alpha_{i+1}$; and a variable "dihedral angle", $\beta_i$, of counterclockwise rotation about the $C^\alpha_{i-1}$, $C^\alpha_i$, $C^\alpha_{i+1}$ frame.

thermostat[54] was used to maintain constant temperature. The trajectories selected for detailed analysis corresponded to those near the folding transition temperature, namely, $T$ = 310 K for 1BDD ($T_f$ = 320 K)[22] and $T$ = 330 K for 1E0L ($T_f$ = 339 K),[53] because these are the most favorable temperature regions for folding both proteins. The time step in molecular dynamics simulations was $\delta t = 0.1$ mtu (where 1 mtu = 48.9 fs is the "natural" time unit of molecular dynamics[55]), and the coupling parameter of the Berendsen thermostat was $\tau$ = 1 mtu. For each trajectory, a total of 35 000 000 steps (about 0.175 $\mu$s of MD time) were run for 1BDD, and 120 000 000 steps (about 0.6 $\mu$s of MD time) were run for 1E0L.

**2.2. Principal Component Analysis.** The PCA method[12] is based on the covariance matrix with elements $C_{ij}$ for coordinates $i$ and $j$

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \tag{3}$$

where $x_1$, ..., $x_{3N}$ are the mass-weighted Cartesian coordinates of an $N$-particle system and $\langle\rangle$ represents the average over all instantaneous structures sampled during the simulations. The symmetric $3N \times 3N$ matrix **C** can be diagonalized with an orthonormal transformation matrix **R**

$$\mathbf{R}^T\mathbf{C}\mathbf{R} = \mathrm{diag}(\lambda_1,\ \lambda_2,\ \cdots,\ \lambda_{3N}) \tag{4}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{3N}$ are the eigenvalues and $\mathbf{R}^T$ is the transpose of **R**. The columns of **R** are the eigenvectors, or the principal modes; the trajectory can be projected onto the eigenvectors to give the principal components $q_i(t)$, $i$ = 1, ..., $3N$

$$\mathbf{q} = \mathbf{R}^T[\mathbf{x}(t) - \langle \mathbf{x} \rangle] \tag{5}$$

The eigenvalue $\lambda_i$ is the mean-square fluctuation in the direction of the principal mode. The first few PCs typically describe collective, global motions of the system, with the first PC containing the largest mean-square fluctuation.

Because we are studying the coarse-grained MD trajectories, in PCA, we replaced the Cartesian coordinates by UNRES backbone coordinates ($\theta_i$,$\gamma_j$)

$$\begin{aligned} x_i &= \cos(\theta_i), & x_{i+1} &= \sin(\theta_i) \\ x_j &= \cos(\gamma_j), & x_{j+1} &= \sin(\gamma_j) \end{aligned} \tag{6}$$

where $i$ = 1, ..., $N - 2$, and $j$ = 1, ..., $N - 3$, are the numbers of $\theta$ and $\gamma$ angles, respectively, with $N$ being the number of amino acid residues in the chain. As shown by Mu et al.[24] and Altis et al.,[26] such a transformation from the space of backbone angles to a linear metric coordinate space enables potential problems due to the periodicity of the angles to be avoided.

## 3. Results and Discussion

**3.1. Determination of Least-Dimensional FEL, Correctly Describing Folding Dynamics.** Based on the results [rmsd vs time, free energy profile (FEP) as a function of rmsd, and FEL as a function of rmsd and $R_g$] shown in Figures 1 and 2, both proteins seem to fold following a two-state model with low-energy non-native and native states separated by a single energy barrier. The one-dimensional FELs (i.e., FEPs) suggest a simple picture containing the "unfolded" (high-rmsd) and "folded" (low-rmsd) states. The 2-D FELs reveal a more complex picture because the high-rmsd mimina correspond to low radii of gyration ($R_g$). Consequently, the high-rmsd states should be regarded as misfolded rather than unfolded states, indicating that both systems can get trapped in metastable conformations during folding. The loose unfolded conformations are present only during a few thousand initial steps of the simulations, and then both proteins collapse rapidly to either roughly folded or misfolded conformations. The complexity of the FELs obtained from the simulations is consistent with the experimentally observed multiple-exponential kinetics of both proteins.[2,56]

Whereas the folded state is unique, the misfolded one does not have to be, and consequently, the description provided by the 2-D rmsd$-R_g$ FEL plot might be oversimplified and misleading. We, therefore, employed a PCA to study the folding dynamics of 1BDD and 1E0L, particularly an internal-coordinate PCA, because FELs of small systems constructed by traditional Cartesian PCA can contain artifacts arising from strong mixing of overall and internal motions.[24-26] This issue is addressed in subsection 3.3.

As mentioned above, the first few PCs can capture more than half of the total fluctuation in the system; however, it

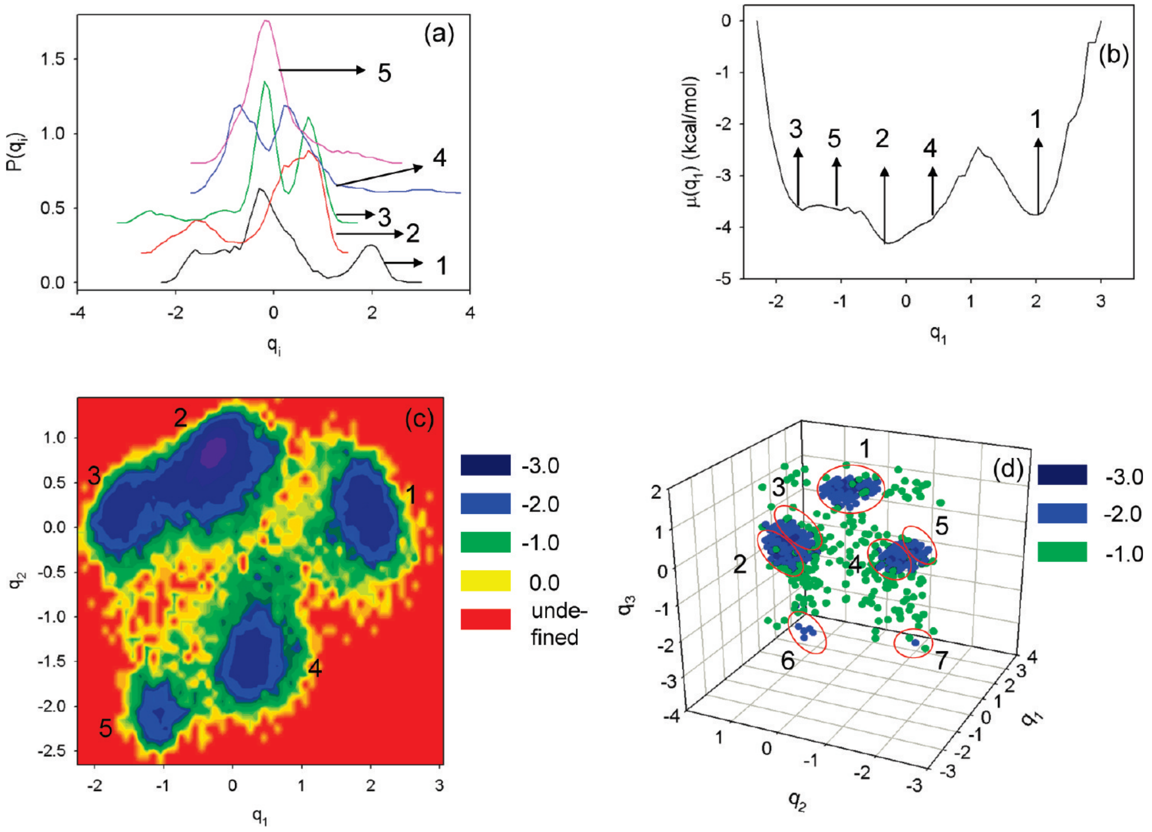


*Figure 4.* (a) Probability distribution functions for the first five internal-coordinate PCs of 1BDD and (b) 1-D, (c) 2-D, and (d) 3-D FELs (in kcal/mol) along internal-coordinate PCs.

is important to specify the criterion for selecting the PCs along which an FEL can be constructed. Based on the facts that multiply hierarchical PCs are a main contributor to the total fluctuations and that the subspace formed by multiply hierarchical PCs contains the most important molecular conformations,[57] Hegger et al.[58] defined the dimension of the free energy landscape by the fewest number of multiply hierarchical PCs. Figures 4 and 5 illustrate the probability distribution functions $P(q)$ of the first five PCs; the FEP, $\mu(q_1) = -k_BT \ln P(q_1)$, along the first PC; the 2-D FEL along the first two PCs, $\mu(q_1,q_2) = -k_BT \ln P(q_1,q_2)$; and the 3-D FEL along the first three PCs, $\mu(q_1,q_2,q_3) = -k_BT \ln P(q_1,q_2,q_3)$ for 1BDD and 1E0L, respectively. In these expressions, $T$ and $k_B$ are the absolute temperature and the Boltzmann constant, respectively.

As in our previous study,[15] carried out with a seemingly more complex folding pathway of 1BDD, the shapes of the probability distribution functions (panel a in Figure 4) suggest that the first four PCs of 1BDD clearly belong to the multiply hierarchical category, which means that, for a correct representation of the folding dynamics of 1BDD, a 4-D FEL is required. This observation is further corroborated by the 1-D, 2-D, and 3-D FELs depicted in panels b−d, respectively, of Figure 4, which show how much information is hidden in low-dimensional FELs. Although five minima are indicated in the 1-D FEP (panel b Figure 4), in reality, this FEP has only two pronounced minima (1 and 2), which represent two conformational states, and a slightly pronounced minimum (3) in one of the states. Aside from the wide basinlike shape (minima 2−5), the conformational state on the left-hand side does not reveal any complexity (ruggedness).

The number of minima increases with the dimensionality of the FEL: five and seven distinct minima can be identified in the 2-D FEL (panel c in Figure 4) (minima 2 and 3 belong to the same sub-basin and have a barely distinguishable low barrier) and in the 3-D FEL (panel d in Figure 4), respectively. It should be noted that, because of strong overlapping of points corresponding to diverse energies, the 3-D FEL (panel d in Figure 4) is represented by the clusters of only the lowest free energy points. Because the 4-D FEL, which is a complete representation, cannot be plotted, we present it in tabular form (Table 1). As expected, one new minimum (number 8) is observed in the 4-D FEL, which was hidden in the low-dimensional FELs. Because of its Gaussian shape (panel a in Figure 4), the fifth PC belongs to a harmonic category, which does not contribute significantly to the total fluctuation and corresponds to local motions.[57] Consequently, the 5-D FEL (Table 1) does not show any new minima; only slight rearrangements of the coordinates of some minima are observed. The minima in the high-dimensional FELs (3-D and higher) were determined by clustering the points with free energies within predefined intervals. It should be noted that, once a PC exhibits a harmonic shape, all higher-indexed PCs are also harmonic.

The shapes of the probability distribution functions (panel a in Figure 5) for 1E0L are quite different from those of 1BDD. Only the first PC can be assigned to the multiply hierarchical category; it should be noted, however, that one peak clearly dominates $P(q_1)$, unlike the case for 1BDD

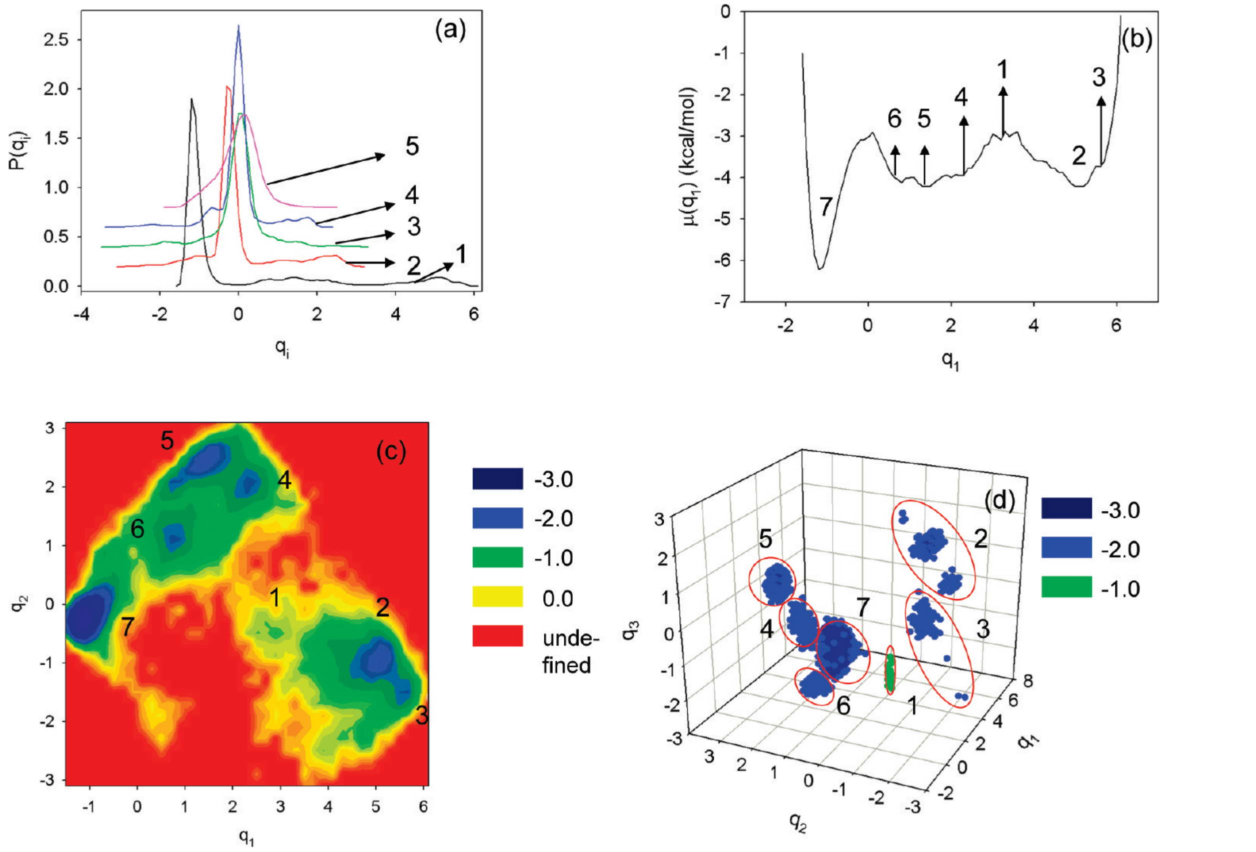


***Figure 5.*** (a) Probability distribution functions for the first five internal-coordinate PCs of 1E0L and (b) 1-D, (c) 2-D, and (d) 3-D FELs (in kcal/mol) along internal-coordinate PCs.

(panel a in Figure 4). Because of the Gaussian-like shape with a single peak, the second, third, and fourth PCs belong to the singly hierarchical category,[57] and the fifth PC belongs to the harmonic category, as in 1BDD. Unlike the FEP of 1BDD (panel b in Figure 4), the FEP along the first PC of 1E0L (panel b in Figure 5) clearly illustrates not only all conformational states (three-state folding), but also all conformational substates (local minima 2, 3, and 4−6) of each conformational state that can be less-clearly identified. Because the free energy profile along a singly hierarchical PC is characterized by a number of local minima arranged within a single coarse-grained minimum,[57] neither the 2-D and 3-D FELs (panels c and d in Figure 5) nor the 4-D FEL of 1E0L (Table 2) reveals any new conformational state. Also, except for making the local minima more distinguishable with slight rearrangements of the coordinates than obtained in the 1-D FEL, no further changes are observed in these FELs. Because the fifth PC (panel a in Figure 5) belongs to a harmonic category,[57] there are no major changes in the 5-D FEL, represented in tabular form, except for slight rearrangements of the coordinates of some minima (see Table 2). Thus, the folding dynamics of 1E0L can, in principle, be described by the 1-D FEP, although, for a clear illustration of all minima, the 2-D representation of the FEL is necessary.

Because the first few PCs capture most of the total fluctuation for both proteins, we have calculated the percentages of the total fluctuations captured by the PCs (panel a for 1BDD and panel b for 1E0L in Figure 6) for both proteins. It turns out that the percentages of total fluctuations captured by the PCs that were necessary for correct description of the folding dynamics (the first four PCs for 1BDD, and first PC for 1E0L) are almost the same, at ∼40%. Thus, the FEL constructed along PCs is correct if these PCs can capture at least 40% of the total fluctuations. This can be considered as another criterion for the determination of the minimal dimensionality for a correct FEL. To ensure that this finding was not accidental, we examined several more trajectories of 1BDD and 1E0L and obtained similar results.

Based on the results illustrated in Figures 4−6, it is clear that 1BDD exhibits more complex dynamics than 1E0L; that is, the former has a rugged FEL and requires a multidimensional FEL. The PCA works more efficiently for 1E0L trajectories than for 1BDD trajectories, by capturing almost half (∼40%) of the fluctuations in the first PC and illustrating the correct dynamics in the 1-D representation. Because of a loose nativelike structure, the amplitude of the fluctuations is large in the 1BDD trajectories, and the native state is quite broad, with several deep minima. Hence, the average values of the full width at half-maximum (fwhm) for $P(q)$ of the rmsd of the nativelike structures for 1BDD (310 K) and 1E0L (330 K) trajectories are 1.56 and 0.61 Å, respectively. To capture the main motions in the 1BDD trajectory, at least three to four PCs are required, whereas the FEP along the first PC was sufficient for 1E0L. Thus, for a correct description of the folding dynamics of largely fluctuating proteins, multidimensional FELs are required.

Based on the results of the computed single trajectory of the 1BDD protein, it should be noted that the definition of Hegger et al.,[58] regarding the dimensionality of an FEL obtained for peptides, needs some revision for some proteins.

***Table 1.*** PCs of the Minima of Basins Found in 1-D, 2-D, 3-D, 4-D, and 5-D FELs of 1BDD[a]

| PC[b] | 1-D | 2-D | 3-D | 4-D | 5-D |
|---|---|---|---|---|---|
| $q_1$ (1) | 1.90 | 1.90 | 1.90 | 2.10 | 2.10 |
| $q_1$ (2) | −0.30 | −0.30 | −0.30 | −0.30 | −0.30 |
| $q_1$ (3) | −1.70 | −1.70 | −1.70 | −1.70 | −1.70 |
| $q_1$ (4) | 0.30 | 0.30 | 0.30 | 0.30 | 0.30 |
| $q_1$ (5) | −1.10 | −1.10 | −1.10 | −1.10 | −1.10 |
| $q_1$ (6) | | | −0.10 | −0.10 | 0.10 |
| $q_1$ (7) | | | 0.50 | 0.50 | 0.50 |
| $q_1$ (8) | | | | −0.90 | −0.90 |
| $q_2$ (1) | | 0.30 | 0.30 | 0.10 | 0.10 |
| $q_2$ (2) | | 0.90 | 0.90 | 0.90 | 0.90 |
| $q_2$ (3) | | 0.20 | 0.30 | 0.30 | 0.30 |
| $q_2$ (4) | | −1.50 | −1.50 | −1.70 | −1.70 |
| $q_2$ (5) | | −2.10 | −2.10 | −2.10 | −2.10 |
| $q_2$ (6) | | | 0.50 | 0.50 | 0.70 |
| $q_2$ (7) | | | −1.70 | −1.70 | −1.70 |
| $q_2$ (8) | | | | 0.50 | 0.50 |
| $q_3$ (1) | | | 0.90 | 0.90 | 0.90 |
| $q_3$ (2) | | | −0.30 | −0.30 | −0.30 |
| $q_3$ (3) | | | 0.70 | 0.70 | 0.70 |
| $q_3$ (4) | | | −0.10 | −0.10 | −0.10 |
| $q_3$ (5) | | | 0.70 | 0.70 | 0.90 |
| $q_3$ (6) | | | −2.50 | −2.70 | −2.50 |
| $q_3$ (7) | | | −2.50 | −2.50 | −2.90 |
| $q_3$ (8) | | | | 0.50 | 0.50 |
| $q_4$ (1) | | | | 0.30 | 0.30 |
| $q_4$ (2) | | | | −0.70 | −0.90 |
| $q_4$ (3) | | | | 0.70 | 0.70 |
| $q_4$ (4) | | | | −0.90 | −0.90 |
| $q_4$ (5) | | | | 0.30 | 0.10 |
| $q_4$ (6) | | | | 0.90 | 0.90 |
| $q_4$ (7) | | | | 0.30 | 0.30 |
| $q_4$ (8) | | | | 0.30 | 0.30 |
| $q_5$ (1) | | | | | −0.10 |
| $q_5$ (2) | | | | | 0.10 |
| $q_5$ (3) | | | | | −0.70 |
| $q_5$ (4) | | | | | 0.10 |
| $q_5$ (5) | | | | | −0.30 |
| $q_5$ (6) | | | | | −0.90 |
| $q_5$ (7) | | | | | −0.90 |
| $q_5$ (8) | | | | | −0.10 |

[a] Numbers in the first column correspond to the conformational states in Figure 4. [b] Indicated PC, with the number of the minimum in parentheses.

***Table 2.*** PCs of the Minima of Basins Found in 1-D, 2-D, 3-D, 4-D, and 5-D FELs of 1E0L[a]

| PC[b] | 1-D | 2-D | 3-D | 4-D | 5-D |
|---|---|---|---|---|---|
| $q_1$ (1) | 3.10 | 2.90 | 2.90 | 2.70 | 2.70 |
| $q_1$ (2) | 4.90 | 4.90 | 5.10 | 5.10 | 5.10 |
| $q_1$ (3) | 5.30 | 5.30 | 5.10 | 5.10 | 5.10 |
| $q_1$ (4) | 2.30 | 2.30 | 2.30 | 2.30 | 2.30 |
| $q_1$ (5) | 1.50 | 1.50 | 1.50 | 1.50 | 1.50 |
| $q_1$ (6) | 0.70 | 0.70 | 0.70 | 0.90 | 0.70 |
| $q_1$ (7) | −1.10 | −1.10 | −1.10 | −1.30 | −1.30 |
| $q_2$ (1) | | −0.50 | −0.30 | −0.50 | −0.30 |
| $q_2$ (2) | | −0.90 | −0.90 | −0.90 | −0.90 |
| $q_2$ (3) | | −1.30 | −0.90 | −0.90 | −0.90 |
| $q_2$ (4) | | 2.10 | 1.90 | 1.90 | 1.90 |
| $q_2$ (5) | | 2.50 | 2.50 | 2.50 | 2.50 |
| $q_2$ (6) | | 1.10 | 1.10 | 1.10 | 1.10 |
| $q_2$ (7) | | −0.30 | −0.30 | −0.30 | −0.30 |
| $q_3$ (1) | | | −2.10 | −2.10 | −1.90 |
| $q_3$ (2) | | | 1.50 | 1.50 | 1.50 |
| $q_3$ (3) | | | −1.10 | −0.90 | −0.90 |
| $q_3$ (4) | | | −0.90 | −0.90 | −0.90 |
| $q_3$ (5) | | | 0.70 | 0.70 | 0.70 |
| $q_3$ (6) | | | −1.90 | −1.90 | −1.90 |
| $q_3$ (7) | | | 0.10 | 0.10 | 0.10 |
| $q_4$ (1) | | | | 0.70 | 0.70 |
| $q_4$ (2) | | | | 1.70 | 1.70 |
| $q_4$ (3) | | | | −2.30 | −2.30 |
| $q_4$ (4) | | | | 1.30 | 1.30 |
| $q_4$ (5) | | | | −0.70 | −0.70 |
| $q_4$ (6) | | | | 1.50 | 1.50 |
| $q_4$ (7) | | | | −0.10 | −0.10 |
| $q_5$ (1) | | | | | 0.30 |
| $q_5$ (2) | | | | | −0.50 |
| $q_5$ (3) | | | | | −0.90 |
| $q_5$ (4) | | | | | 0.70 |
| $q_5$ (5) | | | | | −0.70 |
| $q_5$ (6) | | | | | −0.30 |
| $q_5$ (7) | | | | | 0.10 |

[a] Numbers in the first column correspond to the conformational states in Figure 5. [b] Indicated PC, with the number of the minimum in parentheses.

The point is that, according to Hegger et al.,[58] each peak of the probability distribution function of a multiply hierarchical PC corresponds to a different conformational state of the peptide. However, we have shown that, for some proteins with complex dynamics, not all peaks of the probability distribution functions of multiply hierarchical PCs correspond to conformational states; they might also correspond to conformational substates in a large basin. Therefore, careful examination of the structures in each minimum is necessary.

**3.2. Folding Pathways of 1BDD and 1E0L.** The FELs of both proteins, especially those of 1BDD, are quite complex, with several minima present. Consequently, it is unclear what kinetic model can be used for the description of the folding dynamics of these proteins. Therefore, to examine the folding pathways of the two proteins, we selected representative structures corresponding to all of the minima and transition states of the FELs. These structures are shown in Figure 7 for both 1BDD (panel a) and 1E0L (panel b).

An analysis of the selected trajectory of 1BDD shows that, after ~3 ns, it folds from a fully unfolded conformation to the mirror image of the native structure, where it remains for quite a long time (about 30 ns). This metastable state corresponds to a kinetic trap (minimum 1 in panel a of Figure 7). Any of these misfolded mirror images has energies comparable to those of nativelike structures and high rmsd values (8−10 Å). They have been observed in several different studies with different all-atom force fields for various α-helix bundles,[59,60] including 1BDD.[35] At low temperatures, the metastable mirror-image conformation is observed quite frequently (e.g., at 290 K, in 8 trajectories out of 10); however, it is encountered less and less frequently and finally disappears with increasing temperature. This is not surprising, because construction of an equilibrium free energy landscape requires much longer simulations at low temperatures (glassy-type state) than at higher temperatures.

After remaining in the mirror-image conformation for ~30 ns (at $T = 310$ K), the N-terminal helix forms a separate linear portion of the middle helix (the structure in the transition state), and the protein overcomes the barrier of the metastable state and jumps to the native basin, particularly in minimum 7. For ~8 ns, the system jumps back and forth between the native-basin minima 7 and 6. After that, the system starts the interconversions among ensembles of conformations in a back-and-forth progression between the minima of the native basin (minima 2−5 in panel a of Figure 7) until the end of the trajectory. The most nativelike representative structure (rmsd = 2.7 Å) is observed in minimum 4. The presence of six minima in the native basin means that the native state of 1BDD is quite dynamic. This finding is in agreement with an earlier result obtained by

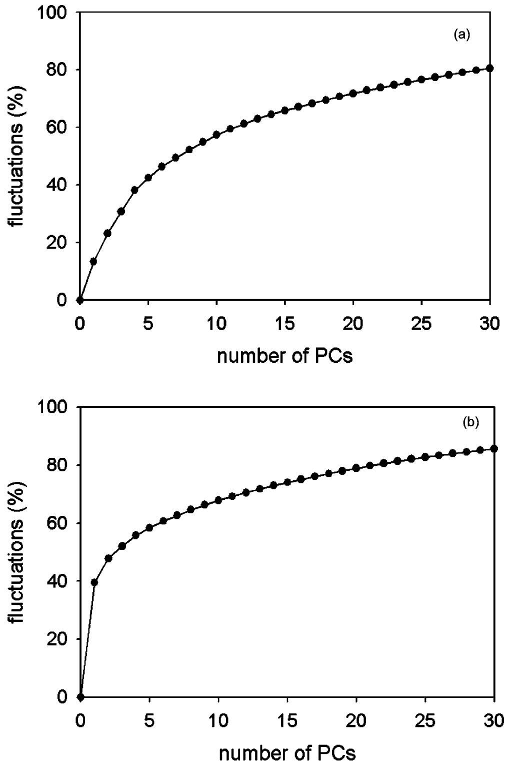


***Figure 6.*** Percentages of total fluctuations captured by internal-coordinate PCs for (a) 1BDD and (b) 1E0L.

Alonso and Daggett[30] who studied the unfolding of 1BDD. Also, in comparison with the results of our earlier study,[40] the FEL of 1BDD obtained here is more rugged in internal principal component space; however, the folding pathways and models are similar to those observed previously.[40]

Thus, the folding pathway and folding mechanism described in panel a of Figure 7 were quite unexpected because of several deep, distinct minima in the FEL. The reason for such behavior is a loose nativelike structure of 1BDD that, with increasing temperature, turns into a loose molten globule.

All FELs of 1E0L in Figure 5 clearly indicate three-state folding. Panel b of Figure 7, in which the 3-D FEL is plotted with representative structures in each minimum, illustrates how 1E0L folds at $T = 330$ K. At the beginning of the trajectory starting from the fully extended conformation, before forming a non-native conformational state (minima 2 and 3 in panel b of Figure 7), the protein forms quite a shallow minimum (minimum 1 in panel b of Figure 7), the representative structure (rmsd = 9.3 Å) of which is not fully or partially unfolded but does not show any sign of formation of strands or loops. The representative structures in the minima of the non-native state do not contain any strands or loops, and moreover, the representative structure of minimum 3 forms a partial helix at the C-terminus. As expected, these structures have quite a high rmsd (~8.9 Å).

After remaining in the non-native state for ~69 ns, the protein overcomes a barrier and jumps to an intermediate basin. On the way, in the transition state, the system loses the helical structure at the C-terminus. The intermediate basin contains three distinct minima (4−6), the representative structures of which are characterized by low rmsd values (between 3.7 and 4.3 Å) and exhibit $\beta$-sheet structural features. Particularly, loop 1 and partially strands 1 and 2 are formed in minima 4 and 6 of an intermediate basin. The representative structure of minimum 5 exhibits loop 1 and fully formed strands 1 and 2. Although the representative structures of these minima, characterized by low rmsd values, illustrate the structural features of a $\beta$-sheet, they are not correctly folded. The protein remains in an intermediate basin and interconverts back and forth between only these minima for ~20 ns; it then jumps to the native state (minimum 7) and starts the interconversion between the native state and an intermediate basin for ~356 ns. After that, the protein remains in the native state until the end of the trajectory.

Thus, the folding pathway and kinetic model of two trajectories, similar by visual inspection of the time dependence of the rmsd;s (panels labeled b in Figures 1 and 2), differ completely from each other. However, to understand the folding pathways of the system (which is not the main goal of this work), the results based on the study of one trajectory cannot be sufficiently representative. Therefore, we combined 10 trajectories at the same temperature and analyzed them by internal-coordinate PCA. Figure 8 illustrates FELs as functions of $q_1$ and $q_2$ for a collection of 10 trajectories of 1BDD at 310 K and 1E0L at 330 K.

Judging from the rmsd as a function of time for 1BDD (not shown), there are four different types of folding trajectories: (1) The protein folds instantly and stays in the native state until the end of the simulation. (2) The protein folds instantly but unfolds and encounters a kinetic trap at the end of the trajectory. (3) Before jumping to the native state, the protein becomes trapped in a metastable state. (4) The protein undergoes folding/unfolding events several times during the MD simulation. Because of this diversity of folding pathways, the FEL for a collection of trajectories does not resemble that for an individual trajectory. In other words, in none of these trajectories does the protein fold in the way shown in the FEL of a collection of trajectories (panel a in Figure 8). However, Figure 8 (panel a) illustrates the percentage of total time spent in each minimum, which describes the general "picture" of a folding pathway. The details of the minima are as follows: Minimum 2 contains only mirror-image conformations; minima 3−7 belong to the native basin; and minimum 1 contains mainly mirror-image conformations, although numerous structures with low rmsd values are found as well. Thus, this protein folds with two probable folding pathways. One of them, the folding through the kinetic trap, formed by the mirror image, is less probable than the other (i.e., direct downhill folding).[40] Also, it should be noted that the folding becomes effectively downhill as the temperature increases because the barrier between the mirror image and the native state decreases.

Unlike the FEL of 1BDD, the FEL of a collection of 10 trajectories for 1E0L (panel b in Figure 8) is quite similar to the FEL of the studied single trajectory (panel c in Figure 5). This indicates that all 10 trajectories at $T = 330$ K are similar to each other and that the folding pathway shown in panel b of Figure 8 is representative of each trajectory. In other words, after starting from the fully extended unfolded conformation, the protein immediately assumes a compact

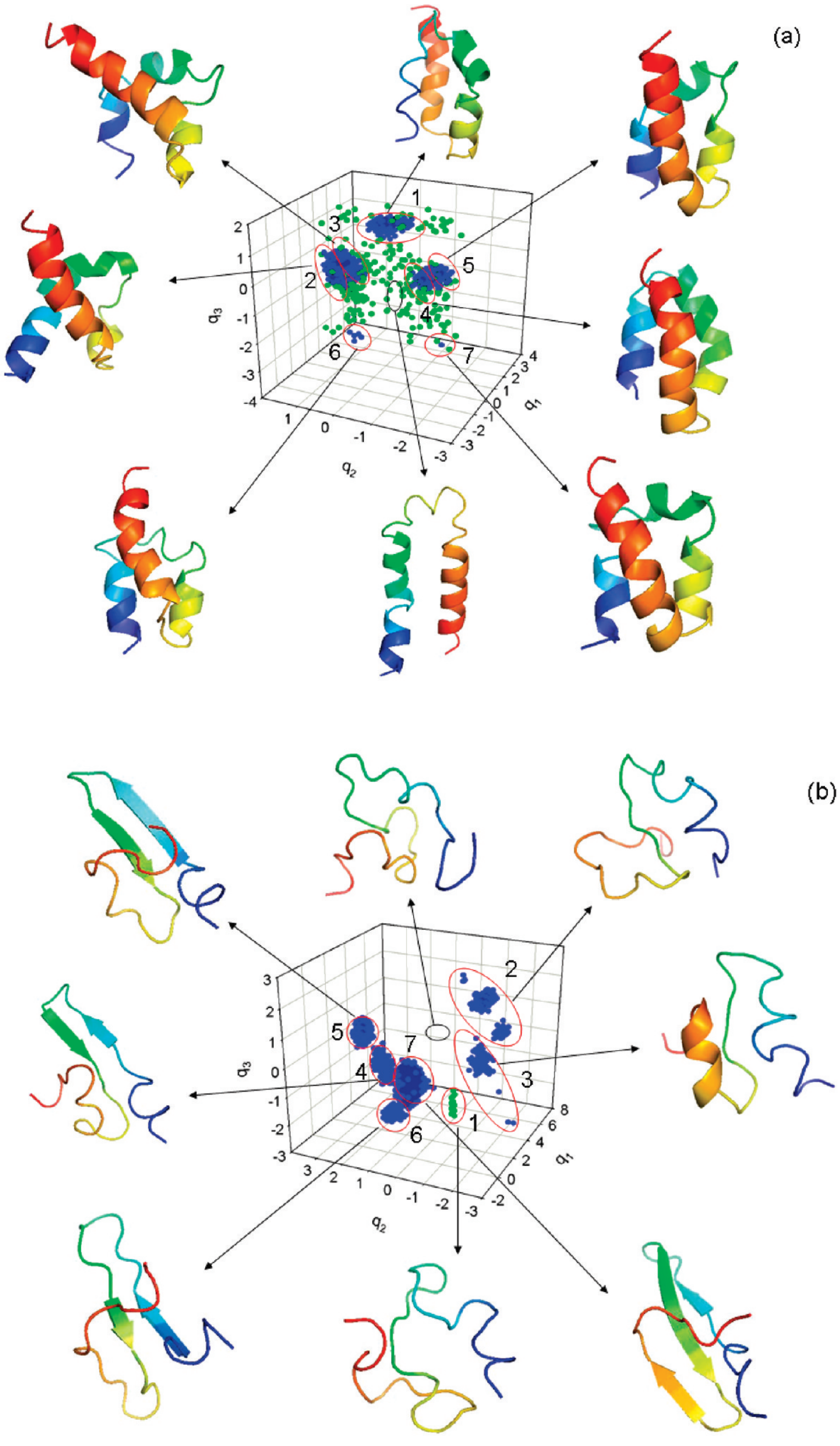


***Figure 7.*** Three-dimensional free energy landscapes (in kcal/mol) along internal-coordinate PCs for (a) 1BDD and (b) 1E0L with representative structures at the minima and transition states. The structures are colored from blue to red from the N- to the C-terminus. Each minimum in both a and b is in blue, circled by a red line and numbered, and the transition is in a white unnumbered cluster, circled by a black line.

shape and remains in shallow minimum 1 for a very short time; it then jumps to the non-native basin (minimum 2), forming two minima there. After spending ~20% of the total time in the non-native basin, it proceeds to the intermediate basin (minima 3−6), in which it interconverts among minima 3−6 for ~19% of total time, and finally jumps to the native state (minimum 7).

**3.3. FEL in Cartesian- and Internal-Coordinate Principal Component Space.** As mentioned in the Methods section and subsection 3.1, the trajectories were analyzed

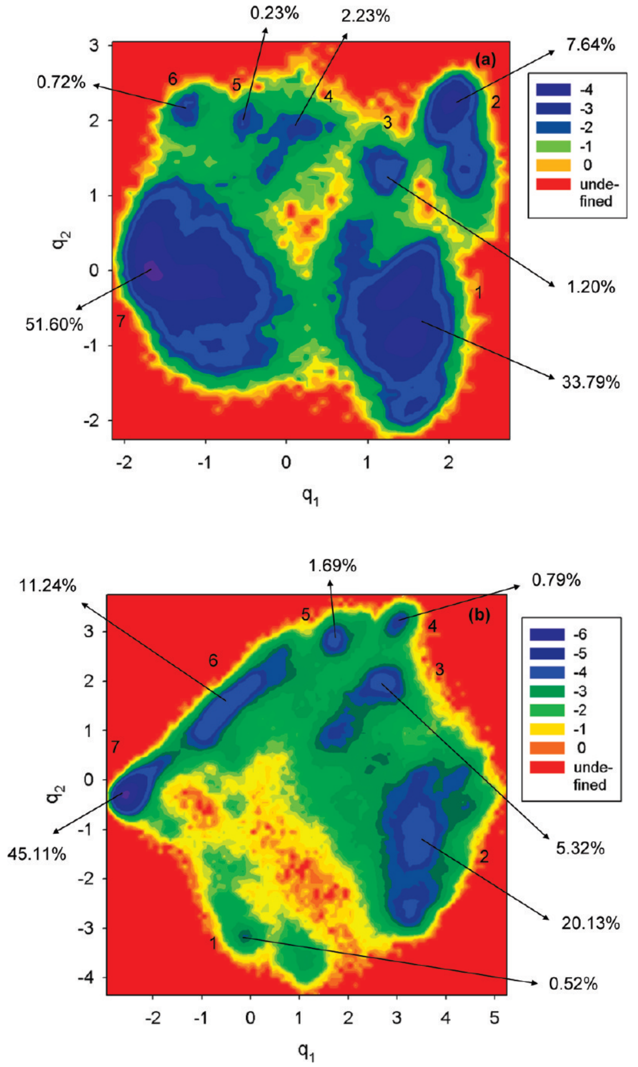


***Figure 8.*** Two-dimensional free energy landscapes (in kcal/mol) of a collection of 10 trajectories along internal-coordinate PCs for (a) 1BDD and (b) 1E0L. The numbers at the ends of the arrows indicate the percentages of total time spent in the corresponding minima.

by internal-coordinate PCA, which normally reveals much more rugged FELs than Cartesian PCA. Our preference for internal-coordinate PCA is based on the facts that the true free energy landscape is actually quite rugged[24−26] and that its smooth appearance in Cartesian PCA represents an artifact of the mixing of internal and overall motions. However, the conclusions about the ruggedness of the FEL obtained by internal-coordinate PCA (particularly dihedral PCA) were drawn from all-atom MD studies performed on peptides.[24−26] Because it is still not easy to fold proteins by all-atom MD simulations, to the best of our knowledge, we do not know whether a comparison of the FELs of the folding trajectories of proteins, rather than peptides, obtained by internal-coordinate PCA and Cartesian PCA was ever carried out. Therefore, we analyzed the trajectory of 1BDD by Cartesian PCA. Figure 9 illustrates $P(q)$ for the first five PCs, the FEP along the first PC, the FEL along the first two PCs, and the percentage of total fluctuations captured by PCs.

The results shown in Figure 9 are quite different from those obtained by internal-coordinate PCA for the same trajectory (Figure 4). First, the shapes of $P(q)$ (panel a in Figure 9) are quite different in Cartesian PCA. Only the first PC belongs to the multiply hierarchical category.[57] Based on the above-mentioned criteria for minimal dimensionality of an FEL, the 1-D FEP (panel b in Figure 9) constructed along Cartesian PCs should be sufficient for the correct representation of folding dynamics. However, in addition to the 1-D FEP, the 2-D FEL (panel c in Figure 9) also does not show any complexity or ruggedness of the FEL. The native state in both representations has one smooth deep minimum, and the FEP along $q_1$ (panel b in Figure 9) resembles that along the rmsd (panel c of Figure 1). Thus, the conclusions drawn in an earlier work[24−26] regarding some drawbacks of Cartesian PCA for small peptides seem to be correct for small proteins, as well.

Moreover, the fluctuations captured by Cartesian PCs (panel d in Figure 9) converge faster than those corresponding to the internal-coordinate PCA (panel a in Figure 6), which conforms with the results obtained for small peptides.[24]

Finally, we computed the average mean first passage times (MFPTs, the times at which the native structures were encountered first) at temperatures near the folding transition for both proteins. The MFPTs can be considered crude estimates of folding times. The values calculated for 1BDD (at $T = 310$ K) and 1E0L (at $T = 335$ K) are 16 and 284 ns, respectively, compared to the experimental folding times of 30 and 900 $\mu$s for 1BDD[56] and 1E0L,[2] respectively. As already pointed out in our earlier work,[9] the folding times calculated by UNRES/MD are orders of magnitude greater than the experimental folding times, because of averaging out of the fast degrees of freedom. Additionally, in this study, we carried out Berendsen and not Langevin dynamics, which makes the calculated times even shorter. Nevertheless, the calculated ratio of the MFPTs of 1E0L and 1BDD is 18 compared to the ratio of experimental folding times equal to 30; consequently, the UNRES simulations correctly reproduce the experimental observation that the folding time of 1E0L is more than an order of magnitude greater than that of 1BDD.

## 4. Conclusions

Using PCA, we have examined the MD trajectories of protein folding, generated with the coarse-grained UNRES force field, for the B-domain of staphylococcal protein A and the triple $\beta$-strand WW domain from the formin binding protein 28 (FBP). The results demonstrate how different the folding dynamics (FELs, folding pathways, folding models, etc.) of the trajectories can be even when the trajectories are very similar by visual inspection of the time dependence of the rmsd.

The ways to determine the minimal dimensionality of an FEL that would be sufficient for a correct description of protein folding dynamics were shown. We found that the fluctuations captured by multiply hierarchical PCs, required for a correct FEL, represent at least ~40% of the total fluctuations. Further, there is a correlation between the amplitude of the fluctuations of a trajectory and the dimensionality of the correct FEL. In other words, we demonstrated

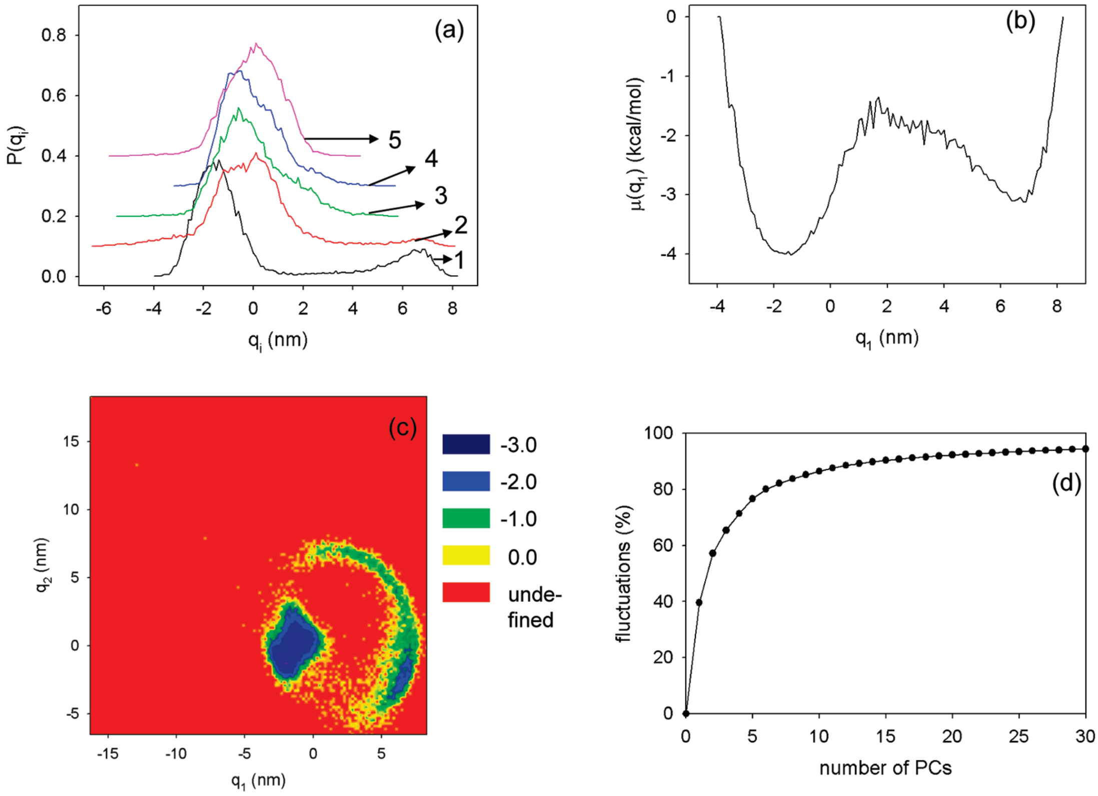


***Figure 9.*** (a) Probability distribution functions for the first five Cartesian PCs of 1BDD, (b) 1-D and (c) 2-D FELs (in kcal/mol) along the Cartesian PCs for 1BDD, and (d) the percentage of total fluctuations captured by Cartesian PCs for 1BDD.

that trajectories with large amplitudes of fluctuation require a multidimensional FEL for a correct description of the folding dynamics, because the first several PCs can exhibit a multiply hierarchical shape, and the percentages of the captured fluctuations by each successive multiply hierarchical PC are comparably small and do not differ very much from each other. Also, we showed that, for some trajectories with large amplitudes of fluctuation, not all peaks of the $P(q)$ of multiply hierarchical PCs correspond to conformational states, as was stated by Hegger et al.;[58] instead, they might correspond to conformational substates in a large basin, and therefore, care must be taken in examining structures in each minimum.

Finally, we demonstrated that, for small proteins, internal-coordinate PCA provides a more descriptive FEL than Cartesian PCA. The relatively simple, smooth FEL constructed by Cartesian PCA does not describe the folding dynamics correctly and represents an artifact of the mixing of internal and overall motions.[24−26]

**Acknowledgment.** This work was supported by grants from the National Institutes of Health (GM-14312), the National Science Foundation (MCB05-41633), and the Polish Ministry of Science and Education (0490/B/H03/2008/35). This research was conducted using the resources of (a) our 880-processor Beowulf cluster at the Baker Laboratory of Chemistry and Chemical Biology, Cornell University; (b) the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center; (c) the John von Neumann Institute for Computing at the Central Institute for Applied Mathematics, Forschungszentrum Jülich, Jülich, Germany; (d) the Beowulf cluster at the Department of Computer Science, Cornell University; (e) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, Poland; and (f) the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) at the University of Warsaw.

## References

(1) Poland, D. C.; Scheraga, H. A. Statistical mechanics of noncovalent bonds in polyamino acids. IX. The two-state theory of protein denaturation. *Biopolymers* **1965**, *3*, 401–419.

(2) Nguyen, H.; Jäger, M.; Moretto, A.; Gruebele, M.; Kelly, J. W. Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3948–3953.

(3) Peng, L.; Oliva, F. Y.; Naganathan, A.; Munoz, V. Dynamics of one-state downhill protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 103–108.

(4) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The energy landscapes and motions of proteins. *Science* **1991**, *254*, 1598–1603.

(5) Brooks, C. L., III; Onuchic, J. N.; Wales, D. J. Taking a walk on a landscape. *Science* **2001**, *293*, 612–613.

(6) Wales, D. J. *Energy Landscapes*; Cambridge University Press: Cambridge, U.K., 2003; p 681.

(7) Mitsutake, A.; Sugita, Y.; Okamoto, Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers* **2001**, *60*, 96–123.

(8) Boczko, E. M.; Brooks, C. L., III. First principles calculation of the free energy surface for folding of a three helix bundle protein. *Science* **1995**, *269*, 393–396.

(9) Liwo, A.; Khalili, M.; Scheraga, H. A. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2362–2367.

(10) Noe, F.; Fischer, S. Transition network for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154–162.

(11) Strodel, B.; Wales, D. J. Free energy surfaces from an extended harmonic superposition approach and kinetics from alanine dipeptide. *Chem. Phys. Lett.* **2008**, *466*, 105–115.

(12) Jolliffe, I. T. *Principal Component Analysis*; Springer: New York, 2002; p 487.

(13) Krivov, S. V.; Karplus, M. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766–14770.

(14) Altis, A.; Otten, M.; Nguyen, P. H.; Hegger, R.; Stock, G. Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. *J. Chem. Phys.* **2008**, *128*, 245102(1−11).

(15) Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. How adequate are one- and two-dimensional free energy landscapes for protein folding dynamics. *Phys. Rev. Lett.* **2009**, *102*, 238102(1−4).

(16) Gouda, H.; Torigoe, H.; Saito, A.; Sato, M.; Arata, Y.; Shimada, I. Three-dimensional solution structure of the B domain of staphylococcal protein A: Comparisons of the solution and crystal structures. *Biochemistry* **1992**, *31*, 9665–9672.

(17) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scherega, H. A. Prediction of protein conformation on the basis of a search for compact structures: Test on avian pancreatic polypeptide. *Protein Sci.* **1993**, *2*, 1715–1731.

(18) Liwo, A.; Ołdziej, S.; Pincus, M. R.; Wawak, R. J.; Rackowsky, S.; Scheraga, H. A. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.* **1997**, *18*, 849–873.

(19) Liwo, A.; Ołdziej, S.; Czaplewski, C.; Kozlowska, U.; Scheraga, H. A. Parametrization of backbone-electrostatic and multibody contributions to the UNRES force field for protein-structure prediction from ab initio energy surfaces of model systems. *J. Phys. Chem. B* **2004**, *108*, 9421–9438.

(20) Ołdziej, S.; Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. Optimization of the UNRES force field by hierarchical design of the potential-energy landscape. 2. Off-lattice tests of the method with single proteins. *J. Phys. Chem. B* **2004**, *108*, 16934–16949.

(21) Ołdziej, S.; Lagiewka, J.; Liwo, A.; Czaplewski, C.; Chinchio, M.; Nanias, M.; Scheraga, H. A. Optimization of the UNRES force field by hierarchical design of the potential-energy landscape. 3. Use of many proteins in optimization. *J. Phys. Chem. B* **2004**, *108*, 16950–16959.

(22) Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Ołdziej, S.; Wachucik, K.; Scheraga, H. A. Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *J. Phys. Chem. B* **2007**, *111*, 260–285.

(23) Macias, M. J.; Gervais, V.; Civera, C.; Oschkinat, H. Structural analysis of WW domains and design of a WW prototype. *Nat. Struct. Biol.* **2000**, *7*, 375–379.

(24) Mu, Y.; Nguyen, P. H.; Stock, G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins* **2005**, *58*, 45–52.

(25) Maisuradze, G. G.; Leitner, D. M. Free energy landscape of a biomolecule in dihedral principal component space: Sampling convergence and correspondence between structures and minima. *Proteins* **2007**, *67*, 569–578.

(26) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. Dihedral angle principal component analysis of molecular dynamics simulations. *J. Chem. Phys.* **2007**, *126*, 244111(1−10).

(27) Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. Principal component analysis for protein folding dynamics. *J. Mol. Biol.* **2009**, *385*, 312–329.

(28) Kolinski, A.; Skolnick, J. Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. *Proteins* **1994**, *18*, 353–366.

(29) Zhou, Y.; Karplus, M. Interpreting the folding kinetics of helical proteins. *Nature* **1999**, *401*, 400–403.

(30) Alonso, D. O. V.; Daggett, V. Staphylococcal protein A: Unfolding pathways, unfolded states, and differences between the B and E domains. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 133–138.

(31) Berriz, G. F.; Shakhnovich, E. I. Characterization of the folding kinetics of a three-helix bundle protein via a minimalist Langevin model. *J. Mol. Biol.* **2001**, *310*, 673–685.

(32) Kussell, E.; Shimada, J.; Shakhnovich, E. I. A structure-based method for derivation of all-atom potentials for protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 5343–5348.

(33) Ghosh, A.; Elber, R.; Scheraga, H. A. An atomically detailed study of the folding pathways of protein A with the stochastic difference equation. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 10394–10398.

(34) Garcia, A. E.; Onuchic, J. N. Folding a protein in a computer: An atomic description of the holding/unfolding of protein A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13898–13903.

(35) Vila, J. A.; Ripoll, D. R.; Scheraga, H. A. Atomically detailed folding simulation of the B domain of staphylococcal protein A from random structures. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 14812–14816.

(36) Karanicolas, J.; Brooks, C. L. III. The structural basis for biphasic kinetics in the folding of the WW domain from a formin-binding protein: Lessons for protein design. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3954–3959.

(37) Karanicolas, J.; Brooks, C. L. III. Integrating folding kinetics and protein function: Biphasic kinetics and dual binding specificity in a WW domain. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 3432–3437.

(38) Khalili, M.; Liwo, A.; Jagielska, A.; Scheraga, H. A. Molecular dynamics with the united-residue model of polypeptide chains. II. Langevin and Berendsen-bath dynamics and tests on model α-helical systems. *J. Phys. Chem. B* **2005**, *109*, 13798–13810.

(39) Mu, Y.; Nordenskiold, L.; Tam, J. P. Folding, misfolding, and amyloid protofibril formation of WW domain FBP28. *Biophys. J.* **2006**, *90*, 3983–3992.

(40) Khalili, M.; Liwo, A.; Scheraga, H. A. Kinetic studies of folding of the B-domain of staphylococcal protein A with molecular dynamics and a united-residue (UNRES) model of polypeptide chains. *J. Mol. Biol.* **2006**, *355*, 536–547.

(41) Jagielska, A.; Scheraga, H. A. Influence of temperature, friction, and random forces on folding of the B-domain of Staphylococcal Protein A: All-atom molecular dynamics in implicit solvent. *J. Comput. Chem.* **2007**, *28*, 1068–1082.

(42) Bottomley, S. P.; Popplewell, A. G.; Scawen, M.; Wan, T.; Sutton, B. J.; Gore, M. G. The stability and unfolding of an IgG binding protein based upon the B domain of protein A from Staphylococcus aureus probed by tryptophan substitution and fluorescence spectroscopy. *Protein Eng.* **1994**, *7*, 1463–1470.

(43) Bai, Y.; Karimi, A.; Dyson, H. J.; Wright, P. E. Absence of a stable intermediate on the folding pathway of protein A (B domain). *Protein Sci.* **1997**, *6*, 1449–1457.

(44) Myers, J. K.; Oas, T. G. Preorganized secondary structure as an important determinant of fast protein folding. *Nat. Struct. Biol.* **2001**, *8*, 552–558.

(45) Dimitriadis, G.; Drysdale, A.; Myers, J. K.; Arora, P.; Radford, S. E.; Oas, T. G.; Smith, D. A. Microsecond folding dynamics of the F13W G29A mutant of the B domain of staphylococcal protein A by laser-induced temperature jump. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 3809–3814.

(46) Sato, S.; Religa, T. L.; Daggett, V.; Fersht, A. R. Testing protein-folding simulations by experiment: B domain of protein A. *Proc Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6952–6956.

(47) Serpell, L. C. Alzheimer's amyloid fibrils: Structure and assembly. *Biochim. Biophys. Acta* **2000**, *1502*, 16–30.

(48) Pruisner, S. B. Prions. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 13363–13383.

(49) Scheraga, H. A.; Liwo, A.; Ołdziej, S.; Czaplewski, C.; Pillardy, J.; Ripoll, D. R.; Vila, J. A.; Kazmierkiewicz, R.; Saunders, J. A.; Arnautova, Y. A.; Jagielska, A.; Chinchio, M.; Ninias, M. The protein folding problem: Global optimization of force fields. *Front. Biosci.* **2004**, *9*, 3296–3323.

(50) Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *J. Chem. Phys.* **2001**, *115*, 2323–2347.

(51) Kubo, R. J. Generalized cumulant expansion method. *Phys. Soc. Jpn.* **1962**, *17*, 1100–1120.

(52) Shen, H.; Liwo, A.; Scheraga, H. A. An Improved Functional Form for the Temperature Scaling Factors of the Components of the Mesoscopic UNRES Force Field for Simulations of Protein Structure and Dynamics. *J. Phys. Chem. B* **2009**, *113*, 8738–8744.

(53) Liwo, A. ; Czaplewski, C. ; Ołdziej, S. ; Kozłowska, U. ; Makowski, M. ; Kalinowski, S. ; Kazmierkiewicz, R. ; Shen, H. ; Maisuradze, G. ; Scheraga, H. A. Optimization of a physics-based united-residue force field (UNRES) for protein folding simulations. In *NIC Symposium, Jülich, Germany, 2008*; Münster, G., Wolf, D., Kremer, M., Eds.; NIC-Directors: Jülich, Germany, 2008; pp 63−70.

(54) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(55) Khalili, M.; Liwo, A.; Rakowski, F.; Grochowski, P.; Scheraga, H. A. Molecular dynamics with the united-residue model of polypeptide chains. I. Lagrange equations of motion and tests of numerical stability in the microcanonical mode. *J. Phys. Chem. B* **2005**, *109*, 13785–13797.

(56) Vu, D. M.; Myers, J. K.; Oas, T. G.; Dyer, R. B. Probing the folding and unfolding dynamics of secondary and tertiary structures in a three-helix bundle protein. *Biochemistry* **2004**, *43*, 3582–3589.

(57) Kitao, A.; Hayward, S.; Gō, N. Energy landscape of a native protein: Jumping-among-minima model. *Proteins* **1998**, *33*, 496–517.

(58) Hegger, R.; Altis, A.; Nguyen, P. H.; Stock, G. How complex is the dynamics of peptide folding. *Phys. Rev. Lett.* **2007**, *98*, 028102(1−4).

(59) Hubner, I. A.; Deeds, E. J.; Shakhnovich, E. I. High-resolution protein folding with a transferable potential. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 18914–18919.

(60) Yang, J. S.; Chen, W. W.; Skolnick, J.; Shakhnovich, E. I. All-atom ab initio folding of a diverse set of proteins. *Structure* **2007**, *15*, 53–63.

CT9005745