JCTC Journal of Chemical Theory and Computation

Reducing the Secondary Structure Bias in the Generalized Born Model via R6 Effective Radii

Boris Aguilar,[†] Richard Shadrach,[‡] and Alexey V. Onufriev*[,§]

Department of Computer Science, Virginia Tech, Blacksburg, Virginia 24060, United States, Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States, and Departments of Computer Science and Physics, Virginia Tech, Blacksburg, Virginia 24060, United States

Received July 14, 2010

Abstract: The generalized Born model (GB) provides a reasonably accurate and computationally efficient way to compute the electrostatic component (ΔG_{el}) of the solvation free energy. In this work, we have developed a method to compute effective Born radii, which is intended to address the known secondary structure bias of the GB model reported earlier (Roe et al. *J. Phys. Chem. B*, **2007**, *111*, 1846−1857). Our analytical approach, termed AR6, is based on the $|\mathbf{r}|^{-6}$ (R6) integration over an approximation to molecular volume. Within the approach, several computationally efficient corrections to the pairwise VDW−volume integration are combined to closely approximate the true molecular volume in the vicinity of each atom. The accuracy of the AR6 model in predicting relative ΔG_{el} is tested on four conformational states of alanine decapeptide. Changes in ΔG_{el} estimated by AR6 between various pairs of conformational states have the same RMS error relative to the explicit solvent, as do the corresponding numerical PB values; at the same time, the RMS error of the proposed model is 2 times lower than that of the popular GB_OBC model from the AMBER package. Tests against the PB treatment on 22 biomolecular structures including proteins and DNA show that the relative error of ΔG_{el} is 0.58%; the RMS error of ΔG_{el} computed by AR6 is 3 times lower than the corresponding value for GB_OBC. However, the computational efficiencies of the AR6 and GB_OBC models are comparable. A variant of the R6 model, NSR6, based on numerically exact integration over triangulated molecular surface is tested on a "challenge" set of small drug-like molecules (Nicholls et al. *J. Med. Chem.* **2008**, *51*, 769−779). When augmented with cavity and VDW terms to account for the nonpolar part of solvation energy, the model with only one free parameter is capable of predicting the total solvation free energy to within 1.73 kcal/mol RMS error relative to experimental data. Within the NSR6 formulation, computation of the nonpolar contribution is particularly efficient because its VDW part depends on the same $|\mathbf{r}|^{-6}$ integrals.

1. Introduction

An accurate description of solvent is essential for modeling and simulation of biological macromolecules. Currently, the most rigorous procedure for modeling the effect of aqueous solvent is to explicitly model every water molecule surrounding the macromolecule. For many applications though, this method is computationally too intense. Implicit solvent models, in which solvent molecules are represented by a continuum function, have become a popular alternative to explicit solvent methods, as they are more computationally efficient.[1−7] Within the framework of implicit solvent models, macromolecules are treated as a low dielectric

* To whom correspondence should be addressed. E-mail: alexey@cs.vt.edu.

[†] Department of Computer Science, Virginia Tech.

[‡] Michigan State University.

[§] Departments of Computer Science and Physics, Virginia Tech.

10.1021/ct100392h
Published on Web 11/24/2010

medium (ε_{in}), surrounded by a high dielectric medium (ε_{out}). The effect of the solvent is represented by the solvation free energy: ΔG_{solv}. The solvation free energy is typically divided into polar (ΔG_{el}) and nonpolar (ΔG_{nonpol}) terms. In this work, we will focus on the calculation of the polar part of the solvation free energy.

Within the linear response continuum implicit solvent framework, solving the Poisson−Boltzmann equation (PB) is theoretically the most rigorous way to compute ΔG_{el}.[1−3,6,8−10] However, the PB model may become quite time-consuming, especially if applied to a large set of conformations of a macromolecule, or if it is incorporated into molecular dynamics (MD) simulations where its practical implementation faces several other challenges. The generalized Born model (GB) has become popular as an alternative to the PB model for the computation of ΔG_{el},[11−34] especially in MD.

The GB model approximates ΔG_{el} using the following formula:

$$\Delta G_{el} \approx \Delta G_{GB} = -\frac{1}{2}\sum_{ij}\frac{q_i q_j}{f^{GB}(r_{ij}, R_i, R_j)}\left(\frac{1}{\varepsilon_{in}} - \frac{1}{\varepsilon_{out}}\right) \quad (1)$$

where r_{ij} is the distance between atoms i and j, q_i is the partial charge of atom i, R_i is the so-called *effective Born radius* of atom i, and the most widely used functional form[12] of f^{GB} is $f^{GB} = [r_{ij}^2 + R_i R_j \exp(-r_{ij}^2/4R_iR_j)]^{0.5}$, although other similar expressions have been tried.[18,35]

Recently, it has been shown that eq 1 produces a systematic error (with respect to PB results) when applied to systems with finite values of ε_{in} and ε_{out}.[36] Sigalov et al.[37] have proposed a modified GB model (ALPB) that eliminates this systematic error while keeping the computational efficiency of Still's original equation:

$$\Delta G_{el} \approx -\frac{1}{2}\left(\frac{1}{\varepsilon_{in}} - \frac{1}{\varepsilon_{out}}\right)\frac{1}{1+\beta\alpha}\sum_{ij} q_i q_j\left(\frac{1}{f^{GB}} + \frac{\alpha\beta}{A}\right) \quad (2)$$

where $\beta = \varepsilon_{in}/\varepsilon_{out}$, $\alpha = 0.571412$, and A is the electrostatic size of the molecule, which is essentially the overall size of the structure, that can be computed analytically.[37] The ALPB model is currently implemented in AMBER,[38] and it will be used throughout this work to compute ΔG_{el}.

Much of the efforts of recent studies aimed at improving the accuracy of the GB model focused on the computation of the effective Born radii R_i, because it is the computation of R_i that, to a large extent, determines the accuracy and efficiency of the entire GB model. One procedure to compute R_i, the so-called "perfect" effective Born radii, is to derive them directly from the self-energies computed with the PB model. It was shown that if the "perfect" effective Born radii are used in eq 1, the GB ΔG_{el} are in close agreement with those of the PB.[35] The computationally expensive "perfect" effective Born radii are commonly used for benchmarking and testing different GB "flavors"—approximations that compute R_i.

Many existing practical GB "flavors" are based on the so-called "Coulomb field approximation" (CFA) in which the effective Born radius of atom i is computed by

$$R_i^{-1} = \rho_i^{-1} - \frac{1}{4\pi}\int_{|\mathbf{r}-\mathbf{r}_i|>\rho_i}^{solute} |\mathbf{r} - \mathbf{r}_i|^{-4}\, d\mathbf{V} \quad (3)$$

where ρ_i is the intrinsic radius of atom i and the integration is over the volume inside the molecule (solute) but outside the atom i. $\mathbf{r}_i$ is the position of atom i with respect to some fixed frame. Among the methods based on CFA, the GB_OBC[32] flavor, available in the AMBER package, has become quite popular, especially in molecular dynamics simulations. This is due to a reasonable compromise between accuracy and speed offered by GB_OBC. Nevertheless, recent comparisons between implicit and explicit models applied to a deca-alanine (Ala10) molecule have shown that the GB_OBC method (and other GB models tested in ref 39) has a clear bias in the free energies of solvation—hence in the relative population—of four different conformational states of Ala10; please refer to ref 39 for details. At the same time, ΔG_{el} values computed with the numerical PB model were in considerably closer agreement with the explicit solvent results, suggesting that the GB accuracy can still be improved by achieving a closer match with the underlying PB model.

A different expression to compute the effective Born radii (R6 radii), which will be called here "R6 integration", was proposed by Svrcek-Seiler[40] and independently by Grycuk[41] as an alternative to the CFA:

$$R_i^{-1} = \left(\frac{3}{4\pi}\int_{ext}\frac{d\mathbf{V}}{|\mathbf{r}-\mathbf{r}_i|^6}\right)^{1/3} = \left(\rho_i^{-3} - \frac{3}{4\pi}\int_{r>\rho_i}^{solute}|\mathbf{r}|^{-6}\, d\mathbf{V}\right)^{1/3} = (\rho_i^{-3} - \mathbf{I}_i^{tot})^{1/3} \quad (4)$$

where in the first expression the integral (ext) is taken over the region outside the molecule. In the second integral, the origin is moved to the center of atom i. Unlike the CFA radii in eq 3, the "R6 radii" are exact for any location of a charged atom within a perfect spherical solute in the $\varepsilon_{out}/\varepsilon_{in} \gg 1$ limit. Recently, Mongan et al.[42] have shown that when the "R6 radii" are computed by essentially exact numerical integration of eq 4, the resulting effective radii and ΔG_{el} are in very close agreement with the PB reference for realistic biomolecular shapes. Thus, the use of "R6 radii" in eq 1 or 2 can potentially eliminate some of the deficiencies of the methods based on CFA. Although the R6 radii potentially offer advantages over the CFA-based methods, analytical methods that compute the "R6" effective Born radii over a physically realistic molecular (Lee−Richards[43]) volume do not yet exist to the best of our knowledge. Analytical, differentiable expressions for the computation of effective Born radii are preferred to their numerical counterparts, as the former are easily extended to calculate solvation forces needed by MD simulations and are often more computationally efficient.

Recently, Tjong and Zhou[44] and Labute[45] have reported analytical methods to compute "R6 radii" in which eq 4 is integrated over the van der Waals (VDW) volume of the solute. These are important steps in the development of the "R6" flavor. However, the use of VDW volume creates multiple interstitial regions of unphysical high dielectric pockets that are smaller than the water molecule. In contrast,

PB calculations generally use the Lee−Richards molecular surface as a dielectric boundary, defined by rolling a solvent sphere over the surface of the molecule. This definition was shown to produce consistently better agreement with the explicit solvent than the VDW based one.[46,47] This point will be visited later in this work, using deca-alanine (Ala10) as an example.

The GBMV2 (generalized Born using molecular volume) model developed by Lee et al.[48] is perhaps the best example of a GB flavor in which the effective radii are obtained through integration over a very close approximation of the Lee−Richard molecular volume. The model has been one of the most successful GB flavors in the ability to reproduce the "perfect" effective Born radii and total solvation free energies of proteins. Nonetheless, GBMV2 is substantially more computationally expensive than comparable VDW-like GB models such as GBSW[49] in CHARMM or AMBER GB variants.[50] The relative computational expense of the GBMV2 model becomes even more noticeable if one also factors in the relative speed of conformational sampling. Here, GB flavors based on "smooth" molecular volume may lead up to several orders of magnitude of speedup in the conformational search.[51] Finally, methods based on a sharp molecular surface definition such as GBMV2 can produce unstable or infinity forces and lead to energy conservation problems when used in MD simulations[52]

In this work, we have developed a new analytical method to compute the effective Born radii based on the R6 integration. Although the method starts with a computationally efficient pairwise approximation over the VDW volume, it includes several molecular volume corrections terms designed to approximate the "true" molecular volume in the vicinity of the atom in question, thus improving the accuracy of the calculations but at the same time avoiding problems associated with the use of a sharp Lee−Richards molecular surface. We show that the proposed method keeps the computational efficiency and stability of the previous GB models implemented in AMBER, such as GB_OBC.

2. Theory

2.1. Numerically Exact Computation of the R6 Radii: NSR6. The inverse of the R6 effective Born radius of atom i can be computed numerically using the surface formulation outlined in Mongan et al.[42] Within this formulation, R_i is calculated by the following equation:

$$R_i^{-1} = \left(-\frac{1}{4\pi} \oint_{\partial V} \frac{\mathbf{r} - \mathbf{r_i}}{|\mathbf{r} - \mathbf{r_i}|^6} \cdot d\mathbf{S} \right)^{1/3} \quad (5)$$

which according to the Gauss−Ostrogradski theorem, is equivalent to eq 4. Here, ∂V represents the molecular surface of the molecule, and $d\mathbf{S}$ is the infinitesimal surface vector. After a triangulation of the surface, R_i is approximated by

$$R_i^{-1} \approx \left(-\frac{1}{4\pi} \sum_k \frac{(\mathbf{c}_k - \mathbf{r}_i) \cdot \hat{\mathbf{n}}_k S_k}{|\mathbf{c}_k - \mathbf{r}_i|^6} \right)^{1/3} \quad (6)$$

where the summation is performed over the surface triangles. For each surface triangle k, $\mathbf{c}_k$ represents the position of its center, S_k its area, and $\mathbf{n}_k$ is a unit vector orthogonal to the triangle k pointing toward the inside of the solute.

In this work, the surface triangulation is carried out over the Lee−Richards molecular surface, which is computed and triangulated by using the MSMS package,[53] see the Methodological Details section for details. Since this procedure, which will be called "numerical surface R6 integration" or "NSR6", gives a numerically exact value of R_i, it will be subsequently used for accuracy benchmarking. Computationally, NSR6 is still much faster than the brute force numerical integration[42] over the molecular volume in eq 4; the reason for the relative inefficiency of the numerical volume-based approach in the context of the R6 is mentioned below.

2.2. Approximate Analytical Computation of the R6 Radii: AR6. In this section, we propose an analytic approach to approximate the R6 radii on the basis of integration of eq 4 over an approximation of the true molecular volume. A reliable and useful model for computing effective Born radii should strive for a balance between being reasonably accurate, computationally efficient, and capable of avoiding the problems of sharp molecular boundary definitions. In order to fulfill all of these requirements, we have designed a methodology that consists of several components which will be described in the following subsections.

2.2.1. Overall Approach. In order to analytically compute R6 radii (eq 4), we propose an approach based on the integration of several geometrical approximations aimed to effectively represent different regions of the true molecular volume. It is important to note that due to the sixth power in eq 4, the R6 approach is very sensitive to inaccuracies in the immediate vicinity of the atom in question. For this reason, our approximation to the R6 integral over molecular volume was designed to deliver maximum accuracy in the region closest to the focus atom. First, for every atom i of the molecule, we separate a predefined small group of covalently linked atoms, including atom i, over which the R6 integration is precomputed numerically. This group of atoms will be referred to as the "chunk" of atom i. The second approximation consists of the R6 integration over "neck" regions defined as solvent-inaccessible spaces between atom i and nearby atoms not belonging to the "chunk" of atom i. The integration over the "necks" is approximated by an empirical and simple pairwise function, following the same strategy described in Mongan et al.[54] in which "necks" were originally introduced in the context of $|r|^{-4}$ integrals. Finally, atoms outside the "chunk" region (arguably the region where eq 4 is least sensitive to inaccuracies) are treated very efficiently as VDW spheres whose contribution to the total R6 integration are analytically derived. Thus, the molecular volume that surrounds atom i is approximated by the union of three distinct regions (Figure 1): (1) the essentially exact molecular volume of the "chunk" of atom i, (2) the "neck" regions between atom i and its nearby atoms, which accounts albeit approximately for the interstitial low dielectric regions present in the true molecular volume, (3) the atomic VDW volume, excluding atoms inside the chunk of atom i. The second volume integral in eq 4 is approximated by:

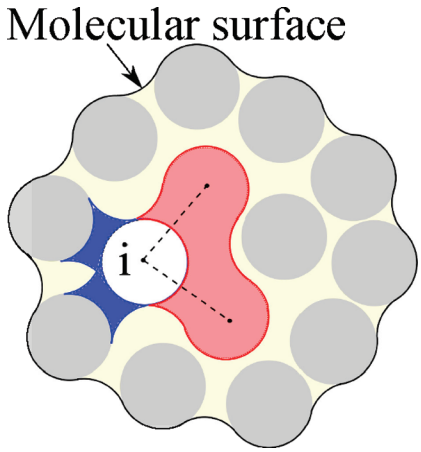


Figure 1. Illustration of the three regions of integration in eq 7 that are combined to approximate the molecular volume: VDW volume (light gray spheres), neck regions (dark blue), and "chunk" molecule (red). The open sphere represents atom *i*, and the dashed lines represent covalent bonds used to define which atoms belong to the chunk molecule.

$$\mathbf{I}_i^{\text{tot}} = \frac{3}{4\pi}\int_{r>\rho_i}^{\text{solute}} |\mathbf{r}|^{-6}\, dV \approx \mathbf{I}_i^{\text{vdw}} + \mathbf{I}_i^{\text{neck}} + \mathbf{I}_i^{\text{chunk}} \quad (7)$$

where $\mathbf{I}_i^{\text{vdw}}$ represents the R6 integration over the van der Waals volume outside the "chunk" of atom *i*, $\mathbf{I}_i^{\text{neck}}$ represents the R6 integration over the "neck" regions (see ref 54 for details), and $\mathbf{I}_i^{\text{chunk}}$ is the R6 integration over the molecular volume of the "chunk". In Figure 1, the regions of integration of $\mathbf{I}_i^{\text{vdw}}$, $\mathbf{I}_i^{\text{neck}}$, and $\mathbf{I}_i^{\text{chunk}}$ are represented by light gray, blue, and red colors, respectively.

The above approximation will overcount overlapping regions between necks and atoms outside the "chunks". Therefore, the contribution of $\mathbf{I}_i^{\text{vdw}}$ and $\mathbf{I}_i^{\text{neck}}$ are reduced in an appropriate manner; this procedure introduces two adjusting parameters, S_{vdw} and S_{neck}, in the overall procedure. One additional integer parameter, "chunk depth", is used to control the sizes of the "chunk" region.

The previous approach provides good results for small molecules of at most a couple hundred atoms. In the case of large structures though, the methodology described above produces a systematic underestimation of the volume of integration, because the model does not account for the interstitial space between atoms far from the vicinity of atom *i*, seen as yellow space in Figure 1. To address this underestimation, we use an additional volume correction which requires the use of two additional parameters.

2.2.2. Integration over van der Waals Volume: $\mathbf{I}_i^{vdw}$. Here, we compute the $\mathbf{I}_i^{\text{vdw}}$ integral in eq 7 over the individual VDW atomic spheres that make up the molecule; the $|\mathbf{r}|^{-6}$ integral contribution of the VDW sphere of atom *j* to the effective Born radius of atom *i* was analytically calculated previously.[44,55] Let ρ_i and ρ_j be the VDW radii of atoms *i* and *j*, respectively, and let r_{ij} be the distance between their centers. Then, the contribution of atom *j* to $\mathbf{I}_i^{\text{vdw}}$ is described by the following function $\mathbf{F}_6$, which is divided into four cases according to the mutual position of both atoms:

Case I. There is no overlap between atoms *i* and *j*: $r_{ij} \geq \rho_i + \rho_j$

$$\mathbf{F}_6(\rho_i, \rho_j, r_{ij}) = \frac{\rho_j^3}{(r_{ij}^2 - \rho_j^2)^3} \quad (8)$$

Case II. Atoms *i* and *j* overlap: $(r_{ij} > |\rho_i - \rho_j|) \wedge (r_{ij} < \rho_i + \rho_j)$

$$\mathbf{F}_6(\rho_i, \rho_j, r_{ij}) = \frac{1}{16 r_{ij}}\left(\frac{r_{ij} + 3\rho_j}{(r_{ij} + \rho_j)^3} + \frac{3(\rho_j^2 - \rho_i^2 - (r_{ij} - \rho_i)^2) + 2r_{ij}\rho_i}{\rho_i^4}\right) \quad (9)$$

Case III. Atom *j* "swallows" *i*: $(\rho_i < \rho_j) \wedge (r_{ij} \leq \rho_j - \rho_i)$

$$\mathbf{F}_6(\rho_i, \rho_j, r_{ij}) = \frac{1}{\rho_i^3} + \frac{\rho_j^3}{(r_{ij}^2 - \rho_j^2)^3} \quad (10)$$

Case IV. Atom *i* "swallows" *j*: $(\rho_j < \rho_i) \wedge (r_{ij} \leq \rho_i - \rho_j)$

$$\mathbf{F}_6(\rho_i, \rho_j, r_{ij}) = 0 \quad (11)$$

It is worth noting that cases III and IV never occur in biological macromolecules; we list them here for the sake of completeness. In practical implementations, e.g., in AMBER, the VDW radius of atom *j* is multiplied by a scaling factor $S_{\text{vdw}}^j < 1$, to correct for overcounting of the volume due to possible overlaps between VDW spheres of neighboring atoms. Then, the total contribution of VDW spheres is

$$\mathbf{I}_i^{\text{vdw}} = \sum_{j \notin \text{"chunk"} i} \mathbf{F}_6(\rho_i, (S_{\text{vdw}}^j)\rho_j, r_{ij}) \quad (12)$$

where the summation is performed over all of the atoms of the molecule not included in the "chunk" of atom *i*. Compared to the methods currently implemented in AMBER, we use a simplified version of the rescaling, in which $S_{\text{vdw}}^j = S_{\text{vdw}}$ is constant for all atoms of the molecule (we have found that $S_{\text{vdw}} = 0.6211$ gives the best results, see below).

2.2.3. Integration over Neck Regions: $\mathbf{I}_i^{neck}$. Here, we consider a correction term which accounts for the integration of $|\mathbf{r}|^{-6}$ over the "neck" space between pairs of atoms (represented by their VDW spheres). This correction term was first introduced by Mongan et al.[54] in the context of the CFA; here, we extend it to the computation of the R6 radii. The "neck" region between atoms *i* and *j*, represented by the blue region in Figure 2, is completely determined by their VDW radii ρ_i and ρ_j, the distance r_{ij} between them, and the water probe radius ρ_w. Moreover, the "neck" exists only if the distance between atoms *i* and *j* is less than $\rho_i + \rho_j + 2\rho_w$. To approximate the integral of $|\mathbf{r}|^{-6}$ over the "neck" region, we use the following analytical and empirical function:

$$\text{neck\_integral}(r_{ij}, \rho_i, \rho_j) \approx A_{ij}(r_{ij} - B_{ij})^4(\rho_i + \rho_j + 2\rho_w - r_{ij})^4 \quad (13)$$

for interatomic distances (r_{ij}) less than $\rho_i + \rho_j + 2\rho_w$ and greater than B_{ij}. Otherwise, neck_integral(r_{ij},ρ_i,ρ_j) is set to zero. Thus, the actual computation is performed only for those atoms that are within the above distance from the atom in question. The corresponding computational complexity is thus $O(N)$, in contrast to the computation of the VDW contribution that scales as $O(N^2)$, where *N* is the total number

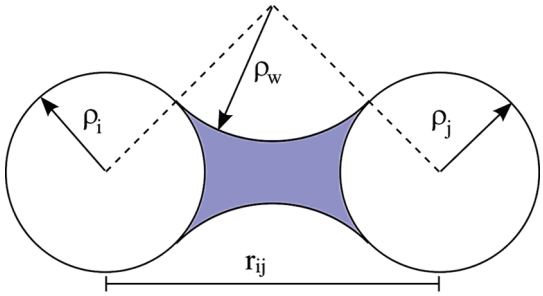


Figure 2. Neck region (blue) between two atoms with radii ρ_i and ρ_j and a water probe radius ρ_w. r_{ij} represents the distance between atoms *i* and *j*.

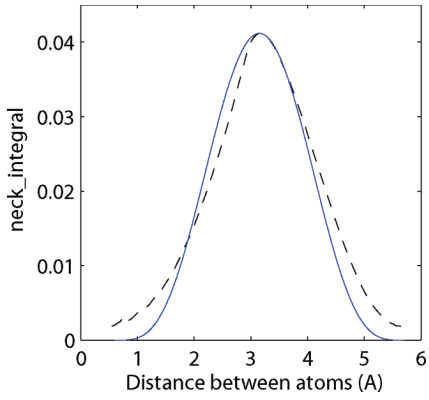


Figure 3. The numerical integration over the "neck" region (dashed black) compared with the analytical approximation (solid blue) used here. In this example, we have used $\rho_i = 1.7$, $\rho_j = 1.2$, and probe radius $\rho_w = 1.4$ Å.

of atoms in the molecule. The neck_integral(r_{ij},ρ_i,ρ_j) function is parametrized by A_{ij} and B_{ij}, which depend on ρ_i, ρ_j, and ρ_w. Following a similar procedure to that of ref 54, we tabulate the optimum values of A_{ij} and B_{ij} for different values of ρ_i, ρ_j, and ρ_w. To obtain optimum values of A_{ij} and B_{ij}, we compute the integral of $|\mathbf{r}|^{-6}$ over the "neck" region by using the NSR6 procedure applied to a diatomic molecule composed of the atoms *i* and *j* located at various distances (different values of r_{ij}, Figure 2). We then store the distance r_{ij}^{max} at which the integration over the "neck" region reaches its maximum $neck_{max}$. The value of B_{ij} is calculated by $B_{ij} = 2r_{ij}^{max} - (\rho_i + \rho_j + 2\rho_w)$, and the value of A_{ij} is computed such that neck_integral(r_{ij}^{max},ρ_i,ρ_j) = $neck_{max}$. The values of A_{ij} and B_{ij} for a range of ρ_i and ρ_j values are available in the Supporting Information. Figure 3 illustrates that eq 13 is a reasonable approximation of the "R6 integration" over the "neck" region. By construction, eq 13 is differentiable in the entire domain of r_{ij}.

Finally, the total integral over neck regions is approximated by

$$\mathbf{I}_i^{neck} = \frac{3}{4\pi} S_{neck} \sum_{j \notin \text{"chunk"} i} \text{neck\_integral}(r_{ij}, \rho_i, \rho_j) \quad (14)$$

where S_{neck} is a free parameter used to correct for the volume overcounting due to overlaps between adjacent "neck" regions, and overlaps between atoms and necks (we have found that $S_{neck} = 0.4058$ gives the best results, see below).

Integration over Chunk Regions: $\mathbf{I}_i^{chunk}$. Since the integrand $|\mathbf{r}|^{-6}$ is very large in the vicinity of atom *i*, it is critical to

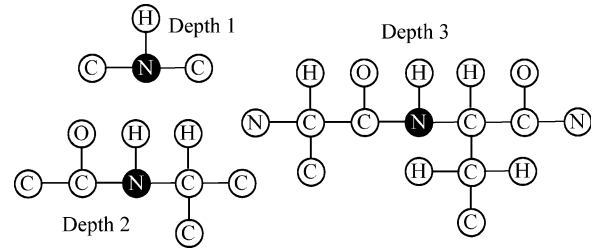


Figure 4. Examples of "chunk" molecules of depths 1, 2, and 3 used for the computation of the effective Born radius of a nitrogen atom (black circles).

treat the nearby regions of molecular volume particularly carefully, ideally exactly. Compared to the relatively lower power $|\mathbf{r}|^{-4}$ of the CFA integrand, this problem becomes especially critical in the case of the R6. In our previous work that focused on foundations of the R6[42] rather than its practical implementation, the required accuracy was achieved by brute force via inefficient numerical volume integration over a very fine 3D mesh in the vicinity of *i*. Since here we are set to develop an efficient analytical model, we take a completely different approach. We isolate a small set of neighboring atoms covalently connected to the atom of interest *i*; see the exact definition below. The geometrical configuration of this small set of atoms, which will be called "chunk", is not expected to change substantially during dynamics. Thus, the contribution of the "chunk" to the effective Born radius of atom *i*, $\mathbf{I}_i^{chunk}$, can be computed essentially exactly by the NSR6 procedure at the setup stage and then subsequently reused at all other steps.

The neighbor atoms that form the "chunk" molecule for a given atom *i* are determined by setting the "chunk depth", which is defined as the maximum possible integer distance (in the graph-theoretic sense where atoms are the vertices and covalent bonds are edges) between atom *i* and any other atom in the "chunk". In Figure 4, we show examples of "chunks" of depths 1, 2, and 3 for the computation of the effective Born radius of a nitrogen atom located in the protein backbone. The "R6 radius" of each atom is computed with the same specified "chunk depth", except for the atoms with only one bonded neighbor, such as hydrogen atoms. For these atoms, the specified "chunk depth" is increased by 1. This way, atoms with only one bonded neighbor and atoms with multiple covalent neighbors are processed using chunks of the same size. For example, when the "chunk depth" is set to 1, the "chunk" used for the hydrogen atom of the molecule labeled "Depth 1" in Figure 4 is composed of all of the atoms of this molecule, which is the same as the "chunk" of Depth =1 for the nitrogen atom.

Note that:

(a) The set of atoms that form the "chunk" do not change during the classical dynamics of a molecule. If the chunk depth is small enough, the chunk's overall shape is maintained during dynamics.

(b) The contribution of the chunk to the effective Born radius of atom *i* can be calculated essentially exactly by the NSR6 procedure described above.

To take into account possible variations (presumably still small) in the chunk geometry during dynamics, we augment

the computation of $\mathbf{I}_i^{\text{chunk}}$ as follows. The idea is to use a fast analytical expression for $\mathbf{I}_i^{\text{chunk}}$ but correct it at every step by a constant factor which accounts for the discrepancy between the approximate analytical and the exact numerical values of the $|\mathbf{r}|^{-6}$ integral over the "chunk". To this end, we define a correction factor, λ_i, as the ratio between the numerically computed and the analytically computed values of $\mathbf{I}_i^{\text{chunk}}$; the constant λ_i is estimated once at the setup stage (e.g., at time = 0). For all other steps, $\mathbf{I}_i^{\text{chunk}}$ is computed analytically on the basis of the current geometry of the "chunk", multiplied by the rescaling factor λ_i previously computed, which compensates for the discrepancy between the analytical and numerical results. The following two equations define the procedure:

$$\lambda_i = \frac{\rho_i^{-3} - (\alpha_i^{\text{chunk}})^3}{\sum_{k\neq i}^{M} \mathbf{F}_6(\rho_i, \rho_k, r_{ik}^{o})} \tag{15}$$

$$\alpha_i^{\text{chunk}} = \left(-\frac{1}{4\pi}\oint_{\partial V_{\text{chunk}}} \frac{\mathbf{r} - \mathbf{r}_i}{|\mathbf{r} - \mathbf{r}_i|^6} \cdot d\mathbf{S}\right)^{1/3} \tag{16}$$

where M is the number of atoms in the "chunk", r_{ik}^{o} is the distance between atoms i and k found in the structure used to set up the computation (e.g., at time = 0). $\partial V_{\text{chunk}}$ represent the surface of the "chunk" molecule, ρ_i is the intrinsic radius of atom i, and F_6 is the same function used for VDW integration. The value of α_i^{chunk} in eq 16, which is just the effective Born radius of the "chunk", is computed by the NSR6 procedure.

Once the values of λ_i are computed at the setup stage for each atom, the values of $\mathbf{I}_i^{\text{chunk}}$ for all of the following steps are computed by

$$\mathbf{I}_i^{\text{chunk}} = \lambda_i \sum_{k\neq i}^{M} \mathbf{F}_6(\rho_i, \rho_k, r_{ik}) \tag{17}$$

The "neck" regions of atoms that belong to the "chunk" or that are covalently bonded to at least one atom of the "chunk" are not considered, as they are very likely to overlap with the "chunk" region (the corresponding neck integrals, eq 13, are not computed). This restriction greatly reduces the number of "necks" needed for each atom. For example, the average number of possible necks per atom for thieredoxin (2TRX) is 60. However, once the "chunks" are defined and their atoms excluded from the neck computation, the average number of necks per atom reduces to 40 (30% reduction); for small structures such as Ala10, the reduction can approach 50%. It is important to note that the necks are still present between atoms that are close in real space and far in bond graph space, for example, those that form hydrogen bonds. So we expect that the recapitulation of the first peak in the PFMs—signature of the use of true molecular volume—presented in ref 54 in which necks were originally defined will still be maintained.

2.2.4. Rescaling the Effective Born Radius. In order to achieve the same computational benefits of the GB_OBC model, such as numerical stability and efficiency, and to obtain better accuracy for deeply buried atoms, we use a similar radii rescaling procedure, which is determined by the following equations that yield R_i^{-1}:

$$\tilde{\rho}_i^{-3} = \rho_i^{-3} - \mathbf{I}_i^{\text{chunk}} \tag{18}$$

$$c_i = 1 - \frac{1}{A^3 \tilde{\rho}_i^{-3}} \tag{19}$$

$$\Psi = (\mathbf{I}_i^{\text{vdw}} + \mathbf{I}_i^{\text{neck}}) \tag{20}$$

$$\beta_0 = 1/c_i \tag{21}$$

$$R_i^{-1} \approx (\tilde{\rho}_i^{-3} - c_i \tilde{\rho}_i^{-3} \tanh(\beta_0 \Psi \tilde{\rho}_i^3 - \beta_1 (\Psi \rho_i^3)^2 + \beta_2 (\Psi \rho_i^3)^3))^{1/3} + B \tag{22}$$

Here, A is the electrostatic size of the molecule, which is essentially its "global" size, see ref 37 for details. Simple and robust routines for computing this parameter are available; in practical MD simulations, it can be approximated by a constant. The rescaling process in eqs 18−22 was built such that if $\Psi \to \infty$, then $R_i \to A$. Thus, the effective Born radius is upper-bounded by the molecular size A. On the other hand, if $\Psi \ll 1$, then $R_i^{-1} \approx (\tilde{\rho}_i^{-3} - \mathbf{I}_i^{\text{vdw}} - \mathbf{I}_i^{\text{neck}})^{1/3}$: the effective Born radii of surface atoms (with small effective radii) are not affected by the rescaling process.

The constant offset parameter B was defined in ref 42 and has a value of 0.028 Å^{-1}. This parameter was introduced to minimize the difference between the computed R6 radii and the "perfect" effective Born radii for a molecular surface computed with a water probe = 1.4 Å; β_1 and β_2 are adjustable parameters to be optimized.

2.2.5. Additional Volume Correction. When eq 22 is applied to relatively large macromolecules such as lysozyme or thioredoxin, we observe that while the computed effective Born radii of solvent-exposed atoms are accurately estimated, the effective Born radii of deeply buried atoms are systematically underestimated, relative to the "perfect" effective Born radii. To correct this underestimation, we further rescale the values of Ψ, eq 20, such that they are increased for buried atoms but unaffected for solvent-exposed atoms. The rescaling is achieved by multiplying Ψ by a function V_i that is proportional to the degree of burial of atom i. This function is similar to that of the "measure of the volume" introduced by the FACTS[56] analytical model of solvation:

$$V_i = \frac{\sum_{j=1, j\neq i}^{N} \rho_j^3 \Theta_{ij}}{R_s^3} \tag{23}$$

where

$$\Theta_{ij} = \begin{cases} \left(1 - \left(\frac{r_{ij}}{R_s}\right)^2\right)^2 & r_{ij} \leq R_s \\ 0 & r_{ij} > R_s \end{cases} \tag{24}$$

The parameter R_s is set to 10 Å, which is the same value used in the FACTS method.[56] The inverses of the effective Born radii are then computed by the following differentiable expression:

$$R_i^{-1} \approx (\tilde{\rho}_i^{-3} - c_i\tilde{\rho}_i^{-3}\tanh(\beta_0\Psi\tilde{\rho}_i^3 - \beta_1(V_i\Psi\rho_i^3)^2 + \beta_2(V_i\Psi\rho_i^3)^3))^{1/3} + B \quad (25)$$

which is the formula that defines the AR6 (Analytical R6) GB flavor to be used throughout the rest of this work.

2.3. Parametrization. There are four parameters to be optimized in the AR6 procedure, S_{vdw}, S_{neck}, β_1, and β_2. In the absence of a unique accepted strategy for such optimizations, a short discussion is due on the logic behind the approach we take. Generally, one can consider two extreme cases. On one end of the spectrum is the purely geometric approach which aims only at achieving the closest agreement between the approximate analytical and the "perfect"(exact) effective Born radii. This approach is expected to work well in a situation where the approximate analytical effective radii can be made "uniformly" near-perfect via a suitable parametrization. When substituted into the "canonical" GB (Still's) formula, eq 1, these would give ΔG_{el} values very close to those that can be obtained with the perfect (exact) radii without any danger of overfitting, that is, without exceeding the inherent accuracy limitations of Still's formula itself. Such an approach was taken in ref 42 to arrive at the optimal value of a small constant offset parameter B (see above) that gave the best agreement between the numerical R6 and perfect (PB) radii. However, if the agreement between the optimal approximate and the perefect radii is expected to be nonuniform, for example, if the largest radii are expected to be consistently underestimated, the approach is likely to be suboptimal in terms of the accuracy of ΔG_{el} since it places equal weights on different effective radii (small radii contribute more to the solvation energy). On the other end of the spectrum is the approach, often taken, where parameters of the GB flavor are optimized to give the most accurate values of ΔG_{el}, or other energetic quantities, relative to some appropriate reference such as the PB or explicit solvent energies. The obvious advantage of the approach is a more accurate ΔG_{el} for the training set. The danger is overfitting. A good agreement between approximate and reference ΔG_{el} along with poor agreement between the approximate and perfect radii is an indicator of the problem; it was seen in earlier GB flavors.[57] In this work, we take a middle ground between these two extremes: the four parameters of AR6 are optimized against ΔG_{el} obtained via Still's equation with NSR6 radii, not the PB solvation energies. Note that the energies obtained by the GB model using numerically computed R6 radii are in good agreement with those obtained by PB.[42] We also test agreement with the corresponding perfect radii, see below. To reduce the possibility of overfitting further, we fit the two sets of parameters, $\{S_{vdw},S_{neck}\}$ and $\{\beta_1,\beta_2\}$, independently.

The rescaling factors S_{vdw} and S_{neck}, eqs 12 and 14, are optimized such that the total electrostatic solvation energies ΔG_{el} obtained by AR6 (through eq 2) match the ΔG_{el} of the NSR6 procedure for four conformational states of an alanine decapeptide (Ala10) represented in Figure 5. For the optimization, each of the four conformational states of Ala10 was represented by 10 MD snapshots.[39] The ΔG_{el} corresponding to each conformational state is computed by averaging the values of ΔG_{el} of each of their corresponding MD snapshots. We have chosen the NSR6 ΔG_{el} rather than the available TIP3P or PB numbers for optimization to avoid overfitting. At this stage, the optimization is carried out with $\beta_1 = \beta_2 = 0$, as these parameters are intended to correct the underestimation of the effective Born radius of deeply buried atoms, not found in the relatively small Ala10. Moreover, fitting only two parameters at a time reduces the likelihood of overfitting and allows for an exhaustive exploration of the parameter domain.

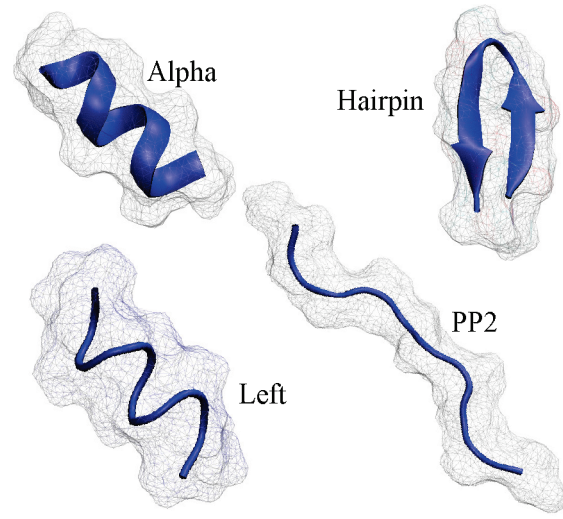


Figure 5. Cartoon representation of the four conformational states of alanine decapeptide, Ala10, used in this work.

We have used the Nelder−Mead[58] simplex algorithm for optimization. The objective function to be minimized was the RMS deviation of total ΔG_{el} between the NSR6 and AR6. The "chunk" contribution used in AR6 can be computed from any of the four conformational states of Ala10; this results in four different values of ΔG_{el} for each conformational state of Ala10. The ΔG_{el} for each conformation used for optimization is computed as the average of these four values. The optimization was carried out using chunks of depth 3, as they are the smallest chunks that provide correct ordering of the values of $\Delta\Delta G_{el}$ between the four conformational states of Ala10, see Table 1. Although the accuracy of the approximation (determined by the RMSD values of Table 1) increases with the chunk depth, the larger the "chunk", the less accurate is our assumption that the "chunk" does not change substantially during dynamics: depth = 3 appears to be an optimum compromise between these two opposite trends. This important point will be discussed in more detail below. For the rest of the analysis presented here, we use only the depth = 3 model.

The energies obtained by using AR6 with optimized parameters S_{vdw} and S_{neck} are in good agreement with the energies obtained by using NSR6. It may be possible though that this is the result of a fortuitous compensation between the inherent errors in Still's equation of the GB model (eq 1) and the errors due to the approximation of the effective Born radii. Figure 6 shows the correlation plots between the effective Born radii computed with the AR6 and NSR6 methods for the four different conformational states of Ala10. The best agreement is obtained for the most solvent-exposed conformational state "pp2", with a correlation coefficient of 0.9968. For more compact structures such as "alpha" and

Table 1. Free Energies of Solvation for Different Conformations of Ala10 (kcal/mol) Obtained with the AR6 and the NSR6 Procedures[a]

		AR6			
	NSR6	depth 1	depth 2	depth 3	depth 4
			(A) ΔG_{el}		
alpha	−45.73	−44.10	−46.53	−45.84	−45.51
PP2	−77.85	−73.37	−79.97	−78.30	−78.24
left	−50.91	−47.81	−50.16	−51.13	−50.86
hairpin	−54.59	−52.98	−57.07	−54.95	−54.28
RMSD	0.0	2.96	1.71	0.31	0.27
			(B) $\Delta\Delta G_{el}$		
PP2-alpha	−32.12	−29.27	−33.44	−32.46	−32.73
PP2-left	−26.94	−25.56	−29.81	−27.17	−27.38
PP2-hairpin	−23.26	−20.39	−22.90	−23.35	−23.96
alpha-left	5.18	3.71	3.63	5.29	5.35
alpha-hairpin	8.86	8.88	10.54	9.11	8.77
left-hairpin	3.68	5.17	6.91	3.82	3.42

[a] Solvation energies were computed using $\varepsilon_{out} = 80$, $\varepsilon_{in} = 1$, and $\kappa = 0$. The parameters used are $S_{vdw} = 0.6211$, $S_{neck} = 0.4058$, and $\beta_1 = \beta_2 = 0$. The values of RMSD are relative to the NSR6 procedure.

"left", AR6 also shows a good agreement with that of NSR6 with correlation coefficients of 0.9802 and 0.9799, respectively. These results show that although the parameters were optimized using total solvation energies, there is also a good agreement between the effective Born radii obtained by AR6 and NSR6 for all of the conformational states of Ala10, and thus the amount of possible error cancellation is not much different from what one can expect from exact R6 used in Still's formula, eq 1.

Parameters β_1 and β_2 are meant to control the rescaling process for large radii in eq 25, such that the rescaling is large for deeply buried atoms and small for the exposed ones. These parameters have little effect on effective radii of small structures such as Ala10. Again, we used the Nelder−Mead algorithm for the optimization. The objective function that was minimized in this case is the RMSD between the ΔG_{el} obtained by the GB and PB models for a training set consisting of 11 proteins and two snapshots of the denaturing trajectory of apo-myoglobin; the PDB codes of the 11 proteins of the training set are presented in Table 7 (bold letters). We chose this strategy to be consistent with previous work, particularly the optimization of GB_OBC.[32] A complete description of the training set is presented in the Methodological Details section. The optimized values of the four parameters are presented in Table 2; these values were used for all of the calculations presented in the Results section.

Table 2. Optimized Parameters

parameter	value
S_{vdw}	0.6211
S_{neck}	0.4058
β_1	18.4377
β_2	313.7171

2.4. Analysis of the Different Geometric Contributions to AR6. In this section, we analyze the relative contribution of the different geometrical approximations used in AR6, namely, the VDW spheres, the "necks", the "chunks", and the additional volume correction. Figure 7 shows the correlation between the effective Born radii computed by AR6 and NSR6, for the most solvent-exposed conformation of Ala10, "pp2", and for the compact conformation, "alpha". Here, AR6 effective radii were computed with one or more of the geometrical contributions to the molecular volume, Figure 1, "switched off". These results shows that it is the combination of the necks' contribution and the approximation of the R6 in the "chunk" regions that contributes most to the good approximation to the numerically exact R6 integration for small molecules such as Ala10.

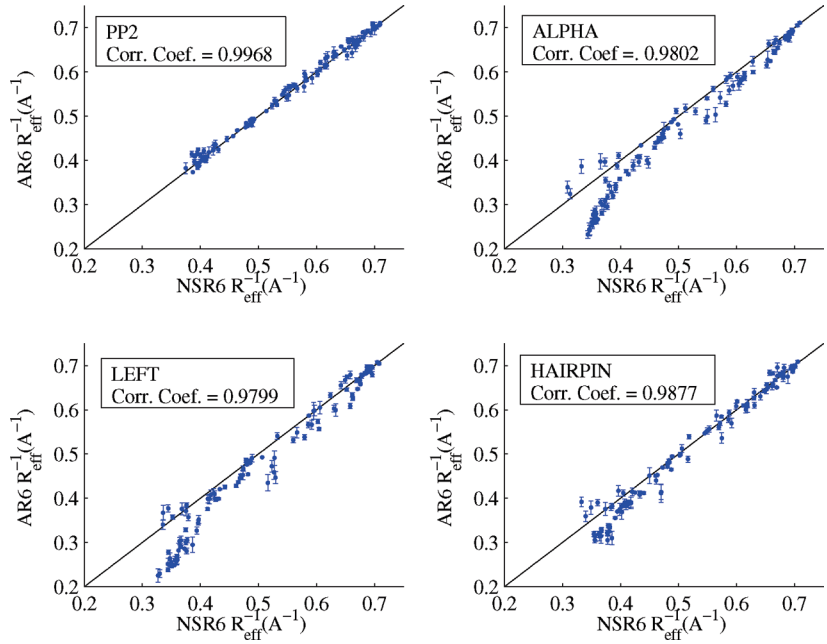


Figure 6. Comparison of the inverse of the approximated R6 effective Born radii (AR6) with the exact R6 effective Born radii (NSR6) for the four conformational states of Ala10. Every point represents the average Born radius over four possible "chunks", with the error bars representing standard deviations.

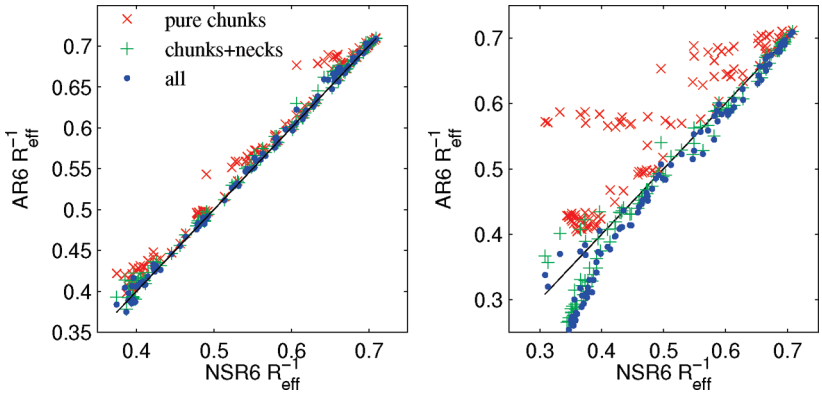


Figure 7. Comparison of the inverse of the approximated R6 effective Born radii (AR6) with the exact R6 effective Born radii (NSR6) for the pp2 (left) and alpha (right) conformational states of Ala10. Red × marks, AR6 with only the chunks contribution ($S_{vdw} = S_{neck} = 0$). Green + marks, AR6 with chunks and neck contribution ($S_{vdw} = 0$, $S_{neck} = 0.4058$). Blue circles, AR6 with all of the contributions ($S_{vdw} = 0.6211$, $S_{neck} = 0.4058$). In all cases, we have used $\beta_1 = \beta_2 = 0$, and a chunk depth of 3.

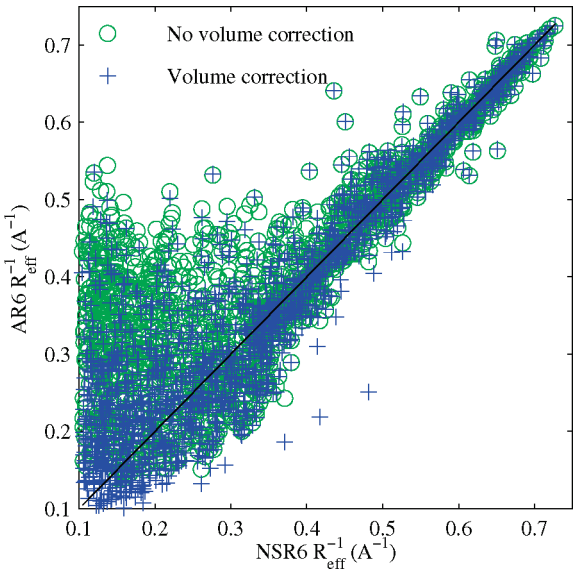


Figure 8. Comparison of the inverse of the approximated R6 effective Born radii (AR6) with the "exact" R6 effective Born radii (NSR6) for thieredoxin (2TRX). Green circles, AR6 with no additional volume correction ($\beta_1 = \beta_2 = 0$). Blue plus marks, AR6 with optimized parameters from Table 2.

Once the "chunks" and "necks" are properly taken care of, the contribution of VDW spheres is almost negligible for small molecules, but it becomes more noticeable in larger structures.

The contribution of the additional volume correction, eq 23, is almost negligible for small structures such as Ala10. However, the contribution of this correction is more evident when the method is applied to a relatively large structure such as thioredoxin. Figure 8 shows that when no volume correction is applied ($\beta_1 = \beta_2 = 0$), the effective Born radii of buried atoms (located in left-most side) are systematically underestimated. When the additional volume correction is activated, the effective Born radii of buried atoms are substantially shifted down toward the correct values of NSR6. Notably, atoms with a small effective Born radius (located in left-most side of Figure 8) are almost unaffected by the rescaling via β_1, $\beta_2 > 0$.

3. Results

Below, we give a brief summary of the accuracy of the AR6 compared with the explicit solvent and the numerical PB model. A detailed description of the results is provided in the following subsections.

One of the problems with current AMBER GB methods was reported recently by Roe et al.[39] They have demonstrated that these methods show a clear bias in the free energies of solvation—hence in the relative populations—of four conformations of a small Ala10 molecule, Figure 5. In Figure 9, we show the error, with respect to explicit solvent, of the $\Delta\Delta G_{el}$ computed by numerical PB, GB_OBC, and AR6, between the four conformational states of Ala10. The $\Delta\Delta G_{el}$ is defined as the difference in ΔG_{el} between two conformational states. Clearly, AR6 is in better agreement with the explicit solvent model than the GB_OBC, having a maximum deviation of 2 kcal/mol. The maximum deviation is 3.9 and 2.3 kcal/mol for GB_OBC and PB, respectively. In fact, on average, AR6 appears to be at least as accurate as the PB in this test. In this summary, we compare AR6 only with GB_OBC, as other GB methods tested by Roe et al. were less accurate.

The accuracy of AR6 is also tested by computing the ΔG_{el} for a set of 22 biomolecular structures and comparing the corresponding numerical PB numbers. The set of structures consists of 19 small proteins, thioredoxin, lysozyme, and a B-DNA molecule, see the Methodological Details section for more details. Table 3 shows the RMSD between ΔG_{el} from the AR6 and the PB model. The RMSD values of the NSR6 and GB_OBC models are also presented in Table 3 for comparison.

Finally, the agreement in the computed $\Delta\Delta G_{el}$ values between numerical PB and AR6 is also verified on the denaturation trajectories of apo-myoglobin and protein-A, see the results in Table 4. In the following subsections, these results are explored in more detail.

3.1. Accuracy of ΔG_{el}: Detailed Analysis. Comparison with explicit solvent models is arguably the most rigorous way to test the performance of any GB model, second only to direct comparisons with experimental results. [However, the latter may not be as clean since GB only computes ΔG_{el},

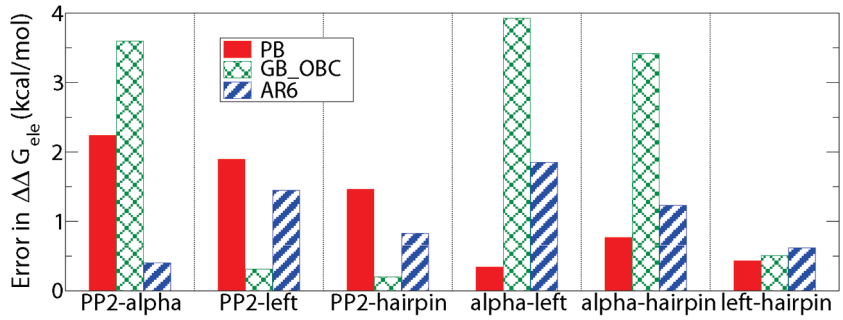


Figure 9. Absolute error in $\Delta\Delta G_{el}$, relative to the explicit solvent model, between four different conformational states of Ala10 (alpha, PP2, left, and hairpin). The energies were obtained using PB (solid red bars), GB_OBC (cross-hatched green bars), and AR6 (striped blue bars). The $\Delta\Delta G_{el}$ for conformational states A and B is defined as $\Delta\Delta G_{el}(A - B) = \Delta G_{el}(A) - \Delta G_{el}(B)$.

Table 3. RMSD of the Solvation Energies (kcal/mol), Relative to the PB Reference of Three GB Flavors[a]

	NSR6	AR6	GB_OBC
RMS	9.98	16.72	50.49

[a] The computation was carried out on a set of 22 structures using optimized parameters from Table 2 and a "chunk" depth of 3.

Table 4. Change in the Electrostatic Part of the Solvation Free Energy, $\Delta G_{el}(N) - \Delta G_{el}(U)$ [kcal/mol], of Apo-Myoglobin and Protein-A on Going from the Unfolded (U) to the Native (N) State Computed with PB and GB Models

	PB	AR6	GB_OBC
(apo)myoglobin, pH = 2	−2087	−2088.2	−2089.9
protein-A, pH = 7	143.37	144.02	145.1

not the total solvation energy, ΔG_{solv}, available from the experiments.] Table 5 shows the results of Roe et al. for TIP3P, PB, GB_HCT, GB_OBC, and GBNeck, plus the results obtained here for the new R6 "flavors" AR6 and NSR6. For the values of ΔG_{el} computed by AR6 and NSR6, each conformational state was represented by 100 MD snapshots.[39] The ΔG_{el} for each conformational state is computed by averaging the values of ΔG_{el} of each of their corresponding MD snapshots. Similar to the optimization process, there are four possible values of ΔG_{el} for each conformational state of Ala10, corresponding to the four possible conformational states used to set up "chunks". The final ΔG_{el} presented in Table 5 for each conformational state is obtained by averaging these four values. An analysis of the sensitivity of ΔG_{el} to the choice of initial structure to set up "chunks" is presented below.

The results in Table 5 show that compared to the other analytical GB flavors tested, the $\Delta\Delta G_{el}$'s obtained with AR6 are in closer agreement to the ΔG_{el} obtained by TIP3P. AR6 also shows a good agreement with the explicit solvent model in the computation of difference in solvation energy ($\Delta\Delta G_{el}$). Table 5 shows that, relative to TIP3P, the values of $\Delta\Delta G_{el}$ between PP2 and alpha are underestimated by −6.64, −3.632, and +2.01 kcal/mol by GB_HCT, GB_OBC, and GBNeck, respectively. Notably, AR6 is almost an exact match; it underestimates the $\Delta\Delta G_{el}$ by only −0.4 kcal/mol relative to the TIP3P. This suggests that AR6 is not biased toward the alpha conformation in contrast to GB_OBC. The AR6 model overestimates TIP3P values by only 1.45 kcal/mol for the $\Delta\Delta G_{el}$ between PP2 and left, and by 0.8 kcal/mol for the $\Delta\Delta G_{el}$ between PP2 and hairpin. Overall, the $\Delta\Delta G_{el}$ obtained by AR6 is in good agreement with the explicit solvent method, with an RMSD of 1.18 kcal/mol. This error is smaller than that in all GB flavors tested by Roe et al.,[39] and essentially the same as the PB result.

Table 5. Free Energies of Solvation between Different Conformations of Ala10 (kcal/mol)[a]

	TIP3P	PB	GB_HCT	GB_OBC	GBNeck	NSR6	AR6
				(A) ΔG_{el}			
alpha	−44.08	−47.97	−51.69	−49.38	−43.26	−45.76	−45.94
PP2	−76.39	−78.05	−77.35	−78.07	−77.59	−77.50	−77.85
left	−51.30	−54.85	−55.05	−52.67	−48.19	−51.12	−51.31
hairpin	−54.16	−57.28	−57.48	−56.03	−52.85	−54.46	−54.79
				(B) $\Delta\Delta G_{el}$			
PP2-alpha	−32.31	−30.07	−25.67	−28.69	−34.33	−31.73	−31.91
PP2-left	−25.09	−23.19	−22.31	−25.40	−29.40	−26.37	−26.54
PP2-hairpin	−22.23	−20.77	−19.87	−22.03	−24.73	−23.04	−23.06
alpha-left	7.22	6.88	3.36	3.29	4.93	5.35	5.37
alpha-hairpin	10.08	9.31	5.80	6.66	9.60	8.69	8.85
left-hairpin	2.86	2.43	2.43	3.37	4.67	3.34	3.48
				(C) $\Delta\Delta G_{el}$ Root Mean Square Deviation			
overall		1.39	3.89	2.60	2.51	1.17	1.18
PP2		1.89	4.37	2.10	3.11	0.94	0.99
non-PP2		0.55	3.34	3.02	1.71	1.37	1.33

[a] The data of TIP3P, GB_HCT, GB_OBC, GBNeck, and PB were taken from Roe et al.[39] Solvation energies were calculated using $\varepsilon_{out} = 80$, $\varepsilon_{in} = 1$, and $\kappa = 0$.

Table 6. RMSD ($Å^{-1}$) between the Inverse of Effective Born Radii Computed by the GB_OBC and AR6, Relative to the Perfect Born Radii

	GB_OBC	AR6
small proteins	0.061	0.046
thioredoxin	0.128	0.077
lysozyme	0.114	0.064
B-DNA	0.051	0.054

When using the original Still's equation instead of eq 2 used throughout this work, the overall RMSDs of ΔG_{el} and $\Delta\Delta G_{el}$ between AR6 and TIP3P results are 1.59 and 1.21 kcal/mol, respectively, which are almost the same as the values present in Table 5. Thus, the improvement showed in Table 5 is mostly due to the use of AR6 for effective radii computation rather than the use of eq 2 instead of the original Still's equation.

3.2. Accuracy of the Effective Born Radii. The "perfect" (obtained via numerical PB calculations) effective Born radii are often used as benchmarks for the accuracy of different GB flavors, as such comparisons can help identify sources of error in the computation of the approximate effective radii.[42,54,59] In Table 6, we show the RMSD of the inverse of the effective radii obtained by AR6 and GB_OBC, relative to the "perfect" effective Born radii. We have chosen to analyze inverse effective Born radii because they directly represent the contribution of effective Born radii to the energy in eq 2. These results show a significant improvement in the accuracy of inverse effective radii computed by the AR6 compared to those computed by GB_OBC in all of the cases except for B-DNA, in which AR6 shows a deviation from the PB reference that is slightly greater than the one produced by GB_OBC. However, the ΔG_{el} of B-DNA produced by GB_OBC is in fact less accurate than the corresponding AR6 number, see the next subsection for details. A more detailed comparison between the two sets of effective radii is presented in Figure 10, which compares the inverse of the effective Born radii computed by AR6 and the inverse of the "perfect" effective Born radii. We see that AR6 shows improvement over GB_OBC in the entire range of the effective radii. Particularly, AR6 agrees well with the perfect radii in the region of small effective radii. It is worth noting that it is this region that contributes most to the energy in eq 2. AR6 is also, on average, more accurate than GB_OBC in regions of large effective Born radii that correspond to atoms deeply buried inside the protein.

3.3. Accuracy of ΔG_{el} Relative to the PB. Here, the electrostatic part of the solvation energy is calculated by the PB, AR6, GB_OBC, and NSR6 methodologies, on a data set composed of 19 small proteins, thioredoxin, lysozyme, and a B-DNA molecule, see the Methodological Details section for details. These structures were used earlier for parametrization of GB models.[42] The results of this comparison are shown in Table 7. AR6 has an overall RMSD of 16.7 kcal/mol relative to the PB reference compared to 50.5 kcal/mol for the GB_OBC model. The percent errors of the GB models shown in Table 7 were calculated as the arithmetic mean of $100(\Delta G_{el}(GB) - \Delta G_{el}(PB))/|\Delta G_{el}(PB)|$ over all 22 molecular structures. Interestingly, the results show that, on average, both GB_OBC and AR6 models produce a relative error close to zero. Thus, just like GB_OBC, AR6 does not appear to have a systematic bias relative to the PB.

3.4. Sensitivity to the Choice of "Chunks". If two or more conformational states are available for a given molecule, then it is possible to use any of those conformational sates to compute the "chunks" contribution, λ_i, which can result in different values of ΔG_{el}. Here, we test the sensitivity of AR6 to the choice of the structure used to set up "chunks". The values of ΔG_{el} for AR6 in Table 5 (upper block) for each conformational state of Ala10 were obtained by averaging the four ΔG_{el} values corresponding to each of the four conformational states of Ala10 used to set up "chunks". The corresponding standard deviations, considering the four possibilities of "chunks", are 0.44, 0.62, 0.63, and 0.59 kcal/mol for alpha, PP2, left, and hairpin, respectively. Thus, for small molecules such as Ala10, the variation in ΔG_{el} due to the choice of structure for setting up "chunks" is very small relative to the absolute values of ΔG_{el}. To further analyze this sensitivity in larger molecules, we compare the $\Delta\Delta G_{el}$ between the PB and AR6 for the denaturation trajectories of apo-myoglobin and protein A. The results are summarized

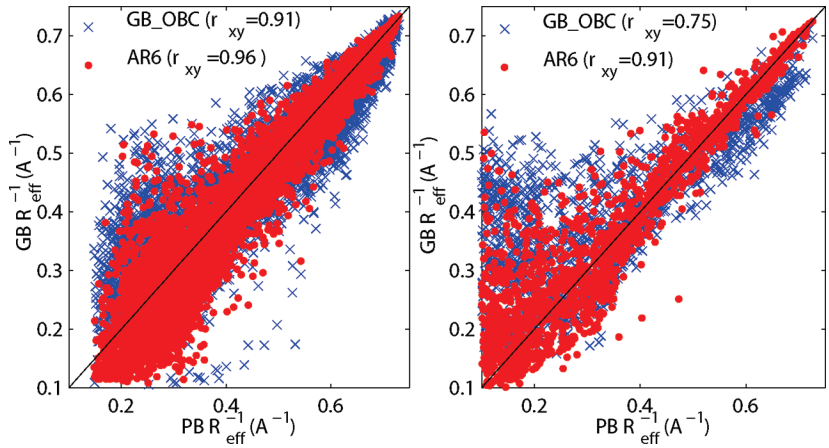


Figure 10. Comparison of the inverse of the approximated effective Born radii (GB R_{eff}^{-1}) with the "perfect" effective Born radii (PB R_{eff}^{-1}) for 19 small proteins (left) and thioredoxin (right). Approximated effective radii were computed by AR6 (red) and GB_OBC (blue). Correlation coefficients r_{xy} are indicated in parentheses.

Table 7. Electrostatic Solvation Energies (kcal/mol) for a Set of 22 Structures[a]

PDB	PB	NSR6	AR6	GB_OBC
1az6	−364.73	−353.36	−358.65	−369.87
1byy	−619.13	−618.88	−625.78	−597.41
1eds	−499.77	−488.10	−489.4	−492.05
1g26	−551.49	−539.00	−549.08	−532.18
1qfd	−539.09	−527.90	−541.72	−526.8
1bh4	−473.11	−463.30	−460.28	−437.49
1cmr	−744.44	−739.29	−789.11	−762.21
1fct	−853.06	−854.41	−860.69	−836.43
1ha9	−669.2	−668.81	−669.79	−646.26
1qk7	−606.12	−600.87	−620.21	−607.56
1bku	−660.81	−657.31	−669.51	−674.11
1dfs	−757.76	−756.22	−802.15	−797.66
1fmh	−1482.9	−1493.00	−1501.5	−1481.5
1hzn	−577.02	−569.69	−584.38	−598.37
1scy	−626.19	−612.96	−609.52	−625.12
1brv	−437.28	−435.38	−443.58	−466.15
1dmc	−894.03	−890.10	−901.63	−848.77
1fwo	−788.95	−774.14	−790.44	−774.33
1paa	−1401.2	−1411.30	−1401.4	−1397.4
2trx	−1602.4	−1595.90	−1603.2	−1608.9
2lzt	−2121	−2100.80	−2099.3	−2100.5
bdna	−4774.7	−4790.10	−4790	−4558.3
percent error		−0.90%	0.58%	−0.67%
unsigned percent error		1.07%	1.55%	2.67%
RMSD		9.69	16.72	50.49

[a] The solvation energies were calculated using $\varepsilon_{out} = 1000$, $\varepsilon_{in} = 1$, and $\kappa = 0$. In all cases, we used the optimized parameters on Table 2 and depth = 3 for AR6. The structures in bold were used in the optimization process as a training set. The errors are computed relative to the numerical PB reference.

Table 8. Change in the Electrostatic Part of Solvation Free Energy, $\Delta\Delta G = \Delta G_{el}(N) - \Delta G_{el}(U)$ [kcal/mol], of Apo-Myoglobin and Protein-A on Going from the Unfolded (U) to the Native (N) State Computed with the PB and AR6 Models[a]

		AR6	
	PB	chunk N	chunk U
(apo)myoglobin, pH = 2	−2087	−2088.2	−2083.8
protein-A, pH = 7	143.37	144.02	144.27

[a] The computations were carried out using "chunks" from one snapshot of the native state (chunk N) and from one snapshot of the unfolded state (chunk U).

in Table 8. Having two protein conformations (in the native and unfolded states), it is possible to compute "chunk" contributions from two completely different sources, one from a snapshot of the native state (chunk "N" in Table 8), and the other from a snapshot of the unfolded state (chunk "U" in Table 8). Ideally, $\Delta\Delta G_{el}$ computed using such different "chunks" should be identical. According to our procedure, the "chunks" contribution (and thus ΔG_{el}) depends slightly on the initial configuration. The results in Table 8 show that the change in $\Delta\Delta G_{el}$ is relatively small between the use of different "chunks": 4.5 kcal/mol for apo-myoglobin and 0.25 kcal/mol for protein-A. Moreover, these results show that AR6 is in good agreement with PB in the computation of $\Delta\Delta G_{el}$.

In order to further extend the analysis of the sensitivity of energy to the use of different "chunks", we show in Figure 11 the solvation energy along the unfolding trajectory of protein-A produced by AR6 using different "chunk" sets. These results show that the variation of energy due to the use of two different "chunks" is smaller (with a standard deviation of 2 kcal/mol) than the error between the PB method and the AR6 method when chunks are calculated numerically for each snapshot of the protein-A unfolding trajectory (standard deviation 6 kcal/mol). Thus, although chunks of depth 3 may undergo conformational changes during dynamics, the variation in energy produced by these changes are "safely" smaller than the overall error produced by the GB model using AR6, relative to the reference PB model.

3.5. Further Optimization: The Tabulated Chunks. The most expensive stage in the AR6 method is the computation of the chunk contributions, λ_i, as it requires a surface triangulation over N chunk molecules, N being the number of atoms. This one-time expense is not critical if AR6 is used in MD simulations or to compute the ΔG_{el} of one structure at different conformational states, because the values of λ_i are computed only once at the initial stage and then reused for all subsequent calculations. However, if the goal is to quickly compute ΔG_{el} once, for a set of different structures, the computation may become expensive, especially for large sets of structures, as this requires computing λ_i for every atom of the set of structures. Moreover, the values of λ_i depend (though slightly) on the choice of the conformational state used to set up the "chunks". This ambiguity has the potential drawback of generating path-dependent energy values during MD simulations. While harmless for an ergodic trajectory, it may present a certain inconvenience under some circumstances.

One way to speed up the setup stage of the AR6 method, and at the same time eliminate the ambiguity in the selection of conformational states to set up the "chunks", is to tabulate an optimum or an average value of λ_i, eq 17, for every atom type within a specific amino acid or nucleotide and save it in a lookup table for all future computations. Within this protocol, the setup stage will consist only of reading "chunk" contributions from a lookup table, which is inexpensive. We test this strategy in Table 9, where we present the values of ΔG_{el} and $\Delta\Delta G_{el}$ for the four conformations of Ala10, obtained by the AR6 method in which the same values of λ_i were used for every distinct atom type in alanine residue. The set of pretabulated $\{\lambda_i\}$ is obtained by averaging the λ_i of the central residues of the four conformational states (chunk depth = 3). The results show an insignificant deviation from the original results shown in Table 5: they are still in better agreement with TIP3P than the GB methods tested by Roe et al. Thus, the use of tabulated λ_i is a promising way to speed up the setup process, introducing little deviation from the original procedure in which λ_i is numerically computed for every atom of the molecule, at the setup stage.

3.6. Molecular or VDW Surface As Dielectric Boundary? Traditionally, numerical PB calculations have used the Lee−Richards molecular surface to define the solute/solvent dielectric boundary. This definition is supported by various studies that compared the PB ΔG_{el} with those from the explicit solvent.[46,47] On the other hand, the use of the van der Waals surface in this context has also been

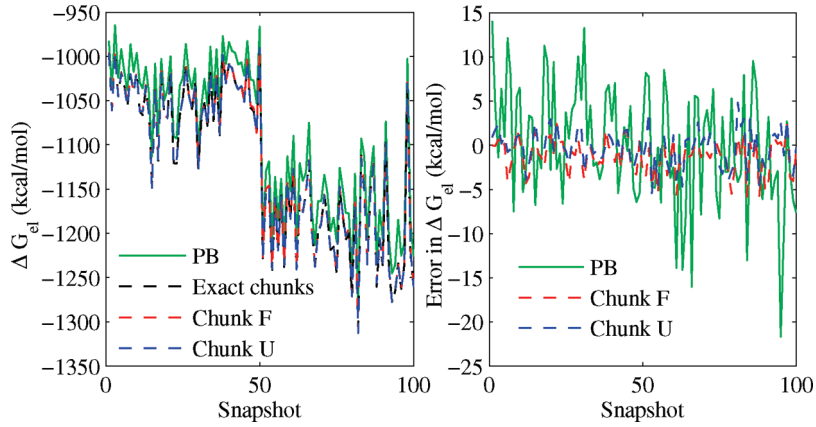


Figure 11. Left: solvation energy along the MD unfolding trajectory of protein-A (PDB ID: 1BDD) obtained by PB (solid lines) and AR6 (dashed lines). The AR6-based energies were obtained using "chunks" from one snapshot of the folded (chunk F, red) and unfolded (chunk U, blue) states, and from "chunks" computed numerically for each snapshot of the MD trajectory ("exact chunks"). Right: Difference in energy after elimination of the systematic constant deviation between GB and PB. Green, difference between PB and AR6 "exact chunks" computed for each snapshot. Dashed blue and red lines, difference between AR6 using different chunks (chunk F or chunk U) and AR6 using "exact chunks".

Table 9. Free Energies of Solvation between Different Conformations of Ala10 (kcal/mol)[a]

			AR6	
	TIP3P	PB	original chunks	pretabulated chunks
		(A) ΔG_{el}		
alpha	−44.08	−47.97	−45.94	−46.21
PP2	−76.39	−78.05	−77.85	−78.11
left	−51.30	−54.85	−51.31	−51.55
hairpin	−54.16	−57.28	−54.79	−54.96
		(B) $\Delta\Delta G_{el}$		
PP2-alpha	−32.31	−30.07	−31.91	−31.9
PP2-left	−25.09	−23.19	−26.54	−26.56
PP2-hairpin	−22.23	−20.77	−23.06	−23.15
alpha-left	7.22	6.88	5.37	5.34
alpha-hairpin	10.08	9.31	8.85	8.75
left-hairpin	2.86	2.43	3.48	3.41
		(C) $\Delta\Delta G_{el}$ root mean square deviation		
overall		1.39	1.18	1.21
PP2		1.89	0.99	1.03
non-PP2		0.55	1.33	1.37

[a] The data of TIP3P and PB were taken from Roe et al.[39] Solvation energies were calculated using $\varepsilon_{out} = 80$, $\varepsilon_{in} = 1$, and $\kappa = 0$.

advocated,[60,61] including some recent implementations of the R6 flavor.[44,45] While the precise nature of the physically realistic dielectric boundary is still an open and complex issue[62] clearly outside of the scope of this work, it is still appropriate to ask a very focused question here: between the VDW and molecular surface based definitions of the dielectric boundary, which one leads to a better agreement with the explicit solvent ΔG_{el} for the set of representative conformation states of alanine decapeptide?

The unambiguous answer is presented in Figure 12, which shows the error in the electrostatic part of the solvation free energy computed by the numerical PB relative to the corresponding TIP3P values as a function of the probe radius used to determine the molecular boundary. Geometrically, as the probe radius decreases, the molecular volume used in the PB computation approaches the VWD volume. The results in Figure 12 show that the error always increases as

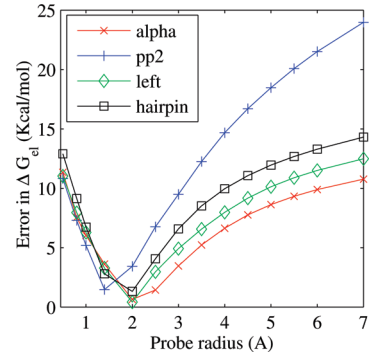


Figure 12. Absolute error of the numerical PB ΔG_{el}, relative to the explicit solvent (TIP3P) reference, as a function of the probe radius used to set the dielectric boundary in the PB calculations. The computations are performed for the four conformational states of alanine decapeptide shown in Figure 5.

the probe radius goes to zero and the dielectric boundary becomes the VDW surface. This means that at least for the set of representative shapes of a small peptide, Figure 5, the use of the Lee−Richards molecular surface for the dielectric boundary in PB calculations results in consistently better agreement with TIP3P solvent model than do the VDW-based definitions. Since the GB model is essentially an approximation of the PB model, these results suggest that in order to obtain more accurate electrostatic solvation free energies relative to the explicit solvent, the dielectric boundary used in the computation of the effective Born radii should strive to approximate the Lee−Richards molecular surface, not the VDW surface.

3.7. Total solvation Free Energies of Small Molecules. Here, we compute the total solvation energy, $\Delta G_{solv} = \Delta G_{el} + \Delta G_{nonpol}$, for a "challenge" set of small molecules and compare the results with the experimentally available ΔG_{solv}. The structures of the test molecules were taken from a recent study by Nicholls et al.[63] in which a set of 17 "challenging" drug-like small molecules were proposed for

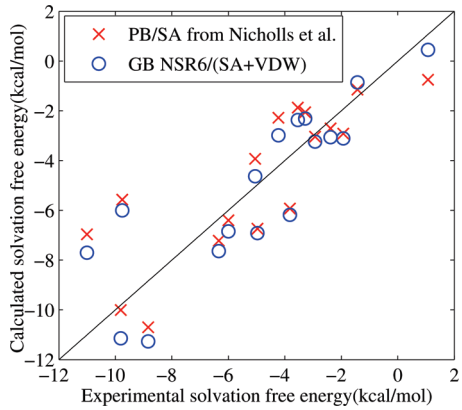


Figure 13. Comparison of computed solvation energies with the experimentally determined solvation energies for a "challenging" data set of small drug-like molecules. Blue: PB plus the cavity term (PB/SA) approach from Nicholls et al.[63] Red: GB NSR6 plus cavity and van der Waals terms (GB NSR6/(SA+VDW)).

the purpose of testing different computational methods aimed at prediction of solvation free energies.

Here, the electrostatic part of the solvation energy is computed by eq 2 and the NSR6 method, eq 6. [For single-point calculations, AR6 has no computational advantage over NSR6 unless pretabulated chunks are available.] The nonpolar component of ΔG_{solv} is divided into the "cavity" and the "solute−solvent van der Waals interactions" terms:

$$\Delta G_{nonpol} = \Delta G_{cav} + \Delta G_{vdW} \tag{26}$$

We compute ΔG_{cav} by applying an atom-independent surface tension γ to the solvent-accessible surface area (SASA): $\Delta G_{cav} = \gamma(\text{SASA})$, where γ is set to 0.065 kcal/(mol Å^2), which is the value used in ref 63. We compute ΔG_{vdW} by the following expression proposed by Gallicchio and Levy:[33]

$$\Delta G_{vdW} = \mu \sum_i \frac{a_i}{(R_i + \rho_w)^3} \tag{27}$$

where μ is a dimensionless adjustable parameter, ρ_w is the water probe radius, R_i is the effective Born radius of atom i, and a_i depends on the number density of water and the Lenard-Jones parameters for each atom i. The methodology to compute a_i proposed by Gallicchio and Levy[33] is described in the Methodological Details section. In the original formulation of Gallicchio and Levy, μ is an atom-dependent parameter. For simplicity, here we use a constant atom-independent parameter instead. Thus, μ is the only adjustable parameter used for the computation of ΔG_{solv}. We found that its optimum value for the "challenge" set is $\mu = 1.838$. A comparison plot between the computed and experimentally determined ΔG_{solv} is presented in Figure 13. The results are at the same level of accuracy as those reported in ref 63. The RMSD values between the experimentally and computationally determined ΔG_{solv} are 1.73 kcal/mol for our method and 1.88 kcal/mol for that of Nicholls et al., which is based on the PB model.

These results are encouraging considering the relative computational efficiency of the approach: after the computation of ΔG_{el}, the potentially equally expensive ΔG_{vdW} in eq 27 is obtained at virtually no additional cost because the R_i values have already been computed. In contrast, PB-based methods would require an independent computation of R_i in eq 27, in addition to numerically solving the PB equation.

4. Conclusion

In this work, we have developed a new analytical method, AR6, to compute the effective Born radii. We were motivated by a recently reported deficiency of a set of currently available GB models that were shown to produce a clear energy bias among representative conformations of a small deca-alanine peptide. Our proposed model is based exclusively on the $|\mathbf{r}|^{-6}$ (R6) integration, which was shown earlier to produce a good approximation to the PB model when applied to protein structures. The R6 approach advocated here is simple—based on a single integral—and has a solid theoretical basis. Since it was already shown that the R6 effective radii can, in principle, deliver electrostatic solvation energies as accurate as those based on the "perfect" PB-based radii, we chose the R6 flavor as the best candidate to improve the accuracy performance of the GB. Our goal was to lay a foundation for an efficient, robust analytical R6 routine that can in the future be used in MD simulations. However, we found that in the R6 case, high accuracy integration over the physically realistic molecular volume is much more difficult than in the case of the still widely used, but less accurate CFA approximation where the singularity of the integrand, $|\mathbf{r}|^{-4}$, is lower: 4 instead of 6. Essentially, the R6 approach is much less forgiving to small integration inaccuracies in the vicinity of the atom in question. To achieve the required accuracy, we perform the integration over an approximation to molecular volume that adds several computationally efficient corrections to the pairwise VDW-based integration to closely approximate the true molecular volume in the vicinity of each atom. One of the key elements of the proposed approximation is the use of predefined groups of atoms, "chunks", over which the integration is performed numerically exactly, at the setup stage. The "chunk" contributions to the total integral are then reused. A "chunk" is a small set of atoms around the atom in question. The set is chosen using the known covalent connectivity of the atom to its neighbors in such a way that the geometry of the chunk is not expected to change substantially during dynamics.

Several additional approximations developed earlier by this group were also used, including those employed in the popular GB_OBC model in AMBER. Apart from the computation setup costs, the resulting analytical R6, or "AR6", model is at least as efficient as GB_OBC. The proposed model uses a number of simplifications relative to many other GB flavors; for example, it has only a single adjustable parameter to account for volume overcounting due to atoms overlapping, as opposed to one for each atom type. In all, AR6 has four fitting parameters separated into two groups of two parameters that can be fitted independently. The latter property has allowed a nearly exhaustive search in the parameter space and lowered chances for overfitting.

We have performed a fairly extensive set of accuracy tests for AR6. These included comparing electrostatic solvation free energies (ΔG_{el}) against the numerical PB and explicit solvent simulations where available. In particular, we tested the accuracy of AR6 on four conformational states of alanine decapeptide that were used previously to reveal the energetic bias of several GB models, in particular, AMBER's GB_OBC. We have found that, relative to the explicit solvent, the RMS error of changes in ΔG_{el} between various pairs of conformational states computed via AR6 equals that of the numerical PB treatment, and it is 2 times lower than that of GB_OBC. Tests against the PB treatment on 22 biomolecular structures including proteins and DNA have shown that the RMS error in ΔG_{el} is 3 times lower than the corresponding value for GB_OBC. When used to compute the difference in ΔG_{el} over unfolding trajectories of apo-myoglobin and protein-A, AR6 shows similar accuracy to GB_OBC, which was originally parametrized using apo-myoglobin folding/unfolding snapshots. Sensitivities of ΔG_{el} to several key approximations have been tested as well. We have also explored a variant of the approach to eliminate the setup costs via the use of pretabulated chunks. The accuracy of this variant, which carries no setup costs, is virtually the same as that of the original. While a difference in the setup efficiency is probably not critical in MD simulations, where the setup time is only a tiny fraction of the whole simulation time, the pretabulated approach may be found easier to implement. To summarize, the analytical AR6 flavor to compute the effective Born radii offers a clear improvement in accuracy over a set of popular pairwise methods based on the CFA, without apparent sacrifices in computational complexity. This makes the approach a promising candidate for applications that require repetitive computations of ΔG_{el} such as molecular dynamics. While it was developed with MD in mind, and robustness, stability, and differentiability were strictly enforced, extensive further testing directly in MD is needed, and is planned to be done in the future.

Two other points not directly related to the analytical R6 model, but relevant to continuum electrostatics and GB models, were also investigated. We have tested a version of the R6 flavor, NSR6, which is based on a direct surface integration over a numerically triangulated molecular surface. While NSR6 is mathematically equivalent to the molecular volume integration approach, which was explored earlier, the surface-based routine is much faster. To assess its potential in a practical setting, we used it on a recently published "challenge" set of small drug-like molecules. In this endeavor, the total solvation free energy was computed as the sum of the polar part from NSR6 and the nonpolar part estimated via the cavity and VDW terms as proposed earlier by Gallicchio and Levy.[33] With only one fitting parameter, we were capable of predicting the total solvation free energy to within 1.73 kcal/mol RMS error relative to the experiment, which is at least as accurate as the recently reported PB-based estimates. Note that within the R6 formulation, computation of the nonpolar contribution is particularly efficient because its VDW part depends on the same $|\mathbf{r}|^{-6}$ integrals. We stress, however, that this little excursion into the realm of small molecule free energy estimates serves only one purpose: to demonstrate promise of the R6 approach for this field. In our view, the results warrant further investigation of this promise by interested parties.

We have also touched upon a still debated issue of which surface definition better approximates the molecular boundary in the context of continuum solvent electrostatics: the Lee−Richards (molecular surface) or the van der Waals surface? For the four conformational states of alanine decapeptide used in this and previous works, the answer we have found is unambiguous (and not unexpected): the molecular surface yields ΔG_{el} in much closer agreement with the explicit solvent results.

All of the software developed during this work is available from http://people.cs.vt.edu/~onufriev/software.php.

5. Methodological Details

The structures of the four conformational states of Ala10 were kindly provided by Daniel Roe. A detailed description of the Ala10 structures and the methods used to compute ΔG_{el} for these structures can be found in Roe et al.[39] The remainder of this paragraph is a brief summary of these procedures. The trajectories of the four conformations of Ala10 were obtained from REMD simulations using TIP3P as a solvent model. The values of ΔG_{el} were then calculated by thermodynamic integration using the trajectories of the REMD simulation. The PB reference energies of the Ala10 snapshots were calculated with DELPHI, version 2.0,[64] with a grid spacing of 0.25 Å. The GB results (except for NSR6 and AR6) were obtained with the AMBER package with igb = 1 for GB_HCT, igb = 5 for GB_OBC, and igb = 7 for GBNeck. In both models, GB and PB, $\varepsilon_{out} = 78.5$, $\varepsilon_{in} = 1$, and the ionic strength was set to zero.

The data set of structures used for optimization and testing of AR6 was randomly selected from a larger data set of representative proteins structures from Feig et al.,[65] the selection criterion being that the compounds are small enough to allow for high-resolution grid computations. Their PDB IDs are presented in Table 7, in which the PDB IDs in bold were used as the training set. Chain "A" or "model 1" has been chosen when appropriate. The assignment of partial charges, protonation states, etc. are described in ref 65. In addition, a canonical B-DNA 10 base pair structure from ref 26 has been used. The Bondi radii set was used for all molecules of this data set. The random selection has resulted in a fairly representative sampling of various structural classes and charge state. The total charge of the structures varies from −18(B-DNA) to +9 (lysozyme) with most of the structures (17) falling in the range from −4 to +4. The structural composition of the proteins is as follows: seven mostly α helical, four mostly β sheet, five roughly equal mix of α/β, and five mostly disordered. The size of most of these proteins is about 30 amino acids, although two of them are larger: 2trx (thioredoxin) and 2lzt (lysozyme) have 108 and 129 residues, respectively.

The "perfect" effective Born radii were calculated using numerical PB treatment as implemented in APBS 0.4.0.[66] A separate calculation was performed for each atom of each molecule. In each calculation, the partial charge of the atom

of interest was set to 1, while partial charges of all other atoms were set to zero. A 129-point cubic grid centered on the atom of interest was used to discretize the problem. Multiple Debye−Huckel boundary conditions were used for the initial grid, which was sufficiently large so that no portion of the molecule was closer than 4 Å to the edge of the grid. Each focusing step halved the grid spacing, while maintaining the same number of grid points. Focusing step boundary conditions were derived from the potential calculated on the immediately preceding grid. Focusing continued until the grid spacing reached 0.1 Å. Except where otherwise indicated, all calculations used a nonsmoothed molecular surface definition with a probe radius of 1.4 Å and a surface probe point density of 50. A four-level finite-difference multigrid solver was employed in conjunction with the linearized Poisson−Boltzmann equation (which reduces to the Poisson equation since ion concentrations were zero). Charge was discretized using cubic B-splines. All solvated calculations used a dielectric constant of $\varepsilon_{out} = 1000$ to mimic the conductor limit $\varepsilon_{out} \rightarrow \infty$ and, therefore, avoid masking the geometry-specific deficiencies of the standard GB model by its inaccuracies arising from finite ε_{out}.[36] The dielectric constant of the solute region was set to 1; a parallel set of reference calculations was performed with a spatially uniform dielectric constant of 1 to determine the gas-phase charge discretization reference energy. The self-energy of each atom was calculated by subtracting the reference energy from the solvated energy from the most focused grid. Radii were calculated from self-energies using the Born equation. MEAD 2.2.5 with double precision and otherwise default parameter settings is used as the reference PB solver in Table 7. The dielectrics are as described above. Six focusing steps are used with the coarsest cubic grid having 81 points in each direction and 3.2 Å grid spacing, and the finest grid of 315 points in each direction and 0.1666 Å spacing.[67]

The set of apo-myoglobin structures was prepared from the holo-Mb coordinate set [Protein Data Bank (PDB) ID: 2mb5] by heme removal and simulated acid unfolding in explicit solvent, as described elsewhere.[68] The native state is represented by 50 consecutive snapshots (2 ps apart from each other) with near-native radius of gyration, ~16 Å taken from the beginning of the acid-unfolding simulation. The unfolded state is represented by 50 consecutive snapshots from the end of that simulation, at which point the radius of gyration has approached ~30 Å—as is experimentally observed in the unfolded state.[69] Protein-A structures were prepared from the NMR average coordinate set (PDB ID: 1BDD, residues 10−55). The native-state ensemble is represented by 50 consecutive snapshots (2 ps apart from each other) from the implicit solvent simulation protocol described below, and deviations from the native coordinates are less than 2 Å for Cα atoms. The unfolded state was prepared by heating the protein to 450 K for 1 ns in an implicit solvent environment (Onufriev, unpublished data), and 50 consecutive snapshots with average RMSD from the native structure of no less than 15 Å were chosen to represent this state. The PB solvation energies of the denaturation process of apo-myoglobin and protein A were computed using DELPHI-II[64] with a cubic box and a grid spacing of 0.5 Å. The dielectric constant for the protein interior is 1, and the ionic strength is zero.

Table 10. Lennard-Jones Parameters Used for the Computation of ΔG_{vdW}

	σ_i (Å)	ε_i (kcal/mol)
H	1.4870	0.0157
C	1.9080	0.1094
N	1.8240	0.1700
O	1.6612	0.2100
S	2.0000	0.2500
Br	2.2200	0.3200
Cl	1.9480	0.2650
F	1.7500	0.0610

The surface triangulation used in the NSR6 procedure and the computation of "chunks" contribution were carried out using the MSMS package[53] using a probe radius of 1.4 and triangle density of 10.

The structures of the 17 "challenging" small molecules were taken from the supporting material of ref 63. The R6 radii of this molecules were obtained with the NSR6 procedure, and the values of SASA for each structure were computed by using the MSMS package. In both cases, a triangle density of 15 and probe radius of 1.4 have been used. The values of ΔG_{el} are calculated by eq 2 with $\varepsilon_{out} = 80$, $\varepsilon_{in} = 1$, and the ionic strength set to zero. For the computation of ΔG_{vdW}, the values of a_i in eq 27 are computed by the following expression:[33]

$$a_i = -\frac{16}{3}\pi d_w \varepsilon_{iw} \sigma_{iw}^6 \tag{28}$$

where $d_w = 0.033428$ Å^{-3} is the number density of water at standard conditions; ε_{iw} and σ_{iw} are computed by

$$\sigma_{iw} = \sqrt{\sigma_i \sigma_w} \tag{29}$$

$$\varepsilon_{iw} = \sqrt{\varepsilon_i \varepsilon_w} \tag{30}$$

where $\sigma_w = 1.7683$ Å and $\varepsilon_w = 0.1520$ kcal/mol are the Lennard-Jones parameters of the TIP3P water oxygen. σ_i and ε_i are the Lennard-Jones parameters for atom i. The values of σ_i and ε_i for each atom type were taken from AMBER 8 and are presented in Table 10.

Acknowledgment. We thank Daniel Roe for providing the Ala10 data and for helpful discussions. We thank Igor Tolokh for useful comments on the manuscript. Support from NIH R01-GM076121 is acknowledged.

Supporting Information Available: Two additional tables containing the values of A_{ij} and B_{ij} used in eq 13, for a range of ρ_i and ρ_j. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161–2200.

(2) Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144–1149.

(3) Beroza, P.; Case, D. A. *Methods Enzymol.* **1998**, *295*, 170–189.

(4) Madura, J. D.; Davis, M. E.; Gilson, M. K.; Wade, R. C.; Luty, B. A.; McCammon, J. A. *Rev. Comp. Chem.* **1994**, *5*, 229–267.

(5) Gilson, M. K. *Curr. Opin. Struct. Biol.* **1995**, *5*, 216–223.

(6) Scarsi, M.; Apostolakis, J.; Caflisch, A. *J. Phys. Chem. A* **1997**, *101*, 8098–8106.

(7) Luo, R.; David, L.; Gilson, M. K. *J. Comput. Chem.* **2002**, *23*, 1244–1253.

(8) Gilson, M. K.; Davis, M. E.; Luty, B. A.; McCammon, J. A. *J. Phys. Chem.* **1993**, *97*, 3591–3600.

(9) Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J. M.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; McCammon, J. A. *Comput. Phys. Commun.* **1995**, *91*, 57–95.

(10) Baker, N. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137–143.

(11) Feig, M.; Brooks, C. L. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.

(12) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(13) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *246*, 122–129.

(14) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824–19839.

(15) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578–1599.

(16) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

(17) Edinger, S. R.; Cortis, C.; Shenkin, P. S.; Friesner, R. A. *J. Phys. Chem. B* **1997**, *101*, 1190–1197.

(18) Jayaram, B.; Liu, Y.; Beveridge, D. L. *J. Chem. Phys.* **1998**, *109*, 1465–1471.

(19) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983–10990.

(20) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.

(21) Lee, M. S.; Salsbury, F. R.; Brooks, C. L. *J. Chem. Phys.* **2002**, *116*, 10606–10614.

(22) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins* **2004**, *56*, 310–321.

(23) Romanov, A. N.; Jabin, S. N.; Martynov, Y. B.; Sulimov, A. V.; Grigoriev, F. V.; Sulimov, V. B. *J. Phys. Chem. A* **2004**, *108*, 9323–9327.

(24) Dominy, B. N.; Brooks, C. L. *J. Phys. Chem. B* **1999**, *103*, 3765–3773.

(25) David, L.; Luo, R.; Gilson, M. K. *J. Comput. Chem.* **2000**, *21*, 295–309.

(26) Tsui, V.; Case, D. A. *J. Am. Chem. Soc.* **2000**, *122*, 2489–2498.

(27) Calimet, N.; Schaefer, M.; Simonson, T. *Proteins* **2001**, *45*, 144–158.

(28) Spassov, V. Z.; Yan, L.; Szalma, S. *J. Phys. Chem. B* **2002**, *106*, 8726–8738.

(29) Simmerling, C.; Strockbine, B.; Roitberg, A. E. *J. Am. Chem. Soc.* **2002**, *124*, 11258–11259.

(30) Wang, T.; Wade, R. C. *Proteins* **2003**, *50*, 158–169.

(31) Nymeyer, H.; Garcia, A. E. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13934–13939.

(32) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383–394.

(33) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479–499.

(34) Lee, M. C.; Duan, Y. *Proteins* **2004**, *55*, 620–634.

(35) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297–1304.

(36) Sigalov, G.; Scheffel, P.; Onufriev, A. *J. Chem. Phys.* **2005**, *122*, 094511.

(37) Sigalov, G.; Fenley, A.; Onufriev, A. *J. Chem. Phys.* **2006**, *124*, 124902.

(38) Case, D. A.; Darden, T.; Cheatham, T. E., III; Simmerling, C.; Wang, J.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Duke, R. E.; Crowley, M.; Brozell, S.; Luo, R.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Caldwell, J. W.; Ross, W. S.; Kollman, W. S. *AMBER 9*; University of California: San Francisco, 2006.

(39) Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C. *J. Phys. Chem. B* **2007**, *111*, 1846–1857.

(40) Svrcek-Seiler, A. Personal communication, 2001.

(41) Grycuk, T. *J. Chem. Phys.* **2003**, *119*, 4817–4826.

(42) Mongan, J.; Svrcek-Seiler, A.; Onufriev, A. *J. Chem. Phys.* **2007**, *127*, 185101–185101.

(43) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379.

(44) Tjong, H.; Zhou, H. X. *J. Phys. Chem. B* **2007**, *111*, 3055–3061.

(45) Labute, P. *J. Comput. Chem.* **2008**, *29*, 1693–1698.

(46) Lee, M. S.; Olson, M. A. *J. Phys. Chem. B* **2005**, *109*, 5223–5236.

(47) Swanson, J. M. J.; Mongan, J.; McCammon, J. A. *J. Phys. Chem. B* **2005**, *109*, 14769–14772.

(48) Lee, M. S.; Feig, M.; Salsbury, F. R.; Brooks, C. L. *J. Comput. Chem.* **2003**, *24*, 1348–1356.

(49) Im, W.; Lee, M. S.; Brooks, C. L. *J. Comput. Chem.* **2003**, *24*, 1691–702.

(50) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

(51) Onufriev, A. In *Continuum Electrostatics Solvent Modeling with the Generalized Born Model*, 1st ed.; Feig, M., Ed.; Wiley: New York, 2010; pp 127−165.

(52) Chocholousová, J.; Feig, M. *J. Comput. Chem.* **2006**, *27*, 719–729.

(53) Sanner, M. F.; Olson, A. J.; Spehner, J. C. *Biopolymers* **1996**, *38*, 305–320.

(54) Mongan, J.; Simmerling, C.; Mccammon, J. A.; Case, D. A.; Onufriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156–169.

(55) Svrcek-Seiler, W. A. Ph.D. thesis, University of Vienna: Vienna, Austria, 2003.

(56) Haberthür, U.; Caflisch, A. *J. Comput. Chem.* **2007**, *29*, 701–715.

(57) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297–1304.

(58) Nelder, J. A.; Mead, R. *Comput. J.* **1965**, *7*, 308–315.

(59) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383–394.

(60) Qin, S.; Zhou, H.-X. *Biopolymers* **2007**, *86*, 112–118.

(61) Dong, F.; Zhou, H.-X. *Proteins* **2006**, *65*, 87–102.

(62) Dzubiella, J.; Swanson, J. M.; McCammon, J. A. *J. Chem. Phys.* **2006**, *124*, 084905.

(63) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. *J. Med. Chem.* **2008**, *51*, 769–779.

(64) Nicholls, A.; Honig, B. *J. Comput. Chem.* **1991**, *12*, 435–445.

(65) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 265–284.

(66) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037–10041.

(67) Bashford, D. In *An Object-Oriented Programming Suite for Electrostatic Effects in Biological Molecules*, 1st ed.; Ishikawa, Y., Oldehoeft, R. R., Reynders, J. V. W., Tholburn, M., Eds.; Springer: Berlin, 1997; Vol. 1343, pp 233−240.

(68) Onufriev, A.; Case, D. A.; Bashford, D. *J. Mol. Biol.* **2003**, *325*, 555–567.

(69) Eliezer, D.; Yao, J.; Dyson, H. J.; Wright, P. E. *Nat. Struct. Biol.* **1998**, *5*, 148–155.

CT100392H

JCTC Journal of Chemical Theory and Computation

Comparing the Predictions of the Nonlinear Poisson−Boltzmann Equation and the Ion Size-Modified Poisson−Boltzmann Equation for a Low-Dielectric Charged Spherical Cavity in an Aqueous Salt Solution

Alexander R. J. Silalahi,[†] Alexander H. Boschitsch,[‡] Robert C. Harris,[†] and Marcia O. Fenley*,[†]

Department of Physics and Institute of Molecular Biophysics, Florida State University, 315 Keen Building, Tallahassee, Florida 32306-3408, United States and Continuum-Dynamics Inc., 34 Lexington Avenue, Ewing, New Jersey 08618-2302, United States

Received May 27, 2010

Abstract: The ion size-modified Poisson−Boltzmann equation (SMPBE) is applied to the simple model problem of a low-dielectric spherical cavity containing a central charge in an aqueous salt solution to investigate the finite ion size effect upon the electrostatic free energy and its sensitivity to changes in salt concentration. The SMPBE is shown to predict a very different electrostatic free energy than the nonlinear Poisson−Boltzmann equation (NLPBE) due to the additional entropic cost of placing ions in solution. Although the energy predictions of the SMPBE can be reproduced by fitting an appropriately sized Stern layer, or ion-exclusion layer to the NLPBE calculations, the size of the Stern layer is difficult to estimate a priori. The SMPBE also produces a saturation layer when the central charge becomes sufficiently large. Ion competition effects on various integrated quantities, such the total number of ions predicted by the SMPBE, are qualitatively similar to those given by the NLPBE and those found in available experimental results.

Introduction

Numerous processes involving the folding, bending, melting, and binding of highly charged biopolyelectrolytes, which are vital for biological function, are strongly influenced by changes in the ionic solvent environment. Nonspecific salt-mediated electrostatic interactions play an important role in these biomolecular processes because of their long-range influence, and such interactions largely govern the complex salt-dependent behavior of the above-mentioned processes. Therefore, physically realistic models of these long-range and salt-mediated electrostatic interactions are essential to predict the physiochemical behavior of charged biomacromolecules in ionic solutions.

Because of its simplicity and ability to accurately predict many thermodynamic properties, the nonlinear Poisson−Boltzmann equation (NLPBE)[1−4] has been extensively used to model the ionic solvent environment of biomolecules. Despite this success, it is subject to well-known approximations, such as omitting finite ion size and ion−ion correlation effects, which prohibit its application to systems where these effects become pronounced such as ionic layering, overcharging, or charge inversion, like-charge attraction, and ion selectivity inversion, all of which are associated with highly charged macroions, usually under high-salt conditions or in the presence of multivalent ions.[5−11]

Addressing these approximations within the Poisson−Boltzmann approach while retaining its simplicity is therefore desirable. As a result, several investigators have formulated various corrections to the NLPBE in order to include the effects of ion−ion correlations,[12] a dipolar solvent with variable density,[13] and other effects[14] in an effort to account

* Address correspondence to mfenley@sb.fsu.edu.
[†] Florida State University.
[‡] Continuum-Dynamics Inc..

10.1021/ct1002785
Published on Web 11/19/2010

for the above-mentioned intriguing experimentally observed phenomena.

Experimental studies have shown that even for small inorganic ions at intermediate- to high-salt concentrations, ion size effects cannot be ignored in quantitative predictions of nanopore selectivity at high surface charge densities.[15] In light of such experimental observations, several methods have been developed to address the effects of finite ion size; one of the first being the use of an ion-exclusion region or "Stern layer" surrounding the charged biomolecule where ions are not permitted to penetrate.[16] Others include using a cutoff on the maximum local salt concentration,[17] Coulomb gas with finite size,[18] modified Poisson−Boltzmann based on the generalized Poisson Fermi formalism,[19] lattice statistics models,[20−22] equation of state coupled to a function integral representation for a hard sphere fluid mixture approach,[23] and Bogoliubov−Bom−Green−Yvon hierarchy by Outhwaite and co-workers.[24,25]

In this paper, we chose to investigate the lattice gas-based method because it provides a physically realistic model that can be extended to the case of nonuniform ion sizes, without increasing the computational complexity of the original Poisson−Boltzmann solution. The predictions of one such theory, the uniform ion size-modified Poisson−Boltzmann equation (SMPBE)[20,21,26] are compared with those of the traditional NLPBE with and without an ion-exclusion region. The uniform ion size MPBE is implemented for a low-dielectric spherical cavity with a central charge embedded in an aqueous salt solution. The mathematical details of this uniform ion SMPBE model are described in the Appendix, and its numerical implementation is discussed in the Methods section.

The most popular method of accounting phenomenologically for the excluded volume or steric effects in the NLPBE relies on using a Stern layer, or ion-exclusion region. Its use is motivated by reasoning that ions will not come within their van der Waal's radius of the biomolecule, an observation that has been confirmed by different Monte Carlo simulations[27,28] of ionic distributions around biomolecules. However, this approximate model provides a limited representation of finite ion size effects and, as will be shown later, does not reproduce the predictions of the SMPBE. Some investigators[29] have used a Stern layer in conjunction with the SMPBE but doing so is an effective overcounting of the observed lack of ion centers within their atomic radius of the molecular surface because in the SMPBE an ion with its atomic center at that distance would still contribute charge density at the molecular surface because of the ion's finite size.

Unfortunately, as will be discussed below, current experimental data[29−32] are insufficient to conduct a decisive validation of these models since they have generally measured global or integrated (over the entire space) properties, such as the number of bound counterions, which are essentially identical for all of the candidate models. To determine which of these competing methods is best, experimental probes of local properties must be devised to measure, say, the forces between charged biomolecules at short separation distances, whose predictions would be sensitive to modeling choice.

Methods

Nonlinear Poisson−Boltzmann Equation. In the NLPBE, the normalized electrostatic potential $\phi = e\varphi/k_bT$ of a biomolecule immersed in an ionic solvent obeys

$$\nabla \cdot \varepsilon(r)\nabla\varphi + \frac{4\pi e}{k_bT}\rho(r) = 0 \tag{1}$$

where $\varepsilon(r)$ is the dielectric constant, e is the protonic charge, k_b is the Boltzmann constant, and T is the absolute temperature of the ionic solution. The charge density $\rho^{NLPBE}(r)$ in the solute, ion-exclusion, and solvent regions (respectively, Ω_1, Ω_2, and Ω_3) is given by

$$\rho^{NLPBE}(r) = \begin{cases} \rho^f = \sum_i Q_i\delta(r - r_i), & r \in \Omega_1 \\ 0, & r \in \Omega_2 \\ \rho^{ion} = e\sum_{k=1}^{nsalt}\{z_k^+c_{bk}^+\exp(-z_k^+\varphi) - z_k^-c_{bk}^-\exp(+z_k^-\varphi)\}, & r \in \Omega_3 \end{cases} \tag{2}$$

In eq 2 Q_i is the discrete charge of the solute at position r_i, and $z_k^{\pm}$ and $c_{bk}^{\pm}$ are the valence and the bulk concentration for both the co-ion and the counterion, respectively. The region Ω_2 is frequently modified by setting ρ_{ion} to 0 in the vicinity of the molecular surface, creating the Stern layer or ion-exclusion region.

In the case of small electrostatic potentials, $\phi \ll 1$, eq 1 in region Ω_3 reverts to the well-known linearized PBE:

$$\nabla^2\varphi = \kappa^2\varphi \tag{3}$$

where the Debye−Hückel screening parameter, κ, is given by $\kappa^2 = (4\pi e^2/k_bT\varepsilon_{ext})\sum_{k=1}^{nsalt}\{c_{bk}^+(z_k^+)^2 + c_{bk}^-(z_k^-)^2\}$, and ε_{ext} is the dielectric constant of the solvent.

Uniform Ion Size-Modified Poisson−Boltzmann Equation. To account for the uniform finite size of ions, the NLPBE can be modified as in the lattice gas model.[20] In this model, each ion is assumed to occupy a cube of volume a^3, where a is twice the radius of the ion, r_{ion}, and all of the ions in the solution have equal radii. The mobile ion density in the exterior region Ω_3 is therefore modified to[20]

$$\rho^{SMPBE}(r) = \frac{\rho^{NLPBE}(r)}{1 + \xi(r)} \tag{4}$$

where $\xi(r) = \sum_k\xi_k(r)$, and the volume exclusion factor of the ith ion is given by

$$\xi_i(r) = a^3\{c_{bi}^+\exp(-z_i^+\varphi(r)) + c_{bi}^-\exp(z_i^-\varphi(r)) - (c_{bi}^+ + c_{bi}^-)\} \tag{5}$$

In both the NLPBE and the SMPBE, the electrostatic free energy F^{elec} can be expressed as the sum of three terms. In the SMPBE, this can be written as

$$F^{\mathrm{SMPBE}} = \int_V \mathrm{d}^3 r\Big(- \frac{\varepsilon|\nabla\phi(r)|^2}{8\pi} + \rho^f\phi(r) - \frac{k_\mathrm{b}T}{a^3}\ln(1 + \xi(r))\Big) \quad (6)$$

where the first term is the electrostatic stress, and the third term, denoted by $\Delta\Pi^{\mathrm{SMPBE}}$, is the osmotic pressure whose NLPBE counterpart is

$$\Delta\Pi^{\mathrm{NLPBE}} = k_\mathrm{b}T\int_V \mathrm{d}^3 r\sum_k c_{bk}\{z_k^- \exp(-z_k^+\varphi(r)) + z_k^+\exp(-z_k^-\varphi(r)) - (z_k^+ + z_k^-)\} \quad (7)$$

which is readily obtained in the limit $a \to 0$. Note that $\Delta\Pi^{\mathrm{SMPBE}}$, unlike $\Delta\Pi^{\mathrm{NLPBE}}$, is not separable into contributions from individual ions because these contributions combine nonlinearly in the logarithm. This means that the well-known relationship between $\Delta\Pi$ and the derivative of the electrostatic free energy with respect to $\log(c_b)$ available for the NLPBE:

$$\frac{\mathrm{d}F^{\mathrm{NLPBE}}}{\mathrm{d}\log(c_{bk})} = -\Delta\Pi_k^{\mathrm{NLPBE}} \quad (8)$$

does not hold in the SMPBE. However, an expression for this derivative is nevertheless available and is given by (see Appendix)

$$\frac{\mathrm{d}F^{\mathrm{SMPBE}}}{\mathrm{d}\log(c_{bk})} = -\frac{k_\mathrm{b}T}{a^3}\int_V \mathrm{d}^3 r\frac{\xi_k(r)}{(1 + \xi(r))} \quad (9)$$

A simple relationship can be found between F^{SMPBE} and F^{NLPBE} in the limit of small ξ by writing:

$$\frac{\mathrm{d}F^{\mathrm{SMPBE}}}{\mathrm{d}a} = \frac{3}{a}\left(\sum_k c_{bk}\frac{\mathrm{d}F^{\mathrm{SMPBE}}}{\mathrm{d}c_{bk}} + \Delta\Pi^{\mathrm{SMPBE}}\right) \quad (10)$$

In the limit of $\xi(r) \ll 1$, eq 10 simplifies to

$$\log(F^{\mathrm{SMPBE}} - F^{\mathrm{NLPBE}}) = 3\log(a) + \log\left(\frac{1}{2}k_\mathrm{b}T\int_V f^2(r)\mathrm{d}^3 r\right) \quad (11)$$

where this integral is taken over Ω_3, and it is assumed that $\xi(r) = a^3 f(r)$, where $f(r)$ is independent of a to the first order (see Appendix). This approximation does not hold when a saturation layer of counterions forms near the surface of the sphere, so eq 11 will not apply in these cases.

In ionic solutions, a charged biomolecule alters the distribution of counter- and co-ions by attracting a cloud of counterions and repelling co-ions. The number of excess counterions attracted to the biomolecule, ν_k, is experimentally accessible.[29,30] Given a solution to either the NLPBE or the SMPBE, this number can be obtained from the following expression:

$$\nu_k^{\pm} = \int_V c_{bk}^{\pm}\left(\frac{\exp(\mp z_k^{\pm}\varphi)}{1 + \xi} - 1\right)\mathrm{d}^3 r \quad (12)$$

Numerical Implementation of the 1D Uniform Ion SMPBE. Here the case of a low-dielectric spherical cavity containing a central charge and surrounded by ionic solution is considered. This simple spherical configuration is directionally invariant, with the solution depending solely on the radial coordinate r and with the origin at the center of the low-dielectric spherical cavity. Therefore, eq 1 reduces to the 1D NLPBE equation:

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(\varepsilon(r)r^2\frac{\mathrm{d}\varphi}{\mathrm{d}r}\right) + \frac{4\pi e}{k_\mathrm{b}T}r^2\rho(r) = 0 \quad (13)$$

A discrete approximation to this equation is easily developed using either finite difference or finite element methods and results in high-resolution solutions. The following results use an approximation[1] previously developed for solving the NLPBE, extended here by modifying the source term to account for uniform finite ion size as in eq 4 and by incorporating the boundary treatment described in Boschitsch and Fenley,[33] which applies to both the linear and the nonlinear PBE and requires only that the potential at the outer boundary be small, $|\phi| \ll 1$. Correction terms are also added to the calculated electrostatic energies and salt gradients to account for contributions from outside the computational domain. These terms can become important, especially at lower salt concentrations.

The calculations presented here used 5000 nodes and an outer boundary placed at 200 Å from the surface of the sphere of radius 20 Å. The solution was considered to have converged when the maximum change in potential at a grid node was less than 10^{-5}*min{1, $|\phi_\mathrm{s}|$} where ϕ_s is the surface potential. To verify that results were both converged and accurate, additional calculations were conducted for randomly selected cases to ensure that no significant changes occurred when: (i) the mesh resolution was halved and (ii) a smaller convergence criterion was used. The derivatives of the electrostatic free energy with respect to bulk concentration and to ion size were also verified by using finite difference methods.

Results

Unless specified otherwise all results were generated for a spherical cavity of radius 20 Å containing a central charge of −50 e. The ion radius is equal to one-half of the size of the lattice spacing (a) used in the SMPBE. It is set to values between approximately 1 to 8 Å. The Stern layer is not invoked for the SMPBE calculations. However, as noted in the text below for some NLPBE calculations, a Stern layer is employed. The dielectric constant was set to 78.5 in the exterior region ($\Omega_2 \cup \Omega_3$) and 4 inside the molecule (Ω_1). All calculations were performed at 298.15 K.

Ion Competition Effects on Global Thermodynamic Properties. The behavior of many highly charged biomolecules, such as nucleic acids, is highly influenced by changes in ionic conditions due to the presence of neutralizing counterions that form a characteristic ionic cloud surrounding the charged biomolecule. Moreover, many vital biological processes occur in salt mixtures and involve changes in competition effects between monovalent and multivalent ions.[34,35]

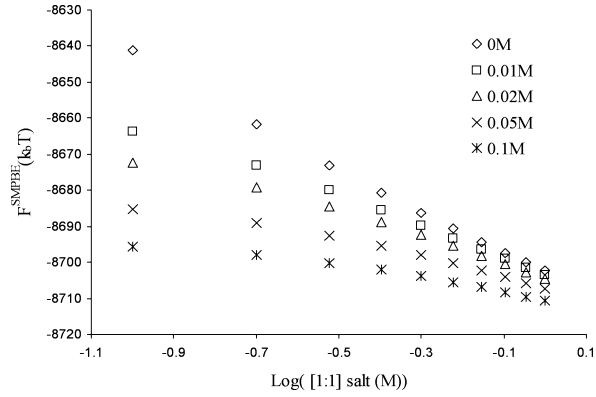


Figure 1. The electrostatic free energy (in k_bT) computed with the uniform ion size-modified Poisson−Boltzmann equation, F^{SMPBE}, of a low-dielectric spherical cavity of radius 20 Å and a central charge of −50 e at concentrations of 2:1 salt, [$MgCl_2$], of 0, 0.01, 0.02, 0.05, and 0.1 M and an ion radius of 1.5 Å is plotted against the logarithm of the concentration of 1:1 salt, [NaCl].

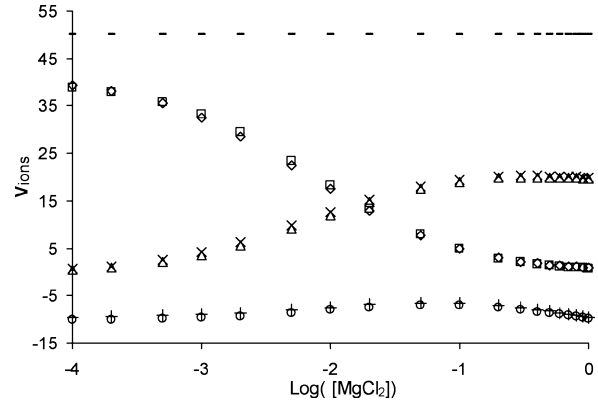


Figure 2. The number of bound Mg^{2+}, Na^+, and Cl^- ions (v_{Mg}, v_{Na}, and v_{Cl}) for a mixed salt solution, with an ion radius of 1.4 Å and a 1:1 NaCl salt concentration fixed at 0.1 M, calculated with both the nonlinear Poisson−Boltzmann equation (Δ, □, ○) and the size-modified Poisson−Boltzmann equation (×, ◇, +) is plotted as a function of $MgCl_2$ salt concentration ([$MgCl_2$]). The NLPBE calculations were performed with a Stern layer of 1.4 Å thickness.

Because the SMPBE models ions differently than the NLPBE, its predictions of competition effects between different ion species may differ from those of the classical NLPBE. This is investigated in Figure 1 which plots F^{SMPBE} against [NaCl] for different values of [$MgCl_2$]. For increasing [$MgCl_2$], the resulting curves exhibit gradual flattening with decreased slopes. This reflects the Mg^{2+} counterions competing the Na^+ ions away from near the surface of the molecule. These curves have the same qualitative form as similar curves for the NLPBE, as shown by Shen and Honig,[2] Boschitsch and Fenley,[1] and Figure A.1. in the Supporting Information, and therefore there is no clear difference between the ion competition effects predicted by the NLPBE with or without the Stern layer and those predicted by the SMPBE, at least for the integrated quantities considered here.

That the SMPBE produces similar predictions for ion competition effects on global thermodynamic properties of the low-dielectric spherical cavity to those of the NLPBE can be further demonstrated by examining Figure 2, where v_{Mg}, v_{Na}, and v_{Cl} are plotted against log([$MgCl_2$]) for NaCl = 0.1 M. These curves are qualitatively similar to the comparable curves generated by Lipfert and co-workers,[26,30] which included the 3D structure of actual biomolecules. This behavior indicates that, as they discussed, the inclusion of accurate 3D structures does not significantly change the predictions of ion competition effects for such integrated quantities. Once again, the ion competition effects of thermodynamic parameters predicted by the SMPBE are in close agreement with the NLPBE's predictions. In these calculations and others we have carried out (though by no means exhaustive), we have generally found that both the NLPBE's and SMPBE's predictions of global or integrated quantities are in good agreement. However, the ion sizes considered here (sodium and magnesium), are comparatively small (radii are on the order of 1 Å), whereas, as was pointed out by Chu and co-workers,[29] if a much larger ion radius is considered (on the order of 10 Å), then the equivalent ion competition curves of global properties predicted by the SMPBE are found to differ significantly from those of the NLPBE. Exploring the validity of the SMPBE may therefore be possible by conducting experiments with much larger ions, such as polyamines (spermine and spermidine).[19] Note that increasing the charge on the biomolecule does not seem to increase the difference between the predictions of the traditional NLPBE and those of the SMPBE, as is evident from the Figure A.2. in the Supporting Information. Increasing the charge of the biomolecule therefore does not appear to be a viable experimental method for investigating the effects of the SMPBE. Although the NLPBE results in this figure use a Stern layer, the figure appears nearly identical when a Stern layer is not used, so the presence or absence of a Stern layer does not alter these conclusions.

Electrostatic Free Energy and its Salt Sensitivity. As mentioned before, any process (e.g., folding, stability, and binding) involving highly charged biopolyelectrolytes is affected by changes in salt concentration.[36] Here we examine how the electrostatic free energy and its salt sensitivity obtained with the SMPBE differ from that of the standard NLPBE.

The electrostatic free energy F^{SMPBE} differs from F^{NLPBE} due to the additional entropic energy cost of displacing solvent when placing ions in solution and the different electrostatic potentials predicted by the SMPBE. Equation 11 indicates that to first order the difference, $F^{SMPBE} - F^{NLPBE}$ is proportional to a^3 under conditions where no saturation layer forms. This is illustrated in Figure 3 where $\log(F^{SMPBE} - F^{NLPBE})$ for the low-dielectric spherical cavity in a 1:1 salt solution is plotted against $\log(r_{ion})$ ($a = 2r_{ion}$) for different concentrations of [NaCl]. The curves in this plot are linear, and their slopes are very close to the expected value of 3. Interestingly, although F^{SMPBE} depends upon [NaCl], $F^{SMPBE} - F^{NLPBE}$ is essentially independent of [NaCl] in the absence of a saturation layer. Despite the high charge of the sphere, the difference between the electrostatic energies predicted by SMPBE and NLPBE is less than k_bT. This indicates that distinguishing between SMPBE and

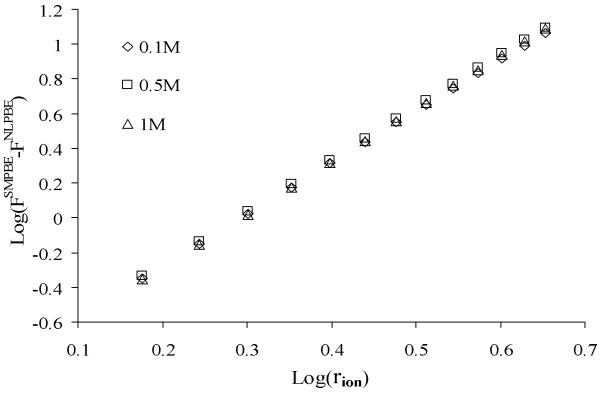


Figure 3. The logarithm of the difference between the electrostatic free energy (in units of k_bT) given by the uniform ion size-modified Poisson−Boltzmann equation and that given by the standard nonlinear Poisson−Boltzmann equation without a Stern layer, $\log(F^{SMPBE} - F^{NLPBE})$, is plotted against $\log(r_{ion})$ for three different values of NaCl concentration. The slopes of the curves are 3 to within 0.67%.

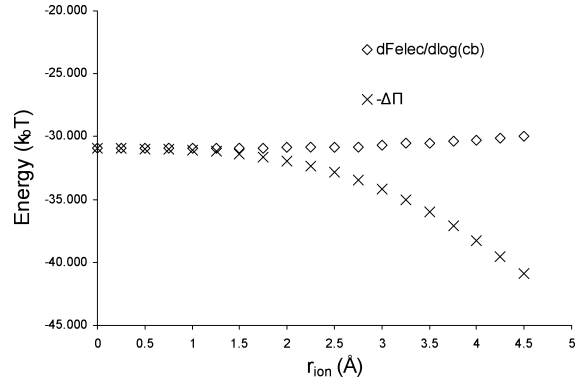


Figure 4. The excess osmotic pressure, $\Delta\Pi^{SMPBE}$ computed using the uniform ion size-modified Poisson−Boltzmann equation, SMPBE, and its derivative with respect to the logarithm of the bulk 1:1 salt concentration, $-dF^{SMPBE}/d\log(c_b)$, are plotted against ion size r_{ion}, where r_{ion} is equal to one-half of a, the size of the lattice spacing used in the SMPBE theory.

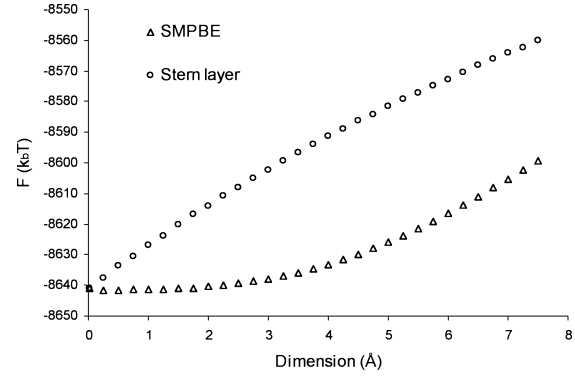


Figure 5. The electrostatic free energy predicted by the uniform ion size-modified Poisson−Boltzmann equation, F^{SMPBE}, as a function of the ion radius, r_{ion}, and the electrostatic free energy predicted by the standard nonlinear Poisson−Boltzmann equation, F^{NLPBE}, as a function of the thickness of the Stern layer.

NLPBE predictions on the basis of experimental solvation free energies is unlikely.

One important difference between the NLPBE and the SMPBE can be seen in the osmotic pressure term, $\Delta\Pi$. In the NLPBE, $\Delta\Pi$ is equal to $-dF^{NLPBE}/d\log(c_b)$, as shown by Sharp and co-workers.[37] This relationship does not hold in the SMPBE because adding an ion to the solvent displaces solvent molecules, introducing an additional entropic cost to $\Delta\Pi$. This additional entropic cost of placing an ion in solution explains the result, discussed by Chu and co-workers,[29] that the NLPBE underestimates the preference that biomolecules have for magnesium counterions over sodium counterions. Because the magnesium ions have roughly the same size as the sodium ions, placing one magnesium ion costs less entropy than placing two sodium ions, thereby increasing the predicted affinity of biomolecules for magnesium in the SMPBE. The additional entropic cost of placing ions introduced in the SMPBE means that, although $dF^{SMPBE}/d\log(c_b)$ depends only weakly on r_{ion}, $\Delta\Pi$ is very sensitive to changes in this parameter. This is clear from Figure 4, where $\Delta\Pi^{SMPBE}$ and $-dF^{SMPBE}/d\log(c_b)$ are plotted against ion size r_{ion}. $\Delta\Pi^{SMPBE} = -dF^{SMPBE}/d\log(c_b)$ when $r_{ion} = 0$ but quickly diverges for larger values of r_{ion}, while $dF^{SMPBE}/d\log(c_b)$ remains relatively constant.

Traditionally, a Stern layer has been added to the NLPBE to approximate ion size effects by reasoning that the main effect of a finite ion size is to exclude ions from the immediate vicinity of the charged biomolecule. To test this assertion, both F^{SMPBE} as a function of r_{ion} and F^{NLPBE} as a function of the thickness of the Stern layer are plotted in Figure 5. The two electrostatic energies diverge with increasing r_{ion}, because the NLPBE with a Stern layer overestimates the change in electrostatic free energy with ion size. The predictions of the salt dependence of the free energy by the two models are also different, as is clear from Figure 6, where the dependence $-dF^{SMPBE}/d\log(c_b)$ on ion size is given with the same parameters as in Figure 5. While it may be possible to adjust the Stern layer size to match the SMPBE's predictions, how to perform this adjustment in an a priori manner for general bimolecular configurations and ionic conditions is not obvious. Note further that the disparities in the slopes of the curves in Figures 5 and 6 at $a = 0$ point to a fundamental physical modeling discrepancy in how finite ion size is accounted for. Specifically, the SMPBE model indicates an $O(a^3)$ dependence on F near $a = 0$, whereas the NLPBE model with a Stern layer reflects an $O(a)$ behavior. The former variation is expected on the basis of a volume-based exclusion effect, whereas the latter behavior reflects the distance-based exclusion characteristic of the Stern layer.

It will be important to examine if the salt dependence of various thermodynamic properties of biopolyelectrolytes in salt mixtures can be better explained or predicted using the more accurate SMPBE, with nonuniform ion size, as opposed to using the NLPE with the Stern layer. Future studies should address this point for realistic biopolyelectrolytes for which thermodynamic salt-dependent data are available in the literature.

Saturation Layer. One feature that the SMPBE is able to capture is the presence of a saturation layer around the

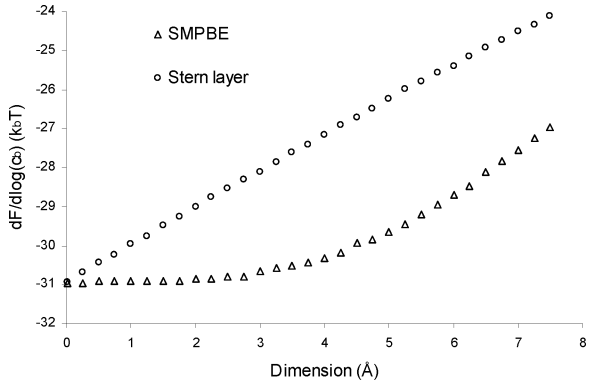


Figure 6. The derivative of the electrostatic free energy predicted by the uniform ion size-modified Poisson−Boltzmann equation with respect to the logarithm of the bulk concentration of 1:1 salt, $-dF^{SMPBE}/d\log(c_b)$, as a function of the ion radius, r_{ion}, and the equivalent quantity predicted by the classical nonlinear Poisson−Boltzmann equation, $-dF^{NLPBE}/d\log(c_b)$, as a function of the thickness of the Stern layer.

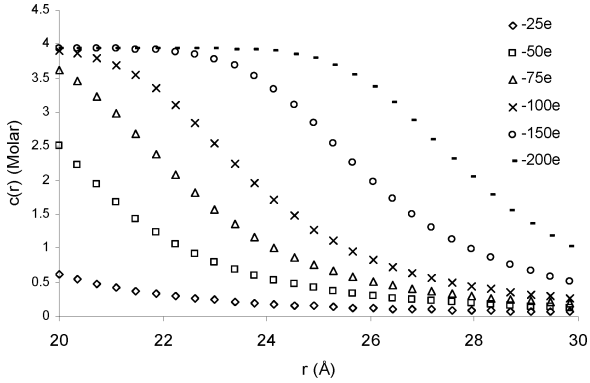


Figure 7. The local concentration of Na^+, $c(r)$, outside a low-dielectric spherical cavity in a 1:1 salt solution with [NaCl] = 0.01 M and an ion radius of 3.75 Å for different central charge values.

biomolecule. Essentially, as the local potential increases, the density of ions reaches a nonzero saturation concentration, which it cannot exceed because of the finite size of the ions. This contrasts with the standard NLPBE, with or without the use of Stern layer, which allows the concentration of ions to increase without bound and reaches physically implausible values in the vicinity of the charged interface. This behavior is illustrated in Figure 7, where the local concentration of counterions is plotted as a function of the distance from the surface of the low-dielectric spherical cavity. In this case the 1:1 salt solution has concentration [NaCl] = 0.01 M. The ion radius is 3.5 Å, and the central charge of the spherical cavity is varied. For a sufficiently large central charge, the counterions form a layer with a saturated ion density before falling off monotonically and rapidly to the bulk ion concentration. This same behavior was captured by Outhwaite and co-workers using a more sophisticated theory.[38] Interestingly, this figure indicates that the saturation layer will form even at very low salt concentrations (0.01M) provided the charge of the sphere is sufficiently large, and indeed, a saturation layer will develop at all nonzero salt concentrations for sufficiently high central charges. This behavior agrees with observations that have been made in other studies for other geometries, including charged cylinders and parallel plates.[21,22,39] Future experimental, simulation, and theoretical studies should confirm the existence of this saturation layer and examine its implications to other physical properties of arbitrary shaped biopolyelectrolytes under varying ionic conditions, including salt mixtures.

Conclusions

The influence of finite ion size was assessed on the basis of the size-modified Poisson−Boltzmann equation (SMPBE) applied to a low-dielectric spherical cavity containing a central charge and compared against the standard nonlinear Poisson−Boltzmann equation (NLPBE) in the context of competition effects between ion species and the salt dependence of the electrostatic free energy. It is found that the predictions of ion competition effects given by the SMPBE do not differ significantly from those of the NLPBE for the global quantities considered here. However, the SMPBE predictions may differ from those of the NLPB for ion distribution profiles of highly charged 3D biomolecules in salt mixtures, especially at and near to the highly charged 3D biomolecular surface.

Although SMPBE does not yield significantly different ion competition effects than NLPBE, it does produce different electrostatic free energy predictions. In particular, although $dF^{SMPBE}/d\log(c_b)$ closely matches the predictions of the classical NLPBE, $\Delta\Pi$ gains an additional entropic energy term from the displacement of solvent upon the placement of ions in solution that is very sensitive to changes in the ion radius. This means that, unlike in NLPBE, $\Delta\Pi$ predicted by the SMPBE is not equal to $-dF^{SMPBE}/d\log(c_b)$, as shown in the Appendix. Unfortunately, the electrostatic free energy is not directly accessible to experiment, making it impossible to examine the difference between the SMPBE's predictions of the free energy and those of the NLPBE.

The SMPBE predicts the formation of a saturation layer. When the central charge on the biomolecule is large enough to create a field that would normally cause the local ion concentration predicted by the NLPBE to exceed the saturation concentration, the SMPBE instead shows the presence of a saturation layer for various geometries including the charged sphere, cylinder, and infinite plate.[20,22,39,40]

We also sought to determine whether these effects of the SMPBE could be approximated by the NLPBE with a Stern layer, but as we demonstrated, these two models predict very different behaviors. Although the SMPBE does not include such effects as ion−ion correlations, it has a more rigorous foundation in statistical mechanics than the NLPBE with and without a Stern layer, and it therefore appears to be a preferable method for including the effects of ion size. The ion distributions can be different to that of the uniform SMPBE when ions with different charges and radii are present in the ionic solution. By using a nonuniform ion SMPBE based on the generalized Poisson−Fermi approach, Tresset and co-workers[41] predicted ion stratification around a uniformly charged plane. The proper modeling of biopolyelectrolytes in salt mixtures will require nonuniform ion SMPB solvers. In order to verify the robustness of such

nonuniform SMPB solvers, it will require the availability of pertinent experimental and simulation data. Unfortunately, comparing the predictions of the SMPBE to experimental findings was impossible in the present paper because the code used could not use a realistic 3D molecular surface. This capability will be implemented shortly, and the results will be presented in a future publication. However, the results presented in this paper should also pertain to calculations performed on realistic molecular surfaces.

Acknowledgment. We thank Dr. Hugh Nymeyer for useful discussions. One of the authors (M.O.F.) acknowledges the support from NIH-GM078538-01 (PI: Dr. Michael S. Chapman, OHSU, Co-PI: MOF). Both authors (M.O.F. and A.H.B.) acknowledge support from SBIR NIH 1R43GM079056.

Appendix

Electrostatic Free Energy. The electrostatic free energy, F^{SMPBE}, of the SMPBE model can be expressed as the integral over the entire domain:

$$F^{\mathrm{SMPBE}} = \int_V \mathrm{d}^3 r\left(- \frac{\varepsilon|\nabla\phi|^2}{8\pi} + \rho^f\phi\right) - \Delta\Pi \qquad (14)$$

where the first term is the electrostatic stress, the third term is the excess osmotic pressure term, $\Delta\Pi$, ε is the dielectric constant, φ is the electrostatic potential, ρ^f is the biomolecule charge distribution, and $\Delta\Pi$ is the excess osmotic pressure of the mobile ion cloud. In the SMPBE, $\Delta\Pi$ is given by:[20]

$$\Delta\Pi^{\mathrm{SMPBE}} = -\int_{\Omega_3} \frac{k_\mathrm{b}T}{a^3}\ln\left(\frac{1-C}{1-C_0}\right)\mathrm{d}^3 r \qquad (15)$$

where the domain of integration in eq 15 is the region Ω_3. C is the fraction of the local volume taken up by ions in the presence of the biomolecule, and C_0 is the same quantity in the absence of the biomolecule:

$$C = a^3\sum_{k=1}^{N\mathrm{salt}}(c_k^+ + c_k^-) \qquad (16)$$

$$C_0 = C(\varphi = 0) = a^3\sum_{k=1}^{N\mathrm{salt}}(c_{bk}^+ + c_{bk}^-) \qquad (17)$$

Here c_k^+ and c_k^- are the local concentrations of the kth ion species in the presence of the biomolecule, and c_{bk}^+ and c_{bk}^- are the local concentrations in the absence of the biomolecule.

In the SMPBE, $c_k^\pm$ is given by:[20]

$$c_k^\pm = c_{bk}^\pm\frac{\exp(\mp z_k^\pm\varphi)}{1+\xi(r)} \qquad (18)$$

where $\xi(r) = \sum_k\xi_k(r)$ and

$$\xi_k(r) = a^3\{c_{bk}^+\exp(-z_k^+\varphi) + c_{bk}^-\exp(z_k^-\varphi) - (c_{bk}^+ + c_{bk}^-)\} \qquad (19)$$

where z_k^+ and z_k^- are the valences of ions A and B.

Salt Gradient of the Electrostatic Free Energy. The salt gradient of the electrostatic free energy, $(\mathrm{d}F^{\mathrm{SMPBE}})/(\mathrm{d}c_{bi})$ is defined as the derivative of the free energy with respect to the bulk concentration of salt. To compute this quantity, first take the variation of the electrostatic free energy term in eq 14:

$$\begin{aligned}\delta F^{\mathrm{SMPBE}} &= \int_V -\nabla(\delta\phi)\cdot\frac{\varepsilon}{4\pi}\nabla\phi + \rho^f\delta\phi + \frac{k_\mathrm{b}T}{a^3}\delta\ln\left(\frac{1-C}{1-C_0}\right)\mathrm{d}^3 r \\ &= \int_V -\nabla\cdot\left(\frac{\varepsilon}{4\pi}\delta\phi\nabla\phi\right) + \delta\phi\nabla\cdot\left(\frac{\varepsilon}{4\pi}\nabla\phi\right) + \\ &\qquad \rho^f\delta\phi + \frac{k_\mathrm{b}T}{a^3}\delta\ln\left(\frac{1-C}{1-C_0}\right)\mathrm{d}^3 r\end{aligned} \qquad (20)$$

From this equation, the salt gradient of the electrostatic free energy is:

$$\frac{\mathrm{d}F^{\mathrm{SMPBE}}}{\mathrm{d}c_{bi}} = S_i + \int_V \frac{\mathrm{d}\phi}{\mathrm{d}c_{bi}}\left(\nabla\cdot\frac{\varepsilon}{4\pi}\nabla\phi + \rho^f + \rho^{\mathrm{ion}}\right) - \rho^{\mathrm{ion}}\frac{\mathrm{d}\phi}{\mathrm{d}c_{bi}} + \frac{k_\mathrm{b}T}{a^3}\frac{\mathrm{d}}{\mathrm{d}c_{bi}}\ln\left(\frac{1-C}{1-C_0}\right)\mathrm{d}^3 r \qquad (21)$$

where the ion distribution ρ^{ion} is given by:

$$\rho^{\mathrm{ion}} = e\sum_{k=1}^{N\mathrm{salt}} c_k^+ z_k^+\exp(-z_k^+\varphi) - c_k^- z_k^-\exp(z_k^-\varphi) \qquad (22)$$

and the surface integral term is given by:

$$S_i = -\int_{S_\infty}\frac{\varepsilon}{4\pi}\frac{\mathrm{d}\phi}{\mathrm{d}n}\frac{\mathrm{d}\phi}{\mathrm{d}c_{bi}}\mathrm{d}S \qquad (23)$$

The second term of eq 21 is 0 because it is a restatement of the PBE. The fourth term can be rewritten by combining eqs 16−19 as:

$$\ln\left(\frac{1-C}{1-C_0}\right) = -\ln(1+\xi(r)) \qquad (24)$$

Therefore the salt gradient of the electrostatic free energy for the ith ion species is given by:

$$\frac{\mathrm{d}F^{\mathrm{SMPBE}}}{\mathrm{d}c_{bi}} = S_i + \int_V -\rho^{\mathrm{ion}}\frac{\mathrm{d}\phi}{\mathrm{d}c_{bi}} - \frac{k_\mathrm{b}T}{a^3(1+\xi(r))}\frac{\mathrm{d}\xi(r)}{\mathrm{d}c_{bi}}\mathrm{d}^3 r \qquad (25)$$

By assuming that the charge neutrality condition for salt A_xB_y requires that $c_{bi}^+ z_i^+ = c_{bi}^- z_i^-$ and that the salt completely dissociates in the solvent, we get $c_{bi}:c_{bi}^+:c_{bi}^- = 1:x:y$.

By calculating the derivative of $\xi(r)$ with respect to c_{bi} and using eq 22, further simplification is possible:

$$\begin{aligned}\frac{dF^{SMPBE}}{dc_{bi}} &= S_i + \int_V d^3r\left(-\rho^{ion}\frac{d\phi}{dc_{bi}} - \frac{k_bT}{a^3(1+\xi(r))}\frac{d\xi(r)}{dc_{bi}}\right)\\ &= S_i + \int_V d^3r\left(-\rho^{ion}\frac{d\phi}{dc_{bi}} - \frac{k_bT}{a^3(1+\xi(r))}\frac{d\xi(r)}{dc_{bi}}\right)\\ &= S_i + \int_V d^3r\left(\begin{array}{c}-\rho^{ion}\frac{d\phi}{dc_{bi}} - \frac{k_bT}{a^3(1+\xi(r))}\{xz_i^+\exp(-z_i^+\varphi) +\\ yz_i^-\exp(z_i^-\varphi) - (z_i^+ + z_i^-)\}\\ +\rho^{ion}\frac{d\varphi}{dc_{bi}}\end{array}\right)\\ &= S_i - \int_V d^3r\left(\begin{array}{c}\frac{k_bT}{a^3(1+\xi(r))}\{xz_i^+\exp(-z_i^+\varphi) +\\ yz_i^-\exp(z_i^-\varphi) - (z_i^+ + z_i^-)\}\end{array}\right)\\ &= S_i - \frac{k_bT}{a^3c_{bi}}\int_V d^3r\frac{\xi_i(r)}{(1+\xi(r))}\end{aligned} \tag{26}$$

where the salt gradient of the electrostatic free energy contains a normalization constant $1 + \xi(r)$.

At finite salt concentrations, the surface integral term, S_i, is 0 because the exponential decay in the potential guarantees that the integrand is 0 at ∞, but in the limit of zero salt concentration, this is not true, and S_i diverges. This difficulty is avoided by considering instead the derivative of the electrostatic free energy with respect to $\log(c_b)$ ($c_b^{1/2}$ is another useful choice yielding finite derivatives) to obtain:

$$\frac{dF^{SMPBE}}{d\log(c_{bi})} = c_{bi}S_i - \frac{k_bT}{a^3}\int_V d^3r\frac{\xi_i(r)}{(1+\xi(r))} \tag{27}$$

The product $c_{bi}S_i$ now vanishes as $c_{bi} \to 0$. To prove this, note that at a large distance, R, the potential is sufficiently small so that the local behavior in the region $r > R$ is governed by the linear PBE. Thus, the potential solution behaves as:

$$\phi(r) \sim \frac{B\exp(-\kappa r)}{r}, \qquad r > R \tag{28}$$

where B is a constant and

$$\begin{aligned}\kappa^2 &= \frac{8\pi e^2}{k_bT\varepsilon}\sum_k\frac{1}{2}\{c_{bk}^+(z_k^+)^2 + c_{bk}^-(z_k^-)^2\}\\ &= \frac{4\pi e^2}{k_bT\varepsilon}\sum_k c_{bk}z_k^-z_k^+\{z_k^+ + z_k^-\}\end{aligned} \tag{29}$$

The derivative of φ with respect to κ is given by:

$$\frac{d\phi}{d\kappa} = -B\exp(-\kappa r) \tag{30}$$

and its derivative with respect to r is given by:

$$\frac{d\phi}{dr} = -\frac{B(\kappa r+1)\exp(-\kappa r)}{r^2} \tag{31}$$

Inserting into eq 23:

$$S_i = -\lim_{R\to\infty}\frac{\varepsilon}{4\pi}\left(-\frac{B(\kappa R+1)\exp(-\kappa R)}{R^2}\right)\left(-B\exp(-\kappa R)\frac{d\kappa}{dc_{bi}}\right)4\pi R^2 = -\varepsilon B^2\left(\frac{d\kappa}{dc_{bi}}\right)\lim_{R\to\infty}\{(\kappa R+1)\exp(-2\kappa R)\} \tag{32}$$

The last term in braces is bounded by 1 so that:

$$c_{bi}|S_i| \le \varepsilon B^2\left(c_{bi}\frac{d\kappa}{dc_{bi}}\right) \tag{33}$$

The derivative of κ with respect to c_{bi} is:

$$\frac{d\kappa}{dc_{bi}} = \frac{2\pi e^2}{\kappa k_bT\varepsilon}z_i^-z_i^+\{z_i^+ + z_i^-\} \tag{34}$$

so that:

$$c_{bi}\frac{d\kappa}{dc_{bi}} \le \sqrt{c_{bi}}D \tag{35}$$

where $D = \sqrt{\pi e^2 z_i^+ z_i^-(z_i^+ + z_i^-)\,/\,k_bT\varepsilon}$

Substituting into the result for $c_{bi}|S_i|$ confirms the convergence with zero concentration, c_{bi}.

Dependence of the Electrostatatic Free Energy on Ion Size. At small salt concentrations, the difference between the free energy predictions of the SMPBE and those of the NLPBE are dominated by the additional entropic cost of placing an ion in solution in the SMPBE model. By considering the derivative of electrostatic free energy with respect to a, dF^{SMPBE}/da, a formula relating F^{SMPBE} to F^{NLPBE} in the limit of small salt concentrations can be derived as follows:

$$\frac{dF^{SMPBE}}{da} = M + \int_V d^3r\left(\frac{d\phi}{da}\left(\nabla\cdot\frac{\varepsilon}{4\pi}\nabla\phi + \rho^f + \rho^{ion}\right) - \rho^{ion}\frac{d\phi}{da} - k_bT\frac{d}{da}\frac{\ln(1+\xi(r))}{a^3}\right) \tag{36}$$

where $M = -\int_{S_\infty}\varepsilon/4\pi\, d\phi/dn\, d\phi/da dS$.

Because the electrostatic potential has the asymptotic form $\exp(-\kappa r)/r$, $M = 0$ and the second term on the right-hand side of eq 36 is 0 because it is the SMPBE equation. By calculating the derivative of $\xi(r)$ with respect to a and using eq 15, eq 36 can be further simplified to:

$$\frac{dF^{SMPBE}}{da} = \left(\frac{3}{a}\right)\frac{k_bT}{a^3}\int_V d^3r\left(\ln(1+\xi(r)) - \frac{\xi(r)}{1+\xi(r)}\right) = \left(\frac{3}{a}\right)\left(\sum_k\frac{dF^{SMPBE}}{d\log(c_{bk})} + \Delta\Pi^{SMPBE}\right) \tag{37}$$

By considering the limit of small $\xi(r)$, we can derive a relationship between F^{SMPBE} and F^{NLPBE}. The Taylor series expansion of eq 37 for small ξ is:

$$\frac{dF^{SMPBE}}{da} = \frac{3k_bT}{a^4}\int_V d^3r\left(\frac{1}{2}\xi^2(r) + O(\xi^3(r))\right) \tag{38}$$

Expanding $\xi(r)$ to lowest order gives $\xi(r) = a^3f(r)$, where $f(r)$ does not depend on a (see Supporting Information Materials). By retaining only the $\xi^2(r)$ term, this equation becomes:

$$\frac{dF^{SMPBE}}{da} = \frac{3a^2}{2}k_bT\int_V d^3r(f^2(r)) \tag{39}$$

Integrating:

$$F^{\mathrm{SMPBE}}(a) = F^{\mathrm{SMPBE}}(0) + \left(\frac{1}{2}k_{\mathrm{b}}T\int_{V} f^{2}(r)\mathrm{d}^{3}r\right)a^{3} \quad (40)$$

demonstrating that ($F^{\mathrm{SMPBE}} - F^{\mathrm{NLPBE}}$) is proportional to a^3.

Supporting Information Available: Graph showing the electrostatic free energy of a low-dielectric spherical cavity in a mixed salt solution computed with the nonlinear Poisson−Boltzmann equation. Graphs illustrating the effect of increasing the charge of the low-dielectric charged spherical cavity embedded in ionic solution for the example in Figure 2. The volume integration used in the derivation of the salt gradient of the electrostatic free energy. The Taylor series of electrostatic potential used in estimating the dependence of electrosatic free energy on ion size. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Boschitsch, A. H.; Fenley, M. O. *J. Comput. Chem.* **2004**, *25*, 935–955.

(2) Shu-wen, W. C.; Honig, B. *J. Phys. Chem. B* **1997**, *101*, 9113–9118.

(3) Misra, V. K.; Draper, D. E. *J. Mol. Biol.* **2000**, *299*, 813–825.

(4) Draper, D. E.; Grilley, D.; Soto, A. M. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 221–243.

(5) Allahyarov, E.; Gompper, G.; Löwen, H. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2004**, *69*, 041904.

(6) Wilson, R. W.; Bloomfield, V. A. *Biochemistry* **1979**, *18*, 2192–2196.

(7) Grosberg, A. Y.; Nguyen, T. T.; Shklovskii, B. I. *Rev. Mod. Phys.* **2002**, *74*, 329–345.

(8) Tresset, G.; Davy Cheong, W. C.; Shireen Tan, Y. L.; Boulaire, J.; Ming Lam, Y. *Biophys. J.* **2007**, *93*, 637–644.

(9) Gelbart, W. M.; Bruinsma, R. F.; Pincus, P. A.; Parsegian, V. A. *Phys. Today* **2000**, *53*, 38–44.

(10) Bhuiyan, L. B.; Vlachy, V.; Outhwaite, C. W. *Int. Rev. Phys. Chem.* **2002**, *21*, 1–36.

(11) Lamperski, S.; Bhuiyan, L. B. *J. Electroanal. Chem.* **2003**, *540*, 79–87.

(12) Netz, R. R.; Orland, H. *Eur. Phys. J. E: Soft Matter Biol. Phys.* **2000**, *1*, 203–214.

(13) Azuara, C.; Orland, H.; Bon, M.; Koehl, P.; Delarue, M. *Biophys. J.* **2008**, *95*, 5587–5605.

(14) Pincus, P. A.; Safran, S. A. *Europhys. Lett.* **1998**, *42*, 103–108.

(15) Cervera, J.; Ramírez, P.; Manzanares, J. A.; Mafé, S. *Microfluid. Nanofluid.* **2009**, *9*, 41–53.

(16) Stern, O. *Z. Elektrochem* **1924**, *30*, 508–516.

(17) Grochowski, P.; Trylska, J. *Biopolymers* **2008**, *89*, 93–113.

(18) Coalson, R. D.; Walsh, A. M.; Duncan, A.; Ben-Tal, N. *J. Chem. Phys.* **1995**, *102*, 4584–4594.

(19) Tresset, G.; Cheong, W. C. D.; Lam, Y. M. *J. Phys. Chem. B* **2007**, *111*, 14233–14238.

(20) Borukhov, I.; Andelman, D.; Orland, H. *Phys. Rev. Lett.* **1997**, *79*, 435–438.

(21) Kralj-Iglic, V.; Iglic, A. *J. Phys. II* **1996**, *6*, 477–491.

(22) Manciu, M.; Ruckenstein, E. *Langmuir* **2002**, *18*, 5178–5185.

(23) Lue, L.; Zoeller, N.; Blankschtein, D. *Langmuir* **1999**, *15*, 3726–3730.

(24) Outhwaite, C. W.; Bhuiyan, L. B. *J. Chem. Soc., Faraday Trans. 2* **1983**, *79*, 707–718.

(25) Bhuiyan, L. B.; Outhwaite, C. W. *J. Colloid Interface Sci.* **2009**, *331*, 543–547.

(26) Chu, V. B.; Bai, Y.; Lipfert, J.; Herschlag, D.; Doniach, S. *Biophys. J.* **2007**, *93*, 3202–3209.

(27) Boda, D.; Fawcett, W. R.; Henderson, D.; Sokołowski, S. *J. Chem. Phys.* **2002**, *116*, 7170–7176.

(28) Jayaram, B.; Swaminathan, S.; Beveridge, D. L.; Sharp, K.; Honig, B. *Macromolecules* **1990**, *23*, 3156–3165.

(29) Chu, V. B.; Bai, Y.; Lipfert, J.; Herschlag, D.; Doniach, S. *Biophys. J.* **2007**, *93*, 3202–3209.

(30) Lipfert, J.; Chu, V. B.; Bai, Y.; Herschlag, D.; Doniach, S. *J. Appl. Crystallogr.* **2007**, *40*, s229–s234.

(31) Das, R.; Mills, T. T.; Kwok, L. W.; Maskel, G. S.; Millett, I. S.; Doniach, S.; Finkelstein, K. D.; Herschlag, D.; Pollack, L. *Phys. Rev. Lett.* **2003**, *90*, 188103.

(32) Qiu, X.; Andresen, K.; Kwok, L. W.; Lamb, J. S.; Park, H. Y.; Pollack, L. *Phys. Rev. Lett.* **2007**, *99*, 038104.

(33) Boschitsch, A. H.; Fenley, M. O. *J. Comput. Chem.* **2007**, *28*, 909–921.

(34) Fu, H.; Chen, H.; Koh, C. G.; Lim, C. T. *Eur. Phys. J. E: Soft Matter Biol. Phys.* **2009**, *29*, 45–49.

(35) Grilley, D.; Misra, V.; Caliskan, G.; Draper, D. E. *Biochemistry* **2007**, *46*, 10266–10278.

(36) Owczarzy, R.; Moreira, B. G.; You, Y.; Behlke, M. A.; Walder, J. A. *Biochemistry* **2008**, *47*, 5336–5353.

(37) Sharp, K. A.; Friedman, R. A.; Misra, V.; Hecht, J.; Honig, B. *Biopolymers* **1995**, *36*, 245–262.

(38) Outhwaite, C. W.; Lamperski, S. *Condens. Matter. Phys.* **2001**, *4*, 739–748.

(39) Bohinc, K.; Gimsa, J.; Kraljiglic, V.; Slivnik, T.; Iglic, A. *Bioelectrochemistry* **2005**, *67*, 91–99.

(40) Kralj-Iglič, V.; Iglič, A. *J. Phys. II* **1996**, *6*, 477–491.

(41) Tresset, G. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2008**, *78*, 061506.

CT1002785

JCTC Journal of Chemical Theory and Computation

Comparing the Efficiency of Biased and Unbiased Molecular Dynamics in Reconstructing the Free Energy Landscape of Met-Enkephalin

Ludovico Sutto, Marco D'Abramo, and Francesco Luigi Gervasio*

Spanish National Cancer Research Center (CNIO), Structural Biology and Biocomputing Programme, Melchor Fernandez Almagro, 3 E-28029 Madrid, Spain

Received July 23, 2010

Abstract: All-atom unbiased molecular dynamics simulations are now able to explore the microsecond to millisecond time scale for simple biological macromolecules in an explicit solvent. This allows for a careful comparison of the efficiency and accuracy of enhanced sampling methods versus long unbiased molecular dynamics in reconstructing conformational free energy surfaces. Here, we use an equilibrium microsecond-long molecular dynamics simulation as a reference to analyze the convergence properties of well-tempered metadynamics with two different sets of collective variables. In the case of the small and very diffusive Met-enkephalin pentapeptide, we find that the performance strongly depends on the choice of the collective variables (CVs). Using a set of principal component analysis derived eigenvectors, the convergence of the FES is faster than with both hand-picked CVs and unbiased molecular dynamics.

1. Introduction

Nowadays, thanks to specialized hardware[1] and grid computing,[2,3] it is increasingly possible to study relevant biological events including folding, molecular recognition, and conformational plasticity with long unbiased all-atom molecular dynamics (MD) simulations. Assuming ergodicity of the systems under study, these long runs can be used to calculate both kinetics and thermodynamics observables in simple systems as fast-folding proteins or small peptides. So far, due to the time-scale problem, these physical properties could only be obtained, with meaningful statistics, by enhanced sampling or coarse-grained methods.[4−20] This is still true for more complex systems having higher free-energy barriers and characteristic times of multiple milliseconds to seconds. Enhanced sampling methods are generally based on some assumptions on the underlying system, such as the choice of meaningful collective variables or the definition of an initial and final state. Thus, a comparison of the relative performance, e.g., in terms of the convergence and accuracy of the reconstructed free energy surfaces, obtained with the two approaches is now important and increasingly urgent.

The goal of the present paper is to compare well-tempered metadynamics,[17,21] a widely used enhanced sampling method, with a long unbiased MD in terms of the convergence of the reconstructed free energy surface (FES). We will look at the relative computational efficiency, and at the effect of the choice and number of the collective variables (CVs) on it. In particular, we will see how a general set of collective variables automatically generated leads to a fast and accurate reconstruction of the free energy surface.

Our model system of choice is the Met-enkephalin, a pentapeptide involved in regulating nociception in the body by binding to the opioid receptors. Due to its extreme flexibility, the native structure of Met-enkephalin has been elusive experimentally.[22,23] A variety of computational approaches have been used to explore its conformational landscape.[24−28] Our choice was guided both by its small size and by the nature of its conformational free-energy surface. The peptide is sufficiently small to allow for an extensive unbiased all-atom explicit solvent MD simulation. Still, its free-energy surface is complex and well-structured. What is more, the small height of the free energy barriers and the

* To whom correspondence should be addressed. E-mail: flgervasio@cnio.es.

10.1021/ct100413b
Published on Web 11/16/2010

diffusive behavior of its dynamics pose an additional challenge to free-energy-based methods like well-tempered metadynamics that generally perform better in the presence of high free energy barriers and ballistic dynamics.

2. Computational Methods

The simulations were performed using version 4 of the molecular dynamics program GROMACS[29] with the Amber03 force field.[30] The Met-enkephalin, with sequence YGGFM, is solvated in 1256 tip3p water molecules[31] enclosed in a cubic box of 38.9 nm^3 under periodic boundary conditions. The van der Waals interactions were cut off at 1.4 nm, and the long-range electrostatic interactions were calculated by the particle mesh Ewald algorithm[32] with a mesh spaced 0.12 nm. The neighbor list for the nonbonded interactions was updated every 0.02 ps. The system evolves in the canonical ensemble, coupled with a Nosé-Hoover thermal bath[33,34] at $T = 300$ K and a time step of 2 fs.

The solvated system was prepared using the following steps: (1) a steepest descent energy minimization; (2) equilibration of the system for 10 ps; (3) a density equilibration with a 2 ns dynamics at 300 K and constant pressure, coupling the system to a Parrinello−Rahman barostat;[35] and finally (4) another 2 ns dynamics at 300 K with a Berendsen thermostat[36] at constant volume to thermalize the system.

We used the well-tempered variant of the metadynamics enhanced sampling technique[37] in which the evolving bias $V(s,t)$ at time t and CV value s is built by the sum of Gaussian-shaped potentials with decreasing height. Such an algorithm, introduced by Barducci et al.[21] for the alanine dipeptide, has been proven more efficient than original metadynamics, guaranteeing the convergence of the free energy surface at a long time limit. According to the well-tempered metadynamics prescription, a Gaussian is deposited every 4 ps with height $W = W_0\, e^{-V(s,t)/fT}$, where $W_0 = 2$ kJ/mol is the initial height, T is the temperature of the simulation, and $f = 1.5$ is the bias factor. The width of the Gaussians is set to 0.02 in units of the respective CV and determines the resolution of the recovered free energy surface.

The collective variables biasing the 250 ns dynamics performed are either one or more atomic distances (d_i) or the projection of the heavy atom positions along one or more eigenvectors (v_i). Those eigenvectors are obtained applying the essential dynamics technique,[38] a principal-component-based analysis (PCA), to the 40 heavy atoms of the whole unbiased trajectory leading to a 120 × 120 covariance matrix. This symmetric matrix represents the correlation between atomic motions in Cartesian coordinate space and the eigenvectors obtained from the diagonalization of the collective motions. The eigenvectors corresponding to the largest eigenvalues contain the largest fluctuations and hence hold the most important motions of the system. The four largest eigenvalues of the covariance matrix describe 32.0%, 13.6%, 12.3%, and 10.0% of the fluctuations.

The same analysis repeated using only the first 210 ns or 21 ns (10% and 1% of the simulation time, respectively) of the unbiased trajectory showed similar results. In particular, the overlap between the subspace generated by the first two eigenvectors calculated after the first 21 ns and the reference subspace generated by the first two eigenvectors of the whole 2.1 μs trajectory is 0.72, where 1 is a complete overlap. The overlap increases up to 0.94 using the eigenvectors obtained after 210 ns. The overlap is calculated as the root-mean-square inner product of the first two PCA eigenvectors.[39]

To perform the metadynamics run, we modified the PLUMED[40] plug-in for GROMACS, introducing the eigenvectors as new collective variables. The reweighting procedure described by Bonomi et al.[41] was used to calculate the projection of the FES on other CVs than the ones used to bias the dynamics.

To quantitatively compare a metadynamics reconstructed FES $F_i(s)$ to the reference unbiased molecular dynamics FES $F_{ref}(s)$ expressed in terms of the CVs defined in a region Ω, we used two different parameters: (i) The distance measure introduced by Alonso and Echenique is used:[42] $d_A(F_i,F_{ref}) = [(\sigma_i^2 + \sigma_{ref}^2)(1 - r_{i,ref}^2)]^{1/2}$ where σ_x, with x denoting either i or ref, is the statistical variance of the free energy F_x defined by $\sigma_x = 1/N\int_\Omega (F_x(s) - \langle F_x \rangle)^2\, ds$ where $\langle F_x \rangle = 1/N\int_\Omega F_x(s)\, ds$ is the average value of F_x in the region Ω and $N = \int_\Omega ds$ is the normalization. The variances σ_x set the physical scale of the measure and confer the energy units to the distance. $r_{i,ref}$ is the Pearson correlation coefficient and measures the degree of correlation between the two energy surfaces. It is defined by $r_{i,ref} = \mathrm{cov}(F_i,F_{ref})/(\sigma_i\sigma_{ref})$ where $\mathrm{cov}(F_i,F_{ref})$ is the covariance between the two free energies. Globally, the d_A measure is convenient since it is expressed in energy units and can be directly compared to the thermal fluctuations. (ii) The Kullback−Leibler divergence is used:[43] $\mathrm{KLdiv}(F_i,F_{ref}) = \int_\Omega \exp(-F_{ref}(s)/k_BT)(F_i(s) - F_{ref}(s))/k_BT\, ds$ expressed in terms of the free energies rather than the probabilities P_i, making use of the relation $F_i = -k_BT \ln(P_i)$, where k_B is Boltzmann's constant and T is the temperature. It should be noted that it is dimensionless and not symmetric. Such a parameter gives an interesting and slightly different measure of the FES similarity, weighting the relevant regions more, i.e., the free energy minima and valleys of the reference surface, as compared to the high free energy ones.

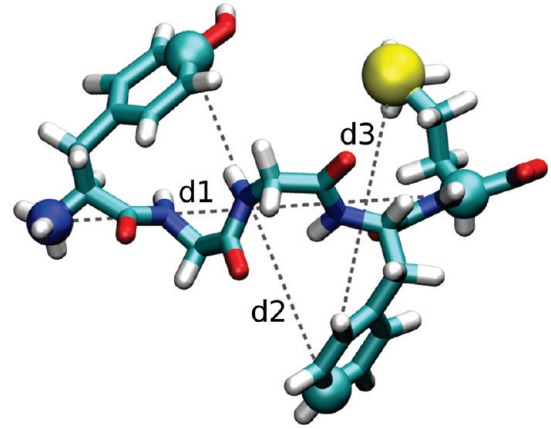


Figure 1. Met-enkephalin with the three distances used as structural observables: d_1 = dist(TYR1:N,MET5:CA), d_2 = dist(TYR1:CZ,PHE4:CZ), d_3 = dist(MET5:SD,PHE4:CZ).

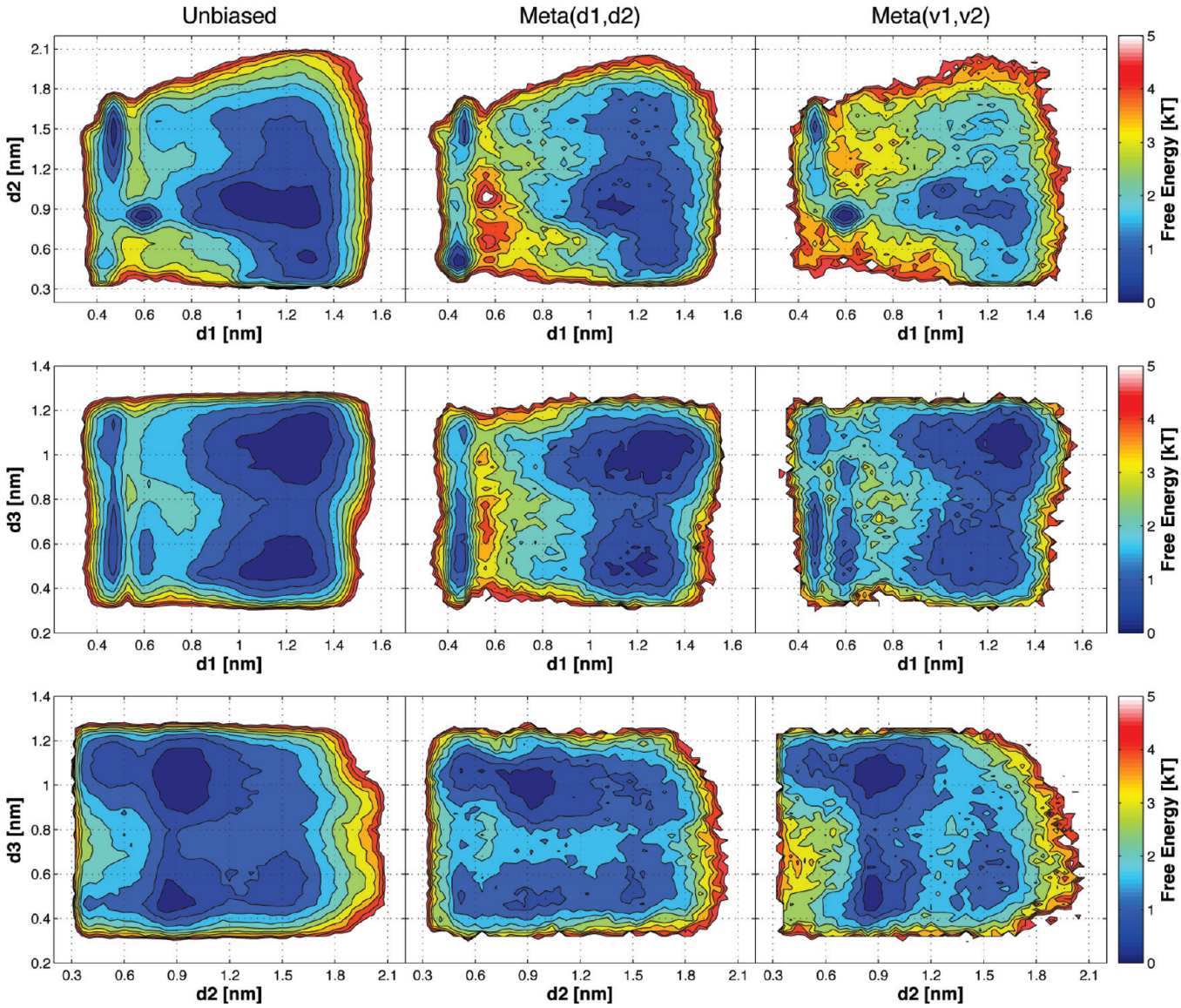


Figure 2. Free energy surfaces as a function of the distances d_1, d_2, and d_3 for the reference unbiased simulation and two metadynamics simulations using respectively (d_1, d_2) and (v_1, v_2) as biases. The contour lines are drawn every 0.5 k_BT.

3. Results and Discussion

The reference FES is obtained from a 2.1 μs unbiased simulation, whose convergence with respect to the FES was checked by block analysis (see Supporting Information). The metadynamics runs are performed using two different sets of collective variables, namely, the distances between atoms shown in Figure 1 and the projections along the first PCA eigenvectors (see Computational Methods). The former CVs have been chosen to intuitively describe the conformational space of the peptide backbone and side chains. Such a set was extracted from a visual inspection of the structures sampled in the unbiased trajectory. The latter CVs take advantage of the essential dynamics technique[38,44] and have been shown to be much more efficient than standard clustering in reducing data and reproducing salient features of protein folding.[45] They have also been used to compress MD trajectories[46] and, with standard metadynamics, to explore the free energy landscape of dialanine and SH3 domain.[47]

The two-dimensional FESs obtained as a function of (d_1, d_2), (d_1, d_3), and (d_2, d_3) are shown in Figure 2. The unbiased 2.1 μs reference simulation (Figure 2, first column) presents barriers smaller than 2 k_BT as expected for a diffusive system.

The same free energy projections are overall well reconstructed already after 100 ns in the two 2-CVs metadynamics runs (Figure 2, center and right columns). Nevertheless, some discrepancies emerge in a more detailed comparison. In particular, the reconstructed landscape in the case of the distance metadynamics (Figure 2, center column, first row) completely misses the minimum at (d_1,d_2) = (0.6,0.9), showing in its place a high barrier. On the contrary, the eigenvector metadynamics (Figure 2, right column) recovers within 0.5 k_BT all four minima, and the overall pattern of the valleys similarly agrees. This holds also for FES(d_1, d_3) and FES(d_2, d_3), in which the eigenvector metadynamics performs better than the metadynamics using d_1 and d_2 as CVs. In the case of the metadynamics using d_1, d_2, and d_3 as CVs, 100 ns is not enough to correctly reproduce the free

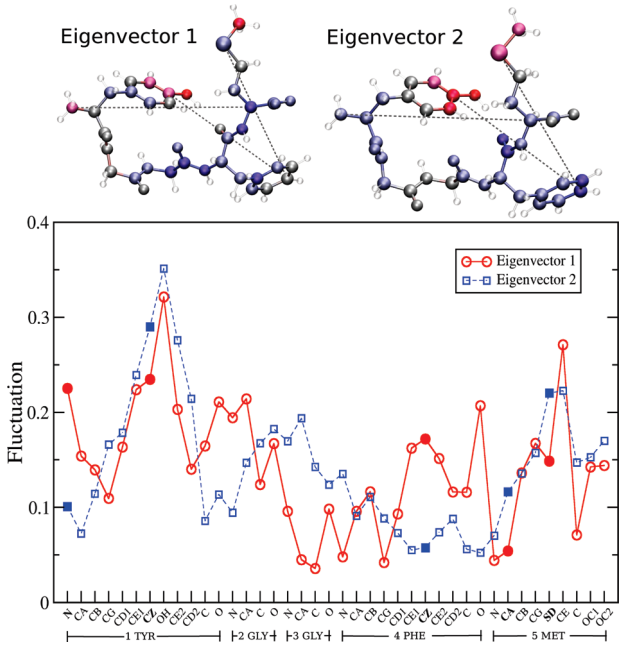


Figure 3. The 40 components of the first two PCA eigenvectors v_1 and v_2. Two identical balls-and-sticks Met-enkephalin conformations are drawn with the 40 heavy atoms colored by their average fluctuation on a scale from highest value (red) to lowest (blue) for both of the two principal eigenvectors. The distances d_1, d_2, and d_3 are depicted as dashed lines between the atoms. In the graph, the same information is represented quantitatively. The five filled symbols represent the atoms involved in either the d_1, d_2, or d_3 distance definition.

energies FES(d_1, d_3) and FES(d_2, d_3) (see the Supporting Information), while the simulation with v_1, v_2, and v_3 performs as well as the metadynamics with v_1 and v_2 (see the Supporting Information and Figure 2).

The reason underlying the worse performance of hand-picked distance CVs over the projection over the principal components can be better understood checking the fluctuation content of each atom. In Figure 3 are shown the components of the first two eigenvectors v_1 and v_2 for each one of the 40 heavy atoms. The five atoms involved in the three chosen distances link highly fluctuating regions even though they are not the most fluctuating ones. Clearly, any set of a few atoms cannot fully describe the conformational motions since their distances only take into account a limited number of degrees of freedom. This suggests that the choice of the eigenvectors as natural CVs has two main advantages. First, the eigenvectors intrinsically describe the largest conformational changes and can be easily obtained from a preliminary unbiased simulation. Second, such a choice avoids the frustrating task of the quest of the most efficient CVs.

Hence, in the cases where the interest is to focus on slow collective motions, like in protein conformational changes, or where the choice of the CV is not trivial, the use of eigenvectors provides both a simple and physically meaningful alternative. It is worth it to mention that the set of PCA eigenvectors always represents a complete and orthogonal basis, avoiding *a priori* a linear correlation among CVs which would lead to inefficient sampling. The number of eigenvectors used as CVs can be increased at will, eventually reaching a complete representation of the degrees of freedom of the system. This is mathematically exact when the number of eigenvectors equals the number of degrees of freedom. However, in practice, a very accurate representation can be reached with a much smaller number of eigenvectors, permitting a dramatic reduction in the size of the CV space.[45,46]

The ability of these "natural" CVs to exhaustively explore the conformational space is also reflected by the accuracy of the reconstructed FES projected along two different and global observables as the root-mean-square deviation from a reference structure (rmsd) and the radius of gyration (rgyr) shown in Figure 4. In fact, the position of the minimum and the overall topology of the free energy landscape of the metadynamics run completely agree with the reference unbiased simulation. Moreover, if we compare the time convergence of the metadynamics FES to the reference FES with respect to the unbiased MD, we observe how these CVs allow the system to explore larger regions of conformational space in less time. In fact, although a fair convergence is achieved within 100 ns with all of the tested CVs (distances and eigenvectors), the eigenvector metadynamics provides a FES that closely resembles the reference FES as early as after 50 ns (see Figure 4). On the other hand, unbiased simulation after neither 50 nor 100 ns explored the region at (rmsd, rgyr) $\approx$ (0.31, 0.40−0.43).

To compare our sampled conformations to the literature values, we extracted the representative geometries for each of the three shallow minima present in the FES(v1,v2) shown in Figure 5. Even though we used different observables than Sanbonmatsu and García,[27] since their PCA eigenvectors were obtained only from the five C-alpha motions, some common features are apparent. In particular, the minimum (A) corresponds to the same U-shaped conformation found by Sanbonmatsu and García[27] and by Henin et al.,[28] in which the phenylalanine and tyrosine side chains are packed. In the other two minima, the backbone is more elongated, and the system explores both extended (B) and helix-like (C) conformations, also in agreement with the aforementioned works. Interestingly, these minima are connected in the essential space, indicating the absence of relevant barriers in agreement with the high flexibility of this peptide.

Finally, in order to assess the convergence of the free energy as a function of the number of collective variables, we used two different parameters: (i) the correlation coefficient introduced by Alonso and Echenique,[42] which allows quantitative measurement of the similarity between different energy potentials, and (ii) the Kullback−Leibler divergence,[43] which weights the free energy minima more with respect to poorly sampled regions (see Computational Methods). These two coefficients are used as different measures of the distance between the reconstructed FES provided by the metadynamics runs and the reference FES(rmsd,rgyr) of the unbiased MD simulation.

In Figure 6, we report these coefficients for both the eigenvector metadynamics and distance metadynamics compared to the unbiased MD run as a function of time. Both coefficients and both sets of CVs give a similar result and show that the metadynamics runs converge faster than the unbiased run. The metadynamics with two and three

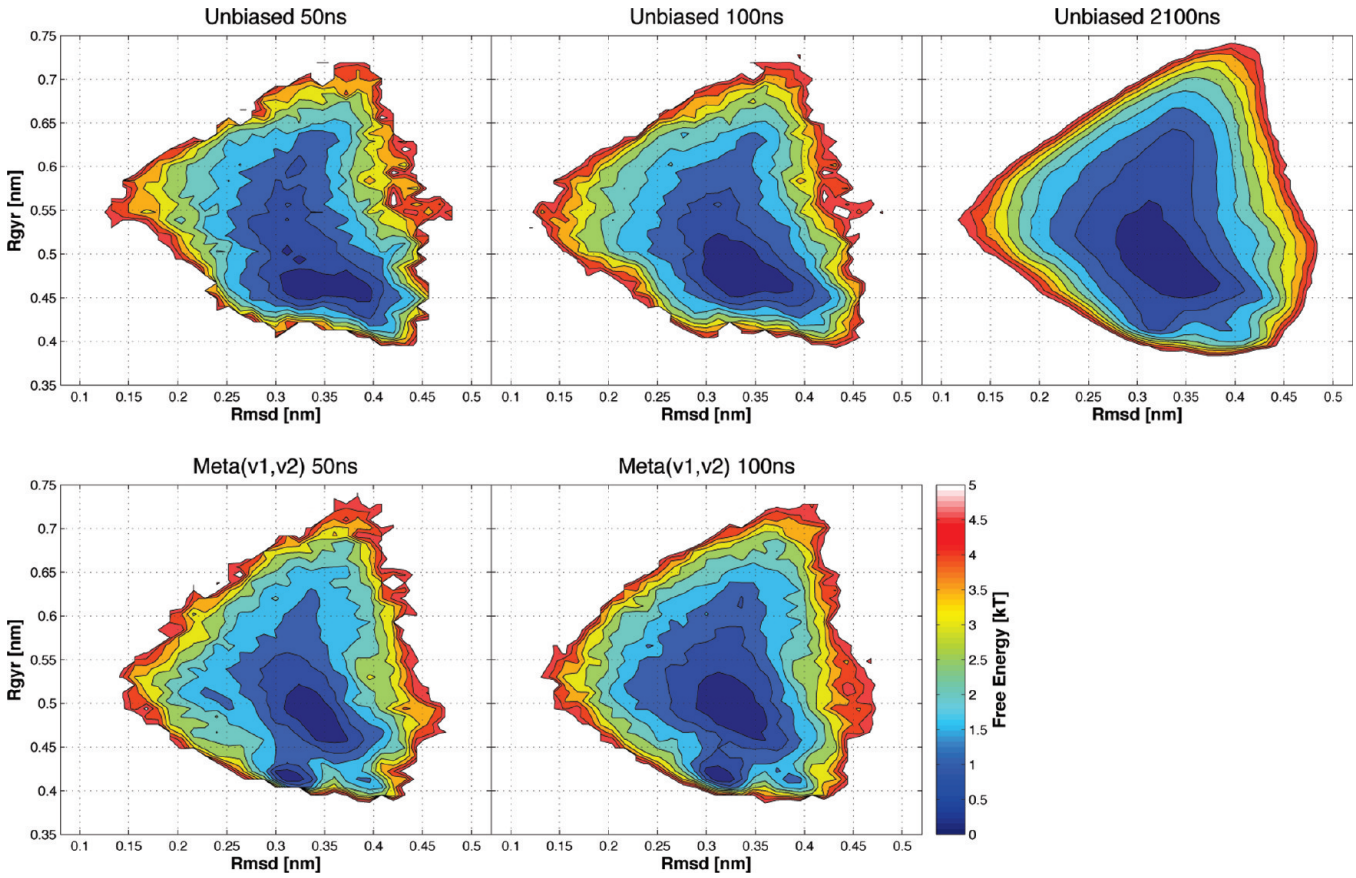


Figure 4. Free energy surfaces for different times as a function of the root-mean-square deviation (rmsd) and the gyration radius (rgyr) for the reference unbiased simulation and the metadynamics simulation using (v_1, v_2) as a bias. The contour lines are drawn every 0.5 $k_B T$.

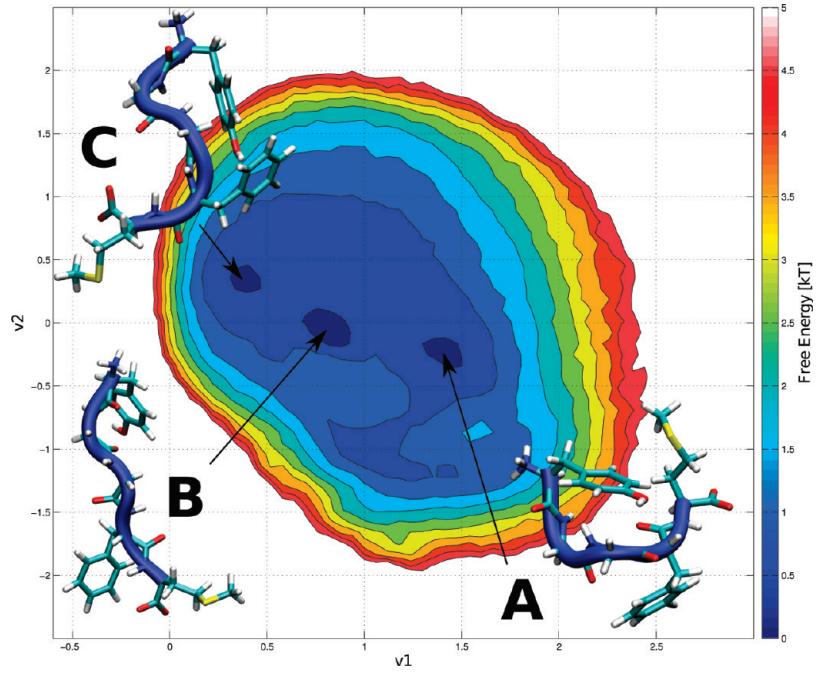


Figure 5. Free energy surface as a function of the projection over the two principal eigenvectors v1 and v2 with contour lines every 0.5 $k_B T$. The representative conformations of the three minima are also shown as insets and correspond to (A) a U-shaped backbone with packed aromatic rings, (B) an elongated conformation, and (C) a helix-like structure.

eigenvectors reach, and stay, below a 1 $k_B T$ reference threshold before any other run. As for the convergence as a function of the number of CVs, we expect that less vectors (or distances) can miss important degrees of freedom, while too many CVs slow down the metadynamics filling time. This is reflected in the longer time

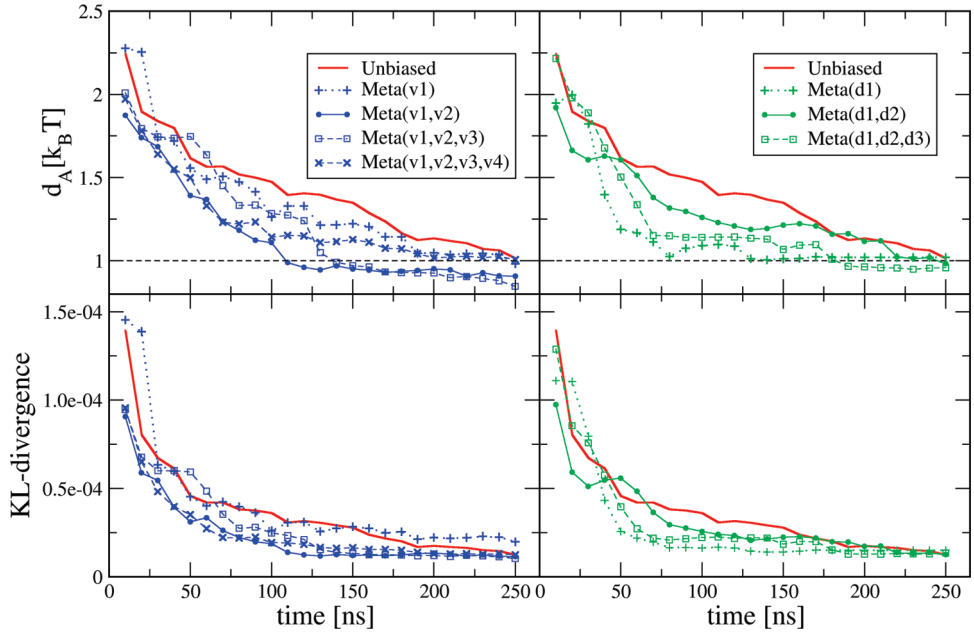


Figure 6. Comparison of the convergence of the free energy two-dimensional surface FES(rmsd,rgyr) for the eigenvector (blue, left-hand-side panels) and distance (green, right-hand-side panels) metadynamics simulations as a function of time. In red is shown the unbiased simulation. The similarities are calculated using as reference the 2.1 μs unbiased simulation FES shown in Figure 4. In the upper panels, the metric d_A is the energy-function distance introduced by Alonso and Echenique[42] and has units of k_BT. A dashed line at 1 k_BT defines the goal accuracy. In the lower panels, the similarity is computed using the Kullback−Leibler divergence.[43]

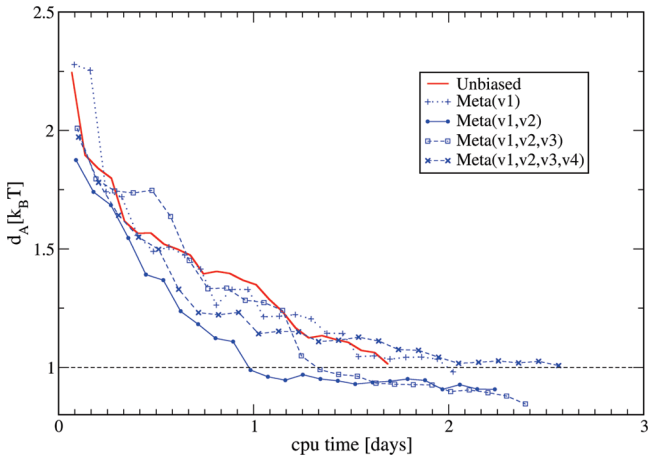


Figure 7. Comparison of the computer simulation time as a function of the distance d_A of the reconstructed FES(rmsd,rgyr) to the reference one, for the unbiased (red) and the eigenvector metadynamics simulations (blue). The computer time refers to a run that uses 24 CPUs on a parallel cluster. The points are plotted every 10 ns of simulation, for a total of 250 ns, time at which all of the simulations have reached the 1 k_BT accuracy threshold represented by the dashed line.

needed by the metadynamics with one and four eigenvectors to reach the goal accuracy.

Clearly, two PCA vectors strike the right balance and converge the fastest, as seen in Figure 7, where we show the accuracy, in terms of the distance d_A to the reference FES, as a function of the actual computer time needed to reach it. After 110 ns, the equivalent of about one day of 24 parallel CPU usage, the metadynamics with v_1 and v_2 reconstructs the FES(rmsd,rgyr) with an accuracy below the 1 k_BT threshold. The same accuracy is reached after 140 ns, or 32 CPU h for the metadynamics with three eigenvectors. The metadynamics with only v_1 and with the four eigenvectors eventually reach the goal after 250 ns, performing worse than the unbiased run.

4. Conclusions

In this paper, we compared the accuracy and computational efficiency of long unbiased MD simulations and well-tempered metadynamics in reconstructing the conformational free energy landscape of a peptide. Given the nature of the test case shown, which has low free energy barriers and diffusive dynamics, the advantage of free energy methods with respect to unbiased MD should be greatly reduced. Nevertheless, we found that, with a rationally built set of collective variables, well-tempered metadynamics is able to reconstruct an accurate free energy surface quicker than unbiased MD. These CVs, i.e., the eigenvectors describing the principal directions of motion, constitute a natural set which can be easily obtained from a preliminary unbiased MD simulation. On the contrary, simple, yet hand-picked, collective variables, as a set of chosen distances, turn out to miss some relevant features of the free energy surface. These results confirm the importance of the choice of the CVs but also provide a simple approach to automatically define a set of CVs that can be applied to reconstruct the FES of more complex systems, where higher free energy barriers make the convergence of unbiased MD simulation problematic.

Acknowledgment. The authors acknowledge S. Piana for useful discussions and advice and the Barcelona Supercomputing Center for a generous allocation of computer resources.

Supporting Information Available: Figures of free energy surface, a table presenting block analysis of the distance d_A and KL-divergence calculated for the unbiased trajectory, and a figure presenting a comparison of the convergence of the free energy two-dimensional surface FES(rmsd,rgyr) for the eigenvector and distance metadynamics simulations as a function of time. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120–127.

(2) Shirts, M.; Pande, V. S. *Science* **2000**, *290*, 1903–1904.

(3) Buch, I.; Harvey, M. J.; Giorgino, T.; Anderson, D. P.; Fabritiis, G. D. *J. Chem. Inf. Model.* **2010**, *50*, 397–403.

(4) Patey, G. N.; Valleau, J. P. *J. Chem. Phys.* **1975**, *63*, 2334–2339.

(5) Grubmüller, H. *Phys. Rev. E* **1995**, *52*, 2893–2906.

(6) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1339–1350.

(7) Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690.

(8) Darve, E.; Pohorille, A. *J. Chem. Phys.* **2001**, *115*, 9169–9183.

(9) Gullingsrud, J.; Braun, R.; Schulten, K. *J. Comput. Phys.* **1999**, *151*, 190–211.

(10) Rosso, L.; Minary, P.; Zhu, Z. W.; Tuckerman, M. E. *J. Chem. Phys.* **2002**, *116*, 4389–4402.

(11) Elber, R.; Karplus, M. *Chem. Phys. Lett.* **1987**, *139*, 375–380.

(12) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. *J. Chem. Phys.* **2006**, *125*, 024106.

(13) Dellago, C.; Bolhuis, P.; Csajka, F. S.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 1964–1977.

(14) Juraszek, J.; Bolhuis, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 15859–15864.

(15) Merlitz, H.; Wenzel, W. *Chem. Phys. Lett.* **2002**, *362*, 271–277.

(16) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(17) Laio, A.; Gervasio, F. L. *Rep. Prog. Phys.* **2008**, *71*, 126601.

(18) Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc.* **2006**, *128*, 13435–13441.

(19) Camilloni, C.; Sutto, L. *J. Chem. Phys.* **2009**, *131*, 245105.

(20) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–50.

(21) Barducci, A.; Bussi, G.; Parrinello, M. *Phys. Rev. Lett.* **2008**, *100*, 020603.

(22) Graham, W. H.; Carter, E. S. I.; Hicks, R. P. *Biopolymers* **1992**, 1755–1764.

(23) D'Alagni, M.; Delfini, M.; Di Nola, A.; Eisenberg, M.; Paci, M.; Roda, L. G.; Veglia, G. *Eur. J. Biochem.* **1996**, *240*, 540–549.

(24) Shen, M.; Freed, K. *Biophys. J.* **2002**, *82*, 1791–1808.

(25) Bartels, C.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 865–880.

(26) Li, Z. Q.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6611–6615.

(27) Sanbonmatsu, K. Y.; García, A. E. *Proteins* **2002**, *46*, 225–34.

(28) Henin, J.; Fiorin, G.; Chipot, C.; Klein, M. L. *J. Chem. Theory Comput.* **2010**, *6*, 35–47.

(29) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(30) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(31) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910–8922.

(32) Essman, U.; Perela, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8592.

(33) Nosé, S. *Mol. Phys.* **1984**, *52*, 255–268.

(34) Hoover, W. G. *Phys. Rev. A.* **1985**, *31*, 1695–1697.

(35) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(36) Berendsen, H. J. C.; Postma, J. P. M.; Di Nola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

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

(38) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. *Proteins: Struct. Funct. Genet.* **1993**, *17*, 412–425.

(39) Amadei, A.; Ceruso, A.; Di Nola, A. M. *Proteins: Struct. Funct. Genet.* **1999**, *36*, 419–424.

(40) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A. *Comput. Phys. Commun.* **2009**, *180*, 1961–1972.

(41) Bonomi, M.; Barducci, A.; Parrinello, M. *J. Comput. Chem.* **2009**, 1615–1621.

(42) Alonso, J. L.; Echenique, P. A. *J. Comput. Chem.* **2006**, *27*, 238–252.

(43) Kullback, S.; Leibler, R. *Ann. Math. Stat.* **1951**, *22*, 79–86.

(44) Amadei, A.; Linssen, A. B. M.; de Groot, B. L.; van Aalten, D. M. F.; Berendsen, H. J. C. *J. Biomol. Struct. Dyn.* **1996**, *13*, 615–625.

(45) Rajan, A.; Freddolino, P. L.; Schulten, K. *Plos One* **2010**, *5*, e9890.

(46) Meyer, T.; Ferrer-Costa, C.; Perez, A.; Rueda, M.; Bidon-Chanal, A.; Luque, F.; Laughton, C.; Orozco, M. *J. Chem. Theory Comput.* **2006**, *2*, 251–258.

(47) Spiwok, V.; Lipovová, P.; Králová, B. *J. Phys. Chem. B* **2007**, *111*, 3073–3076.

CT100413B

JCTC Journal of Chemical Theory and Computation

Continuous Localized Orbital Corrections to Density Functional Theory: B3LYP-CLOC

Michelle Lynn Hall, Jing Zhang, Arteum D. Bochevarov, and Richard A. Friesner*

Department of Chemistry, Columbia University, 3000 Broadway, New York, New York 10027, United States

Received July 26, 2010

Abstract: Our previous works have demonstrated the ability of our localized orbital correction (LOC) methodology to greatly improve the accuracy of various thermochemical properties at the stationary points of the density functional theory (DFT) reaction coordinate (RC). Herein, we extend this methodology from stationary points to the entire RC connecting any stationary points by developing continuous localized orbital corrections (CLOCs). We show that the resultant method, DFT-CLOC, is capable of producing RCs with far greater accuracy than uncorrected DFT and yet requires negligible computational cost beyond the uncorrected DFT calculations. Various post-Hartree−Fock (post-HF) reaction coordinate profiles were used, including a sigmatropic shift, Diels−Alder reaction, electrocyclization, carbon radical, and three hydrogen radical reactions to show that this method is robust across multiple reaction types of general interest.

I. Introduction

Density functional theory (DFT)[1] has proven a very useful theoretical tool for computing atomic and molecular electronic structures. In comparison to post-Hartree−Fock methods, DFT methods are capable of calculating relatively large systems and transition-metal-containing systems and, therefore, are widely used in quantum chemistry and condensed matter physics. The accuracy of DFT methods is essentially dependent on the density functional used, which is always an approximation of the hypothetical exact density functional. During past decades, many attempts to construct a more accurate functional have been undertaken, starting from either first principles or empirical fitting, or both.[2] However, the approximate nature of extant density functionals inevitably weakens the robustness of DFT performance in predicting, in particular, thermodynamic properties.[3]

Two distinct methods can be envisioned to tackle some of the problems that still continue to plague DFT: (a) Many researchers have had remarkable success by developing wholly new density functionals.[4] (b) Alternatively, one can envision creating a new functional by simply taking an existing functional and adding terms on top of it. These terms can be used to target systematic errors endemic to each functional.

In previous publications, we have shown that the accuracy of DFT can be greatly improved for various thermochemical properties with the use of localized orbital corrections, or LOCs.[5] These LOCs have been developed to treat stationary points (i.e., reactants, products, and transition states) and are based on a chemically intuitive dissection of each stationary point's electronic structure into valence bond terms. Further, because the LOCs are applied *a posteriori* using a simple noniterative computational algorithm, they require negligible computational cost beyond standard DFT calculations. With the application of LOCs, atomization energies, ionization potentials, electron affinities, enthalpies of reaction, and barrier heights can all be obtained with very good accuracy for stationary points.

In this work, we extend our methodology beyond the treatment of discrete stationary points with the goal of providing energetics for the entire reaction coordinate (RC) of a chemical reaction. As we already have developed LOCs to treat the reactant, transition state, and product for an arbitrary reaction, an obvious next step is to interpolate these LOCs for all intermediate points and develop what we shall refer to as continuous localized orbital corrections, CLOCs.

* rich@chem.columbia.edu.

10.1021/ct100418n
Published on Web 11/08/2010

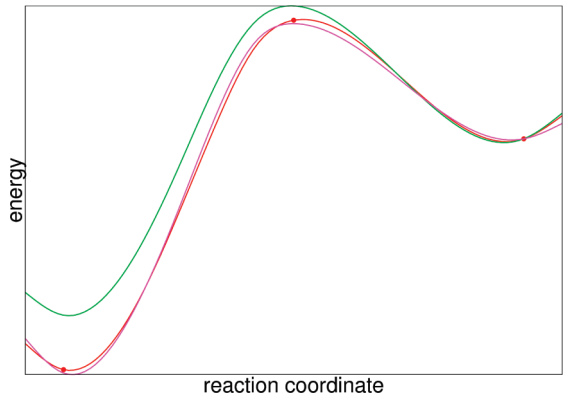


Figure 1. Reaction coordinate for an arbitrary reaction, where B3LYP (green), B3LYP-LOC (red dots), and B3LYP-CLOC (red line) are compared to an accurate benchmark (pink).

This is depicted schematically in Figure 1, where the B3LYP-LOC stationary point energies are shown (red points), connected with the B3LYP-CLOC energy curve (also red). While previous publications defined B3LYP-LOC energies only at the stationary points (red points), B3LYP-CLOC energies are defined all along the reaction coordinate (red curve). The latter is the subject of the present work.

A naming convention implicit from the previous discussion is that we use "LOC" to describe the discrete corrections, those at the stationary points, exclusively. "CLOC" is used to describe the continuous corrections, for all points that are not stationary points. This is represented schematically in Figure 1, where the B3LYP-LOC energies are defined at the stationary points (red dots), whereas the B3LYP-CLOC energies are defined all along the reaction coordinate (red line).

In order for our model to be consistent, any CLOC computed at a stationary point must agree with the LOC for that same stationary point. Importantly, it is not possible to compute a LOC for a point other than a stationary point because LOC parameters have been developed for the stationary points only. To treat points other than stationary points, interpolation is necessary in a continuous fashion, hence the necessity of the CLOC method.

It should be emphasized that LOCs are simple numerical corrections that should improve the accuracy of the DFT-predicted electronic energy. LOCs are added to the DFT energy *a posteriori* and therefore cannot be used to improve DFT-predicted geometries of molecular systems. In order to perform geometry optimizations, and hence have our corrections affect the geometry (at least in theory) and not just the energy, we must be able to calculate gradients of these corrections. In order to be able to perform geometry optimizations, with CLOCs having an effect on the changes in geometry, their first derivatives with respect to nuclear displacements must be computed. This necessitates the extension of these discrete LOCs into a continuous form that connects the various stationary points. The innovations described in this work extend LOCs to not only treat nonstationary points but also to contribute to the optimization of molecular geometries.

Herein, we present the results of our CLOC development and show that it can be applied to seven reaction coordinate profiles with greatly improved results compared to uncorrected B3LYP. We also provide comparison with the M06-2X[6] functional, which has a substantially improved performance for reaction energetics as compared to B3LYP. Note that while we have examined points along the reaction profile exclusively in this paper, this is not a necessary condition for the application of CLOCs. Points off the reaction coordinate can also be treated, as described in more detail in section III.A.

In this publication, as in others, we have focused our efforts on corrections to the well-established B3LYP functional. At the same time, we have previously tested our LOC methodology in combination with other important functionals including the M05-2X and M06-2X[6] functionals of Truhlar and co-workers.[5c] We find that no functional tested to date combines with the LOC method as favorably as B3LYP. Nevertheless, it is still possible for other functionals not yet tested to produce more accurate results in combination with LOCs than B3LYP-LOC itself.

II. Overview of the B3LYP-LOC Methodology

The B3LYP-LOC model has been successfully employed to reduce errors endemic to DFT across a wide range of thermodynamic properties including atomization energies,[5a] ionization potentials and electron affinities,[5b] enthalpies of reaction,[5c,e] and barrier heights,[5e] as shown in Table 1 and Figure 2 below.

We have previously asserted that this impressive reduction in error upon application of LOCs is not fortuitous but rather reflects the systematic nature of the errors that are intrinsic to DFT in general and specialized to the specific errors characteristic of B3LYP. The LOCs dramatically reduce such errors by assigning fitting parameters dependent upon the

Table 1. Performance of B3LYP vs B3LYP-LOC for Various Chemical Properties

	mean unsigned error (MUE) (kcal/mol)		number of parameters (LOCs) employed[a]	size of data set	ref
	B3LYP	B3LYP-LOC			
atomization energies	4.8	0.8	22 (22)	222	5a
ionization potentials and electron affinities	3.2	0.9	45 (23)	134	5b
enthalpy of reactions	4.9	0.9	28 (0)	139	5c, 5e
barrier heights	3.2	1.2	36 (8)	105	5e

[a] The number of new parameters developed in each work is shown in parentheses. Specifically, the same 22 parameters developed initially for atomization energies are used for all other calculations: ionization potentials and electron affinities, enthalpies of reactions, and barrier heights. An additional 23, 0, and 8 parameters are developed specifically for these calculations, respectively. Some of the 23 parameters developed uniquely for ionization potentials and electron affinities were applied to enthalpies of reaction and barrier heights of ionic reactions.

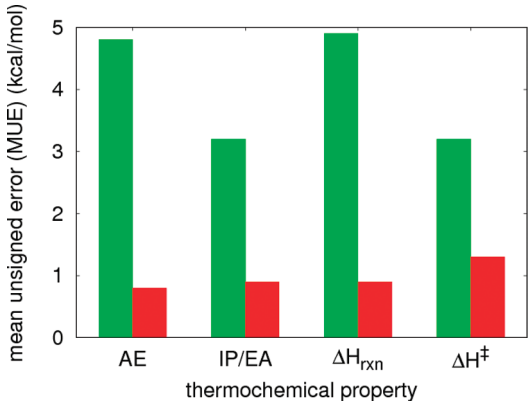


Figure 2. Performance of B3LYP (green) vs B3LYP-LOC (red) for various thermochemical properties including atomization energies (AE), ionization potentials and electron affinities (IP/EA), enthalpies of reaction (ΔH_{rxn}), and barrier heights ($\Delta H^{\ddagger}$). Data shown are taken from publications referenced in Table 1.

local environment of an electron pair or single electron. These parameters are then assumed to be transferable across molecular species. As is discussed in detail in ref 5, the dominant error in B3LYP can be identified as a difficulty in accurately modeling variations in nondynamical electron correlation across different types of chemical bonds, lone pairs, hybridization states, chemical environments, and singly vs doubly occupied orbitals.[7] The LOCs yield a more accurate representation of this variation as a function of local chemical environment.

Assigning LOCs to a particular molecular system is often straightforward. On the basis of the atomic coordinates of the molecule, a valence bond structure can be proposed. Some of the characteristics of valence bond structures have been identified as contributing to DFT's systematic errors. Accordingly, each of these particular characteristics is assigned a LOC to mitigate its error. In previous publications, all LOC values, C_k, were determined using linear regression such that they minimize the deviation between B3LYP and the reference value for many different thermochemical properties computed with several large data sets, as summarized in Table 1.[5] (A complete list of all of the LOCs and their values, C_k, is provided in the Supporting Information.) The total LOC(x) for any system x is then given simply by the sum of all individual LOCs' optimally determined values, C_k, multiplied by their number of occurrences, N_k, i.e., the number of times the particular valence bond characteristic associated with them is counted.

$$\mathrm{LOC}(x) = \sum_k N_k C_k \qquad (1)$$

These LOCs are then used in a straightforward manner to correct the enthalpy of reaction, for example. The expression for B3LYP reaction enthalpy is the difference in the enthalpies of products and reactants:

$$\Delta H_{rxn}^{B3LYP} = \sum_{products} \Delta H_{products}^{B3LYP} - \sum_{reactants} \Delta H_{reactants}^{B3LYP} \qquad (2)$$

Analogously, B3LYP-LOC reaction enthalpy is given by the LOC-corrected enthalpy differences in products and reactants:

$$\Delta H_{rxn}^{B3LYP-LOC} = \sum_{products} \Delta H_{products}^{B3LYP-LOC} - \sum_{reactants} \Delta H_{reactants}^{B3LYP-LOC} \qquad (3)$$

or equivalently,

$$\Delta H_{rxn}^{B3LYP-LOC} = \Delta H_{rxn}^{B3LYP} + \Delta \mathrm{LOC}_{rxn} \qquad (4)$$

where $\Delta \mathrm{LOC}_{rxn}$ is defined as the difference between LOC(products) and LOC(reactants). For example, consider the reaction in Scheme 1. Each species involved is assigned characteristics summarized in Table 2.

The B3LYP-LOC reaction enthalpy may be written as

$$\Delta H_{rxn}^{B3LYP-LOC}(CH_3 + CH_2{=}CH_2 \rightarrow CH_3{-}CH_2{-}CH_2) = \Delta H_{rxn}^{B3LYP}(CH_3 + CH_2{=}CH_2 \rightarrow CH_3{-}CH_2{-}CH_2) + \mathrm{LOC}(CH_3{-}CH_2{-}CH_2) - \mathrm{LOC}(CH_3) - \mathrm{LOC}(CH_2{=}CH_2) \qquad (5)$$

where the LOC(x) terms on the right-hand side of this equation are those given in the last row of Table 2.

Similar formulas can be derived for atomization energies, ionization potentials, electron affinities, and barrier heights, although the last of these involves treating a rather more complex situation. This most recent work[5e] forms the basis for the method described here. Specifically, the accuracy of B3LYP's barrier height prediction was improved with simple numerical corrections to the reactant, product, and transition state energies.[5e] The success of this effort suggests that we can develop a robust description of a potential energy surface by interpolating these corrections between the various stationary points (reactant, transition state, and product) to arrive at corrections for points intermediate between stationary points. This is described in detail in the section that follows.

III. Development of Continuous Localized Orbital Corrections

III.A. An Overview of CLOCs. In this section and throughout the rest of the text, various new terms will be introduced. Therefore, we have defined these terms for convenience in addition to others that will be defined later, in Table 3. In order to develop corrections for the entire B3LYP reaction coordinate profile, it is necessary to first evaluate the accuracy of B3LYP with respect to high-level post-HF benchmarks along the entirety of the reaction coordinate. While a fairly large amount of benchmark data exists for thermochemical properties such as enthalpies of reaction and barrier heights in the literature (wherein only stationary points are required), there is a relative paucity of

Scheme 1. Reaction between Methyl Radical and Ethene to Give Propyl Radical

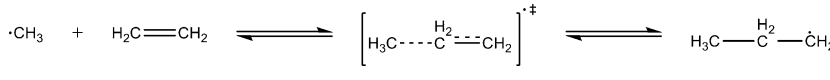


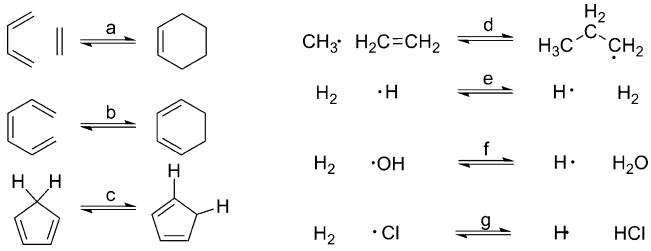
Table 2. Valence Bond Characteristics and Corresponding LOCs for the Reaction of Scheme 1

valence bond characteristic	LOC(x)	C_k[kcal/mol]	N_k methyl	ethene	propyl	transition state
sp^2 carbons	n/a	n/a	1	2	1	
$sp^{2.5}$ carbons	n/a	n/a				2
sp^3 carbons	n/a	n/a			2	1
C—H	$LOC(x)_{NPOLH}$	0.25	3	4	7	7
C—H attached to the radical-containing C	$LOC(x)_{RH}$	0.54	3		2	5
C---C	$LOC(x)_{MSBC/LSBC\_0.5}$	−2.05				1
C—C	$LOC(x)_{MSBC}$	−1.90			2	
C⋯C	$LOC(x)_{AA\_1.5}$	−0.88				1
C═C	$LOC(x)_{DBC}$	−1.00		1		
C—C adjacent to another C—C	$LOC(x)_{ESBC}$	−0.50			2	1/2
total LOC(x) [kcal/mol]			2.37	0.00	−1.97	1.27

Table 3. New Terms and Definitions

term	definition
active atom	an atom that belongs to at least one active bond
active bond	a bond with an order that changes throughout the reaction coordinate, for example, from a singe bond in the reactant to a double bond in the product
cutoff	the distance at which a bond is considered to have zero bond order according to the distances given in Table 4
inactive atom	an atom that belongs to no active bonds
inactive bond	a bond with an order which does not change throughout the reaction coordinate
product-side bond	description given to any bond which is intermediate in length between the transition state and product bonds
product-side structure	description given to any structure where the majority of the bonds are assigned as "product-side bonds"
reactant-side bond	description given to any bond which is intermediate in length between the transition state and reactant lengths
reactant-side structure	description given to any structure where the majority of the bonds are assigned as "reactant-side bonds"

Scheme 2. Reactions Employed in This Study[a]



[a] (a) Diels−Alder, (b) electrocyclic, (c) sigmatropic shift, (d) carbon radical, and (e−g) hydrogen transfer.

published data on complete reaction coordinates for systems larger than a few atoms. To address this problem, we produced our own curves computed with coupled cluster with single, double, and iterative triple excitations [CCSD(T)] for reactions a−f in Scheme 2. Specifically, single-point calculations along the reaction path were performed at the RCCSD-(T)/cc-pVTZ//B3LYP/6-31+G** level. CCSD(T) energies and B3LYP geometries were obtained with MolPro 2006.1[8] and Jaguar 7.6,[9] respectively. The data for reaction g were available in the literature.[10] The reactions were chosen from our latest DFT-LOC publication[5e] and represent a broad range of chemistries: (a) cycloaddition, (b) electrocyclization, (c) sigmatropic shift, (d) carbon radical, and (e−g) hydrogen radical reactions. Although hardly exhaustive, we argue that this set represents an acceptable starting point sufficient to evaluate the accuracy of our method as it applies to systems of general interest.

A subsequent examination of the B3LYP and post-HF profiles for each reaction gives a qualitative picture of what functional form the CLOC corrections should take. Specifically, the ideal CLOC is one that minimizes the error along the B3LYP-CLOC reaction profile in comparison with the post-HF profile for each arbitrary point x and hence is given by the following equation:

$$E_{\text{post-HF}}(x) = E_{\text{B3LYP-CLOC}}(x) = E_{\text{B3LYP}}(x) + \text{CLOC}(x) \tag{6}$$

When x is a stationary point, LOC(x) necessarily equals CLOC(x). Note that this equality only holds for stationary points, as LOC(x) is undefined for nonstationary point structures.

$$E_{\text{B3LYP-CLOC}}(x) = E_{\text{B3LYP-LOC}}(x) = E_{\text{B3LYP}}(x) + \text{CLOC}(x) = E_{\text{B3LYP}}(x) + \text{LOC}(x) \tag{7}$$

if and only if x is a stationary point. However, everywhere where x is not a stationary point, one must define CLOC(x).

A simple examination of the extant LOCs shows that they may be divided into those aimed at treating bonds, hybridization, radicals, hypervalency, environment, and charge transfer of any system x.

$$\text{LOC}(x) = \text{LOC}(x)_{\text{bond}} + \text{LOC}(x)_{\text{hyb}} + \text{LOC}(x)_{\text{radical}} + \text{LOC}(x)_{\text{hyperval}} + \text{LOC}(x)_{\text{environ}} + \text{LOC}(x)_{\text{CT}} \tag{8}$$

The continuous implementation necessarily takes the same form:

$$\text{CLOC}(x) = \text{CLOC}(x)_{\text{bond}} + \text{CLOC}(x)_{\text{hyb}} + \text{CLOC}(x)_{\text{radical}} + \text{CLOC}(x)_{\text{hyperval}} + \text{CLOC}(x)_{\text{environ}} + \text{CLOC}(x)_{\text{CT}} \tag{9}$$

Each term of eq 9 will be discussed in its own subsection directly following this one.

To begin the calculation of CLOC(x) for an arbitrary molecule x, we first require the availability of all relevant stationary points (reactant, transition state, and product) for reference. Specifically, the Cartesian coordinates of all of these structures obtained with the same level of theory as the arbitrary point (here, B3LYP/6-31+G**) must be provided.

In order to perform the interpolation between the stationary points, we must know where along the reaction coordinate profile the arbitrary structure lies with respect to the input structures. To this end, the arbitrary structure is analyzed against these input structures to determine whether it is reactant-side or product-side (i.e., whether it lies along the reaction coordinate connecting reactant to transition state, or transition state to product, respectively). Because of this, the quality of the user-provided stationary points is critical. Each bond of the structure x is analyzed individually with respect to its bond length l_x and receives its own assignment: either reactant- or product-side. A reactant-side bond is intermediate in length between the reactant and transition state lengths, i.e., $l_r \leq l_x < l_{ts}$ or $l_r \geq l_x > l_{ts}$. Similarly, a product-side bond is intermediate in length between the product and transition state lengths, i.e., $l_p \leq l_x < l_{ts}$ or $l_p \geq l_x > l_{ts}$. A structure that lies strictly along the reaction coordinate will have all bonds fall into the same category; however, this is not necessary for our algorithm to function, as each bond is interpolated independently. In spite of the ability to treat points that do not lie strictly along the reaction coordinate, in this work we have restricted ourselves to the study of structures that lie along the reaction coordinate exclusively. While we have high confidence in the ability of our method to treat these points, treatment of points that do not lie exactly along the reaction coordinate is feasible where these points lie at least close to the reaction coordinate. Because the integrity of the method outlined is dependent upon the choice of reaction coordinate, meaningful results may not be obtained for cases where the choice of most appropriate reaction coordinate is not straightforward. However, we leave an assessment of the accuracy of the model for such structures to a future publication.

Once a bond in x is determined to be either product-side or reactant-side, we use its bond length, l_x, to determine quantitatively where along that half of the reaction coordinate it lies. Each l_x is compared to the nearest equilibrium bond lengths, l_{eq} (reactant if it is a reactant-side bond, product if it is a product-side bond), and the transition-state bond length, l_{ts}, to obtain δ_x.

$$\delta_x = \frac{l_x - l_{eq}}{l_{ts} - l_{eq}} \tag{10}$$

From eq 10, δ_x approaches zero for bond lengths close to the equilibrium lengths (reactant or product) and approaches one for bond lengths similar to the transition state lengths. Equation 10 is undefined, however, where l_{eq} is infinitely long, i.e., when the bond is completely broken, in either the reactant or product structure. This difficulty is encountered in all intermolecular reactions. To circumvent this problem, we have defined an effective cutoff length, such that any bond with a length exceeding the cutoff is instead assigned the cutoff value. At this length, the bond is assigned a bond order of zero, and hence no bond corrections, CLOC(x)$_{bond}$ or LOC(x)$_{bond}$, are assigned to it, as bond corrections are only assigned for bonds with nonzero bond orders. For the reactions depicted in Scheme 2, we have arrived empirically at the cutoff lengths given in Table 4. However, an inspection of Scheme 2 shows that only a limited number of bond types are studied: H−X, where X = H, O, Cl, and C, and C−C. Therefore, we are forced to define cutoff lengths for bonds heretofore not studied. To do so, we note that the cutoff lengths given in Table 4 for any bond i correspond to roughly twice the transition state bond length for that same bond, $2l_{ts}$. Specifically, the average transition state bond length, l_{ts}^{avg}, for all H−X bonds (X = H, O, Cl, C) studied herein, is l_{ts}^{avg} = 1.2 Å, and the empirically determined cutoff of 2.2 Å = $1.9l_{ts}^{avg}$. Likewise, the average transition state bond length for all C−C bonds studied herein is l_{ts}^{avg} = 2.3 Å, and therefore the cutoff of 4.0 Å = $1.7l_{ts}^{avg}$. Therefore, all cutoff lengths for systems heretofore not studied are taken as 1.8 times the length of the bond in the transition state, $1.8l_{ts}$, assuming transferability of the empirically determined cutoff length trend. We are not barring the possibility of refinement of these cutoff values when more reaction profiles are explored in the future.

Table 4. Cutoff Lengths for l_{eq}

bond type	cutoff length (Å)
H−X (X = H, O, Cl, C)	2.2
C−C	4.0
all other bonds	$1.8l_{ts}$

Equipped with our estimate of how far along the reaction coordinate the arbitrary structure lies with respect to each bond length (δ_x), we only need the equilibrium LOCs, to proceed with the interpolations between LOC(reactant), LOC(ts), and LOC(product). These are determined using an automated script that gives the corrections already described in previous publications[5c,e] with one minor exception described in section III.F.

The obtained equilibrium LOCs are then used to calculate each component i of CLOC(x) (see eq 9) according to the equation

$$\mathrm{CLOC}(\delta_x)_i = \mathrm{LOC(eq)}_i + f(\delta_x)_i \times \Delta\mathrm{LOC}_i \tag{11}$$

where

$$\Delta\mathrm{LOC}_i = \mathrm{LOC(ts)}_i - \mathrm{LOC(eq)}_i \tag{12}$$

and LOC(eq)$_i$ is set to be LOC(reactant)$_i$ for a reactant-side interpolation or LOC(product)$_i$ for a product-side interpolation. Therefore, we are only left with the task of choosing the appropriate $f(\delta_x)_i$ for each component i of eq 9, where i can be bond, hyb, etc.

In eq 11, the i subscript is used to emphasize that we have chosen to interpolate each CLOC term individually, each term receiving its own unique $f(\delta_x)_i$. Hypothetically, the interpolations could be performed instead on the basis of just one value of $f(\delta_x)$ that reflects where along the reaction coordinate the structure lies in its entirely. However, in spite

of its simplistic appeal, to obtain meaningful results with this method, all bond lengths and hybridization states etc. must fall at the same place along the reaction coordinate. By using the formulation presented in eq 11, where each term is interpolated individually, no such restriction is applied. Therefore, we have chosen to interpolate on a term-by-term basis to allow for increased flexibility and accuracy.

Clearly, the ability to assign LOCs to stationary points, and hence interpolate CLOCs for all intermediate structures, is dependent upon the ability to accurately assign Lewis structures to the former. All assignments of Lewis structures in this work were performed using an automated script (also used in other works described above[5c,e]). When this automatic assignment fails, human intervention might be necessary to provide information about the formal charges and/or spins in the same input file with the input structures, thereby preventing misassignment of more complicated systems. It is possible that the CLOC approach will be inapplicable to some systems with a poorly understood or badly defined Lewis structure. For the vast majority of systems of practical interest, no difficulty is encountered in this respect whatsoever. Specifically, large systems, such as those of interest to organic chemists and biochemists, are regularly studied using DFT for its excellent balance of performance and accuracy.[11] These same systems generally have well-defined Lewis structures and can therefore be treated easily with our CLOC methodology, as shown by the successful treatment of various organic chemistry reactions in our latest work.[5e]

III.B. CLOCs for Bonds, CLOC$(x)_{bond}$. As is discussed in detail in our previous publications,[5] the DFT-LOC model provides improvements to the estimation of nondynamical electron correlation by a specific DFT functional for localized electron pairs. The DFT-LOC bond corrections, or LOC$(x)_{bond}$, rest upon the assumption that the localized nuclear framework supporting an electron pair is a principal factor controlling the deviations in value of the nondynamical correlation from the "average" value within global hybrid functionals. Therefore, empirical corrections are applied on the basis of these localized frameworks. Consider, for example, the corrections applied to single bonds between heavy atoms of various lengths, when the 6-311++G(3df,3pd) basis set is used: short (−1.36 kcal/mol), medium (−1.90 kcal/mol), and long (−2.57 kcal/mol). These values become appreciably more negative with increasing bond length. This reflects the physically intuitive notion that as bond length increases, nondynamical correlation becomes more negative (as the electrons have more room to avoid each other), and B3LYP systematically underestimates this two-particle correlation effect with increasing severity.

The total LOC$(x)_{bonds}$ is given as the sum of all the various terms' bond LOCs:

$$\mathrm{LOC}(x)_{\mathrm{bonds}} = \sum_i \mathrm{LOC}(x)_i \tag{13}$$

where i runs over the 14 LOCs unique to bonds. The rationale for each correction is described in detail in ref 5, while the optimized value for each correction is given in the Supporting Information.

Analogously, the total CLOC$(x)_{bonds}$ is given by

$$\mathrm{CLOC}(x)_{\mathrm{bonds}} = \sum_i \mathrm{CLOC}(x)_i \tag{14}$$

where i again runs over the 14 LOCs unique to bonds.

All LOCs for bonds, LOC$(x)_{bonds}$, are designed to treat bonds of order 0.5, 1, 1.5, 2, 2.5, and 3. Yet we desire the ability to treat all bond orders and, thereby, transform LOC$(x)_{bond}$ to CLOC$(\delta_x)_{bond}$. As stated previously, CLOC$(\delta_x)_i$ is given by eq 11, which is modified such that it is specific to CLOC$(\delta_x)_{bond}$ (i = bond) and is written as

$$\mathrm{CLOC}(\delta_x)_{\mathrm{bond}} = \mathrm{LOC(eq)}_{\mathrm{bond}} + f(\delta_x)_{\mathrm{bond}} \times \Delta\mathrm{LOC}_{\mathrm{bond}} \tag{15}$$

where

$$\Delta\mathrm{LOC}_{\mathrm{bond}} = \mathrm{LOC(ts)}_{\mathrm{bond}} - \mathrm{LOC(eq)}_{\mathrm{bond}} \tag{16}$$

Therefore, we are only left with the task of choosing a proper form for $f(\delta_x)_{bond}$ such that it satisfies the appropriate boundary conditions:

$$f(\delta_x)_{\mathrm{bond}} = \begin{cases} 0, & \text{if } \delta_x = 0; \\ 1, & \text{if } \delta_x = 1 \end{cases} \tag{17}$$

Inspection shows that these boundary conditions are designed to ensure that CLOC$(\delta_x)_{bond}$ = LOC(eq)$_{bond}$ at δ_x = 0 (i.e., at the reactant or product) and that CLOC$(\delta_x)_{bond}$ = LOC(ts)$_{bond}$ at δ_x = 1 (i.e., at the transition state). Put simply, we are ensuring agreement between the previously developed LOCs and the continuous version, CLOCs, in the reactant, product, and transition state; i.e., LOC(x) = CLOC(x) where x is a stationary state.

It is reasonable to define $f(\delta_x)_{bond}$ as either a linear, Gaussian, or power function to satisfy these boundary conditions:

$$f(\delta_x)_{\mathrm{bond}} = \delta_x \tag{18}$$

$$f(\delta_x)_{\mathrm{bond}} = \mathrm{e}^{-\gamma(1-\delta_x)^2} \tag{19}$$

$$f(\delta_x)_{\mathrm{bond}} = 1 - (1 - \delta_x)^n \tag{20}$$

One can envision using other functions as well to perform the interpolations. For example, $f(\delta_x) = \sin(\delta_x \times \pi/2)$ could also be employed. We are not barring the possibility of adopting this or other interpolating functions in the future.

While eqs 18−20 all satisfy the necessary boundary conditions, it is also necessary that any $f(\delta_x)$ be everywhere differentiable such that its gradients can be defined. It is easy to see how linear interpolations based upon eq 18 would lead to nondifferentiable cusps at the transition state (δ_x = 1), where the reactant-side and product-side linear interpolations intersect, giving a curve with a shape similar to a triangle wave. For this reason, linear interpolations were discarded in spite of their simplicity. Among the power functions described by eq 20, we found cubic functions to best mimic the qualitative shape of the B3LYP error (for at least the shorter-bond-length half of the reaction coordinate) and therefore to give the best results. Unfortunately, testing has revealed that for the application of the CLOC method, cubic functions do not decay quickly enough to zero as δ_x

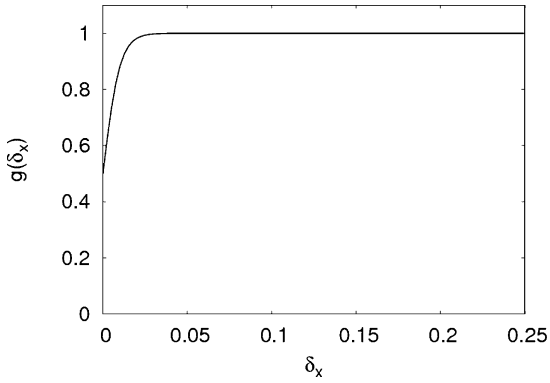


Figure 3. $g(\delta_x)$ vs δ_x as defined by eq 21.

approaches zero. To solve this problem, we multiply our cubic function by a function of the following form:

$$g(\delta_x) = \frac{1}{1 + e^{-\beta\delta_x}} \tag{21}$$

where β is an adjustable coefficient which controls the rate of decay, here chosen to be 200, and δ_x is defined by eq 10.

Inspection of Figure 3 shows that $g(\delta_x)$ decays rapidly as $\delta_x \rightarrow 0$. Therefore, multiplication of the cubic function given in eq 20 with $g(\delta_x)$ gives a new function that decays to zero with the proper rate as $\delta_x \rightarrow 0$. This function now has a desirable analytical behavior and can be used to interpolate between our limiting stationary point LOCs according to eq 11 for all bonds. Further, Gaussian functions, given by eq 19, can also be used, without modification, to the same end.

Interestingly, we have found that a combination of the two functions, modified cubic and Gaussian, serves as an even better match for the DFT B3LYP error as a function of intrinsic reaction coordinate. To understand how the two functions are combined, first consider how an active bond changes along the reaction coordinate. The characteristics of a bond in a transition state structure along the reaction coordinate change from bond order s_r with length l_r in the reactant to bond order s_p with length l_p in the product. For such a bond, we assume that the bond order in the transition state is an average of these two bond orders, $s_{ts} = (s_r + s_p)/2$, with corresponding length l_{ts}.

If $l_r > l_p$, any bond in the arbitrary structure with bond order s_x and bond length l_x has an interpolated LOC obtained from eq 11 where $f(\delta_x)_{bond}$ is given by

$$f(\delta_x)_{bond} = \begin{cases} e^{-\gamma(1-\delta_x)^2}, & \text{if } l_r > l_x \geq l_{ts}; \\ \dfrac{1-(1-\delta_x)^3}{1+e^{-\beta\delta_x}}, & \text{if } l_{ts} > l_x \geq l_p \end{cases} \tag{22}$$

Alternatively, if $l_r < l_p$, we have

$$f(\delta_x)_{bond} = \begin{cases} \dfrac{1-(1-\delta_x)^3}{1+e^{-\beta\delta_x}}, & \text{if } l_r > l_x \geq l_{ts}; \\ e^{-\gamma(1-\delta_x)^2}, & \text{if } l_{ts} > l_x \geq l_p \end{cases} \tag{23}$$

In both eqs 22 and 23, γ is an adjustable parameter to modulate the width of the Gaussian curve, set here to 5. β is used as in eq 21, and δ_x is defined by eq 10.

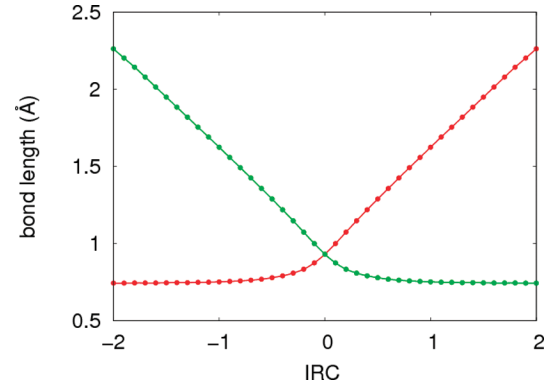


Figure 4. Bond lengths vs intrinsic reaction coordinate (IRC) for $H^1{-}H^2 + H^3 \rightarrow H^1 + H^2{-}H^3$. $H^1{-}H^2$ bond lengths are shown in red, while $H^2{-}H^3$ bond lengths are shown in green. Note that $H^1{-}H^2$ bond length changes rapidly for IRC > 0, yet slowly for IRC < 0. The opposite is true for $H^2{-}H^3$ according to symmetry.

Although empirically derived, this differential treatment, i.e., cubic interpolations for the shorter-bond-length half of the reaction coordinate (i.e., $l_r < l_x < l_{ts}$ or $l_{ts} > l_x > l_p$) and Gaussian interpolations for the longer, is founded upon the dependence of the reaction coordinate on bond length. Consider the reaction $H^1{-}H^2 + H^3 \rightarrow H^1 + H^2{-}H^3$, where each hydrogen has been marked with a unique superscript for the purpose of the argument. Inspection of Figure 4 shows that the $H^1{-}H^2$ bond length changes insignificantly with reaction coordinate on the reactant side, from 0.74 Å in the reactant to 0.93 Å in the transition state. Conversely, the $H^1{-}H^2$ bond length changes considerably with reaction coordinate on the product side, from 0.93 Å in the transition state to essentially infinite bond length in the product. Therefore, a Gaussian is employed on the product side for $H^1{-}H^2$ interpolations to ensure that the CLOC decays rapidly along the reaction coordinate, whereas a cubic function is used on the reactant side for $H^1{-}H^2$ interpolations for the opposite reason.

III.C. Defining CLOCs for Hybridization, CLOC$(x)_{hyb}$. In previous publications, we have also defined LOCs to describe various hybridization states, LOC$(x)_{hyb}$. While it is true that DFT in general benefits from significant cancellation of intra-atomic error as one goes from atoms to a molecule, i.e., as bonds are formed and atomic electronic structure is changed, these parameters were developed to address errors that remain in spite of this cancellation. These LOC$(x)_{hyb}$ parameters address the relatively large changes in orbital sizes, shapes, and occupancies that accompany bond formation and hence cause variations in nondynamical electron correlation for an electron pair contained in an orbital with a particular hybridization.

Each of these LOCs has its own unique purpose. For example, LOC$(x)_{N/P_sp2}$ and LOC$(x)_{N/P_sp3}$ are assigned for each nitrogen or phosphorus atom with sp^2 or sp^3 hybridization, respectively. Extensive definitions and optimized values (C_k) for all hybridization LOCs [LOC$(x)_{hyb}$] can be found in our previous publications[5] or the Supporting Information.

If we wish to treat hybridization states other than sp, $sp^{1.5}$, sp^2, $sp^{2.5}$, and sp^3, we must transform our LOC$(x)_{hyb}$ to the continuous CLOC$(x)_{hyb}$. To simplify the calculation of

$\mathrm{CLOC}(x)_{\mathrm{hyb}}$, we have split each of the $\mathrm{CLOC}(x)_{\mathrm{hyb}}$ terms into "active" and "inactive" terms so that $\mathrm{CLOC}(x)_{\mathrm{hyb}}$ is written as the sum of these two components.

$$\mathrm{CLOC}(x)_{\mathrm{hyb}} = \mathrm{CLOC}(x)_{\mathrm{hyb}}^{\mathrm{active}} + \mathrm{CLOC}(x)_{\mathrm{hyb}}^{\mathrm{inactive}} \quad (24)$$

The total $\mathrm{CLOC}(x)_{\mathrm{hyb}}$ is given by the sum of all hybridization CLOCs, each having an inactive and active component:

$$\mathrm{CLOC}(x)_{\mathrm{hyb}} = \sum_i [\mathrm{CLOC}(x)_{\mathrm{hyb}}^{\mathrm{active}} + \mathrm{CLOC}(x)_{\mathrm{hyb}}^{\mathrm{inactive}}]_i \quad (25)$$

where i runs over the nine LOCs unique to hybridization.

Accordingly, a given reaction profile (with reactant, transition state, and product structures) is processed to classify all of the bonds as either "active" or "inactive", i.e., as bonds with changed or unchanged bond orders along the reaction coordinate, respectively. Similarly, atoms are also classified as either inactive or active. Inactive atoms are those attached exclusively to inactive bonds, whereas active atoms are those attached to one or more active bonds. For example, in the reaction between methanol and the hydrogen atom, $H_3C-O-H + H\cdot \rightarrow H_3C-O\cdot + H-H$, the OH bond and the HH bond are both active, whereas the CO and CH bonds are inactive. Further, the carbon atom and hydrogens attached to it are inactive atoms, while all others are active.

The inactive component of $\mathrm{CLOC}(x)_{\mathrm{hyb}}$ is taken in direct analogy to eq 1 as

$$\mathrm{CLOC}(x)_{\mathrm{hyb}}^{\mathrm{inactive}} = \sum_k N_k C_k \quad (26)$$

where the index k runs over all inactive atoms, assigning LOCs in the same way as if these atoms were part of an equilibrium structure. Because inactive atoms experience no change in hybridization throughout the entirety of the reaction coordinate (as ascertained upon analysis of the reactant, product, and transition state structures input), we treat them as if they were still in their equilibrium states. Instead, we concern ourselves with treating only the active components of the reaction coordinate for hybridization in a dynamic fashion, $\mathrm{CLOC}(x)_{\mathrm{hyb}}^{\mathrm{active}}$. Specifically, $\mathrm{CLOC}(x)_{\mathrm{hyb}}^{\mathrm{active}}$ is taken as an interpolated value between $\mathrm{LOC(eq)}_{\mathrm{hyb}}$ and $\mathrm{LOC(ts)}_{\mathrm{hyb}}$, where $\mathrm{LOC(eq)}_{\mathrm{hyb}}$ is $\mathrm{LOC(r)}_{\mathrm{hyb}}$ for a reactant-side interpolation or $\mathrm{LOC(p)}_{\mathrm{hyb}}$ for a product-side interpolation. Adapting eq 11 to the active hybridization term gives

$$\mathrm{CLOC}(x)_{\mathrm{hyb}}^{\mathrm{active}} = \mathrm{LOC(eq)}_{\mathrm{hyb}}^{\mathrm{active}} + f(\delta_x)_{\mathrm{hyb}} \times \Delta\mathrm{LOC}_{\mathrm{hyb}}^{\mathrm{active}} \quad (27)$$

where

$$\Delta\mathrm{LOC}_{\mathrm{hyb}}^{\mathrm{active}} = \mathrm{LOC(ts)}_{\mathrm{hyb}}^{\mathrm{active}} - \mathrm{LOC(eq)}_{\mathrm{hyb}}^{\mathrm{active}} \quad (28)$$

In order to properly specify $\mathrm{CLOC}(x)_{\mathrm{hyb}}^{\mathrm{active}}$, we must first arrive at a proper definition of hybridization itself. While a bond is defined simply by two atomic centers and the distance between them, l, hybridization of an atom is a more complex characteristic which depends on all atoms surrounding the given atom, as well as the respective bond lengths, l_1, l_2, ..., l_n. Therefore, eqs 22 and 23, which depend only upon one bond length, are not sufficient to define hybridization, and similarly to the situation above, we define an interpolating $f(\delta_x)_{\mathrm{hyb}}$ for eq 27 such that it satisfies the appropriate boundary conditions.

$$f(\delta_x)_{\mathrm{hyb}} = \begin{cases} 0, & \text{for stationary state;} \\ 1, & \text{for transition state} \end{cases} \quad (29)$$

As in section III.B for $\mathrm{CLOC}(x)_{\mathrm{bonds}}$, a simple examination reveals that these boundary conditions ensure that $\mathrm{LOC(ts)}_{\mathrm{hyb}}^{\mathrm{active}} = \mathrm{CLOC}(x)_{\mathrm{hyb}}^{\mathrm{active}}$ where $f(\delta_x)_{\mathrm{hyb}} = 1$, i.e., at the transition state, and that $\mathrm{LOC(eq)}_{\mathrm{hyb}}^{\mathrm{active}} = \mathrm{CLOC}(x)_{\mathrm{hyb}}^{\mathrm{active}}$ where $f(\delta_x)_{\mathrm{hyb}} = 0$, i.e., for the reactant or product. Again, we are simply ensuring that $\mathrm{LOC}(x)_{\mathrm{hyb}}^{\mathrm{active}} = \mathrm{CLOC}(x)_{\mathrm{hyb}}^{\mathrm{active}}$ when x is a stationary state.

To take into account the multiatom dependence of hybridization, we have defined $f(\delta_x)_{\mathrm{hyb}}$ for atomic LOCs as the average of all $f(\delta_x)_{\mathrm{bond}}$'s values over all active bonds connected to that atomic center:

$$f(\delta_x)_{\mathrm{hyb}} = \frac{1}{n}\sum_{i=1}^{n} [f(\delta_x)_{\mathrm{bond}}]_i \quad (30)$$

where i is an index that runs over all active bonds. It is easy to see that as the bonds connected to any particular atom become more transition-state-like, as $f(\delta_x)_{\mathrm{bond}} \rightarrow 1$, on average, the interpolated hybridization also becomes more transition-state-like, that is, $f(\delta_x)_{\mathrm{hyb}} \rightarrow 1$. This also holds in the reverse direction, i.e., as bonds become more reactant- or product-like. In this manner, the interpolated value of $\mathrm{CLOC}(x)_{\mathrm{hyb}}^{\mathrm{active}}$ according to eq 27 is tuned to reflect how reactant-, product-, or transition-state-like the hybridization of an active atom is as a function of how reactant-, product-, or transition-state-like the bonds connected to it are on average.

III.D. Defining CLOCs for Radicals, $\mathrm{CLOC}(x)_{\mathrm{radical}}$. As argued extensively in our previous publications,[5] the self-interaction term in DFT is used to quantitatively model the nondynamical electron correlation of an electron pair. However, this self-interaction term becomes problematic for unpaired electrons. We have previously developed corrections to specifically treat atoms with radicals localized on them, $\mathrm{LOC}(x)_{\mathrm{radical}}$, to remedy systematic overbinding: $\mathrm{LOC}(x)_{\mathrm{RH}}$, $\mathrm{LOC}(x)_{\mathrm{RA}}$, and $\mathrm{LOC}(x)_{\mathrm{RT}}$, to treat atomic centers with localized radicals that have neighboring bonds to hydrogen, single or double bonds to heavy atoms, or triple bonds to heavy atoms, respectively.

The number of unpaired electrons and formal charge on each atomic center are ascertained by assuming a set number of valence electrons, χ, for each atom type, as specified in Table 5. The formal charge (q) and number of unpaired electrons (u) is then a function of the element's group number

Table 5. Valency for Each Atom Type

atom type	χ	example
H, He	2	H_2
Al, B	6	BH_3
Cl, P, or S with $\sum_i s_i > 8 - g$	10	PCl_5
all other 1st and 2nd row atoms	8	CH_4

on the periodic table (g) and the bond order of all bonds connected to it (s_i) according to

$$g - q + u + \sum_{i=1}^{n} s_i = \chi \tag{31}$$

Given the number of unpaired electrons, u, we can compute the radical CLOCs: CLOC$(x)_{RH}$, CLOC$(x)_{RA}$, and CLOC$(x)_{RT}$. [For a complete list of all CLOCs and their definitions and values, including the radical CLOCs, CLOC$(x)_{radical}$, the reader is referred to our previous publications[5] or the Supporting Information.]

According to eq 1, the contribution due to radical LOCs to the total LOC for an equilibrium structure will be given by

$$\mathrm{LOC}(x)_{\mathrm{radical}} = \sum_i \mathrm{LOC}(x)_i = \sum_i N_i \times C_i \tag{32}$$

where, here, i runs over the three LOCs unique to radicals [LOC$(x)_{RH}$, LOC$(x)_{RA}$, and LOC$(x)_{RT}$] and N_k and C_k are the count and LOC value, respectively. Specifically,

$$N_{\mathrm{RH}} = \sum_i u_i[\eta_{\mathrm{RH}}]_i \tag{33}$$

$$N_{\mathrm{RA}} = \sum_i u_i[\eta_{\mathrm{RA}}]_i \tag{34}$$

$$N_{RT} = \sum_i u_i[\eta_{\mathrm{RT}}]_i \tag{35}$$

where u_i is the number of unpaired electrons, $[\eta_{RH}]_i$ is the number of single bonds to hydrogen atoms, $[\eta_{RA}]_i$ is the number of single or double bonds to non-hydrogen atoms, and $[\eta_{RT}]_i$ is the number of triple bonds, all corresponding to center i.

In order to adapt our equilibrium LOC$(x)_{radical}$ contribution to the continuous representation, CLOC$(x)_{radical}$, we must allow for noninteger values of n_{RH}, n_{RA}, and n_{RT}. To accomplish this, we substitute η in eqs 33−35 with some function of the bond orders for the bonds of interest, $f(s)$:

$$N_{\mathrm{RH}} = \sum_i u_i[f_{\mathrm{RH}}(s_{\mathrm{RH}})]_i \tag{36}$$

$$N_{\mathrm{RA}} = \sum_i u_i[f_{\mathrm{RA}}(s_{\mathrm{RA}})]_i \tag{37}$$

$$N_{\mathrm{RT}} = \sum_i u_i[f_{\mathrm{RT}}(s_{\mathrm{RT}})]_i \tag{38}$$

where s_{RH}, s_{RA}, and s_{RT} are the bond orders between the atomic center of interest i and its neighboring hydrogen atom (RH) or neighboring non-hydrogen atom (RA or RT).

Again, we desire $f(s)$ functions that both are differentiable and satisfy the appropriate boundary conditions. LOC$(x)_{RH}$ is applied to any radical-containing atom i with bonds to hydrogen. Therefore, the boundary conditions dictate that there be no LOC$(x)_{RH}$ when atom center i is not bonded to a hydrogen atom. Conversely, there must be one LOC$(x)_{RH}$ applied for each (single) bond to hydrogen from atom center i. The boundary condition for $f_{RH}(s_{RH})$ in eq 36 is hence given by

$$f_{\mathrm{RH}}(s_{\mathrm{RH}}) = \begin{cases} 0, & \text{for } s_{\mathrm{RH}} = 0 \text{ or } s_{\mathrm{RH}} = 2; \\ 1, & \text{for } s_{\mathrm{RH}} = 1 \end{cases} \tag{39}$$

where s_{RH} is the bond order of the bond between atom center i and the neighboring hydrogen atom.

Likewise, LOC$(x)_{RA}$ is applied to any radical-containing atom i with single or double bonds to non-hydrogen atoms. Thus, we desire one LOC$(x)_{RA}$ for each atom center i with a single or double bond to another non-hydrogen atom and no LOC$(x)_{RA}$ for each atom center i with no bond or a triple bond to another non-hydrogen atom. The boundary condition for $f_{RA}(s_{RA})$ in eq 37 is hence given by

$$f_{\mathrm{RA}}(s_{\mathrm{RA}}) = \begin{cases} 0, & \text{for } s_{\mathrm{RA}} = 0 \text{ or } s_{\mathrm{RA}} = 3; \\ 1, & \text{for } s_{\mathrm{RA}} = 1 \text{ or } s_{\mathrm{RA}} = 2 \end{cases} \tag{40}$$

where s_{RA} is the bond order between the radical-containing atom i and the neighboring non-hydrogen atom.

Lastly, LOC$(x)_{RT}$ is applied to any radical-containing atom i with a triple bond to a non-hydrogen atom. Thus, we desire one LOC$(x)_{RT}$ for each atom center i with a triple bond to another non-hydrogen atom and no LOC$(x)_{RT}$ for each atom center i with no triple bond to another non-hydrogen atom. The boundary condition for $f_{RT}(s_{RT})$ in eq 38 is hence given by

$$f_{\mathrm{RT}}(s_{\mathrm{RT}}) = \begin{cases} 0, & \text{for } s_{\mathrm{RT}} = 2 \text{ or } s_{\mathrm{RT}} = 4; \\ 1, & \text{for } s_{\mathrm{RT}} = 3 \end{cases} \tag{41}$$

where s_{RT} is the bond order between the radical-containing atom i and the neighboring non-hydrogen atom.

As discussed above, we can readily employ Gaussian functions to both meet the differentiability requirement and satisfy our various boundary conditions.

$$f_{\mathrm{RH}}(s_{\mathrm{RH}}) = \mathrm{e}^{-\upsilon(s_{\mathrm{RH}}-1)^2} \tag{42}$$

$$f_{\mathrm{RA}}(s_{\mathrm{RA}}) = \begin{cases} \mathrm{e}^{-\upsilon(s_{\mathrm{RA}}-1)^2}, & \text{if } s_{\mathrm{RA}} \le 1; \\ 1, & \text{if } 1 < s_{\mathrm{RA}} \le 2; \\ \mathrm{e}^{-\upsilon(s_{\mathrm{RA}}-2)^2}, & \text{if } s_{\mathrm{RA}} > 2 \end{cases} \tag{43}$$

$$f_{\mathrm{RT}}(s_{\mathrm{RT}}) = \mathrm{e}^{-\upsilon(s_{\mathrm{RT}}-3)^2} \tag{44}$$

where υ is set to 5 to ensure a proper rate of growth and decay for our functions.

Therefore, the CLOC$(x)_{radical}$ term is written as

$$\mathrm{CLOC}(x)_{\mathrm{radical}} = \mathrm{CLOC}(x)_{\mathrm{RH}} + \mathrm{CLOC}(x)_{\mathrm{RA}} + \mathrm{CLOC}(x)_{\mathrm{RT}} = N_{\mathrm{RH}} \times C_{\mathrm{RH}} + N_{\mathrm{RA}} \times C_{\mathrm{RA}} + N_{\mathrm{RT}} \times C_{\mathrm{RT}} \tag{45}$$

where N_{RH}, N_{RA}, and N_{RT} are defined by eqs 36−38 and 42−44.

III.E. CLOCs for Hypervalency, CLOC$(x)_{hyperval}$. We also define LOCs for atoms with more than eight valence electrons or two valence electrons, for hydrogen and helium atoms: LOC$(x)_{hyperval}$. These LOCs are designated LOC-

$(x)_{OCT\_EXP}$ and $LOC(x)_{H\_dival}$, respectively. The total $LOC(x)_{hyperval}$ is thus given by the sum of these two terms:

$$LOC(x)_{hyperval} = LOC(x)_{OCT\_EXP} + LOC(x)_{H\_dival} \quad (46)$$

Analogously, we write the continuous version, $CLOC(x)_{hyperval}$, as

$$CLOC(x)_{hyperval} = CLOC(x)_{OCT\_EXP} + CLOC(x)_{H\_dival} \quad (47)$$

Both of these terms will be discussed in the subsections that follow.

III.E.1. Hydrogen Hypervalency: CLOC(x)$_{H\_dival}$. In ref 5e, 105 transition states and barrier heights were analyzed at the B3LYP level. A thorough analysis of the errors in these B3LYP barrier heights reveals that transition states in which the central hydrogen atom is flanked by at least one nonhydrogen/noncarbon atom (as shown in the transition states of examples i−iii below) all display systematic errors. Presumably, this originates in overestimation of nondynamical electron correlation due to localized high electron density.

In this same study,[5e] we also found that the transition state in which the central hydrogen atom is flanked by two hydrogen atoms (as in example iv below) displays approximately the same error in barrier height. This transition state is highly analogous to the well-described H_2^+ molecule[12] where the self-interaction term described earlier does not serve to model nondynamical electron correlation but instead engenders a clear source of systematic error.[13] Accordingly, $LOC(x)_{H\_dival}$ is assigned to cases where the central hydrogen atom is flanked by two hydrogens to remedy this error as well.

In summary, $LOC(x)_{H\_dival}$ is applied to transition states in which the central hydrogen atom is flanked by at least one noncarbon/nonhydrogen atom ($n_A \geq 1$, in eq 50) or where the central hydrogen atom is flanked by two additional hydrogen atoms ($n_H = 2$ in eq 50).

We may rewrite the above discussion in terms of equations as follows. In the discrete version of the approach, the H_dival correction for any system x, $LOC(x)_{H\_dival}$, is written as

$$LOC(x)_{H\_dival} = N_{H\_dival} \times C_{H\_dival} \quad (48)$$

C_{H_dival} is the correction's numerical value, optimized to reduce the B3LYP error, and N_{H_dival} is given by

$$N_{H\_dival} = \sum_i \eta_i \quad (49)$$

where i is an index that runs over all hydrogen atoms and η is defined by the number of bonds to each hydrogen atom according to

$$\eta = \begin{cases} 0, & \text{if } n_A < 1; \\ 1, & \text{if } n_A \geq 1 \text{ or } n_H = 2 \end{cases} \quad (50)$$

Here, n_A is the number of bonds between hydrogen atom i and noncarbon/non-hydrogen atoms, whereas n_H is the number of bonds between hydrogen atom i and other hydrogens. For example, the transition states of the following reactions each merit $\eta = N_{H\_dival} = 1$:

(i) $H_2O + NH_2 \rightarrow HO + NH_3$ via $[HO{-}H{-}NH_2]^{\bullet\ddagger}$
(ii) $CH_4 + OH \rightarrow CH_3 + OH_2$ via $[CH_3{-}H{-}OH]^{\bullet\ddagger}$
(iii) $H_2 + Cl \rightarrow H + HCl$ via $[H{-}H{-}Cl]^{\bullet\ddagger}$
(iv) $H_2 + H \rightarrow H + H_2$ via $[H{-}H{-}H]^{\bullet\ddagger}$

In each of these transition states, the central hydrogen atom is flanked by either two hydrogen atoms or at least one nonhydrogen/noncarbon atom. Alternatively, the transition states of the following reactions have $\eta = N_{H\_dival} = 0$:

(v) $CH_4 + CH_3 \rightarrow CH_3 + CH_4$ via $[CH_3{-}H{-}CH_3]^{\bullet\ddagger}$
(vi) $CH_3 + H_2 \rightarrow CH_4 + H$ via $[CH_3{-}H{-}H]^{\bullet\ddagger}$

In these transition states, the central hydrogen atom is flanked by either two carbon atoms or one carbon atom and one hydrogen atom. Notice that the case where the central hydrogen atom is flanked by two hydrogen atoms (example iv) still merits $\eta = N_{H\_dival} = 1$ as described in the discussion above.

We assume integer bond orders in reactants and products and integer bond orders in addition to bond orders 0.5, 1.5, and 2.5 in transition states. In eq 50, any of these would be considered "bonds" and hence contribute to n as illustrated in the examples above.

To adopt $LOC(x)_{H\_dival}$ to the continuous case and hence specify the functional form for $CLOC(x)_{H\_dival}$, we redefine η as a function of δ_x. Specifically, eq 50 may be written as η_{X-H-Y}, where the subscripts X and Y indicate the type of atoms flanking the central hydrogen atom. As before, A is defined as any atom other than carbon or hydrogen.

$$\eta_{X-H-Y} = \begin{cases} \sqrt{\delta_{X-H} \times \delta_{Y-H}}, & \text{for } X = A_1 \text{ and } Y = A_2, \\ & \text{or } X = A \text{ and } Y = C, \\ & \text{or } X = A \text{ and } Y = H, \\ & \text{or } X = H_1 \text{ and } Y = H_2; \\ 0, & \text{for } X = C_1 \text{ and } Y = C_2, \\ & \text{or } X = C \text{ and } Y = H \end{cases} \quad (51)$$

In eq 51, δ_{X-H} is dependent upon the length between atom X and the central hydrogen atom, l_{X-H}, according to eq 10. δ_{Y-H} is defined similarly. Note that η_{X-H-Y} is defined for examples i−iv above but is always zero for examples v−vi. This is consistent with the prescriptions detailed at the beginning of this section.

An inspection of eq 51 shows that as both bond lengths, l_{X-H} and l_{Y-H}, approach their transition state lengths, δ_{X-H} and δ_{Y-H}, $\eta_{X-H-Y} \rightarrow 1$ and hence $CLOC(x)_{H\_dival}$ is applied. Likewise, as both bond lengths approach the equilibrium lengths, δ_{X-H} and δ_{Y-H}, $\eta_{X-H-Y} \rightarrow 0$ and hence $CLOC(x)_{H\_dival}$ is not applied. This is consistent with our understanding of hydrogen abstraction reactions, wherein the hydrogen being abstracted is divalent in the transition state, where $CLOC(x)_{H\_dival}$ is applied, but only monovalent in the reactant and product, where $CLOC(x)_{H\_dival}$ is not applied.

In eq 51, a product of δ_{X-H} and δ_{Y-H} is employed to ensure that η_{X-H-Y} is a function of both bond lengths l_{X-H} and l_{Y-H} and that it decays quickly to zero when at least one of the bond lengths is greater than the transition state bond length, i.e., as δ_{X-H} or $\delta_{Y-H} \rightarrow 0$. This reflects the fact that hypervalency on hydrogen is a function of two bond lengths. For example, if only l_{X-H} is near the transition state bond

length, giving $\delta_{X-H} \approx 1$, but l_{Y-H} is near the equilibrium length, giving $\delta_{Y-H} \approx 0$, in fact the central hydrogen is not hypervalent via chemical intuition. Accordingly, we desire $\eta_{X-H-Y} \approx 0$ such that $CLOC(x)_{H\_dival}$ is effectively not applied. This is indeed realized with the functional form of eq 51.

Alternatively, had we defined η_{X-H-Y} as simply the average of δ_{X-H} and δ_{Y-H}, i.e., $\eta_{X-H-Y} = (\delta_{X-H} + \delta_{Y-H})/2$, the proper behavior would not be observed in the illustrative example given above. Specifically, for $\delta_{X-H} \approx 1$ and $\delta_{Y-H} \approx 0$, $\eta_{X-H-Y} = (\delta_{X-H} + \delta_{Y-H})/2 \approx 1/2$ and $CLOC(x)_{H\_dival}$ would be nonzero and hence imply partial hypervalency on the central hydrogen atom, in spite of the fact that we know from chemical intuition that hypervalent character of this central hydrogen atom is negligible.

The necessity for the square root over the product in eq 51 becomes clear upon consideration of an additional illustrative example. Imagine a central hydrogen atom that is only "half-hypervalent", i.e., halfway along the reaction coordinate between the reactant or product and transition state in a standard hydrogen abstraction reaction. Here, $\delta_{X-H} = \delta_{Y-H} = 1/2$, and we desire $\eta_{X-H-Y} = 1/2$ to reflect this "half-hypervalency." Defining η_{X-H-Y} as a simple product, i.e., $\eta_{X-H-Y} = \delta_{X-H} \times \delta_{Y-H}$, would yield $\eta_{X-H-Y} = 1/4$ in this case, and hence this system would not be described as "half-hypervalent" as we desire but instead as only "quarter-hypervalent". Instead, we define η_{X-H-Y} as the square of the product to effect the proper behavior, in this case, $\eta_{X-H-Y} = 1/2$, consistent with our understanding that this system is "half-hypervalent".

Combining eq 49 with eq 51 allows us to define $CLOC(x)_{H\_dival}$ as

$$CLOC(x)_{H\_dival} = N_{H\_dival} \times C_{H\_dival} \tag{52}$$

III.E.2. Heavy-Atom Hypervalency: $CLOC(x)_{OCT\_EXP}$. The motivation driving the definition of this term is analogous to that described for $CLOC(x)_{H\_dival}$ in section III.E.1. Specifically, we assume that there is overestimation of the nondynamical electron correlation for environments with overall higher electron density from neighboring orbitals. These systems are exemplified by the systems ClF_3 and PCl_5 where the central atom has a valence shell expansion beyond the usual octet. This term is also applied to transition states exemplified by the S_N2 reaction $F^- + CH_3Cl \rightarrow Cl^- + CH_3F$. Here, the carbon of the transition state also experiences an increase in electron density that also leads to an overestimation of nondynamical electron correlation. The overbinding of hypervalent structures is manifested in both atomization energies of hypervalent species and in transition states with hypervalent character, as is shown in detail in previous works such as refs 5a and 5e.

This LOC, $LOC(x)_{OCT\_EXP}$, is defined as

$$LOC(x)_{OCT\_EXP} = N_{OCT\_EXP} \times C_{OCT\_EXP} \tag{53}$$

where N_{OCT_EXP}, analogously to N_{H_dival} in eq 49, is

$$N_{OCT\_EXP} = \sum_i \eta_i \tag{54}$$

Here, i is an index that runs over all non-hydrogen atoms, and η_i is given by

$$\eta_i = \begin{cases} 0, & \text{if } n \leq 8 - g; \\ 1, & \text{if } n \geq (8 - g) + 1 \end{cases} \tag{55}$$

In this formula, n is the number of bonds around an atom center i, g is the element's group number on the periodic table, and "bonds" are defined as for eq 50. This equation ensures that atoms bonded to a number of elements that violate their octets are assigned $N_{OCT\_EXP} = 1$, whereas the opposite is true for atoms with a number of bonds that are within their octet.

For example, consider how chlorine is treated in HCl vs ClF_3. Chlorine's group number in the periodic table, g, is 7. In HCl, the number of bonds, n, to chlorine is one, and from the equation above, we have $\eta = 0$. Therefore, $N_{OCT\_EXP} = LOC(x)_{OCT\_EXP} = 0$; i.e., $LOC(x)_{OCT\_EXP}$ is not assigned for HCl. In ClF_3, however, we have $n = 3$, and from the equation above, $\eta = 1$. Therefore, $N_{OCT\_EXP} = LOC(x)_{OCT\_EXP} = 1$; i.e., $LOC(x)_{OCT\_EXP}$ is assigned for the chlorine of ClF_3.

We use these equilibrium values of N_{OCT_EXP} given by the equations above to determine $CLOC_{OCT\_EXP}$. Specifically, $CLOC_{OCT\_EXP}$ takes on the same general form as LOC_{OCT_EXP} in eq 53, except that ζ, which is continuous, is now used in place of N_{OCT_EXP}, which is discrete.

$$CLOC(x)_{OCT\_EXP} = \zeta \times C_{OCT\_EXP} \tag{56}$$

where

$$\zeta = \sum_i [f(N^{eq}_{OCT\_EXP}, N^{ts}_{OCT\_EXP})]_i \tag{57}$$

and i is an index that runs over all non-hydrogen atoms, i.e., those atoms which are eligible to receive $CLOC(x)_{OCT\_EXP}$. We define $f(N^{eq}_{OCT\_EXP}, N^{ts}_{OCT\_EXP})$ as a function of the equilibrium and transition state N_{OCT_EXP} values, $N^{eq}_{OCT\_EXP}$ and $N^{ts}_{OCT\_EXP}$, respectively. In this manner, the interpolated value of $CLOC(x)_{OCT\_EXP}$ for any intermediate structure x is a function of the stationary states' LOCs, $LOC(eq)_{OCT\_EXP}$ and $LOC(ts)_{OCT\_EXP}$. The equilibrium structure is taken as the reactant for a reactant-side arbitrary structure, or product for a product-side arbitrary structure.

$$f(N^{eq}_{OCT\_EXP}, N^{ts}_{OCT\_EXP}) = \begin{cases} 0, & \text{for } N^{eq}_{OCT\_EXP} = N^{ts}_{OCT\_EXP} = 0; \\ 1, & \text{for } N^{eq}_{OCT\_EXP} = N^{ts}_{OCT\_EXP} = 1; \\ f(\delta_x)_{hyb}, & \text{for } N^{eq}_{OCT\_EXP} = 0 \text{ and } N^{ts}_{OCT\_EXP} = 1; \\ 1 - f(\delta_x)_{hyb}, & \text{for } N^{eq}_{OCT\_EXP} = 1 \text{ and } N^{ts}_{OCT\_EXP} = 0 \end{cases} \tag{58}$$

where $f(\delta_x)_{hyb}$ is as defined in eq 30. As discussed earlier, $f(\delta_x)_{hyb}$ takes into account the multiple-bond-length dependency of hybridization. Because $CLOC(x)_{OCT\_EXP}$ also depends upon multiple bond lengths, $f(\delta_x)_{hyb}$ is employed here.

Inspection of this equation shows that where both the equilibrium structure (reactant for a reactant-side interpolation, or product for a product-side interpolation) and transition state do not receive an OCT_EXP correction, $N^{eq}_{OCT\_EXP} = N^{ts}_{OCT\_EXP} = 0$, the interpolated structure, $f(N^{eq}_{OCT\_EXP},$

$N^{ts}_{OCT\_EXP}) = 0$, does not either. This would apply to the reaction $CH_3\cdot + CH_2CH_2 \rightarrow CH_3CH_2CH_2\cdot$ depicted in Scheme 1, for example. The same holds for the opposite case. Namely, where both the equilibrium structure and transition state structure do receive $LOC(x)_{OCT\_EXP}$, so does the interpolated structure. This would apply to the reaction $SO_4^{2-} + H_3O^+ \rightarrow HSO_4^- + H_2O$, for example, where the sulfur atom merits $LOC(x)_{OCT\_EXP}$ throughout the reaction. For a reaction where the equilibrium structures do not merit $LOC(x)_{OCT\_EXP}$, yet the transition state does, the amount of $CLOC(x)_{OCT\_EXP}$ the interpolated structure receives is proportional to $f(\delta_x)_{hyb}$. Therefore, the amount of $CLOC(x)_{OCT\_EXP}$ increases smoothly toward the transition-state value as the structure itself becomes more transition-state-like, as quantified by $f(\delta_x)_{hyb}$, defined by eq 30 above. This applies to the reaction $FCH_3 + Cl^- \rightarrow F^- + CH_3Cl$, for example, where neither equilibrium structure (reactant or product) merits $LOC(x)_{OCT\_EXP}$, yet the transition state does. Lastly, where the equilibrium structure does merit $LOC(x)_{OCT\_EXP}$ but the transition state does not, the amount of $CLOC(x)_{OCT\_EXP}$ decreases smoothly toward the transition-state value, again as a function of $f(\delta_x)_{hyb}$.

None of the reactions in Scheme 2 merit the $CLOC(x)_{OCT\_EXP}$ correction, and hence this specific term has not yet been tested. However, we anticipate that this correction will work well judging from the behavior of all of the other similar terms.

III.F. CLOCs for Environment, $CLOC(x)_{environ}$. We have also argued that the presence of neighboring bonds connected to a particular base bond contributes to systematic error in the quantification of nondynamical correlation of that base bond,[5] and we introduce the $LOC(x)_{environ}$ term, $LOC(x)_{ESBC}$, to capture these effects. This term arises from the fact that an electron in the base bond can make an excursion to a neighboring bond, increasing its nondynamical correlation energy, particularly if it is a long bond, vs a single bond to hydrogen, for example.

We stated in section III.A that it is necessary that $CLOC(x)$ agrees with $LOC(x)$ where x is a stationary point (reactant, product, or transition state). To meet this requirement, we found it necessary to slightly modify the way the previously defined $LOC(x)_{ESBC}$ parameter was extended to transition states in the latest LOC publication.[5e] Let us begin with a detailed explanation of how $LOC(x)_{ESBC}$ was assigned to transition states in the previous work[5e] to understand why a modification was necessary.

While the definition of $LOC(x)_{ESBC}$ for reactants and products is straightforward, formulating an implementation for transition states is less obvious. For example, consider the Diels−Alder reaction between butadiene and ethene (reaction a in Scheme 2). The reactant and product are readily assigned $N_{ESBC} = 0$ and 8, respectively. Yet, the $LOC(x)_{ESBC}$ assignment for the transition state is not immediately obvious.

Because $LOC(x)_{ESBC}$ is applied only for neighboring single bonds, the bond order s_i for each bond is transformed into a value to describe its percent single bond character, $f(s_i)$, on a scale from 0−1; 1 being a pure single bond and 0 being no bond or a pure double bond.

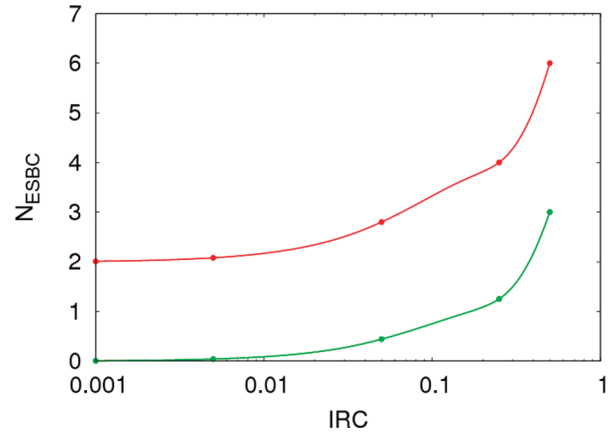


Figure 5. N_{ESBC} vs intrinsic reaction coordinate (IRC) for reactant to transition state of reaction a in Scheme 2. N_{ESBC} as defined in eqs 59−61 is shown in red, while N_{ESBC} as defined by eqs 59 and 62−64 is shown in green.

$$f(s_i) = e^{-\gamma(s_i-1)^2} \tag{59}$$

Here, γ here is chosen to be 3 such that $f(s_i) \approx 0.5$ for $s_i = 0.5$.

The N_{ESBC} for any bond i, $(N_{ESBC})_i$, is then given by the sum of $f(s_j)$ over all neighboring bonds j.

$$(N_{ESBC})_i = \sum_j f(s_j) \tag{60}$$

and the total N_{ESBC} for the system is then the sum over all $(N_{ESBC})_i$ for each bond i.

$$N_{ESBC} = \sum_i (N_{ESBC})_i \tag{61}$$

Using the above formulas, we find that the reactant and transition state of reaction a in Scheme 2 have $N_{ESBC} = 0$ and 6, respectively. In Figure 5, we show the results of using the formulations given in eqs 59−61 to interpolate points intermediate between the reactant and transition state for this reaction in red. Importantly, we see that $N_{ESBC} \rightarrow 2$ as the reaction coordinate $\rightarrow 0$. Recall that for the reactant, $N_{ESBC} = 0$, and therefore, $CLOC(x)$ does not agree with $LOC(x)$ as we approach the reactant. Yet, we stated in section III.A that $CLOC(x)$ must agree with $LOC(x)$ where x is a stationary point.

To force agreement between $CLOC(x)$ and $LOC(x)$ at the reactant and product, we have changed the formulation of $LOC(x)_{ESBC}$ for transition states from that described by eqs 59−61. In the previous publication,[5e] it was only the neighboring bonds and their bond orders, s_j, which determined N_{ESBC} for the bond under consideration, but not the bond order of that bond itself, s_i (see eq 60). In the latest implementation, both the bond order, s_i, of the bond under consideration and that of its neighboring bond, s_j, are considered when determining N_{ESBC}. Specifically, both $f(s_i)$ and $f(s_j)$, defined in eq 59, are multiplied to give a number that reflects cumulative percent single bond character for the pair, π_{ij}

$$\pi_{ij} = f(s_i) \times f(s_j) \tag{62}$$

In this manner, two neighboring single bonds receive the maximum value ($\pi_{ij} = 1$), while a neighboring single bond and "half bond" will receive a lesser value ($\pi_{ij} = 1/2$), and two neighboring "half" bonds will receive a lesser value still ($\pi_{ij} = 1/4$), for example. This reflects the fact that the more single-bond-like the neighboring bonds are, the more excursions are possible from a base bond into these neighboring bonds, and hence the more correction is necessary to account for these excursions.

For each bond i, the sum of all π_{ij} values is taken across all neighboring bonds j, which produces the total N_{ESBC} for that bond, $N_{\mathrm{ESBC}}(\pi_{ij})$.

$$N_{\mathrm{ESBC}}(\pi_{Ij}) = \sum_{i<j} (\pi_{ij}) \tag{63}$$

The increase in π_{ij} with increasing single-bond-character is thus utilized here to also assign larger $N_{\mathrm{ESBC}}(\pi_{ij})$ for systems with more neighboring bonds with high single-bond character.

All $N_{\mathrm{ESBC}}(\pi_{ij})$'s for each unique ij pair are summed to give the final N_{ESBC} for that system.

$$N_{\mathrm{ESBC}} = \sum_{i<j} N_{\mathrm{ESBC}}(\pi_{ij}) \tag{64}$$

The results of the new definition of N_{ESBC}, as given in eqs 59 and 62−64, are shown in Figure 5. Notice that while the previous definition of N_{ESBC} (shown in red in Figure 5) did not have the proper behavior, i.e., N_{ESBC} did not approach zero as the reaction coordinate → 0, this definition of N_{ESBC} (shown in green in Figure 5) does indeed have the proper behavior, i.e., $N_{\mathrm{ESBC}} \rightarrow 0$ as the reaction coordinate → 0. Therefore, we have satisfied the requirement that CLOC(x) agree with LOC(x) where x is a stationary point, at least for x being the reactant or product.

Inspection of Figure 5 also shows that N_{ESBC} for the transition state of this reaction has changed, from $N_{\mathrm{ESBC}} = 6$ in the old definition to $N_{\mathrm{ESBC}} = 3$ in the new definition. As with any parametrization, we are free to change the definition of how parameters are applied (N_k in eq 1) so long as we reoptimize the values of the parameters (C_k in eq 1) in accordance with their new definitions. Therefore, the newly defined application of N_{ESBC} to transition states necessitated reoptimizing the values of the transition-state specific parameters, C_k, to optimally reduce the B3LYP error in barrier heights. The updated values, which are only slightly different than those previously published, and all LOC parameter values and definitions can be found in the Supporting Information. Note that while the individual B3LYP-LOC barrier heights have changed slightly, the overall performance of B3LYP-LOC remains unchanged. That is, the LOCs still produce a dramatic reduction in the B3LYP barrier height errors and predict barrier heights within or near chemical accuracy (traditionally taken as ≤1 kcal/mol) across a broad spectrum of reactions.

III.G. CLOC for Charge Transfer, CLOC(x)$_{\mathrm{CT}}$. In section III.B, we argue that as the bond length increases, nondynamical correlation becomes more negative (as the electrons have more room to avoid each other), and DFT systematically underestimates this effect with increasing severity. An extreme example of this is manifest in systems such as carbon monoxide, $^{-}C{\equiv}O^{+}$, or sodium chloride, $Na^{+}Cl^{-}$, which have zero overall formal charge but nonzero formal charge on individual atoms. In these systems, the localized orbitals are highly ionic in character and hence compactly organized around one of the two atoms, and the bonds are also relatively long in comparison to the size of the orbitals in which the electron pairs are localized. Because this is a severe example of underbinding, this situation when it arises receives its own special parameter, LOC(x)$_{\mathrm{CT}}$, according to

$$\mathrm{LOC}(x)_{\mathrm{CT}} = N_{\mathrm{CT}} \times C_{\mathrm{CT}} \tag{65}$$

where C_{CT} is the optimized value of the LOC(x)$_{\mathrm{CT}}$ parameter (see the Supporting Information) and N_{CT} is given by

$$N_{\mathrm{CT}} = \sum_{i<j} f(q_i) \times f(q_j) \tag{66}$$

Here, i and j are indices that run over all neighboring atom pairs and $f(q)$ is a function of the formal charge q on an atom (as defined in eq 31) given by

$$f(q) = \begin{cases} 0, & \text{for } q = 0; \\ 1, & \text{for } |q| \geq 1 \end{cases} \tag{67}$$

An inspection of this equation shows that N_{CT} and LOC(x)$_{\mathrm{CT}}$, by extension, are nonzero only where two neighboring atoms both have nonzero charge.

The continuous version of LOC(x)$_{\mathrm{CT}}$, CLOC(x)$_{\mathrm{CT}}$, may be written analogously as

$$\mathrm{CLOC}(x)_{\mathrm{CT}} = N_{\mathrm{CT}} \times C_{\mathrm{CT}} \tag{68}$$

where N_{CT} is given still by eq 66 and only the definition of $f(q)$ is modified to allow for continuous representation of partial formal charges.

$$f(q) = \begin{cases} 0, & \text{for } q = 0; \\ e^{-\gamma(|q| - 1)^2}, & \text{for } 0 < |q| < 1; \\ 1, & \text{for } |q| \geq 1 \end{cases} \tag{69}$$

where γ, as before, is chosen to be 3.0 such that $f(|q|) \approx 0.5$ for $|q| = 0.5$. This continuous definition of $f(q)$ is identical to the former discrete version with the exception that it allows for noninteger charges on atoms. Inspection of this equation further shows that as the absolute values of charges on any two neighboring atoms approach 1, $|q| \rightarrow 1$, then $f(q) \rightarrow 1$, and thus $N_{\mathrm{CT}} \rightarrow 1$, its maximal value. Therefore, we are equipped to treat noninteger partial charges in a smooth and continuous fashion and give only the maximal value of N_{CT} to systems with integer values of formal charge on neighboring atoms.

Because none of the reactions in Scheme 2 merit the CLOC(x)$_{\mathrm{CT}}$ correction, this specific term has not yet been tested. However, we anticipate its correct behavior on account of the behavior of all of the other similar terms.

III.H. Total CLOC(x). As stated in section III.A, the total CLOC(x) is given by the sum of its constituents:

$$\mathrm{CLOC}(x) = \mathrm{CLOC}(x)_{\mathrm{bond}} + \mathrm{CLOC}(x)_{\mathrm{hyb}} + \mathrm{CLOC}(x)_{\mathrm{radical}} + \mathrm{CLOC}(x)_{\mathrm{hyperval}} + \mathrm{CLOC}(x)_{\mathrm{environ}} + \mathrm{CLOC}(x)_{\mathrm{CT}} \quad (70)$$

Therefore, to arrive at the total CLOC(x) for any arbitrary x, the individual components of this expression are calculated according to the prescriptions given in sections III.B−III.G, and summed over. This CLOC(x) may then be used directly to obtain more accurate relative energies, as described in section III.A.

Once CLOC(x) is known for any arbitrary x, we may define gradients of the B3LYP-CLOC functional. They are given according to the formula

$$\nabla E_{\mathrm{B3LYP-CLOC}}(x) = \nabla[E_{\mathrm{B3LYP}}(x) + \mathrm{CLOC}(x)] = \nabla E_{\mathrm{B3LYP}}(x) + \nabla\mathrm{CLOC}(x) \quad (71)$$

III.I. Computational Methods. All intrinsic reaction coordinate (IRC) scans were performed at the B3LYP/6-31+G** level using the computational package Jaguar 7.6.[9] (In previous LOC publications,[5] all geometry optimizations and transition state searches were performed at the B3LYP/6-31G* level; however, we found that using this slightly larger basis greatly improved the ease with which stationary points could be located, without substantially increasing computational cost.) These geometries were then used to perform single-point energy calculations at the B3LYP/6-311++G(3df,3pd) and M06-2X/6-311++G(3df,3pd) levels, also within Jaguar; at the RCCSD(T)/cc-pVTZ level using MolPro 2006.1;[8] and at the BW2 post-HF level using the code provided by Hans Joachim-Werner for reaction g.[14] Following Joachim-Werner's precedent for this reaction, a mixed basis was used in which chlorine was treated with the aug-cc-pV5Z[8s7p5d4f3g] basis, and hydrogens were treated with the aug-cc-pVQZ[5s4p3d2f] basis.

The potential energy curves obtained with the post-HF, B3LYP, M06-2X, and B3LYP-CLOC methods were aligned by relative energy. Specifically, the energy of the product(s) was subtracted from the energy at every point along the curve such that the energy of the product(s) for all four curves was strictly zero, and all other energies were given with respect to the product(s) energy.

The B3LYP-CLOC curves were generated directly from the B3LYP curves with the simple addition of the numerical CLOC. For example, to compute the B3LYP-CLOC energy for an arbitrary point x on the reaction profile, $E_{\mathrm{B3LYP\text{-}CLOC}}(x)$, the CLOC at point x, CLOC(x), had to be initially obtained. The B3LYP-CLOC energy at point x is then the sum of the B3LYP energy and CLOC.

$$E_{\mathrm{B3LYP}}(x) + \mathrm{CLOC}(x) = E_{\mathrm{B3LYP-CLOC}}(x) \quad (72)$$

Since all reaction profile curves have to be scaled by the subtraction of the product(s) energy, we must know the B3LYP-CLOC energy of the product(s). This is computed similarly to the description given above.

$$E_{\mathrm{B3LYP}}(\mathrm{product}) + \mathrm{CLOC}(\mathrm{product}) = E_{\mathrm{B3LYP-CLOC}}(\mathrm{product}) \quad (73)$$

Finally, the energies along the B3LYP-CLOC curves are given by the difference in $E_{\mathrm{B3LYP\text{-}CLOC}}(x)$ and $E_{\mathrm{B3LYP\text{-}CLOC}}$(product).

$$E^{\mathrm{relative}}_{\mathrm{B3LYP-CLOC}}(x) = E_{\mathrm{B3LYP-CLOC}}(x) - E_{\mathrm{B3LYP-CLOC}}(\mathrm{product}) \quad (74)$$

These are the final points given in all of the graphs and tables in this work.

All numerical CLOCs can be computed with a simple script.[15] As input, a reaction coordinate (defined by reactant, transition state, and product structures) and an arbitrary structure along that coordinate are required, and CLOCs for all four structures are produced in the output.

IV. Results and Discussion

To assess the effectiveness of the CLOC approach, we surveyed reaction profiles for the seven reactions shown in Scheme 2. Three functionals—B3LYP, M06-2X, and B3LYP-CLOC developed herein—were tested against post-HF level calculations, with the resulting plots shown in Figure 6. Table 6 and Figure 7 show the mean unsigned errors (MUEs) along each reaction profile for all reactions and functionals studied. These numbers reflect the disagreement between post-HF and DFT at every point along the reaction profile. The mean value of all of these MUEs, MMUE(overall), is also given.

While MMUE(overall) reflects the performance across the entire reaction coordinate, the performance at the stationary points for most practical applications is more critical than the performance at intermediate points. Therefore, the differences in relative SCF energies at the stationary points, ΔE_{scf}, are compared to the values at the post-HF level in Table 7. Specifically, the differences in transition state and equilibrium SCF energies, ΔE_{scf}(eq→ts), where the equilibrium structure may be either reactant or product, were tabulated in addition to the differences in reactant and product SCF energies, ΔE_{scf}(r→p). The mean unsigned errors in these two values, ΔE_{scf}(eq→ts) and ΔE_{scf}(r→p), across all seven reactions were then tabulated for each DFT method to give the final values shown in Table 7, MUE[ΔE_{scf}(eq→ts)] and MUE[ΔE_{scf}(r→p)].

An inspection of the plots in Figure 6 not surprisingly shows that while the B3LYP curves have the qualitatively correct shape compared to the post-HF standard in many cases, there still remain large quantitative errors at many points along the reaction coordinate. This is also reflected by the relatively large values of MUE[ΔE_{scf}(eq→ts)] and MUE[ΔE_{scf}(r→p)] given in Table 7. Perhaps fortuitously, there are some reactions for which the B3LYP and post-HF curves are nearly convergent, at least for part of the reaction profile, i.e., reaction d, CH_3· + CH_2CH_2. However, there are other curves where serious quantitative disagreement between B3LYP and the post-HF method is observed, i.e., reaction g, H_2+Cl.

Further inspection of Figures 6 and 7 and Tables 6 and 7 shows that, on average, M06-2X outperforms B3LYP across all stationary and intermediate structures. This is reflected by lower values of MUE[ΔE_{scf}(eq→ts)] and MUE[ΔE_{scf}-(r→p)] in Table 7 and MMUE(overall) in Table 6 for M06-2X vs B3LYP. Yet, we see that B3LYP performs as well

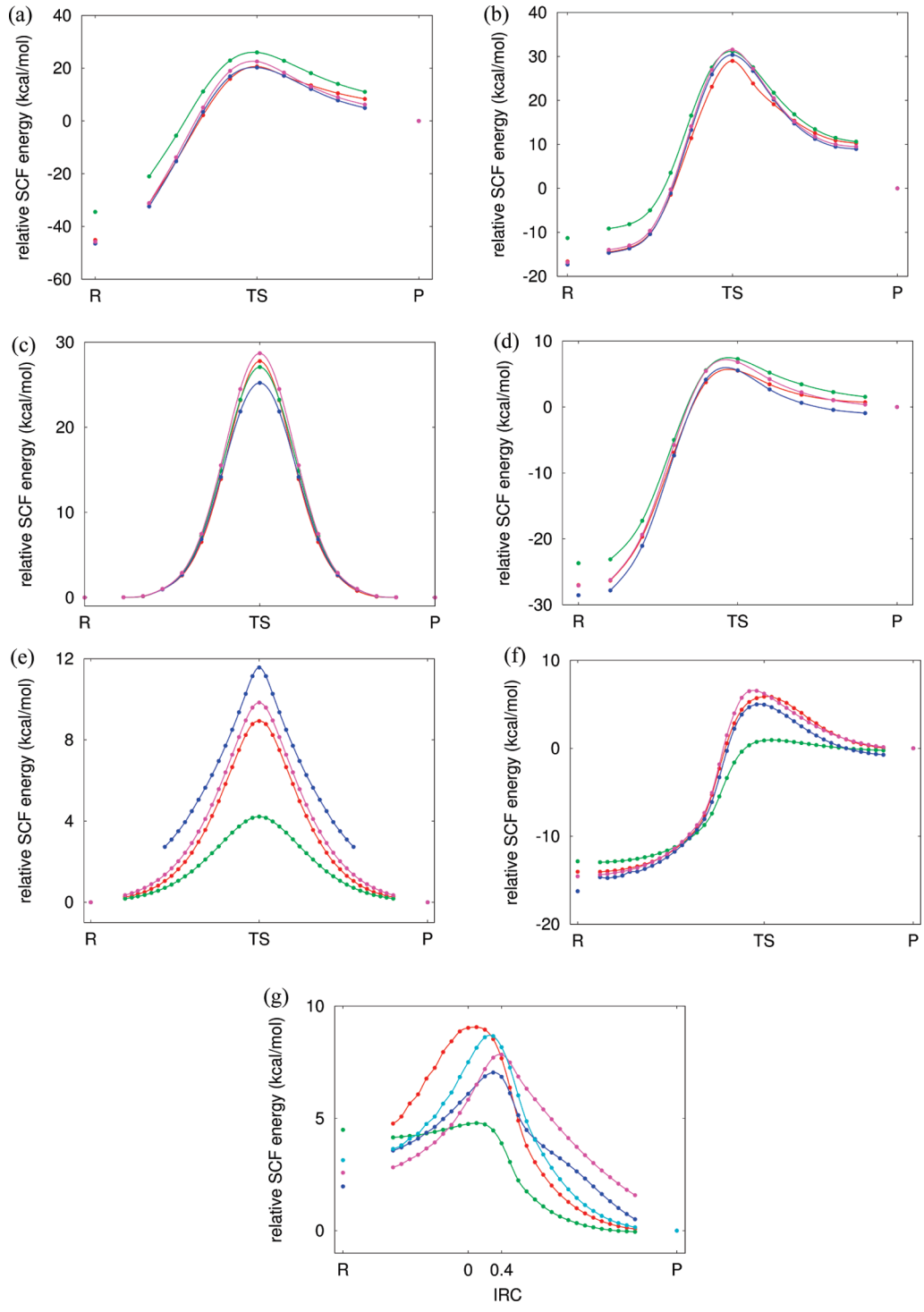


Figure 6. Reaction profile plots for the reactions of Scheme 2. All plots show relative SCF energies with respect to intrinsic reaction coordinate (IRC), where the points on the coordinate are labeled R, TS, and P for reactant, transition state, and product, respectively. The plots are constructed for the following reactions: (a) Diels−Alder, (b) electrocyclic, (c) sigmatropic shift, (d) carbon radical reaction, and hydrogen radical reactions (e) $H_2 + H$, (f) $H_2 + OH$, and (g) $H_2 + Cl$. The curves plotted are B3LYP (green), B3LYP-CLOC (red), post-HF (magenta), and M06-2X (blue). For reaction g, B3LYP-CLOC with a shifted IRC for the transition state is also plotted in teal. Note that in this plot, the B3LYP-optimized transition state occurs at IRC = 0.0, while the post-HF transition state occurs at IRC = 0.4. Post-HF plots for reactions a−f are at the RCCSD(T) level, while the post-HF plot for reaction g is at the BW2 level.

as, or better than, M06-2X for both the sigmatropic shift (reaction c) and carbon radical reaction (reaction d) in Table 6 when we consider the MUE along the entire reaction coordinate.

Similarly to M06-2X, B3LYP-CLOC performs better than B3LYP along the whole reaction coordinate for many different reaction types. The same trend holds for both ΔE_{scf}(eq→ts) and ΔE_{scf}(r→p). [Notably, however, all func-

Table 6. Mean Unsigned Errors (MUEs)[a] in kcal/mol along Entire Reaction Profile

reaction	reaction type	B3LYP	M06-2X	B3LYP-CLOC
a	Diels−Alder	5.69	1.27	1.39
b	electrocyclic	2.30	0.62	1.31
c	sigmatropic shift	1.33	1.35	0.57
d	carbon radical	0.36	0.78	0.56
e	$H_2 + H$	2.66[b]	1.43[b]	0.55[b]
f	$H_2 + OH$	2.12	0.89	0.34
g	$H_2 + Cl$	2.34	0.93	2.28 (1.41[c])
MMUE(overall)[d]		2.40	1.04	1.00 (0.88[c])

[a] The deviation between post-HF and DFT energies at every point along the reaction coordinate was computed. The absolute values of these deviations were then averaged to give the mean unsigned error, MUE, along the entire curve. [b] All data computed using only data along the range $-1.4 \leq$ IRC ≤ 1.4 because M06-2X data points outside this range could not be obtained due to convergence difficulties. [c] Computed using the shifted B3LYP-CLOC data for the reaction $H_2 + Cl$. [d] Mean of MUEs for reactions *a*−*g*, i.e., the values given in the rows directly above.

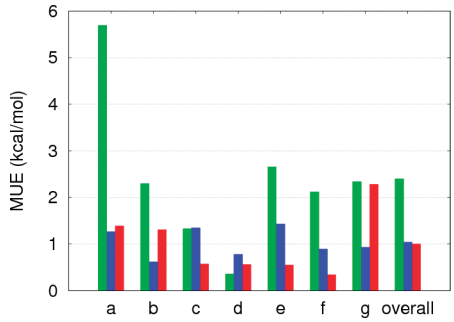


Figure 7. Mean unsigned error (MUE) vs reaction type for all reactions and functionals tested in this study as shown in Table 6. Functionals include B3LYP (green), M06-2X (blue), and B3LYP-CLOC (red). Reactions include (a) Diels−Alder, (b) electrocyclic, (c) sigmatropic shift, (d) carbon radical, (e) $H_2 + H$, (f) $H_2 + OH$, and (g) $H_2 + Cl$.

Table 7. Mean Unsigned Errors (MUEs) in kcal/mol for ΔE_{scf}(eq→ts) and ΔE_{scf}(r→p)

	B3LYP	M06-2X	B3LYP-CLOC
MUE[ΔE_{scf}(eq→ts)][a]	4.1	1.3	1.5
MUE[ΔE_{scf}(r→p)][b]	3.4	0.7	0.3

[a] Mean unsigned error of ΔE_{scf}(eq→ts), where eq may be either equilibrium structure, reactant(s), or product(s), for reactions a−g. Individual errors for ΔE_{scf}(eq→ts) given in Supporting Information. All ΔE_{scf}(eq→ts) were computed using B3LYP geometries. [b] Mean unsigned error of ΔE_{scf}(r→p) for reactions a−g. Individual errors for ΔE_{scf}(r→p) given in Supporting Information. All ΔE_{scf}(r→p) were computed using B3LYP geometries.

tionals tested have difficulty with reaction g, $H_2 + Cl$. This will be discussed further below.] While M06-2X is outperformed by B3LYP for both reactions c and d, B3LYP-CLOC is only outperformed by B3LYP in the case of reaction d. Arguably, this is a case in which B3LYP performs anomalously well. An inspection of Table 6 shows that B3LYP's performance for reaction d is far better than for all the other reactions studied. Additionally, examination of Tables 6 and 7 and Figure 7 reveals that the performance of B3LYP-CLOC rivals that of M06-2X with respect to MUE for the individual cases as well as all cases combined. This is also true for the MUE[ΔE_{scf}(eq→ts)] and MUE[ΔE_{scf}(r→p)], as shown in Table 7.

Table 8. Geometry of $H_2 + Cl$ Transition State

	bond length (Å)	
method	H−H	H−Cl
BW2[a]	1.00	1.46
B3LYP	1.29	1.34
B3LYP-CLOC	1.29	1.34

[a] Taken as the geometry at IRC = 0.4 in Figure 6g.

Overall, the performance of B3LYP-CLOC rivals that of M06-2X. Both methods exhibit similar accuracy over the entire test set, but M06-2X performs appreciably better than B3LYP-CLOC for reactions b and g, whereas B3LYP-CLOC performs appreciably better than M06-2X for reactions c, e, and f. More reactions in addition to those examined in this work will have to be studied before a broad conclusion about the relative performance of these two functionals may be drawn.

Reaction g, as stated previously, is particularly problematic. This exception may be explained at least in part by the fact that the B3LYP and post-HF transition states' geometries differ markedly from one another, as shown in Table 8. In fact, an inspection of the reaction profile in Figure 6g shows that the post-HF transition state actually occurs at IRC = 0.4, where the geometry more closely matches that found by the post-HF transition state search. Interestingly, the same shift of transition state along the reaction coordinate is observed for M06-2X, where the transition state occurs at IRC = 0.3. We observe that all three tested DFT methods—B3LYP, B3LYP-CLOC, and M06-2X—have difficulty in accurately predicting the transition state geometry for this complicated case, where notably M06-2X performs better than B3LYP and B3LYP-CLOC.

Since our algorithm requires reactant, product, and transition state structures as input to handle all intermediate points along the reaction coordinate, and we know that the "true" transition state geometry resembles that at IRC = 0.4, we can instead provide this alternate geometry as the input transition state geometry to our algorithm and therefore shift the location of the transition state on the reaction coordinate accordingly. This is indeed what we have done to produce the teal curve in Figure 6g. Note that all points along the reaction profile now better approximate the post-HF curve. A complete solution to this problem would not require prior knowledge of the post-HF geometry. Specifically, we ultimately seek a method whereby we can produce energy curves as accurate as post-HF methods without prior knowledge of the energy curves or geometries produced by these methods whatsoever.

Disagreement between B3LYP and "true" transition-state geometries is not without precedent. This same behavior is observed for the highly analogous reaction $H_2 + F$, where transition states predicted by B3LYP and post-HF methods differ substantially in bond lengths and angles.[15]

The above analysis suggests that the inability of DFT to treat both these problematic cases stems partly from defective reproduction of geometries predicted with high-level post-HF methods. One possible solution is the alteration of DFT functionals in such a way that they produce geometries more closely matching the post-HF ones, presumably also leading to more accurate energies. This can be achieved via alteration of

the DFT gradients, ∇E_{DFT}. As stated in section III.H, gradients for the B3LYP-CLOC functional are given according to

$$\nabla E_{B3LYP-CLOC}(x) = \nabla[E_{B3LYP}(x) + CLOC(x)] = \nabla E_{B3LYP}(x) + \nabla CLOC(x) \quad (75)$$

In this regard, it is important to consider the relative magnitudes of the two terms on the right-hand side of eq 75. Specifically, $\nabla E_{B3LYP}(x)$ is often much larger than $\nabla CLOC(x)$, which is only ever a few killocalories per mole, such that $\nabla E_{B3LYP\text{-}CLOC}(x) \approx \nabla E_{B3LYP}(x)$. In fact, we see no change whatsoever in the transition state geometries, as $\nabla E_{B3LYP}(x) = \nabla E_{B3LYP\text{-}CLOC}(x)$. This is clear upon inspection of Table 8, where we see that the geometry of the H_2 + Cl transition state is identical for both B3LYP and B3LYP-CLOC. For non-transition-state structures, we expect only very small changes in the B3LYP-CLOC geometries vs B3LYP.

Therefore, the power of the B3LYP-CLOC method does not lie in its ability to produce more accurate geometries, but rather in its ability to produce more accurate energies. This is most useful for reactions where B3LYP already produces reasonably accurate geometries. Fortunately, most reactions of practical interest fall into this category. For example, note that the larger systems employed in this study, such as reactions a−d in Scheme 2, do not suffer from the difficulties encountered with H_2 + Cl and H_2 + F. A particularly attractive characteristic of our method is therefore its ability to deliver highly accurate, yet computationally inexpensive energies for larger systems.

V. Conclusions

In this work, we have shown how simple empirical localized orbital corrections (LOCs) can be generalized to formulate a continuous implementation (CLOC) that is defined throughout a reaction profile. These corrections were applied specifically to the B3LYP functional, as this functional has shown itself most amenable to this correction scheme. The resultant method, B3LYP-CLOC, gives more accurate energetics in comparison to B3LYP, and its accuracy rivals that of M06-2X for the test cases examined. Furthermore, negligible additional computational cost is required over standard B3LYP calculations, and convergence of geometry optimizations is facile. The accuracy is best where B3LYP already produces reasonable geometries and assignment of Lewis structures is straightforward.

Future work will focus on extending this implementation to the treatment of ionic reactions in addition to the neutral reactions studied herein. More reaction profiles should be studied to test the robustness of this method.

Acknowledgment. This work was supported in part by a grant from the National Institute of Health (NIH) training program in Molecular Biophysics (M. L. Hall, T32GM008281) and by the Division of Chemical Sciences, Geosciences, and Biosciences, Office of Basic Energy Sciences of the U.S. Department of Energy through Grant DE-FGO2-903R14162.

Supporting Information Available: Cartesian coordinates for all structures along with their relative energies at the post-HF, B3LYP, M06-2X, and B3LYP-CLOC levels are provided. The complete suite of all LOCs in addition to detailed examples of parameter assignments and their application to a database of over 105 barrier heights is also available. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Kohn, W.; Becke, A. D.; Parr, P. G. *J. Phys. Chem.* **1996**, *100*, 12974–12980.

(2) For a review of DFT, the history of functional development, and an assessment of the performances of various popular functionals, see: Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *J. Phys. Chem. A.* **2007**, *111*, 10439–10452.

(3) (a) Oliveira, G.; Martin, J. M. L.; Proft, F.; Geerlings, P. *Phys. Rev. A* **1999**, *60*, 1034–1045. (b) Izgorodina, E. I.; Brittain, D. R. B.; Hodgson, J. L.; Krenske, E. H.; Lin, C. Y.; Namazian, M.; Coote, M. L. *J. Phys. Chem. A* **2007**, *111*, 10754–10768. (c) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2001**, *105*, 2936–2941.

(4) (a) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103. (b) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. *J. Phys. Chem. A.* **2000**, *104*, 4811–4815. (c) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A.* **2004**, *108*, 6908–6918.

(5) (a) Friesner, R. A.; Knoll, E. H.; Cao, Y. J. *J. Chem. Phys.* **2006**, *125*, 124107. (b) Knoll, E. H.; Friesner, R. A. *J. Phys. Chem. B* **2006**, *110*, 18787–18802. (c) Goldfeld, D. A.; Bochevarov, A. D.; Friesner, R. A. *J. Chem. Phys.* **2008**, *129*, 214105. (d) Rinaldo, D.; Tian, L.; Harvey, N. J.; Friesner, R. A. *J. Chem. Phys.* **2008**, *129*, 164108. (e) Hall, M. L.; Goldfeld, D. A.; Bochevarov, A. D.; Friesner, R. A. *J. Chem. Theory Comput.* **2009**, *5*, 2996–3009.

(6) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(7) (a) Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2004**, *121*, 8187. (b) Polo, V.; Kraka, E.; Cremer, D. *Mol. Phys.* **2002**, *100*, 1771–1790. (c) Grafenstein, J.; Kraka, E.; Cremer, D. *Phys. Chem. Chem. Phys.* **2004**, *6*, 1096–1113. (d) Grafenstein, J.; Kraka, E.; Cremer, D. *J. Chem. Phys.* **2004**, *120*, 524. (e) Cremer, D. *Mol. Phys.* **2001**, *99*, 1899–1940. (f) Polo, V.; Kraka, E.; Cremer, D. *Mol. Phys.* **2002**, *100*, 1771–1790.

(8) *MOLPRO*, version 2006.1; MOLPRO: Cardiff, U.K., 2006.

(9) *Jaguar*, version 7.6; Schrödinger, LLC: New York, 2009.

(10) Bian, W. S.; Werner, H. J. *J. Chem. Phys.* **2000**, *112*, 220.

(11) For reviews of the application of DFT to the study of organic chemistry and biochemistry, see: (a) Roos, G.; Geerlings, P.; Messens, J. *J. Phys. Chem. B.* **2009**, *113*, 13465–13475. (b) Riley, K. E.; Op't Holt, B. T.; Merz, K. M. *J. Chem. Theory Comput.* **2007**, *3*, 407–433.

(12) Zhang, Y. K.; Yang, W. T. *J. Chem. Phys.* **1998**, *109*, 2604.

(13) (a) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372. (b) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(14) Joachim-Werner, H. The $Cl + H_2 \rightarrow HCl + H$ Reaction. http://www.theochem.uni-stuttgart.de/~werner/h2cl/h2cl.html (accessed Jun 1, 2010).

(15) Werner, H.-J.; Kállay, M.; Gauss, J. *J. Chem. Phys.* **2008**, *128*, 034035.

CT100418N

JCTC Journal of Chemical Theory and Computation

Toward an All-Around Semilocal Potential for Electronic Exchange

Micael J. T. Oliveira,*[,†,⊥] Esa Räsänen,*[,‡,§] Stefano Pittalis,[||] and Miguel A. L. Marques[†]

Laboratoire de Physique de la Matière Condensé et Nanostructures, Université Lyon I, CNRS, UMR 5586, Domaine scientifique de la Doua, F-69622 Villeurbanne Cedex, France, European Theoretical Spectroscopy Facility (ETSF), Nanoscience Center, Department of Physics, University of Jyväskylä, FI-40014 Jyväskylä, Finland, and Department of Physics and Astronomy, University of Missouri, Columbia, Missouri 65211, United States

Received August 11, 2010

Abstract: We test local and semilocal approximations of the exchange potential for a variety of systems including atoms, molecules, and atomic chains. In particular, we focus on a recent universal extension of the Becke−Johnson exchange potential [Räsänen, E.; Pittalis, S.; Proetto, C. R. *J. Chem. Phys.* **2010**, *132*, 044112]. It is shown that when this potential is used together with the Becke−Roussel approximation to the Slater potential [Becke, A. D.; Roussel, M. R. *Phys. Rev. A* **1989**, *39*, 3761−3767], a good overall agreement is obtained with experimental and numerically exact results for several systems, and with a moderate computational cost. Thus, this approximation is a very promising candidate in the quest for a simple and all-around semilocal potential.

1. Introduction

Density-functional theory[1,2] (DFT) has become the standard tool both in quantum chemistry and in atomic, molecular, and solid-state physics. The practical applicability of DFT crucially depends on the approximation for the exchange-correlation (xc) energy functional. The "Jacob's ladder" of functionals developed in the past few decades[3] has posed the following well-known problem: by climbing successive rungs of the ladder, one increases the accuracy of the functional, but one also increases substantially the computational burden of the method. Finding a balance between accuracy and efficiency, together with *universality* (which is the ideal ability to deal equally well with any kind of system), has remained a major challenge in DFT.

As the simplest density functionals, occupying the first two rungs of Jacob's ladder, the local density approximation (LDA) and generalized-gradient approximations (GGA) are numerically efficient and surprisingly accurate for many (strongly inhomogeneous) systems. However, both of these families of functionals exhibit well-known failures in the calculation of, e.g., band gaps of semiconductors and insulators,[4] the response to electric fields,[5] etc. The problems are particularly dramatic in systems where long-range interactions play a crucial role, i.e., elongated molecules and atomic chains.[6−13] The main origin for these errors is the wrong (exponential) asymptotic behavior and the lack of derivative discontinuity in the xc potential.

Climbing the ladder further, the optimized-effective-potential (OEP) method[14−16] or its simplification within the Krieger−Li−Iafrate (KLI) approximation[17] provide, in principle, access to the *exact* exchange energy and potential within DFT. Thus, as long as the electronic correlation is not significant, OEP and KLI are free from the failures mentioned above. However, as nonlocal orbital functionals,

* To whom correspondence should be addressed. E-mail: micael@teor.fis.uc.pt (M.J.T.O.); erasanen@jyu.fi (E.R.).
† Université Lyon I.
‡ ETSF.
§ University of Jyväskylä.
|| University of Missouri.
⊥ Present address: Center for Computational Physics, University of Coimbra, Rua Larga, 3004-516 Coimbra, Portugal.

10.1021/ct100448x
Published on Web 11/29/2010

they are computationally demanding and therefore usable only for systems containing a small number of particles.

To bridge the gap between the GGA and OEP, meta-GGAs[18,19] are appealing candidates. They supplement the GGA by further semilocal information through the kinetic-energy density and/or the Laplacian of the density, and, in some cases, also through the paramagnetic current density. Recently, Räsänen, Pittalis, and Proetto[20] (RPP) developed a meta-GGA for the exchange part of the xc potential. The RPP potential introduces a number of important constraints and features (see below) and performs well for, e.g., non-Coulombic systems and atomic chains. It is based on the Becke−Johnson (BJ) potential[21]—a simple meta-GGA close to the OEP accuracy for atoms—but, in contrast to BJ, the RPP potential is fully gauge-invariant, exact for any one-particle system, and has the correct asymptotic behavior for any N-particle system.

Also, other modifications to the BJ potential have been suggested to improve the performance for atomic chains[22] and band gaps.[23] In fact, the latter modification[23] allows the calculation of band gaps of semiconductors and insulators with an error on the same order of GW calculations, but at a very small fraction of the GW computational time.

In this paper, we test the RPP potential,[20] used together with the Becke−Roussel (BR) approximation to the Slater potential,[24] for a large variety of systems. We compare this approximation to the BJ one, also complemented by the BR potential. In order to allow for a comparison to experimental reference data, we have added to the above exchange potentials the correlation within the LDA. We compare the results also against the LB94 potential of van Leeuwen and Baerends[25] (a GGA with correct asymptotic behavior also including correlation). Moreover, for completeness, we include results calculated with standard LDA and GGA functionals. As a reference, we use experimental or high-quality ab initio data. In some cases, the performance of the exchange potentials alone, i.e., without the addition of correlation, is compared to the exact-exchange OEP results. The combination of RPP and BR potentials is found to yield the best overall performance of the tested approximations, and thus it provides a promising step toward an all-around semilocal exchange potential in DFT.

2. Theory

2.1. Exact Exchange. In a majority of atomic, molecular, and solid-state systems, the electronic exchange gives, in absolute terms, a much larger contribution to (most) observables than the correlation. Therefore, in practical applications, the exchange is the most important term to be approximated in the functional. The exact exchange energy in Hartree atomic units (au) is written as

$$E_{\mathrm{x}}[\rho_\sigma] = -\frac{1}{2}\sum_{\sigma=\uparrow,\downarrow}\sum_{j,k=1}^{N_\sigma}\int \mathrm{d}^3r\int \mathrm{d}^3r'\,\frac{\varphi^*_{j\sigma}(\mathbf{r})\,\varphi^*_{k\sigma}(\mathbf{r}')\,\varphi_{j\sigma}(\mathbf{r}')\,\varphi_{k\sigma}(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} \quad (1)$$

and its functional derivative gives the Kohn−Sham (KS) exchange potential as $v_{\mathrm{x}\sigma}(\mathbf{r}) = \delta E_x/\delta\rho_\sigma(\mathbf{r})$. These quantities can be rigorously calculated with the OEP method[14−16] through an integral equation that has to be solved together with the KS equations.

At this point, it is useful to write the (KS) exchange potential as a sum

$$\begin{aligned} v_{\mathrm{x}\sigma}(\mathbf{r}) &= v^{\mathrm{SL}}_{\mathrm{x}\sigma}(\mathbf{r}) + \Delta v^{\mathrm{OEP}}_{\mathrm{x}\sigma}(\mathbf{r}) \\ &= v^{\mathrm{SL}}_{\mathrm{x}\sigma}(\mathbf{r}) + \Delta v^{\mathrm{KLI}}_{\mathrm{x}\sigma}(\mathbf{r}) + \Delta v^{\mathrm{OS}}_{\mathrm{x}\sigma}(\mathbf{r}) \end{aligned} \quad (2)$$

where

$$v^{\mathrm{SL}}_{\mathrm{x}\sigma}(\mathbf{r}) = -\sum_{j,k=1}^{N_\sigma}\int \mathrm{d}^3r'\,\frac{\varphi^*_{j\sigma}(\mathbf{r})\,\varphi^*_{k\sigma}(\mathbf{r}')\,\varphi_{j\sigma}(\mathbf{r}')\,\varphi_{k\sigma}(\mathbf{r})}{\rho_\sigma(\mathbf{r})|\mathbf{r}-\mathbf{r}'|} \quad (3)$$

is the Slater potential, i.e., the average of the Fock potential felt by the electrons, and $\Delta v^{\mathrm{OEP}}_{\mathrm{x}\sigma}(\mathbf{r})$ is the exact (OEP) contribution,[14−16] which can be decomposed into the Krieger−Li−Iafrate[17] (KLI) part and the orbital shifts. Apart from, e.g., atomic chains,[6] the orbital shifts in a ground-state calculation are usually of minor importance and therefore neglected, leading to so-called KLI approximation. This relieves the computational burden of solving the integral equation, but the tedious integrals in the Slater potential are still to be calculated. Therefore, even within the KLI approximation, the efficiency of an OEP calculation is far from that of semilocal functionals.

2.2. Becke−Johnson Potential. The BJ potential[21] is a simple approximation to the OEP contribution in eq 2:

$$\Delta v^{\mathrm{OEP}}_{\mathrm{x}\sigma}(\mathbf{r}) \approx \Delta v^{\mathrm{BJ}}_{\mathrm{x}\sigma}(\mathbf{r}) = C_{\Delta v}\sqrt{\frac{\tau_\sigma(\mathbf{r})}{\rho_\sigma(\mathbf{r})}} \quad (4)$$

where

$$\tau_\sigma(\mathbf{r}) = \sum_{j=1}^{N_\sigma}|\nabla\varphi_{j\sigma}(\mathbf{r})|^2 \quad (5)$$

is (twice) the spin-dependent kinetic-energy density, and $C_{\Delta v} = \sqrt{[5/(12\pi^2)]}$. The BJ potential is exact for the hydrogen atom and for the homogeneous electron gas, and regarding quantum chemistry applications, it has several beneficial properties. First, it yields the atomic step structure in the exchange potential (which was the main motivation for the approximation) very accurately.[21] Second, it has the derivative discontinuity for fractional particle numbers.[22]

To improve the numerical efficiency of this potential, one often also replaces the Slater potential $v^{\mathrm{SL}}_{\mathrm{x}\sigma}(\mathbf{r})$ with the Becke−Roussel potential.[24] This is again a meta-GGA potential, written in terms of $\nabla^2\rho_\sigma$ and of τ_σ, that reproduces to a very high precision the Slater potential for atoms.

2.3. Universal Extension to Becke−Johnson. The main limitations of the BJ potential are that it is not gauge-invariant and that it is not exact for *all* one-particle systems. Both limitations were recently removed in the extension by RPP,[20] which proposed the form

$$\Delta v^{\mathrm{OEP}}_{\mathrm{x}\sigma}(\mathbf{r}) \approx \Delta v^{\mathrm{RPP}}_{\mathrm{x}\sigma}(\mathbf{r}) = C_{\Delta v}\sqrt{\frac{D_\sigma(\mathbf{r})}{\rho_\sigma(\mathbf{r})}} \quad (6)$$

where

$$D_\sigma(\mathbf{r}) = \tau_\sigma(\mathbf{r}) - \frac{1}{4}\frac{|\nabla\rho_\sigma(\mathbf{r})|^2}{\rho_\sigma(\mathbf{r})} - \frac{|\mathbf{j}_{p\sigma}(\mathbf{r})|^2}{\rho_\sigma(\mathbf{r})} \quad (7)$$

describes the local curvature of the exchange (Fermi) hole.[26] This quantity has already been useful in the derivation of several functionals[24,27−31] and is the key ingredient of the electron-localization function,[32−34] a standard tool used to analyze bonding in electronic systems. Finally, the spin-dependent paramagnetic current density is defined as

$$\mathbf{j}_{p\sigma}(\mathbf{r}) = \frac{1}{2i}\sum_{j=1}^{N_\sigma}\{\varphi_{j\sigma}^*(\mathbf{r})[\nabla\varphi_{j\sigma}(\mathbf{r})] - [\nabla\varphi_{j\sigma}^*(\mathbf{r})]\varphi_{j\sigma}(\mathbf{r})\} \quad (8)$$

The RPP approximation is gauge-invariant, and it is exact for *all* one-particle systems. Furthermore, it has a correct asymptotic limit for finite N -electron systems (except on nodal surfaces of the highest occupied orbitals[35,36]). The universality of the approach, whose principles have also been shown to work in two dimensions,[37] is reflected in a resulting potential that can be applied reasonably well to any kind of system. For example, the RPP potential has been seen to reproduce well the KLI potential in hydrogen chains in electric fields and in Hooke's atoms subject to magnetic fields.[20] The present study aims at further evaluating the capability of this approximation for atoms, small molecules, and atomic chains.

3. Numerical Procedure

The evaluation of the Slater part in the BJ[21] and RPP[20] potentials is computationally more demanding than the evaluation of the correction terms $\Delta v_{x\sigma}^{BJ}$ and $\Delta v_{x\sigma}^{RPP}$. Nevertheless, as already pointed out by Becke and Johnson,[21] it is possible to approximate the Slater part by using the semilocal Becke−Roussel (BR) exchange-energy functional.[24] In this way, the cost of evaluating the full BJ and RPP potentials becomes similar to that of a usual LDA or GGA. To avoid any ambiguity, we will hereafter denote the BJ and RPP potentials, where the Slater part was replaced by the BR potential, as BJBR and RPPBR, respectively.

When using experimental results as a reference, it is necessary to add a correlation contribution to the BJBR and RPPBR potentials for a proper comparison. We use the correlation in the LDA level within the Perdew−Wang[38] (PW) form. The results are compared also to the standard LDA—with the PW parametrization for the correlation part; the GGA of Perdew, Burke, and Ernzerhof[39] (PBE); and the GGA of van Leeuwen and Baerends[25] (LB94)—again using the PW parametrization for the LDA part of the potential. In all cases, we have applied the potentials self-consistently in the KS-DFT framework. Although, in the case of PBE, the correlation functional used is not the same as in the other cases, we expect this fact to result in negligible differences in the quantities and systems studied in this work.

In the case of atoms and hydrogen chains, calculations are also performed using *exchange-only* potentials. Results are then compared with exact-exchange OEP data available in the literature. Besides the BJBR and RPPBR potentials, we also performed these calculations using the exchange part of the LDA (xLDA) and of the PBE (xPBE).

It is important to bear in mind that BJ, RPP, and LB94 are such approximations to the exchange (or xc) potential that are not functional derivatives of corresponding exchange (or xc) *energies*.[40] Here, we focus on fairly standard quantities that may be accessed without the computation of total energies. These quantities include ionization potentials and electronic affinities of atoms, ionization potentials and dipole polarizabilities of small molecules, and longitudinal polarizabilities of hydrogen chains. We believe that these benchmarks provide us with a fairly complete view on the properties of different approximations considered in this work.

All of the single-atom calculations are performed with the APE code,[41] while molecules and atomic chains are dealt with the octopus code.[42] In the latter case, the electron−ion interaction is handled through norm-conserving pseudopotentials generated with APE for each functional and approximation studied in this work.

4. Results

4.1. Atoms. First, we consider single atoms and focus on the ionization energies and electron affinities (see Table 1). There are several ways to estimate these quantities within DFT. The most direct one is to calculate the differences in total energy of both the neutral atom and its anion and cation, respectively. In this way, traditional LDA and GGA functionals usually yield quite good ionization potentials. Electron affinities are more complicated as often LDAs and GGAs fail to bind the extra electron.

The other approach, the one used in this work, is to look at the KS eigenenergy of the highest occupied atomic orbital (HOMO), which should be equal to the negative of the ionization potential. The electron affinity is computed simply from the ionization potential of the respective anion. This method samples much better the quality of the potential, and it is particularly sensitive to the asymptotic description of the potential.

As known from previous studies,[25] the LDA and PBE perform poorly for the ionization potential: the mean absolute error (last row of Table 1) is larger than 40% for this set of atoms. The result indicates the crucial role of the correct asymptotic behavior in the exchange potential. The decay of the xc potential is properly described by the LB94 potential showing good performance. For the same reason, good results have been obtained also with KLI-CS—a combination of KLI[17] for the exchange and the Colle−Salvetti[43] functional for the correlation—as reported by Grabo and Gross.[44] It seems that RPPBR-PW is slightly more accurate than the original BJBR-PW potential. When compared against exact-exchange OEP results,[45] xLDA and xPBE perform poorly, while BJBR and RPPBR perform better, the latter again being more accurate.

As noted already by Becke and Johnson,[21] the BJ exchange potential goes asymptotically to a finite (nonzero) constant. In principle, this constant only redefines the zero of orbital energy and should have no implication in the quality of the results, but it has to be taken into account when computing the ionization potential. This can be done by subtracting the

Table 1. Ionization Potentials from the Highest Occupied Kohn−Sham Orbital (in au)[a]

atom	xLDA	xPBE	BJBR	RPPBR	OEP[b]	LDA	PBE	LB94	KLI-CS[c]	BJBR-PW	RPPBR-PW	expt.[d]
He	0.517	0.553	0.857	0.924	0.918	0.570	0.585	0.851	0.945	0.922	0.982	0.903
Li	0.100	0.109	0.254	0.183	0.196	0.116	0.111	0.193	0.200	0.276	0.201	0.198
Be	0.170	0.182	0.355	0.300	0.309	0.206	0.201	0.321	0.329	0.401	0.338	0.343
B	0.120	0.128	0.279	0.226		0.151	0.143	0.296	0.328	0.321	0.260	0.305
C	0.196	0.204	0.399	0.332		0.227	0.218	0.401	0.448	0.440	0.366	0.414
N	0.276	0.285	0.526	0.451	0.571	0.309	0.297	0.510	0.579	0.567	0.486	0.534
O	0.210	0.224	0.391	0.383		0.272	0.266	0.516	0.559	0.472	0.450	0.500
F	0.326	0.339	0.564	0.526		0.384	0.376	0.647	0.714	0.636	0.588	0.640
Ne	0.443	0.456	0.743	0.686	0.851	0.498	0.491	0.788	0.884	0.810	0.745	0.792
Na	0.097	0.103	0.247	0.178	0.182	0.113	0.106	0.205	0.189	0.270	0.197	0.189
Mg	0.142	0.149	0.313	0.252	0.253	0.175	0.168	0.291	0.273	0.357	0.287	0.281
Al	0.086	0.092	0.227	0.160		0.111	0.102	0.216	0.222	0.263	0.188	0.220
Si	0.144	0.150	0.320	0.237		0.170	0.160	0.290	0.306	0.356	0.267	0.300
P	0.203	0.210	0.416	0.324	0.392	0.231	0.219	0.369	0.399	0.453	0.355	0.385
S	0.174	0.182	0.349	0.305		0.229	0.219	0.410	0.404	0.420	0.362	0.381
Cl	0.254	0.262	0.469	0.400		0.305	0.295	0.491	0.506	0.533	0.453	0.477
Ar	0.334	0.343	0.592	0.506	0.591	0.382	0.373	0.577	0.619	0.652	0.557	0.579
Δ(%)	43	41	13.8	8.5		41	42	3.7	5.7	14.4	7.4	

[a] The last row shows the mean absolute error in percentage with respect to exact-exchange and experimental results for exchange potentials and combined exchange and correlation potential, respectively. [b] From the work of Engel and Vosko.[45] [c] From the work of Grabo and Gross.[44] [d] Experimental results taken from Ratzig and Smirnov.[46]

Table 2. Electron Affinities Calculated from the Highest Occupied Kohn−Sham Orbital of the Anion (in au)[a]

atom	LB94	KLI-CS[b]	BJBR-PW	RPPBR-PW	expt.[c]
Li	0.020	0.024		0.036	0.023
B	0.016	0.033			0.010
C	0.049	0.083		0.032	0.046
O	0.077	0.110			0.054
F	0.128	0.208		0.110	0.125
Na	0.023	0.022	0.012	0.036	0.020
Al	0.018	0.024			0.016
Si	0.050	0.065	0.019	0.039	0.051
P	0.061	0.048		0.026	0.027
S	0.098	0.106		0.069	0.076
Cl	0.140	0.174	0.118	0.127	0.133
Δ(%)	29	66	38[d]	28[d]	

[a] The last row shows the mean absolute error in percentage. [b] From the work of Grabo and Gross.[44] [c] Experimental results taken from Ratzig and Smirnov.[46] [d] Mean error calculated for bound solutions only.

value of the constant, which can be obtained from the asymptotic expansion of the density and the kinetic energy density, from the value of the KS eigenenergy of the HOMO. A perfectly equivalent procedure is to shift the BJ exchange potential so that it goes asymptotically to zero. In the case of spin-uncompensated atoms, the constant depends on spin. Then, it is possible to shift the spin-up and spin-down potentials by different amounts, provided that this does not imply a change in the occupancies of the orbitals. In this work, we have chosen to shift the BJ potential when computing ionization potentials and electron affinities. For some selected cases, we also performed calculations without shifting the potential and verified that the results obtained with both methods were identical.

The electron affinities for our set of atoms are given in Table 2. As is well-known, the LDA and most GGAs do not give bound solutions for most negative ions, so we chose not to include them in the table. In most cases, BJBR-PW failed to give bound solutions for the anions, while for RPPBR-PW, this happened only in a few cases. Considering only the cases where RPPBR-PW gave bound solutions, the deviation from the exact values was around 28%. It seems that LB94, having a similar overall accuracy, works better for small ions, whereas RPPBR-PW increases its accuracy for larger systems. For example, for the last three atoms in Table 2 (P, S, Cl), RPPBR-PW has an error of only a few percent. Interestingly, KLI-CS results deviate by more than 60% from the exact values. This might be due to the poor compatibility between the exact nonlocal exchange and the correlation part, when the asymptotic regime is strongly dominated by the ionic HOMO.

Table 3. Ionization Potentials for Molecules Calculated from the Highest Occupied Kohn−Sham Orbital (in eV)[a]

molecule	LDA	PBE	LB94	BJBR-PW	RPPBR-PW	expt.[b]
CS_2	6.93	6.81	11.54	13.08	10.76	10.07
H_2S	6.4	6.3	11.33	12.51	11.05	10.46
C_2H_4	6.92	6.74	11.85	12.71	10.96	10.51
PH_3	6.69	6.64	11.65	12.88	11.62	10.59
NH_3	6.28	6.19	11.55	12.58	11.3	10.8
Cl_2	7.47	7.36	12.3	14.03	11.86	11.48
C_2H_6	8.13	8.15	12.94	15.04	13.33	12
SiH_4	8.53	8.53	13.44	15.44	14.04	12.3
SO_2	8.3	8.09	14.06	15.2	13.29	12.35
H_2O	7.38	7.23	13.2	14.08	12.66	12.62
HCl	8.14	8.04	13.29	14.81	12.83	12.74
N_2O	8.6	8.35	14.48	15.4	13.37	12.89
CH_4	9.46	9.45	14.29	16.69	14.65	13.6
CO_2	9.31	9.05	15.32	16.37	14.2	13.78
CO	9.16	9.09	14.49	16.46	14.47	14.01
H_2	10.28	10.4	15.27	17.92	17.54	15.43
N_2	10.39	10.24	16.94	18.18	16.09	15.58
F_2	9.79	9.54	17.03	17.56	16.18	15.7
HF	9.85	9.65	16.44	17.3	15.69	16.03
Δ(%)	35	36	8.0	19	5.7	

[a] The last row shows the mean absolute error in percentage. [b] Experimental results taken from Grüning et al.[7]

4.2. Molecules. Next, we test the approximations for a large set of small molecules by computing ionization potentials and static (isotropic) dipole polarizabilities. The ionization potentials are obtained from the HOMO as in the previous section, while the polarizabilites are computed as a derivative of the dipole moment of the system with respect to the applied electric field. The ionization potentials are

Table 4. Static (Isotropic) Dipole Polarizabilities for Molecules (in au)[a]

molecule	LDA	PBE	LB94	BJBR-PW	RPPBR-PW	expt.[b]
CS_2	56.50	56.45	51.72	55.44	55.29	55.28
H_2S	26.21	25.91	21.95	24.24	22.51	24.71
C_2H_4	28.71	28.52	24.93	27.71	25.21	27.7
PH_3	32.29	31.72	27.39	29.99	27.47	30.93
NH_3	15.58	15.45	12.41	13.83	12.32	14.56
Cl_2	32.33	32.21	30.92	31.41	32.39	30.35
C_2H_6	30.17	29.73	27.41	28.23	26.52	29.61
SiH_4	34.03	33.07	30.17	31.11	28.47	31.9
SO_2	27.44	27.53	22.97	25.78	23.68	25.61
H_2O	10.74	10.73	8.28	9.49	8.53	9.64
HCl	18.61	18.47	15.85	17.18	16.21	17.39
N_2O	20.7	20.74	17.42	19.46	18.47	19.7
CH_4	17.77	17.45	15.87	16.46	15.41	17.27
CO_2	18.21	18.24	15.66	17.39	16.16	17.51
CO	13.91	13.87	11.6	13.13	12.29	13.08
H_2	5.87	5.64	5.02	5.27	4.56	5.43
N_2	12.64	12.63	10.79	11.9	11.4	11.74
F_2	8.86	8.97	7.23	8.31	7.73	8.38
HF	6.23	6.27	4.8	5.52	4.89	5.6
Δ(%)	6.1	5.3	9.8	2.0	8.9	

[a] The last row shows the mean absolute error in percentage. [b] Experimental results taken from Grüning et al.[7]

listed in Table 3. Interestingly, RPPBR-PW is significantly more accurate than BJBR-PW and deviates less than 6% from the experimental values. LB94 performs also well with a mean absolute error of 8%. In contrast, the LDA and PBE fail in a similar fashion as in the atomic cases considered in the previous section.

For static (isotropic) dipole polarizabilities (see Table 4), the situation is different in the sense that the LDA and PBE perform rather well, which is surprising in view of the fact that the polarization is largely a nonlocal and collective effect. It is noteworthy, however, that the present test set does not include problematic elongated molecules or chains (see next section), for which going beyond LDA (and GGA) is essential.[6−13] For the present cases, BJBR-PW works remarkably well with a mean error of only 2%, whereas RPPBR-PW and LB94 deviate almost 10% from the experiments. Nevertheless, no dramatic failures are obtained by using any of the tested approximations.

4.3. Hydrogen Chains. In Table 5, we show the polarizabilities calculated for hydrogen chains from H_2 up to H_{20}. As the reference results, we use available data from CCSD(T) (coupled-cluster with single and double and perturbative triple excitations) and MP4 (fourth-order Møller−Plesset perturbation theory).[11] This well-studied system has proved to be a remarkable challenge for DFT.[6,11−13,22] For example, LDA severely overestimates the polarizability, as demonstrated also by our results in Table 5. The error of PBE is slightly smaller. The failure of LDA and PBE to capture the electric response is believed to be due to the inherent self-interaction error.[11,48,49] We find that the mean error of LB94 is almost the same as that of LDA, whereas for BJBR-PW it is smaller. RPPBR-PW has the best performance of all of the tested potentials when compared to MP4, although the mean error is still quite large (27.7%). Possible sources of error in RPPBR-PW (and BJBR-PW) results are the ultranonlocal effects in long chains, which might be beyond reach of any semilocal functionals without *ad hoc* modifications, and the use of LDA for the correlation part. This last point seems to be confirmed by the results obtained without adding a correlation part to the exchange potentials: when comparing the polarizabilities obtained from the exchange-only potentials against exact-exchange OEP results,[6] all of the average errors are reduced, while the overall trend remains the same.

5. Summary and Outlook

In summary, we have tested recently constructed meta-generalized-gradient (meta-GGA) functionals for the exchange potential, in particular, the potential of Räsänen, Pittalis, and Proetto (RPP) and that of Becke and Johnson (BJ), when complemented by the Becke−Roussel (BR) approximation to the Slater potential (denoted in total as RPPBR and BJBR, respectively) and by the correlation in the LDA level. These approximations were compared to the van Leeuwen and Baerends potential (LB94), a GGA that shares some properties with these new meta-GGAs, as well as to standard LDA and GGA functionals. As the reference data, we used experimental results whenever available, numerically exact data, and, in the case of comparing the exchange-only results, the exact-exchange results obtained from the optimized-effective-potential method.

Overall, the RPPBR potential fared best in the present test suite consisting of ionization potentials and electronic affinities of atoms, ionization potentials and dipole polarizabilities of small molecules, and longitudinal polar-

Table 5. Longitudinal Polarizabilities of Hydrogen Chains (in au)[a]

chain	xLDA	xPBE	BJBR	RPPBR	OEP[b]	LDA	PBE	LB94	BJBR-PW	RPPBR-PW	CCSD(T)[c]	MP4[c]
H_2	13.1	12.5	12.4	11.2		12.4	12.0	11.2	11.8	10.8		
H_4	39.6	37.2	36.3	33.3	32.2	37.7	36.1	35.5	34.9	32.4	29	29.5
H_6	76.4	70.7	68.6	63.6	65.6	72.9	69.4	70.5	65.8	61.6	50.9	51.9
H_8	120.6	110.2	106.0	99.0	84.2	115.2	108.8	112.9	101.6	95.8	74.4	76.2
H_{10}	169.9	153.2	146.1	137.1		162.2	152.1	160.5	140.8	132.7		
H_{12}	222.4	199.2	188.4	177.2	138.1	212.2	197.8	211.6	182.1	171.1	124	127.3
H_{14}	277.0	246.1	231.9	218.0		264.0	245.2	264.3	224.1	210.6		155
H_{16}	333.0	294.1	277.5	259.5		317.2	293.4	318.6	267.3	250.5		
H_{18}	389.8	342.5	323.0	301.5		371.1	342.2	373.2	309.8	290.8		205.39
H_{20}	447.3	391.4	367.2	343.6		425.4	391.4	425.0	353.9	331.3		
Δ(%)	40.6	28.9	24.0	15.4		56.2	46.5	53.8	36.2	27.7		

[a] The last row shows the mean absolute error in percentage, calculated against OEP and MP4 (when available) for exchange only potentials and combined exchange and correlation potentials, respectively. [b] Results from the work of Kümmel et al.[6] [c] The MP4 and CCSD(T) results have been taken from the work of Ruzsinszky et al.[11] apart from the MP4 result for H_{18} taken from Champagne et al.[47]

izabilities of hydrogen chains. The LB94 potential performed in an appealing fashion in several instances. The BJBR potential gave particularly good results for the calculation of static polarizabilities of small molecules. Desired future developments would include the development of correlation potentials compatible with the RPRBR potential.

In conclusion, the RPPBR potential combines a proper theoretical foundation with very good results for a series of properties of atoms and molecules. Moreover, it is very light from the computational point of view, thus allowing an efficient calculation of large systems. Therefore, we believe that the RPPBR potential is an important step in the quest for a simple and all-around semilocal potential for applications of density-functional theory.

Acknowledgment. This work was supported by the Academy of Finland, and the EU's Sixth Framework Programme through the ETSF e-I3. S.P. acknowledges support by DOE grant DE-FG02-05ER46203. M.J.T.O. thankfully acknowledges financial support from the Portuguese FCT (contract #SFRH/BPD/44608/2008). M.A.L.M. acknowledges partial funding from the French ANR (ANR-08-CEXC8-008-01) and from the program PIR Matériaux—MaProSu of CNRS. Part of the calculations were performed at the LCA of the University of Coimbra and at GENCI (project x2010096017).

References

(1) Dreizler, R. M.; Gross, E. K. U. *Density Functional Theory*; Springer: Berlin, 1990.

(2) von Barth, U. *Phys. Scr.* **2004**, *109*, 9.

(3) Perdew, J. P.; Kurth, S. In *A Primer in Density Functional Theory*; Fiolhais, C., Nogueira, F., Marques, M. A. L., Eds.; Springer: Berlin, 2003; Vol. 620, pp 1−55.

(4) Heyd, J.; Peralta, J. E.; Scuseria, G. E.; Martin, R. L. *J. Chem. Phys.* **2005**, *123*, 174101.

(5) Champagne, B.; Perpete, E. A.; van Gisbergen, S. J. A.; Baerends, E.-J.; Snijders, J. G.; Soubra-Ghaoui, C.; Robins, K. A.; Kirtman, B. *J. Chem. Phys.* **1998**, *109*, 10489–10498.

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

(7) Grüning, M.; Gritsenko, O. V.; van Gisbergen, S. J. A.; Baerends, E. J. *J. Chem. Phys.* **2002**, *116*, 9591–9601.

(8) Mori-Sánchez, P.; Wu, Q.; Yang, W. *J. Chem. Phys.* **2003**, *119*, 11001.

(9) Maitra, N. T.; van Faassen, M. *J. Chem. Phys.* **2007**, *126*, 191106.

(10) Karolewski, A.; Armiento, R.; Kümmel, S. *J. Chem. Theory Comput.* **2009**, *5*, 712–718.

(11) Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I.; Scuseria, G. E.; Vydrov, O. A. *Phys. Rev. A* **2008**, *77*, 060502(R).

(12) Messud, J.; Wang, Z.; Dinh, P. M.; Reinhard, P.-G.; Suraud, E. *Chem. Phys. Lett.* **2009**, *479*, 300–305.

(13) Grüning, M.; Gritsenko, O. V.; Baerends, E. J. *J. Chem. Phys.* **2002**, *116*, 6435–6442.

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

(15) Talman, J. D.; Shadwick, W. F. *Phys. Rev. A* **1976**, *14*, 36–40.

(16) Kümmel, S.; Kronik, L. *Rev. Mod. Phys.* **2008**, *80*, 3–60.

(17) Krieger, J. B.; Li, Y.; Iafrate, G. J. *Phys. Rev. A* **1992**, *45*, 101–126.

(18) Perdew, J. P.; Kurth, S.; Zupan, A.; Blaha, P. *Phys. Rev. Lett.* **1999**, *82*, 2544–2547.

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

(20) Räsänen, E.; Pittalis, S.; Proetto, C. R. *J. Chem. Phys.* **2010**, *132*, 044112.

(21) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2006**, *124*, 221101.

(22) Armiento, R.; Kümmel, S.; Körzdörfer, T. *Phys. Rev. B* **2008**, *77*, 165106.

(23) Tran, F.; Blaha, P. *Phys. Rev. Lett.* **2009**, *102*, 226401.

(24) Becke, A. D.; Roussel, M. R. *Phys. Rev. A* **1989**, *39*, 3761–3767.

(25) van Leeuwen, R.; Baerends, E. J. *Phys. Rev. A* **1994**, *49*, 2421–2431.

(26) Dobson, J. F. *J. Chem. Phys.* **1993**, *98*, 8870–8872.

(27) Becke, A. D. *J. Chem. Phys.* **1988**, *88*, 1053–1062.

(28) Becke, A. D. *Can. J. Chem.* **1996**, *74*, 995–997.

(29) Pittalis, S.; Räsänen, E.; Helbig, N.; Gross, E. K. U. *Phys. Rev. B* **2007**, *76*, 235314.

(30) Pittalis, S.; Räsänen, E.; Proetto, C. R.; Gross, E. K. U. *Phys. Rev. B* **2009**, *79*, 085316.

(31) Räsänen, E.; Pittalis, S.; Proetto, C. R. *Phys. Rev. B* **2010**, *81*, 195103.

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

(33) Burnus, T.; Marques, M. A. L.; Gross, E. K. U. *Phys. Rev. A* **2005**, *71*, 010501(R).

(34) Räsänen, E.; Castro, A.; Gross, E. K. U. *Phys. Rev. B* **2008**, *77*, 115108.

(35) Sala, F. D.; Görling, A. *Phys. Rev. Lett.* **2002**, *89*, 033003.

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

(37) Pittalis, S.; Räsänen, E.; Proetto, C. R. *Phys. Rev. B* **2010**, *81*, 115108.

(38) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244–13249.

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

(40) Gaiduk, A. P.; Staroverov, V. N. *J. Chem. Phys.* **2009**, *131*, 044107.

(41) Oliveira, M. J. T.; Nogueira, F. *Comput. Phys. Commun.* **2008**, *178*, 524–534.

(42) (a) Marques, M. A. L.; Castro, A.; Bertsch, G. F.; Rubio, A. *Comput. Phys. Commun.* **2003**, *151*, 60–78. (b) Castro, A.; Appel, H.; Oliveira, M.; Rozzi, C. A.; Andrade, X.; Lorenzen, F.; Marques, M. A. L.; Gross, E. K. U.; Rubio, A. *Phys. Stat. Sol. B* **2006**, *243*, 2465–2488.

(43) Colle, R.; Salvetti, O. *Theor. Chim. Acta* **1975**, *37*, 329–334.

(44) Grabo, T.; Gross, E. K. U. *Chem. Phys. Lett.* **1995**, *240*, 141–150.

(45) Engel, E.; Vosko, S. H. *Phys. Rev. A* **1993**, *47*, 2800−2811.

(46) Radzig, A. A.; Smirnov, B. M. *Reference Data on Atoms and Molecules*; Springer Verlag: Berlin, 1985.

(47) Champagne, B.; Mosley, D. H.; Vrako, M.; André, J.-M. *Phys. Rev. A* **1995**, *52*, 178–188.

(48) Pemmaraju, C. D.; Sanvito, S.; Burke, K. *Phys. Rev. B* **2008**, *77*, 121204(R).

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

CT100448X

JCTC Journal of Chemical Theory and Computation

Computing Second-Order Functional Derivatives with Respect to the External Potential

Nick Sablon,*[,†,‡] Frank De Proft,[†] Paul W. Ayers,[§] and Paul Geerlings[†]

Eenheid Algemene Chemie, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium (Member of the QCMM Ghent-Brussels Alliance Group), Aspirant of the Research Foundation-Flanders (FWO-Vlaanderen), Egmontstraat 5, 1000 Brussels, Belgium, and Department of Chemistry, McMaster University, Hamilton, Ontario L8S 4M1, Canada

Received August 16, 2010

Abstract: Following the increasing interest in the higher-order (functional) derivatives of conceptual density functional theory, we developed and implemented a method for calculating second-order functional derivatives with respect to the external potential. Our method is theoretically exact but involves two numerical approximations: the functional derivatives are expanded in a basis set, and the values of the corresponding expansion coefficients are determined by probing the molecular environment by a finite set of external potential perturbations. Exact solutions are obtained only in the limit of a complete basis set and an infinite number of distinct perturbations. We use this method to compute the atom-condensed linear response kernel for a series of six molecules and show that the results are comparable to the ones obtained by a previously proposed, approximate approach from second-order perturbation theory. The numerical error of the current implementation is about 0.01 au. Because the present method gives exact or quasi-exact solutions, it can be used as a benchmark against which approximate approaches are assessed.

1. Introduction

The interpretation of chemical reactivity on the basis of response functions is the central concern of conceptual or chemical density functional theory (DFT).[1−5] Concepts of chemical relevance are defined as derivatives of the electronic energy E with respect to the number of electrons N or as functional derivatives of E with respect to the external potential $v(\mathbf{r})$, which is the electron−nuclear potential for isolated systems (in au):

$$v(\mathbf{r}) = -\sum_{\alpha} \frac{Z_\alpha}{|\mathbf{r} - \mathbf{R}_\alpha|} \tag{1}$$

The sum runs over all the atomic nuclei with nuclear charges Z_α and positions $\mathbf{R}_\alpha$. Fundamental chemical reactivity indicators include the electronic chemical potential μ,[6,7] the chemical hardness η,[8,9] the electron density $\rho(\mathbf{r})$, and the Fukui function $f(\mathbf{r})$,[10,11] which have been defined as

$$\mu = \left(\frac{\partial E}{\partial N}\right)_{v(\mathbf{r})} \tag{2}$$

$$\eta = \left(\frac{\partial^2 E}{\partial N^2}\right)_{v(\mathbf{r})} \tag{3}$$

$$\rho(\mathbf{r}) = \left(\frac{\delta E}{\delta v(\mathbf{r})}\right)_{N} \qquad \text{and} \tag{4}$$

$$f(\mathbf{r}) = \left(\frac{\partial}{\partial N}\left(\frac{\delta E}{\delta v(\mathbf{r})}\right)_{N}\right)_{v(\mathbf{r})} \tag{5}$$

These and other derived quantities characterize the chemical behavior of individual molecules by assessing their response to model perturbations without the explicit description of the partner reagents. Such chemical reactivity indicators have been applied in many studies in order to interpret both

* Address correspondence to Nick.Sablon@vub.ac.be.
† Vrije Universiteit Brussel.
‡ Research Foundation−Flanders.
§ McMaster University.

10.1021/ct1004577
Published on Web 11/10/2010

theoretical and experimental data on various types of chemical reactions.[3]

The mixed derivative of eq 5, the Fukui function, can be written in either of two ways:

$$f^{\pm}(\mathbf{r}) = \left(\frac{\partial \rho(\mathbf{r})}{\partial N}\right)^{\pm}_{v(\mathbf{r})} \qquad \text{or} \qquad (6)$$

$$f^{\pm}(\mathbf{r}) = \left(\frac{\delta \mu^{\pm}}{\delta v(\mathbf{r})}\right)_N \qquad (7)$$

The "±" sign indicates that the derivatives with respect to the electron number should be evaluated from the left- or right-hand side. As shown by Perdew et al.[7] in a zero-temperature grand canonical ensemble framework, these derivatives of the electronic energy E and electron density $\rho(\mathbf{r})$ are discontinuous at integer electron number (which is always the case for isolated systems). The most common approach for evaluating the Fukui functions is based on a finite difference approximation of eq 6:

$$f^{-}_{N_0}(\mathbf{r}) = \rho_{N_0}(\mathbf{r}) - \rho_{N_0-1}(\mathbf{r}) \qquad \text{and} \qquad (8)$$

$$f^{+}_{N_0}(\mathbf{r}) = \rho_{N_0+1}(\mathbf{r}) - \rho_{N_0}(\mathbf{r}) \qquad (9)$$

where the Fukui functions of a system consisting of N_0 electrons are obtained in terms of the electron densities of the N_0, $(N_0 - 1)$, and $(N_0 + 1)$ electron systems. Although relations (eqs 8 and 9) are exact for solutions to the Schrödinger equation, they are inexact for most approximate computational methods.[11−13] This observation led the present authors to explore a different route, namely the calculation of the functional derivative of the chemical potential with respect to the external potential (eq 7).[14,15] A general numerical procedure to compute the first-order functional derivative of any quantity with respect to the external potential was developed for this purpose; the basic idea is to use the computed responses of the quantity under consideration upon external potential perturbations to calculate the expansion coefficients for the desired functional derivative. This methodology has been extensively analyzed and applied in a series of papers. In the first paper,[14] the theoretical background to this methodology was presented, focusing on the accurate calculation of the Fukui function for the beryllium atom and formaldehyde molecule. A second contribution[15] dealt with the calculation of atom-condensed Fukui functions for a range of molecules, including monosubstituted benzenes. A detailed study of the locally resolved Fukui function and dual descriptor was provided in a third contribution.[16] The most recent paper in this series[17] employed the approach in the reactivity description of alkaline earth metal oxide clusters, thus avoiding periodic boundary condition calculations for this kind of system, and introduced the concept of the molecular orbital-averaged Fukui function, which takes the reactivity information of various molecular orbitals (MOs) into account.

The purpose of this article is the calculation of second-order functional derivatives with respect to the external potential. Building upon our previous work, we will present the necessary theory for and practical implementation of the methodology. The numerical results in this contribution concentrate on the linear response or polarizability kernel $\chi(\mathbf{r},\mathbf{r}')$,[18,19] defined as

$$\chi(\mathbf{r},\mathbf{r}') = \left(\frac{\delta^2 E}{\delta v(\mathbf{r})\delta v(\mathbf{r}')}\right)_N = \left(\frac{\delta \rho(\mathbf{r})}{\delta v(\mathbf{r}')}\right)_N \qquad (10)$$

and, more specifically, on the direct calculation of the atom-condensed linear response. The computation of other quantities is, however, within our reach. One could think, for example, of the Fukui kernels $f^{\pm}(\mathbf{r},\mathbf{r}')$:[19−21]

$$f^{\pm}(\mathbf{r},\mathbf{r}') = \left(\frac{\delta^2 \mu^{\pm}}{\delta v(\mathbf{r})\delta v(\mathbf{r}')}\right)_N = \left(\frac{\delta f^{\pm}(\mathbf{r})}{\delta v(\mathbf{r}')}\right)_N \qquad (11)$$

which take polarization effects on the Fukui functions into account, as the linear response kernel does for the electron density.

Several authors have been showing an increased interest in the higher-order (functional) derivatives of conceptual DFT.[22−26] In the case of the linear response kernel, this has led to a number of papers devoted to the formulation of its theoretical properties and formal solutions. The fundamental role of the linear response kernel is highlighted by the Berkowitz−Parr equation:[18]

$$\chi(\mathbf{r},\mathbf{r}') = -s(\mathbf{r},\mathbf{r}') + \frac{s(\mathbf{r})s(\mathbf{r}')}{S} \qquad (12)$$

which relates the linear response kernel to the softness kernel $s(\mathbf{r},\mathbf{r}')$, the local softness $s(\mathbf{r})$, and the global softness S. The softness kernel, which is defined as

$$s(\mathbf{r},\mathbf{r}') = -\frac{\delta \rho(\mathbf{r})}{\delta u(\mathbf{r}')} \qquad (13)$$

is, in turn, the inverse of the hardness kernel $\eta(\mathbf{r},\mathbf{r}')$:[18,27]

$$\int s(\mathbf{r},\mathbf{r}')\eta(\mathbf{r}',\mathbf{r}'')\mathrm{d}\mathbf{r}' = \delta(\mathbf{r} - \mathbf{r}'') \qquad (14)$$

which is ultimately connected to the local hardness $\eta(\mathbf{r})$.[28−32] Senet[19] derived exact functional relations between the linear and nonlinear response functions and the ground-state electron density in terms of the universal Hohenberg−Kohn functional $F[\rho]$. Theoretical expressions for and the mutual relations between the linear and nonlinear responses and the softness and hardness kernels have been elaborated within a Kohn−Sham (KS) formalism.[33−35] The linear response kernel can also be obtained as the zero-frequency limit of the dynamic linear response kernel in time-dependent DFT.[1] A recent paper by Liu et al.[36] summarizes the most important mathematical properties of $\chi(\mathbf{r},\mathbf{r}')$.

There are very few numerical results on the linear respone function in the context of conceptual DFT, and they are typically obtained in approximate manners, without explicitly evaluating the second-order functional derivative. We should mention the studies of Baekelandt et al.[37] and Wang et al.,[38] in which atom-condensed linear response matrices are calculated within the context of the electronegativity equalization method (EEM).[39] Numerical data are also given in some papers by Morita and Kato,[40−42] where this quantity is obtained by solving the coupled-perturbed Hartree−Fock

(HF) or KS equations. Their work is based on the ideas of Stone and Alderton,[43] who analyzed dipole and multipole polarizabilities and has been extended by Yang et al.[44] The explicit calculation of the response $\Delta\rho(\mathbf{r})$ of the electron density upon a point charge perturbation in the external potential has been analyzed for various atoms.[45] Related work has also been done by Cedillo et al.[46] Cioslowski and Martinov have calculated approximate atomic softness matrices, which can be interpreted as the negatives of approximate linear response matrices.[47] Some of the present authors calculated the atom-condensed linear response kernel within the context of second-order perturbation theory and proposed a sound basis for its chemical interpretation:[48,49] They have shown that the linear response kernel measures the extent of electron delocalization, providing a way to differentiate between inductive, resonance, and hyperconjugation effects.

This paper is organized as follows: the theoretical background for the methodology to evaluate second-order functional derivatives will be given in Section 2; Section 3 is concerned with a detailed description of the computational algorithm that implements the presented theory; and numerical data are given in Section 4, where some relevant test systems are analyzed in order to illustrate and validate the proposed methodology. Some final theoretical considerations are made in the Appendix.

2. Theoretical Background

One could seek to formulate analytic expressions for second-order functional derivatives of a given property with respect to the external potential corresponding to a specific theoretical level (e.g., HF or KS DFT with a certain exchange-correlation functional). We will, however, develop a numerical procedure that is able to evaluate these functional derivatives of any property P independently from the theoretical level. The starting point for our approach is provided by a functional Taylor series[4] of P for which the initial (or unperturbed) external potential $v(\mathbf{r})$ is perturbed by $w(\mathbf{r})$. This yields the next expression, where third- and higher-order terms in the norm of the perturbation are neglected:

$$P[v(\mathbf{r}) + w(\mathbf{r})] = P[v(\mathbf{r})] + \int \left(\frac{\delta P}{\delta v(\mathbf{r})}\right)_N w(\mathbf{r})\mathrm{d}\mathbf{r} + \frac{1}{2}\iint \left(\frac{\delta^2 P}{\delta v(\mathbf{r})\delta v(\mathbf{r}')}\right)_N w(\mathbf{r})w(\mathbf{r}')\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' + \mathcal{O}(||w(\mathbf{r})||^3) \quad (15)$$

The functional derivatives, $(\delta P/\delta v(\mathbf{r}))_N$ and $(\delta^2 P/\delta v(\mathbf{r})\delta v(\mathbf{r}'))_N$, are evaluated at the unperturbed external potential $v(\mathbf{r})$. So long as the perturbation $w(\mathbf{r})$ is small enough, this second-order truncation does not introduce a significant error. Construction of an analogous equation for an external potential perturbation of $-w(\mathbf{r})$ and addition to eq 15 gives

$$P[v(\mathbf{r}) + w(\mathbf{r})] - 2P[v(\mathbf{r})] + P[v(\mathbf{r}) - w(\mathbf{r})] = \iint \left(\frac{\delta^2 P}{\delta v(\mathbf{r})\delta v(\mathbf{r}\prime)}\right)_N w(\mathbf{r})w(\mathbf{r}')\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \quad (16)$$

We will now expand the second-order functional derivative in a basis set $\{\beta_k(\mathbf{r})\}_{k=1}^K$ as follows

$$\left(\frac{\delta^2 P}{\delta v(\mathbf{r})\delta v(\mathbf{r}\prime)}\right)_N = \sum_{k=1}^{K}\sum_{l=1}^{K} q_{kl}\beta_k(\mathbf{r})\beta_l(\mathbf{r}') \quad (17)$$

This basis set expansion is mathematically rigorous if the quantity P is the electronic energy or one of its first- or higher-order (functional) derivatives with respect to the electron number or external potential.[14,50] If the set $\{\beta_k(\mathbf{r})\}_{k=1}^{\infty}$ spans the function space of electron densities, eq 17 converges toward the exact functional derivative as $K \rightarrow \infty$.[14] To determine the expansion coefficients, q_{kl}, insert eq 17 into eq 16 and consider a set of external potential perturbations $\{w_j(\mathbf{r})\}_{j=1}^J$, $J \geq K$, instead of the single perturbation $w(\mathbf{r})$ in eq 16. A set of simultaneous linear equations results, which can be solved for the expansion coefficients, and thus enables the calculation of the desired second-order functional derivative. This set of equations can be written as

$$\mathbf{D} = \mathbf{BQ} \quad (18)$$

The J-dimensional column matrix $\mathbf{D}$ consists of the responses of quantity P upon the various external potential perturbations:

$$D_j = P[v(\mathbf{r}) + w_j(\mathbf{r})] - 2P[v(\mathbf{r})] + P[v(\mathbf{r}) - w_j(\mathbf{r})], \quad \text{with} \quad j = 1, ..., J \quad (19)$$

A vital requirement for our methodology is that these responses can be obtained, which is rarely an obstacle since most quantum chemical program packages can provide the necessary information. The $J \times K^2$ matrix $\mathbf{B}$ is comprised of the integrals over the various basis functions and the external potential perturbations:

$$\mathbf{B}_{j,(k-1)K+l} = \int \beta_k(\mathbf{r})w_j(\mathbf{r})\mathrm{d}\mathbf{r} \int \beta_l(\mathbf{r}')w_j(\mathbf{r}')\mathrm{d}\mathbf{r}', \quad \text{with } j = 1, ...,J \text{ and } k, l = 1, ..., K \quad (20)$$

These integrals can be evaluated analytically with the chosen $\{\beta_k(\mathbf{r})\}_{k=1}^K$ and $\{w_j(\mathbf{r})\}_{j=1}^J$ (vide infra). The K^2-dimensional column matrix $\mathbf{Q}$, finally, contains the expansion coefficients for the second-order functional derivative:

$$\mathbf{Q}_{(k-1)K+l} = q_{kl}, \quad \text{with } k, l = 1, ..., K \quad (21)$$

The set of eq 18 can be solved through a linear least-squares fitting procedure, as the number of external potential perturbations (J) will exceed the number of expansion coefficients to be determined (K^2). Indeed, a large value for J is required to ensure that enough information about the molecule's responses is collected to calculate a reliable second-order functional derivative.

So far, the general theoretical framework has been outlined. As might be expected, the numerical results are also dependent upon some practical considerations; particularly upon the manner in which the external potential perturbations are modeled and upon the basis set used for the expansion of the second-order functional derivative. These specifications form the subject of the following section.

3. Computational Method

In order to ensure the reproducibility of our results, the practical side of the computational algorithm will be detailed here.

3.1. External Potential Perturbations. One of the key elements is the type of external potential perturbations that are considered. We chose to perturb the external potential by point charges so that the $\{w_j(\mathbf{r})\}_{j=1}^{J}$ is given by

$$w_j(\mathbf{r}) = \frac{-q_j}{|\mathbf{r} - \mathbf{R}_j|} \tag{22}$$

with q_j the charge values and $\mathbf{R}_j$ the positions of the point charges. This type of perturbation has already proven its usefulness for calculating first-order functional derivatives. Moreover, the utility of point-charge perturbations for elucidating a molecule's reactivity is becoming increasingly apparent due to recent theoretical and computational developments.[46,51−53]

Two molecular regions are defined; perturbations are placed on a cubic grid within these regions to ensure complete and uniform sampling. We consider three scaled van der Waals surfaces (with scaling factors: $R_{\min} < R_{\text{middle}} < R_{\max}$) to designate an inner molecular region, extending between the van der Waals surfaces scaled by the factors $R_{\min}$ and R_{middle} and an outer region, determined analogously by R_{middle} and $R_{\max}$. The inner region will typically cover the space lying within the molecular van der Waals surface, excluding the nuclear zone if one is not interested in its accurate description, which is usually the case when atom-condensed properties are studied. The corresponding grid should be relatively fine so that the desired second-order functional derivative is accurately represented in this molecular region, where it fluctuates significantly because of chemical bonding and proximity to the atomic nuclei. Given the considerable electron density within this region, the charge value q_j should not be too high for the second-order truncation of eq 15 to be applicable. The point charges may assume higher values in the outer region; the corresponding grid can also be constructed more coarsely as the fluctuations in the functional derivatives will be less pronounced here. It is important to extend this outer grid rather far (up to several van der Waals radii) in order to capture all the information that is relevant to describe a system's chemical reactivity.

Once the external potential perturbations are defined, the responses of quantity P (as given by eq 19) should be determined. This is readily done by performing single point calculations of the molecule under study in the presence of one of the point charge perturbations w_j and by repeating this for all the other perturbations. For the properties we are interested in (the electronic energy, the chemical potential, etc.), any standard ab initio program can provide the necessary data.

3.2. Basis Set and Atom Condensation. The choice of a basis set for the expansion of the second-order functional derivative (eq 17) is another vital element. We have used s- and p-type Gaussian basis functions, centered on the atomic nuclei. Higher angular momentum functions could be included but are expected to induce only minor variations.[14,15] The exponents can be chosen from standard atomic basis sets, uncontracted into primitive Gaussians, but with a doubled value. This is a consequence of the fact that the second-order functional derivatives of conceptual DFT are related to the first-order functional derivative of the electron density $\rho(\mathbf{r})$ (see, e.g., eq 10) and that $\rho(\mathbf{r})$ decays twice as fast as the associated wave function. The exponents of auxiliary basis sets,[54,55] which are used for the acceleration of the evaluation of the Coulomb integrals in DFT calculations, can also be used; these do not require doubling.

Computation of the atom-condensed[56] variants greatly simplifies the situation. The atom-condensed indices are normally defined as

$$P_{\text{AB}}^{(2)} = \int_{V_\text{A}} \int_{V_\text{B}} P^{(2)}(\mathbf{r}, \mathbf{r}')\text{d}\mathbf{r}\text{d}\mathbf{r}' \tag{23}$$

where $P^{(2)}(\mathbf{r},\mathbf{r}')$ is used as a shorthand notation for the second-order functional derivative of property P and V_A and V_B denote the volumes of atoms A and B, respectively. Despite that, the basis set expansion of eq 17 provides an alternative approach:

$$P_{\text{AB}}^{(2)} = \sum_{\substack{k=1 \\ k\in A}}^{K} \sum_{\substack{l=1 \\ l\in B}}^{K} q_{kl} \int \beta_k(\mathbf{r})\text{d}\mathbf{r} \int \beta_l(\mathbf{r}')\text{d}\mathbf{r}' \tag{24}$$

Instead of integrating the kernel $P^{(2)}(\mathbf{r},\mathbf{r}')$ over the atomic volumes V_A and V_B, the contributions of atom A to the $\mathbf{r}$-dependent part and of atom B to the $\mathbf{r}'$-dependent part of the functional derivative are integrated over the entire space. These contributions originate from the terms in the basis set expansion for which the function $\beta_k(\mathbf{r})$ is centered on atom A and $\beta_l(\mathbf{r}')$ on atom B. It has been shown that the use of one sharp s-type Gaussian function (mimicking a Dirac δ-function) per atomic center yields reliable results for the condensed first-order functional derivatives.[15] The validity of this kind of basis set in the evaluation of condensed second-order indices will be illustrated in Section 4. The requirement of sharp functions is essential for this condensation scheme. It is indeed important that the atomic contributions do not overlap because the integrations over the entire space in eq 24 would lose their relevance. It should be stressed that the use of such a simplified basis set leads to a vast reduction in the number of expansion coefficients (K^2) to be determined and hence in the computational effort.

3.3. Normalization Constraint. As the normalization of the functional derivatives is often a priori known, the user has the option to impose it as a constraint. Let us consider the case of the linear response kernel $\chi(\mathbf{r},\mathbf{r}')$ as an example. We know that this quantity is normalized to 0 and, furthermore, that the integration over one of the position variables should give this value as well:[18,36]

$$\int \chi(\mathbf{r}, \mathbf{r}')\text{d}\mathbf{r}' = 0 \tag{25}$$

Combination of this expression with the basis set expansion of eq 17 yields the following normalization equation:

$$\sum_{k=1}^{K} \sum_{l=1}^{K} q_{kl}\beta_k(\mathbf{r})I_l = 0 \tag{26}$$

where I_l denotes the integral over the l-th basis function. An infinite number of such equations, corresponding to any value for the position variable $\mathbf{r}$, can be constructed and added to

the set of eq 18 as a constraint. In practice, a limited number of well-chosen ones suffices to obtain normalization. Numerical illustrations will be given in the next section. The use of normalization constraints is even superfluous provided that sufficient external potential perturbations are taken into account, and this is in all molecular regions, including the nuclear zone.

3.4. Symmetry. Given the extent of the computational cost of our methodology (a single point calculation for every positive and negative external potential perturbation), the implementation of symmetry properties is paramount for the reduction of this effort. Two levels of symmetry can be distinguished: the symmetry properties of the second-order functional derivative and the point group symmetry of the system under study. The first type is a consequence of the arbitrary order in which the functional derivatives with respect to $v(\mathbf{r})$ and $v(\mathbf{r}')$ are evaluated, implying that

$$P^{(2)}(\mathbf{r},\mathbf{r}') = P^{(2)}(\mathbf{r}',\mathbf{r}) \tag{27}$$

The immediate result is that the number of expansion coefficients in eq 17 can be reduced from K^2 to $K(K+1)/2$ as q_{kl} should equal q_{lk}, which simultaneously leads to a decrease in the required number of external potential perturbations. Inclusion of molecular symmetry properties further diminishes this number. As perturbations that are linked through symmetry operations belonging to the molecular point group give rise to identical responses, it is sufficient to sample a restricted, symmetry unique portion of space and to spatially propagate the obtained responses based on the molecular symmetry elements.

3.5. Computational Procedures. A final word should be said about the computational procedures we have used. Construction of matrix **D** in eq 18 requires the calculation of quantity P in the presence and the absence of the various external potential perturbations. The Gaussian 03 program package[57] will be used to calculate the electronic energy responses, as we will be focusing on the linear response kernel in the next section; obviously, other quantities can be obtained in a similar way. Matrix **B** contains integrals over a basis function and an external potential perturbation. The use of Gaussian basis functions and point charge perturbations enables an analytic evaluation, which essentially needs the computation of the incomplete γ function. This is done by a Fortran 90 numerical recipes routine.[58] The least-squares fitting procedure chosen to solve the set of linear eq 18 is taken from LAPACK[59] and uses a singular value decomposition approach.

4. Numerical Results

In this section, the atom-condensed linear response kernel, defined by eqs 10 and 24, will be calculated for a series of simple test systems. Not only will we show the numerical data that can be obtained, but we will also explain how the various parameters that emerged from our implementation should be chosen. The results will be compared with values obtained in another manner, which was previously studied by some of the present authors.[48,49] They used second-order perturbation theory to derive the next, approximate expression, valid for closed-shell systems described by a single Slater determinant:[26,35,60]

$$\chi_s(\mathbf{r},\mathbf{r}') = 4\sum_{i=1}^{N_0/2}\sum_{a=(N_0/2)+1}^{\infty}\frac{\varphi_i^*(\mathbf{r})\varphi_a(\mathbf{r})\varphi_a^*(\mathbf{r}')\varphi_i(\mathbf{r}')}{\varepsilon_i-\varepsilon_a} \tag{28}$$

The sum over i runs over all the occupied molecular orbitals $\phi_i(\mathbf{r})$ (with associated orbital energies ε_i), while the index a spans the unoccupied ones. It is important to note that a frozen-orbital approximation was made in its derivation and that energy differences between excited and ground states were replaced by orbital energy differences. These approximations are, however, exactly applicable to the KS noninteracting reference system, so that eq 28 can be seen as the exact functional derivative of the electron density with respect to the KS potential. There is a relation between the interacting linear response kernel, $\chi(\mathbf{r},\mathbf{r}')$, and the noninteracting one, $\chi_s(\mathbf{r},\mathbf{r}')$[35,61]

$$\chi(\mathbf{r},\mathbf{r}') = \chi_s(\mathbf{r},\mathbf{r}') + \iint \chi_s(\mathbf{r},\mathbf{x})\left(\frac{1}{|\mathbf{x}-\mathbf{x}'|}+\frac{\delta^2 E_{xc}}{\delta\rho(\mathbf{x}')\delta\rho(\mathbf{x})}\right)\chi(\mathbf{x}',\mathbf{r}')\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \tag{29}$$

which shows that formula (eq 28) is a zeroth-order approximation to the linear response kernel for the interacting system. Eq 28 can be condensed in the sense of eq 23, which was done using Becke's multicenter numerical integration procedure.[62−64] In the absence of reference data for the linear response kernel, we will make use of this approximate approach to assess the validity of our implementation.

We have chosen to study a series of six molecules—formaldehyde, water, ammonia, carbon monoxide, hydrogen cyanide, and nitrous oxide—that present interesting chemical properties and are computationally convenient because of their small size and high symmetry. Their geometries have been optimized on the B3LYP/6-311++G** level of theory,[65−68] whereas the single point calculations for eq 19 and the input for eq 28 have been done on the PBE/6-31+G* level.[69,70] As introduced in the previous section, the user should specify a number of parameters prior to the actual evaluation of the second-order functional derivatives. It is advisible that the specific choices are extensively tested beforehand with respect to convergence (e.g., the number of point charges) and numerical stability of the results. We will provide an example of this at the end of this section, where the numerical error is estimated through variation of the various parameter values. It is, however, interesting to start with a discussion of the optimal parameter values and the corresponding results.

The main group of parameters to be chosen is associated with the construction of the set of external potential perturbations. As suggested in Section 3.1, R_{middle} will be assigned a value of 1.0 van der Waals radii so that the inner sampling region spans the molecular van der Waals volume. The nuclear region will, however, be excluded by choosing an R_{min} value of 0.3 van der Waals radii. This can be done without harm because chemical variations take place in the valence region and because we are interested in the calculation of an atom-condensed quantity, without aiming for a

Table 1. Linear Response Matrix Elements χ_{AB} (in au) between Atoms A and B[a]

molecule		χ_{AB} (a)					χ_{AB} (b)				correlation
		H_1	C	O	H_2		H_1	C	O	H_2	
	H_1	−1.21				H_1	−1.21				
CH_2O	C	1.19	−4.82			C	0.68	−4.35			0.96 (0.99, −0.01)
	O	0.42	2.45	−3.28		O	0.34	2.99	−3.67		
	H_2	−0.40	1.19	0.42	−1.21	H_2	0.19	0.68	0.34	−1.21	
		O	H_1	H_2			O	H_1	H_2		
	O	−3.77				O	−1.76				
H_2O	H_1	1.88	−1.39			H_1	0.88	−0.93			0.96 (1.98, 0.05)
	H_2	1.88	−0.49	−1.39		H_2	0.88	0.05	−0.93		
		N	H_1	H_2	H_3		N	H_1	H_2	H_3	
	N	−5.65				N	−2.73				
NH_3	H_1	1.88	−1.37			H_1	0.91	−1.00			0.97 (1.92, 0.06)
	H_2	1.88	−0.25	−1.37		H_2	0.91	0.05	−1.00		
	H_3	1.88	−0.25	−0.25	−1.37	H_3	0.91	0.05	0.05	−1.00	
		C	O				C	O			
CO	C	−4.33				C	−3.78				1.00 (1.15, 0.00)
	O	4.33	−4.33			O	3.78	−3.78			
		H	C	N			H	C	N		
HCN	H	−0.85				H	−0.96				1.00 (0.92, 0.00)
	C	0.76	−5.24			C	0.60	−5.49			
	N	0.09	4.49	−4.58		N	0.36	4.89	−5.25		
		N_1	N_2	O			N_1	N_2	O		
NNO	N_1	−3.76				N_1	−5.79				0.92 (0.74, −0.04)
	N_2	2.56	−4.26			N_2	2.84	−4.10			
	O	1.20	1.70	−2.90		O	2.95	1.25	−4.21		

[a] Calculated with: (a) our proposed methodology to compute second-order functional derivatives and (b) the approach based on eq 28. Linear regressions between both data sets for each molecule give rise to the correlation coefficients (R^2) shown in the last column; the values in brackets indicate the corresponding slopes and intercepts.

locally resolved description. Furthermore, it has been shown in the context of the first-order functional derivatives that an accurate representation of the nuclear cusps requires many thousands of additional point charges.[14] The R_{max} value will be set equal to 4.0 van der Waals radii; a further extension of the external potential sampling region does not change the results. Point charges in the inner region will have values of ±0.02 e (elementary charge) and the spacing of the corresponding cubic grid will be 0.12 Å. Larger charge values of ±0.10 e can be chosen for the outer region; the perturbations are also placed farther apart, with an associated grid spacing of 0.40 Å.

We have observed that it is often advantageous to impose the normalization constraints (eq 26) for the study of atom-condensed quantities. The implementation of this equation occurs in quite a similar way as the construction of the external potential perturbations: it will be evaluated for a number of positions **r** lying on a cubic grid between two scaled van der Waals surfaces. The points where the normalization constraint should be evaluated are, however, influenced by the basis set used for the expansion of the functional derivative. We will use the simplified basis set of one sharp s-type Gaussian function per atom with exponent values of 10.0 au, ensuring that overlap between two or more atoms is negligible. As a consequence, the normalization constraints should be evaluated within a zone that is close to the atomic nuclei, on which the various basis functions are centered. The sharp basis functions $\beta_k(\mathbf{r})$ of eq 26 will indeed assume values close to 0 if **r** is chosen too far away from any of the atomic nuclei, yielding the trivial expression "0 = 0". The normalization equations will be evaluated in the volume contained within the van der Waals surfaces scaled by the factors 0.1 and 0.3. A grid spacing of 0.2 Å gives satisfactory results. It is interesting to note that the normalization equations are evaluated in the region of space where no external potential perturbations are placed. Information about this zone is thereby indirectly taken into account, while avoiding an extensive sampling to represent the nuclear cusps.

The above-mentioned parameters typically give rise to a number of external potential perturbations (J) of the order of 50 000 and around 200 normalization equations. Such an extensive sampling by external potential perturbations requires a considerable computational effort but should be close to the convergence limit, except for the nuclear zone, which is not explicitly dealt with here. The computational cost will be minimized by inclusion of the molecular symmetry properties, which gives a reduction in the number of required single point calculations by a factor 4 for formaldehyde (C_{2v} point group) and 6 for ammonia (C_{3v} point group), for example. The symmetry of the linear response kernel (eq 27) will also be taken into account.

Table 1 gives the numerical results obtained by our approach to calculate second-order functional derivatives and by the perturbation theoretical methodology based on eq 28. It is encouraging to see how well the data obtained by both methods correlate; the linear regression correlation coefficient varies between 0.92 and 1.00 for the set of chosen molecules.

Table 2. Variation of the Relevant Parameters Around Their Optimal Values and the Corresponding RMSE Induced in the Atom-Condensed Linear Response Values for Formaldehyde[a]

parameter	optimal value	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
R_{max} (vdW radii)	4.0	5.0	3.0						
q_{inner} (e)	0.02					0.04	0.01		
q_{outer} (e)	0.10					0.20	0.05		
a_{inner} (Å)	0.12			0.11	0.15				
a_{outer} (Å)	0.40			0.35	0.50				
exponent basis functions (au)	10.0							20.0	5.0
number of perturbations	12 998	20 250	8632	18 172	6964				
RMSE		0.00021	0.00099	0.011	0.062	0.0021	0.00019	0.0024	0.036

[a] The number of external potential perturbations considered (before symmetry propagation) is indicated where relevant.

Some interesting differences are, nonetheless, visible. While the intercepts of the regression lines consistently stay close to 0, the corresponding slopes assume values from 0.74 to 1.98. Therefore, although the intramolecular trends are similar in either of the approaches, the intermolecular ones may change. This leads, for example, to a higher linear response value for χ_{OO} in water than in formaldehyde with the present methodology, whereas the inverse trend is predicted through application of eq 28. Another point is that the present methodology finds slightly negative values for the off-diagonal H−H elements in formaldehyde, water, and ammonia. This can be interpreted as if an external potential change induced by a positive charge distribution in the atomic volume of one of the hydrogen atoms leads to an electron density increase in the neighboring hydrogens. Although this cannot be excluded because of their spatial proximity, the same observation cannot be made from eq 28, where near-zero values are obtained. A final dissimilarity is seen for nitrous oxide, where the central atom is predicted to be the most polarizable atom by our method, while it seems to be the least polarizable one with the other approach.

It should be emphasized that these differences do not come as a surprise for some reasons: First of all, the atomic condensation procedures used in both approaches are completely different. Second, application of eq 28 involves the theoretical approximations that were explained at the beginning of this section.[48,49] Finally, while our methodology does not involve approximations on the theoretical level, there are numerical approximations. It is, however, true that the present methodology in principle allows the computation of exact solutions if the situation of an infinite number of external potential perturbations and a complete basis set is approached. We therefore assume that the current methodology provides the best representation of chemical reality, even though further research is certainly needed to confirm this statement.

It is now appropriate to make two comments on the chemical interpretation of the linear response kernel. Some of the present authors have shown that this quantity can be used as a measure of electron delocalization.[48,49] Moreover, the linear response kernel contains the necessary information to evaluate a system's polarizability.[60,71] For example, $\chi(\mathbf{r},\mathbf{r}')$ can be related to the dipole polarizability tensor α_{ij},[72] where the indices i and j stand for the Cartesian directions:

$$\alpha_{ij} = -\iint r_i r'_j \chi(\mathbf{r},\mathbf{r}')\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \qquad (30)$$

Unfortunately, this quantity is not accessible with the presented results for the atom-condensed linear response kernel but could straightforwardly be obtained if the locally resolved kernel was calculated using our methodology. One should, however, be careful when interpreting the atom-condensed linear response elements as indicators of the atomic polarizabilites. A consideration of the diatomic molecule CO, for example, shows that these elements do not necessarily equal polarizabilities of atoms in molecules. Indeed, the symmetry properties of the kernel result in identical values for the χ_{CC} and χ_{OO} elements, while the carbon and oxygen atoms should clearly have a different polarizability. A second note concerns the use of the linear response kernel as a tool for the description of intermolecular trends. The fact that both methodologies of Table 1 give similar intramolecular trends, but different intermolecular ones could indicate that the linear response kernel is less suitable for making intermolecular comparisons. An analogous observation has been made for the local reactivity descriptors. The Fukui function is an appropriate property for intramolecular reactivity descriptions, whereas the local softness is preferred for intermolecular comparisons.[3] Future research should verify whether a similar reasoning, based on the Berkowitz−Parr relation (eq 12), leads to the conclusion that the softness kernel $s(\mathbf{r},\mathbf{r}')$ is a better alternative for the description of intermolecular trends.

As a final item, we will estimate the computational error associated with the current implementation and parameter choices. Table 2 gives the root-mean-square errors (RMSE) in the atom-condensed linear response values for formaldehyde induced by variations of the relevant parameters around the optimal values we have put forward. Variations (a) and (b) analyze the effect of an enlargement and reduction of the external potential sampling region. Errors of 0.00021 and 0.00099 au are found when the R_{max} values are changed from 4.0 to 5.0 and 3.0 van der Waals radii, respectively, indicating that our chosen sampling border does not form an obstacle for the numerical stability of the results. We are thus allowed to ignore the large sampling region associated with an R_{max} value of 5.0 van der Waals radii, which concomitantly gives rise to a considerable increase in computational cost (20,250 external potential perturbations instead of 12,998). The number of perturbations can also be varied by altering the spacings of the corresponding grids. Reduction of the inner grid lattice parameter (a_{inner}) from 0.12 to 0.11 Å and of the outer one (a_{outer}) from 0.40 to 0.35 Å, yielding an increase

in the number of external potential perturbations to 18 172, leads to a RMSE of 0.011 au. This result indicates that higher precision values can be obtained by refinement of the grids on which the point charges are placed, however, at substantial computational expense. Our proposed values for the lattice parameters attempt to limit this computational effort and yield values that are precise up to the second decimal. Variation (d) provides an example of this, which shows that a RMSE of 0.062 au is found when the number of perturbations is reduced by a factor of around 2; this dissuades us from using coarser grids. The sizes of the point charges exert only a minor effect on the obtained results. Variations (e) and (f), respectively doubling and halving their values, give rise to errors of 0.0021 and 0.00019 au. The last parameter to be considered is the exponent value of the basis functions. An increase of this value from 10.0 to 20.0 au only leads to a small variation in the linear response elements of 0.0024 au. Reduction of the exponent to 5.0 au, however, gives rise to a considerable RMSE of 0.036 au. This indicates that such an exponent value is too low to prevent the basis functions from spreading over several atoms, which should be avoided. Overall, the presented optimal values yield numerical stability at a minimal computational expense. We have shown that the associated numerical error is of the order of 0.01 au.

5. Conclusion and Prospects

In this paper, the first generally applicable methodology to calculate second-order functional derivatives of arbitrary properties with respect to the external potential is proposed. The central idea is to expand the desired functional derivative in a basis set and to determine the expansion coefficients by probing the molecular environment with external potential perturbations. Although this approach is theoretically rigorous, exact solutions can only be obtained if an infinite basis set and a number of perturbations are considered.

We applied the methodology to evaluate the atom-condensed linear response kernel and estimated that the numerical error of the current implementation is of the order of 0.01 au. The results for a set of six simple molecules were compared with values obtained through an approximate methodology based on second-order perturbation theory. Both approaches are generally in agreement.

The most important disadvantage of the proposed methodology is its substantial computational cost. For the set of molecules considered, a number of 50 000 positive and negative external potential perturbations—each requiring a single point calculation—was typically needed to obtain converged results. Even though this number was reduced by taking molecular symmetry into account, the vast computational effort will probably prevent this method from being the routinely used procedure in the future. Nonetheless, the fact that this methodology does not involve any theoretical approximations and that exact or quasi-exact solutions can be found, if the sets of external potential perturbations and basis functions are chosen large enough, implies that it can be used as a benchmark method against which more easily applicable, approximate methodologies may be assessed. Another advantage of this algorithm, compared to the more conventional approach based on eqs 28 and 29, is that the atom-condensed linear response kernel is accessed directly.

We should, finally, mention that it is possible to enhance the computational efficiency by adding extra equations to the set of eq 18, the nature of which depends upon the property one is interested in. As an illustration, the Appendix will show how the calculation of the linear response kernel could be accelerated by taking responses of the electron density or electrostatic potential into account.

Acknowledgment. N. S. acknowledges the Research Foundation - Flanders (FWO) for a position as research assistant. F. D. P. and P. G. thank the FWO and the Vrije Universiteit Brussel for continuous financial support. P. W. A. would like to thank NSERC, the Canada Research Chairs and SHARCNET for research support.

Appendix

The definition of the linear response kernel (eq 10) provides a means to accelerate the proposed methodology if one is specifically interested in the calculation of this quantity. Not only can the linear response kernel be seen as the second-order functional derivative of the electronic energy E with respect to the external potential, but it can also be considered as the first-order functional derivative of the electron density $\rho(\mathbf{r})$. An analogous reasoning as in Section 2, but this time based on a functional Taylor series of the electron density evaluated at point $\mathbf{x}$, yields the following equations for the expansion coefficients q_{kl}

$$\frac{1}{2}(\rho[v(\mathbf{r}) + w_j(\mathbf{r});\mathbf{x}] - \rho[v(\mathbf{r}) - w_j(\mathbf{r});\mathbf{x}]) = \sum_{k=1}^{K}\sum_{l=1}^{K} q_{kl}\beta_k(\mathbf{x})\int \beta_l(\mathbf{r}')w_j(\mathbf{r}')\mathrm{d}\mathbf{r}', \quad \text{with} \quad j = 1, ..., J \tag{A.1}$$

The advantage is that a series of equations of this type, corresponding to electron density responses evaluated at various points ($\mathbf{x} = \mathbf{x}_1$, $\mathbf{x} = \mathbf{x}_2$, ...) can be constructed for every external potential perturbation j. The fact that just two single point calculations (one for $w_j(\mathbf{r})$ and one for $-w_j(\mathbf{r})$) can give rise to a number of equations to be considered in the least-squares fitting procedure implies a significant computational benefit.

Another possibility lies in the evaluation of electrostatic potential[73] responses. The electrostatic potential is defined as

$$\Phi(\mathbf{r}) = -v(\mathbf{r}) - \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\mathrm{d}\mathbf{r}' \tag{A.2}$$

A functional Taylor series of $\Phi(\mathbf{x})$ for which the external potential is perturbed by $w(\mathbf{r})$ can be written as follows:

$$\Phi[v(\mathbf{r}) + w(\mathbf{r});\mathbf{x}] = -v(\mathbf{x}) - w(\mathbf{x}) - \int \frac{\rho[v(\mathbf{r});\mathbf{r}'] + \int \left(\frac{\delta\rho(\mathbf{r}')}{\delta v(\mathbf{r}'')}\right)_N w(\mathbf{r}'')\mathrm{d}\mathbf{r}'' + ...}{|\mathbf{x} - \mathbf{r}'|}\mathrm{d}\mathbf{r}' \tag{A.3}$$

$$= \Phi[v(\mathbf{r});\mathbf{x}] - w(\mathbf{x}) - \iint \frac{\chi(\mathbf{r}',\mathbf{r}'')w(\mathbf{r}'')}{|\mathbf{x}-\mathbf{r}'|}d\mathbf{r}'d\mathbf{r}'' + \ldots \quad (A.4)$$

Application of the arguments used in Section 2 now leads to the subsequent equations for the expansion coefficients q_{kl}:

$$\frac{1}{2}(\Phi[v(\mathbf{r}) + w_j(\mathbf{r});\mathbf{x}] - \Phi[v(\mathbf{r}) - w_j(\mathbf{r});\mathbf{x}]) + w_j(\mathbf{x}) = -\sum_{k=1}^{K}\sum_{l=1}^{K} q_{kl}\int \frac{\beta_k(\mathbf{r}')}{|\mathbf{x}-\mathbf{r}'|}d\mathbf{r}'\int \beta_l(\mathbf{r}'')w(\mathbf{r}'')d\mathbf{r}'', \quad \text{with } j = 1, \ldots, J \quad (A.5)$$

Analogously to eq A.1, this set of equations could yield a substantial improvement to the computational efficiency. A major advantage of this last approach is that the molecular electrostatic potential has a much smoother behavior than the electron density so that convergence is expected to be more easily attained.

References

(1) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.

(2) Chermette, H. *J. Comput. Chem.* **1999**, *20*, 129–154.

(3) Geerlings, P.; De Proft, F.; Langenaeker, W. *Chem. Rev.* **2003**, *103*, 1793–1873.

(4) Ayers, P. W.; Anderson, J. S. M.; Bartolotti, L. J. *Int. J. Quantum Chem.* **2005**, *101*, 520–534.

(5) Chattaraj, P. K.; Sarkar, U.; Roy, D. R. *Chem. Rev.* **2006**, *106*, 2065–2091.

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

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

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

(9) Pearson, R. G. *Chemical Hardness*; Wiley-VCH: New York, 1997.

(10) Parr, R. G.; Yang, W. T. *J. Am. Chem. Soc.* **1984**, *106*, 4049–4050.

(11) Ayers, P. W.; Levy, M. *Theor. Chem. Acc.* **2000**, *103*, 353–360.

(12) Ayers, P. W.; Melin, J. *Theor. Chem. Acc.* **2007**, *117*, 371–381.

(13) Ayers, P. W. *J. Math. Chem.* **2008**, *43*, 285–303.

(14) Ayers, P. W.; De Proft, F.; Borgoo, A.; Geerlings, P. *J. Chem. Phys.* **2007**, *126*, 224107.

(15) Sablon, N.; De Proft, F.; Ayers, P. W.; Geerlings, P. *J. Chem. Phys.* **2007**, *126*, 224108.

(16) Fievez, T.; Sablon, N.; De Proft, F.; Ayers, P. W.; Geerlings, P. *J. Chem. Theory Comput.* **2008**, *4*, 1065–1072.

(17) Sablon, N.; De Proft, F.; Geerlings, P. *J. Chem. Theory Comput.* **2009**, *5*, 1245–1253.

(18) Berkowitz, M.; Parr, R. G. *J. Chem. Phys.* **1988**, *88*, 2554–2557.

(19) Senet, P. *J. Chem. Phys.* **1996**, *105*, 6471–6489.

(20) Fuentealba, P.; Parr, R. G. *J. Chem. Phys.* **1991**, *94*, 5559–5564.

(21) Contreras, R.; Domingo, L. R.; Andrés, J.; Pérez, P.; Tapia, O. *J. Phys. Chem. A* **1999**, *103*, 1367–1375.

(22) Chamorro, E.; Contreras, R.; Fuentealba, P. *J. Chem. Phys.* **2000**, *113*, 10861–10866.

(23) Ayers, P. W.; Parr, R. G. *J. Chem. Phys.* **2008**, *129*, 054111.

(24) Geerlings, P.; De Proft, F. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3028–3042.

(25) Cárdenas, C.; Echegaray, E.; Chakraborty, D.; Anderson, J. S. M.; Ayers, P. W. *J. Chem. Phys.* **2009**, *130*, 244105.

(26) Senet, P. Reactivity and Polarisability Responses. In *Chemical Reactivity Theory*; Chattaraj, P. K., Ed.; CRC Press: Boca Raton, FL, 2009; pp 331−362.

(27) Chattaraj, P. K.; Cedillo, A.; Parr, R. G. *J. Chem. Phys.* **1995**, *103*, 7645–7646.

(28) Ghosh, S. K.; Berkowitz, M. *J. Chem. Phys.* **1985**, *83*, 2976–2983.

(29) Berkowitz, M.; Ghosh, S. K.; Parr, R. G. *J. Am. Chem. Soc.* **1985**, *107*, 6811–6814.

(30) Langenaeker, W.; De Proft, F.; Geerlings, P. *J. Phys. Chem.* **1995**, *99*, 6424–6431.

(31) Ayers, P. W.; Parr, R. G. *J. Chem. Phys.* **2008**, *128*, 184108.

(32) Chattaraj, P. K.; Roy, D. R.; Geerlings, P.; Torrent-Sucarrat, M. *Theor. Chem. Acc.* **2007**, *118*, 923–930.

(33) Cohen, M. H.; Ganduglia-Pirovano, M. V.; Kudrnovský, J. *J. Chem. Phys.* **1995**, *103*, 3543–3551.

(34) Senet, P. *J. Chem. Phys.* **1997**, *107*, 2516–2524.

(35) Ayers, P. W. *Theor. Chem. Acc.* **2001**, *106*, 271–279.

(36) Liu, S. B.; Li, T. L.; Ayers, P. W. *J. Chem. Phys.* **2009**, *131*, 114106.

(37) Baekelandt, B. G.; Mortier, W. J.; Lievens, J. L.; Schoonheydt, R. A. *J. Am. Chem. Soc.* **1991**, *113*, 6730–6734.

(38) Wang, C. S.; Zhao, D. X.; Yang, Z. Z. *Chem. Phys. Lett.* **2000**, *330*, 132–138.

(39) Mortier, W. J. *Struct. Bonding (Berlin)* **1987**, *66*, 125–143.

(40) Morita, A.; Kato, S. *J. Am. Chem. Soc.* **1997**, *119*, 4021–4032.

(41) Morita, A.; Kato, S. *J. Phys. Chem. A* **2002**, *106*, 3909–3916.

(42) Ishida, T.; Morita, A. *J. Chem. Phys.* **2006**, *125*, 074112.

(43) Stone, A. J. *Mol. Phys.* **1985**, *56*, 1065–1082.

(44) Lu, Z. Y.; Yang, W. T. *J. Chem. Phys.* **2004**, *121*, 89–100.

(45) Langenaeker, W.; Liu, S. *J. Mol. Struct. (THEOCHEM)* **2001**, *535*, 279–286.

(46) Cedillo, A.; Contreras, R.; Galvan, M.; Aizman, A.; Andres, J.; Safont, V. S. *J. Phys. Chem. A* **2007**, *111*, 2442–2447.

(47) Cioslowski, J.; Martinov, M. *J. Chem. Phys.* **1994**, *101*, 366–370.

(48) Sablon, N.; De Proft, F.; Geerlings, P. *J. Phys. Chem. Lett.* **2010**, *1*, 1228–1234.

(49) Sablon, N.; De Proft, F.; Geerlings, P. *Chem. Phys. Lett.* **2010**, *498*, 192–197.

(50) Lieb, E. H. *Int. J. Quantum Chem.* **1983**, *24*, 243–277.

(51) Ayers, P. W.; Parr, R. G. *J. Am. Chem. Soc.* **2001**, *123*, 2007–2017.

(52) Anderson, J. S. M.; Melin, J.; Ayers, P. W. *J. Chem. Theory Comput.* **2007**, *3*, 358–374.

(53) Anderson, J. S. M.; Melin, J.; Ayers, P. W. *J. Chem. Theory Comput.* **2007**, *3*, 375–389.

(54) Eichkorn, K.; Treutler, O.; Ohm, H.; Haser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283–289.

(55) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119–124.

(56) Ayers, P. W.; Morrison, R. C.; Roy, R. K. *J. Chem. Phys.* **2002**, *116*, 8731–8744.

(57) 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.; Kno, 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, 2005.

(58) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in Fortran 90. The Art of Parallel Scientific Computing*, 2nd ed.; Cambridge University Press: New York, 1999.

(59) Anderson, E.; Bai, Z.; Bischof, C.; Blackford, S.; Demmel, J.; Dongarra, J.; Du Croz, J.; Greenbaum, A.; Hammarling, S.; McKenney, A.; Sorensen, D. *LAPACK Users' Guide*, 3rd ed.; Society for Industrial and Applied Mathematics: Philadelphia, PA, 1999.

(60) Ayers, P. W. *Faraday Discuss.* **2007**, *135*, 161–190.

(61) Gross, E. K. U.; Kohn, W. *Phys. Rev. Lett.* **1985**, *55*, 2850–2852.

(62) Becke, A. D. *J. Chem. Phys.* **1988**, *88*, 2547–2553.

(63) Torrent-Sucarrat, M.; Salvador, P.; Geerlings, P.; Sola, M. *J. Comput. Chem.* **2007**, *28*, 574–583.

(64) Torrent-Sucarrat, M.; Salvador, P.; Sola, M.; Geerlings, P. *J. Comput. Chem.* **2008**, *29*, 1064–1072.

(65) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785–789.

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

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

(68) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986; pp 65−88.

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

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

(71) Krishtal, A.; Senet, P.; Van Alsenoy, C. *J. Chem. Phys.* **2009**, *131*, 044312.

(72) Mahan, G. D.; Subbaswamy, K. R. *Local Density Theory of Polarisability*; Springer: New York, 1990.

(73) Politzer, P.; Truhlar, D. *Chemical Applications of Atomic and Molecular Electrostatic Potentials*; Plenum: New York, 1981.

CT1004577

JCTC Journal of Chemical Theory and Computation

An Error and Efficiency Analysis of Approximations to Møller−Plesset Perturbation Theory

Michael S. Marshall,[†] John S. Sears,*[,‡] Lori A. Burns,[†] Jean-Luc Brédas,[‡] and C. David Sherrill*[,†]

Center for Computational Molecular Science and Technology, School of Chemistry and Biochemistry, and School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0400, United States, and Center for Computational Molecular Science and Technology, Center for Organic Photonics and Electronics, and School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia 30332-0400, United States

Received August 18, 2010

Abstract: We present a systematic study of the synergistic effects of popular approximations to Møller−Plesset perturbation theory through the second order (MP2). This work applies the density-fitting (DF) approximation for two-electron integrals, the dual-basis (DB) approximation for the Hartree−Fock reference, and the use of "heavy augmented" Dunning basis sets for basis set reduction, as well as combinations of these, to the S22 benchmark set of weakly bound dimers. For each approach, we report an error analysis as well as relative speedups for the 22 interaction energies in the set. Compared to the MP2/aug-cc-pVTZ level of theory, the DB-DF-MP2/heavy-aug-cc-pVTZ approach achieves an average speedup of 18 with a root-mean-square error of only 0.076 kcal mol^{-1} (2%).

1. Introduction

In the past decade, the field of computational chemistry has demonstrated that high-level calculations on small molecules can in some cases achieve an accuracy comparable to that of experimental results.[1−5] A current challenge lies in the development of approximations to robust levels of theory to address larger systems of interest. Dispersion-dominated interactions, for which dynamic electron correlation has been shown to play an important role,[6−11] have attracted significant recent attention. When applying computational techniques to large-scale problems, long-range interactions can accumulate and must be accounted for properly. While an accurate description is provided by highly correlated methods such as coupled-cluster theory,[12−14] the steep computational cost of such methods constrains their applicability to systems of but a few dozen atoms with modest basis sets. To overcome this problem, ongoing research efforts focus on two fronts: (1) the modification of established methods by adding adjustable parameters fit to experimental results or a higher level of theory or (2) the development of approximations to robust levels of theory that maintain their inherent accuracy while reducing the cost.

The strategy of incorporating *ad hoc* terms with fitted parameters has shown great success for methods such as DFT-D (which adds a scaled, damped dispersion correction to a DFT functional).[15−17] However, there remains no means of systematically improving the accuracy, and such methods sometimes require numerous parameters trained upon specific test sets to produce high-quality results. Correlated wave function methods have also been modified by fitted parameters in spin-component-scaled Møller−Plesset perturbation theory (SCS-MP2),[18−22] spin-opposite-scaled Møller−Plesset perturbation theory (SOS-MP2),[23] and spin-component-scaled coupled-cluster with singles and doubles theory (SCS-CCSD).[24] MP2 tends to give reasonably reliable results for certain types of noncovalent interactions (such as alkane−alkane interactions and H-bonded interactions). In cases where MP2

* Corresponding authors: e-mail: sherrill@gatech.edu; john.sears@gatech.edu.

[†] School of Chemistry and Biochemistry and School of Computational Science and Engineering.

[‡] Center for Organic Photonics and Electronics and School of Chemistry and Biochemistry.

10.1021/ct100468f
Published on Web 11/03/2010

exhibits significant errors (e.g., π-stacking interactions), the scaled MP2 methods such as SCS-MP2 tend to perform well.[21] Even in cases where very accurate binding energies are desired for noncovalent interactions, MP2 remains a critical ingredient in the theoretical procedure. When benchmark-quality results are needed, the current standard procedure is to evaluate the binding energies in the MP2 complete basis set limit and then to correct for higher-order correlation terms by adding a ΔCCSD(T) correction [evaluated as the difference between CCSD(T) and MP2 binding energies in a smaller basis set].[6,7] Thus, whether one uses bare MP2, scaled MP2, or MP2 in conjunction with CCSD(T) corrections, MP2 computations remain important in studies of noncovalent interactions, and it is useful to explore approximations for speeding up these MP2 computations and to assess the associated errors.

In electronic structure theory, the evaluation and storage of four-index integrals is a common bottleneck. Various approaches to this problem have been explored, such as resolution of the identity[25−33] [now commonly referred to as density fitting (DF)], Cholesky decompositions[34−43] (CD), and pseudospectral techniques.[44−46] In the DF treatment, four-index integrals $(\mu\nu|\rho\sigma)$ are approximated by summations over three-index quantities:

$$(\mu\nu|\rho\sigma) \approx \sum_{PQ} (\mu\nu|P)[J^{-1}]_{PQ}(Q|\rho\sigma) \tag{1}$$

where $[J^{-1}]_{PQ}$ is the inverse of the Coulomb metric evaluated in an auxiliary basis set:

$$[J]_{PQ} = \int P(\mathrm{r}_1)\, \frac{1}{\mathbf{r_{12}}}\, Q(\mathbf{r_2})\, \mathrm{d}^3\mathbf{r_1}\, \mathrm{d}^3\mathbf{r_2} \tag{2}$$

The three-index quantity $(\mu\nu|P)$ serves to cast the product $(\mu\nu|$ onto the auxiliary basis via the Coulomb metric

$$(\mu\nu|P) = \int \mu(\mathbf{r_1})\, \nu(\mathbf{r_1})\, \frac{1}{\mathbf{r_{12}}}\, P(\mathbf{r}_2)\, \mathrm{d}^3\mathbf{r_1}\, \mathrm{d}^3\mathbf{r_2} \tag{3}$$

While density-fitting does not lower the asymptotic scaling of MP2, it does reduce the prefactor significantly, with speedups in the range of 2 to 5.5 reported.[32,47] There exist many more methods for speeding up the evaluation of the correlation energy (e.g., local molecular orbital approaches such as local-MP2[32,33]), yet the application of DF alone is often sufficient to reduce the cost of the correlation energy computation to the point that the time needed for the underlying self-consistent field (SCF) becomes the rate-determining step.

Numerous algorithmic advances have been achieved over the past three decades to improve SCF efficiency. These range from Pulay's direct inversion of iterative subspace[48,49] (DIIS), which minimizes the number of SCF iterations, to modern linear scaling methods.[50−54] Two recent, similar advances in SCF theory are dual-basis techniques[55−60] to project the SCF energy from a smaller basis set and perturbative corrections to estimate the SCF complete basis set limit.[61]

The dual-basis (DB) approximation proposed in the work of Steele et al.[60] involves performing an iterative SCF in a small basis and then taking a single Roothan diagonalization step in a larger target basis set. In practice, the small basis is typically a specially designed subset of the target basis set, although this restriction is not imposed by the theory. Once the SCF is converged with the small basis set, the occupied molecular orbital (MO) coefficients are projected onto the larger basis via

$$\mathbf{C}_{\bar{\mu}i} = \sum_{\bar{\mu}\bar{\nu}} \sum_{\lambda} \mathbf{S}^{-1}_{\bar{\mu}\bar{\nu}} \mathbf{S}_{\bar{\nu}\lambda} \mathbf{C}_{\lambda i} \tag{4}$$

where $\mathbf{S}$ is the atomic orbital (AO) overlap matrix, i represents a MO index, Greek letters represent AO indices, and barred indices signify large-basis quantities. Using the newly constructed coefficient matrix, the new density matrix $\mathbf{P}$ is formed, and a single Fock matrix is built and diagonalized. After including some first-order corrections, the DB-SCF energy is shown to be

$$E_{\mathrm{dualbasis}} = E_{\mathrm{smallbasis}} + \sum_{\bar{\mu}\bar{\nu}} \Delta\mathbf{P}_{\bar{\mu}\bar{\nu}} \mathbf{F}_{\bar{\mu}\bar{\nu}} \tag{5}$$

where $\Delta\mathbf{P} = \mathbf{P}' - \mathbf{P}$ is the difference between the postdiagonalization density matrix $\mathbf{P}'$ and the small basis density matrix $\mathbf{P}$ (projected into the large basis). The small truncated basis sets used in the dual-basis methods have already been implemented[62] in the Q-Chem 3.2 program suite for several Pople and Dunning basis sets.

Another broadly employed approximation is the truncation of the aug-cc-pVXZ (X = D, T, Q) basis sets by eliminating diffuse functions from hydrogen atoms. These truncated basis sets are commonly referred to as heavy-aug-cc-pVXZ and are often abbreviated as haXZ (X = D, T, Q). For biological applications and polymer studies, where a large number of hydrogens are present, haXZ can introduce a significant savings. Dropping augmented functions on hydrogen has been shown to have a small effect on properties such as interaction energies for nonbonded complexes.[47] The DF, DB, and haXZ approximations have all been developed independently. In this work, we systematically examine the practicability of combining these approximations and evaluate the magnitude of accumulated errors and attainable speedups. The S22 benchmark set[63] has been adopted because of its focus on noncovalent complexes, which are theoretically challenging.

Recent work by Steele et al. has shown that by combining DB and DF approximations within MP2, one can expect root-mean-square errors (RMSEs) of 0.043 and 0.019 kcal mol^{-1} for the MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ levels of theory, respectively, for the S22 set.[60] Their timings focus mainly on evaluating DB-DF-MP2 analytical gradients. In this work, we present an error and efficiency analysis for each approximation independently, then repeat with the S22 set for the combination of approximations, thereby permitting dissection of any errors incurred in the energy, as well as elucidating the origins of the speedup. We also consider a series of linear alkanes to examine how these approximations behave as a function of increasing system size. Timings are compared to those from some other methods such as density functional theory.

2. Theoretical Approach

2.1. Efficiency Study of Approximate MP2 on Linear Alkanes. To evaluate gains by approximate MP2 methods, we examine a series of linear alkanes (C_nH_{2n+2}). Recent work has considered the effect of RI,[64,65] Cholesky,[66,67] and atomic-orbital-based MP2[68] approximations on linear alkanes. Single-point energy computations were performed using B3LYP, DB-B3LYP, MP2, DB-MP2, DF-MP2, and DB-DF-MP2 with the aug-cc-pVDZ basis set. For all dual-basis approximations, we employed the optimized basis set of Steele et al.[60] referred to as racc-pVDZ, which has been shown to reproduce the target basis set (aug-cc-pVDZ) with minimal error in total energy. For the density-fitting auxiliary basis sets, we employed the basis set from Hättig and co-workers[69,70] referred to as rimp2-aug-cc-pVDZ. In this work, we will only be density-fitting the MP2 contribution, not the underlying SCF, as that capability is not currently implemented in Q-Chem 3.2. The frozen-core approximation was employed for all of the perturbative methods. Alkane geometries were constructed from the following parameters: $r_{CC} = 1.53$ Å, $r_{CH} = 1.09$ Å, and $\theta_{CCC} = 109.5°$. For each level of theory, we report the overall user time as well as a decomposition of SCF and MP2 user times. For the tests performed, I/O time was typically minor; hence user times were very similar to wall times.

All computations were performed without taking advantage of spatial symmetry. All alkane computations used the Q-Chem 3.2[62] program suite on an Altus 1702 server featuring dual AMD Opteron 2378 processors (2.4 GHz, Quad Core), 32 GB of DDR2 RAM, and 2 × 1 TB 7200 rpm RAID-0 local disks. The SCF was converged to 10^{-8} hartree, and the integral threshold was 10^{-13}.

2.2. Performance Analysis of Approximate MP2 on the S22 Set. For a detailed analysis of the performance (both speedups and errors) by various approximate MP2 methods, we chose the S22 benchmark set,[63] which features diverse types of nonbonded interactions over a wide range of system sizes, from a water dimer (six atoms) to a hydrogen-bonded adenine-thymine complex (30 atoms). Benchmark-quality CCSD(T)/CBS reference binding energies are available for this test set.[63,71] For each of the 22 complexes, we report the interaction energy and the total user time for MP2, DB-MP2, DF-MP2, and DB-DF-MP2 with the aug-cc-pVDZ, aug-cc-pVTZ, heavy-aug-cc-pVDZ, and heavy-aug-cc-pVTZ Dunning basis sets [a heavy-aug-cc-pVXZ (X = D,T) basis set consists of cc-pVXZ on hydrogen atoms and aug-cc-pVXZ on all other atoms]. The choice of the DB basis set and DF auxiliary basis is as described above. All interaction energies were corrected for basis-set superposition error (BSSE) using the counterpoise correction scheme outlined by Boys and Bernardi,[72] and individual calculations employed the frozen-core approximation. The benchmark machine for the S22 test set is an Intel Xeon (3.2 GHz, single core), with 4 GB of DDR2 RAM and a 150 GB local disk. The SCF was converged to 10^{-8} hartree, and the integral threshold was 10^{-13}. This work focuses on approximating three computations: MP2/aug-cc-pVDZ, MP2/aug-cc-pVTZ, and MP2/CBS(aDZ,aTZ), where CBS(aDZ,aTZ) refers to a two-point extrapolation as defined by Halkier et al.[73] using aug-cc-pVDZ and aug-cc-pVTZ correlation energies. All computations in this part of the study use one of these three canonical MP2 results as a reference point.

3. Results and Discussion

3.1. Efficiency of Approximations to MP2 for Linear Alkanes. In the following analysis, MP2 will be discussed in terms of two components: (a) the underlying SCF computation and (b) the evaluation of the MP2 correlation energy (including the transformation of the atomic orbital integrals to the molecular orbital basis). Using the alkane test cases, we investigate how the DF, DB, and haXZ approximations affect the speed and accuracy of the computation. For a medium-sized molecule such as $C_{20}H_{42}$, the underlying SCF takes 33% of the total user time as shown in Figure 1a. The MP2 contribution, which formally scales as $O(N^5)$, dominates over the SCF, which formally scales as $O(N^4)$, because of the AO to MO transformation required. (Of course, actual computational scalings with respect to system size will be lower than these formal scalings because of techniques such as integral prescreening; moreover, integral prescreening will work best in one-dimensional systems such as these.) As system size or the basis set increases, an increasing fraction of time will be spent in the MP2 portion of the computation. It is this costly step that DF abates by changing the complexity of the AO to MO transformation from $O(N^5)$ to $O(N^4)$ while increasing the correlation energy evaluation from $O(N^4)$ to $O(N^5)$, but with a much lower prefactor than before. Figure 1b shows how DF-MP2 reduces the time to compute the MP2 correlation energy, shifting the majority of the computation time to the underlying SCF. For $C_{20}H_{42}$, 90% of the time to compute the DF-MP2 energy is spent in the SCF, to obtain an overall speedup of 2.46 relative to traditional MP2.

Now that the majority of the work has been shifted to the underlying SCF, we investigate dual-basis techniques that can drastically reduce the time to compute this stage. Figure 2 shows the DB-MP2 speedup to be only 1.18 relative to MP2 for $C_{20}H_{42}$, but this is only because of the large amount of time spent computing the correlation energy (and accordingly, the small cost of SCF) in the absence of the DF approximation. Figure 2 shows that the combination of the DF and DB approximations can yield speedups of 4.94 for $C_{20}H_{42}$. Within DB-DF-MP2, the bottleneck is the SCF in the small basis, which consumes 54% of the total compute time for $C_{20}H_{42}$. To put these improvements into context, we compared each method to B3LYP and dual-basis B3LYP (DB-B3LYP) in Figure 2, demonstrating that DF-MP2 is competitive with B3LYP and that DB-DF-MP2 is competitive with DB-B3LYP for system sizes upward of $C_{20}H_{42}$. This implies that the overhead in computing the correlation contribution within DFT is comparable to the time to compute the density fitted correlation energy in MP2. Note, however, that DF-MP2 and DB-DF-MP2 are still not competitive with any pure DFT method that lacks Hartree−Fock exchange.

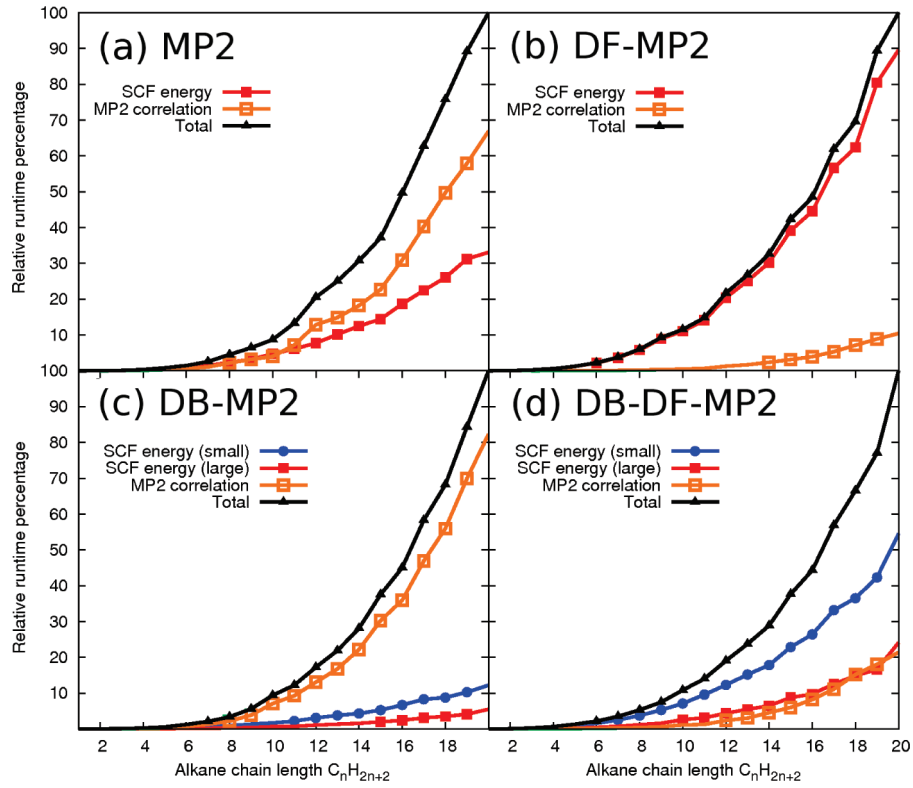


Figure 1. The relative user time and decompositions of (a) MP2, (b) DF-MP2, (c) DB-MP2, and (d) DB-DF-MP2 are shown for the alkane series methane (CH_4) through dodecane ($C_{20}H_{42}$). The percentages are computed by taking the $C_{20}H_{42}$ as a reference, with the total broken into SCF and MP2 correlation components. For the DB approximations, SCF (small) refers to the percentage of time to solve the iterative part, and SCF (large) refers to the percentage of time to perform the single Fock build in the target basis.

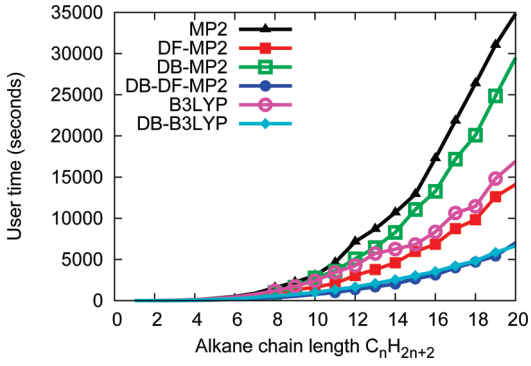


Figure 2. Total user times of MP2, DF-MP2, DB-MP2, DB-DF-MP2, B3LYP, and DB-B3LYP all with aug-cc-pVDZ basis set for the alkanes methane (CH_4) through dodecane ($C_{20}H_{42}$).

For the systems investigated, the errors of all of these approximations scale linearly with system size. DB incurs an average error of 0.027 kcal mol^{-1} per atom, DF incurs an error of 0.006 kcal mol^{-1} per atom, and DB-DF incurs and error of 0.033 kcal mol^{-1} per atom.

3.2. Performance Analysis of Approximate MP2 on the S22 Set. To analyze the error introduced by the DB, DF, and haXZ approximations, we compute interaction energies for each of the 22 complexes in the S22 benchmark test set.[63] Table 1 presents the root-mean-square error (RMSE) for the test set at each level of theory. In trying to reproduce MP2/aug-cc-pVDZ results, we see that the RMSE from DF is 0.003 kcal mol^{-1} while achieving an average speedup of 1.3. The magnitude of this error is very reasonable when compared to other remaining errors such as basis set incompleteness error (BSIE). On the other hand, the DB approximation incurs a RMSE of 0.043 kcal mol^{-1}. While an order of magnitude larger, the DB error is still rather small, especially considering it has a speedup of 1.78. Applying both of the approximations simultaneously shows that they do indeed compound well, achieving a speedup of 3.1, but the errors are additive also, for a RMSE of 0.045 kcal mol^{-1}. The use of heavy-aug-cc-pVDZ in place of aug-cc-pVDZ yields a speedup of 1.74 but at the cost of 0.120 kcal mol^{-1} average error. This average error is still small considering that the S22 MP2/aug-cc-pVDZ interaction energies range from −0.39 to −18.41 kcal mol^{-1}, but it may not be acceptable in some high-accuracy applications. The largest error introduced by neglecting diffuse functions on H atoms is 0.22 kcal mol^{-1} for the ethylene dimer test case, which has four closely packed hydrogens. Heavy-augmented basis sets should be avoided for systems with multiple hydrogen−hydrogen contacts, such as methane and ethene dimers.

To determine how well these approximations perform for larger basis sets, they were also tested against the MP2/aug-cc-pVTZ level of theory. For this larger basis set, RMSEs are reduced for all three approximations: DB has a RMSE of 0.017 kcal mol^{-1} (down from 0.043 kcal mol^{-1}), DF has a RMSE of 0.001 kcal mol^{-1} (down from 0.003 kcal mol^{-1}), and the use of heavy-augmented basis sets has a RMSE of 0.070 kcal mol^{-1} (down from 0.120 kcal mol^{-1}). When all

Table 1. Mean Unsigned Error (MUE), Root Mean Square Error (RMSE), Average Percent Error, and Average Speedup Analysis of Approximating MP2/aug-cc-pVXZ and MP2/CBS(aDZ,aTZ) for the S22 Test Set of Complexes[a]

reference	level of theory	speedup	MUE	RMSE	% error
MP2/aug-cc-pVDZ	MP2/haDZ	1.74	0.113	0.120	4.26
	DB-MP2/haDZ	2.61	0.106	0.120	4.05
	DF-MP2/haDZ	2.29	0.107	0.116	4.20
	DB-DF-MP2/haDZ	4.12	0.104	0.119	3.99
	DB-MP2/aDZ	1.78	0.034	0.043	0.82
	DF-MP2/aDZ	1.31	0.002	0.003	0.04
	DB-DF-MP2/aDZ	3.09	0.036	0.045[b]	0.84
MP2/aug-cc-pVTZ	MP2/haTZ	1.87	0.068	0.070	1.95
	DB-MP2/haTZ	5.76	0.072	0.077	2.01
	DF-MP2/haTZ	2.43	0.066	0.069	1.91
	DB-DF-MP2/haTZ	18.04	0.071	0.076	1.97
	DB-MP2/aTZ	3.25	0.012	0.017	0.20
	DF-MP2/aTZ	1.30	0.001	0.001	0.02
	DB-DF-MP2/aTZ	10.73	0.012	0.017[c]	0.19
MP2/CBS(aDZ,aTZ)	MP2/CBS(haDZ,haTZ)	1.86	0.039	0.044	0.94
	DB-MP2/CBS(haDZ,haTZ)	5.48	0.050	0.057	1.08
	DF-MP2/CBS(haDZ,haTZ)	2.42	0.038	0.043	0.92
	DB-DF-MP2/CBS(haDZ,haTZ)	15.73	0.049	0.056	1.05
	DB-MP2/CBS(aDZ,aTZ)	3.14	0.017	0.023	0.26
	DF-MP2/CBS(aDZ,aTZ)	1.30	0.001	0.001	0.01
	DB-DF-MP2/CBS(aDZ,aTZ)	9.65	0.017	0.022	0.26

[a] All errors in kilocalories per mole. [b] Reference 60 reports a RMSE of 0.043 kcal mol^{-1}. [c] Reference 60 reports a RMSE of 0.019 kcal mol^{-1}.

Table 2. Mean Unsigned Error (MUE), Root-Mean-Square Error (RMSE), and Average Percent Error for the Interaction Energies for Each Subgroup in the S22 Test Set in Kilocalories Per Mole[a]

	H bonding			dispersion			mixed		
level of theory	MUE	RMSE	% error	MUE	RMSE	% error	MUE	RMSE	% error
MP2/CBS(haDZ,haTZ)	0.059	0.061	0.64	0.029	0.033	1.26	0.031	0.032	0.88
DB-MP2/CBS(haDZ,haTZ)	0.057	0.065	0.56	0.051	0.057	1.60	0.043	0.047	1.00
DF-MP2/CBS(haDZ,haTZ)	0.058	0.060	0.63	0.027	0.032	1.21	0.030	0.031	0.86
DB-DF-MP2/CBS(haDZ,haTZ)	0.057	0.064	0.55	0.049	0.055	1.56	0.042	0.046	0.98
DB-MP2/CBS(aDZ,aTZ)	0.020	0.025	0.17	0.018	0.024	0.29	0.015	0.017	0.32
DF-MP2/CBS(aDZ,aTZ)	0.000	0.000	0.00	0.001	0.001	0.02	0.000	0.000	0.01
DB-DF-MP2/CBS(aDZ,aTZ)	0.019	0.025	0.17	0.018	0.023	0.28	0.015	0.017	0.32

[a] All values are relative to MP2/CBS(aDZ,aTZ).

three approximations are combined, speedups of 18.0 are achieved at the cost of 0.076 kcal mol^{-1} RMSE. Considering the large gain in computational efficiency, these errors are tolerable, and DB-DF-MP2/haTZ is recommended for typical studies of nonbonded interactions.

We also examined how complete basis set (CBS) extrapolations affect the error for each approximation. CBS extrapolations consistently reduce the RMSE (shown at the bottom of Table 1) for the approximations considered. The extrapolations particularly abate the error caused by the use of heavy-augmented basis sets, reclaiming 0.026 kcal mol^{-1} on average. The compounding of all three approximations and CBS extrapolations [DB-DF-MP2/CBS(haDZ,haTZ)] yields a RMSE of 0.056 kcal mol^{-1} and a speedup of 15.7. This speedup is not quite as large as that observed for the DB-DF-MP2/haTZ (18.0), because the CBS extrapolations include haDZ computations which have a lesser efficiency gain. We note that MP2/CBS(aDZ,aTZ) has a 0.118 kcal mol^{-1} RMSE compared to MP2/CBS(aTZ,aQZ)[71] for the S22 test set.

The DF-MP2 speedups in our study are not as large as might be expected. We were forced to use a core Hamiltonian guess to be consistent between the DF and DB tests, because in Q-Chem 3.2, one cannot use the DB technique in conjunction with more advanced initial orbital guesses. The core Hamiltonian guess requires more SCF iterations to converge, thereby increasing the time spent in SCF. If superior SCF guesses were used, such as superposition of atomic densities (SAD) or a small basis projection, the overall computation would spend less time in the SCF and more time in the MP2 correlation. This would cause the DF methods to have better overall speedups and DB methods to have slightly smaller speedups.

To better understand the errors incurred through these approximations, a decomposition by binding type is shown in Table 2. Reference 63 defines the division of complexes between hydrogen-bonded, dispersion-bound, and mixed-influence subgroups. As shown in Table 2, dispersion-bound complexes experience a larger mean percent error than the hydrogen-bonded subset for every approximation examined, by a factor of 1.7−3.9, thereby suggesting the approximations examined in this work, particularly haXZ, may have difficulty with longer-range interactions. For the CBS limit, we report errors among dispersion-dominated systems of 0.02%, 0.29%, 0.28%, and 1.26% for DF, DB, DB-DF, and heavy-augmented basis sets, respectively, while the corresponding value for hydrogen-bonded complexes in the last case is only 0.64%.

4. Conclusions

This work demonstrates that with a careful choice of approximations, MP2-quality results can be computationally affordable for systems with a few dozen atoms or larger without introducing significant error. Density fitting reduces the time to compute the MP2 correlation energy. Dual-basis techniques abate the cost of the underlying SCF, and heavy-augmented functions speed up both parts of the computation relative to the fully augmented basis sets. Except for comparisons using the smaller heavy-aug-cc-pVDZ basis set, all of these approximations show a significant speedup while never incurring a RMSE greater than 0.045 kcal mol^{-1} for the S22 test cases. We also demonstrate that all of these approximations do indeed combine very efficiently. In future tests, density fitting will be extended to the SCF stage (currently not implemented in Q-Chem). The use of DF within the DB-SCF framework should be a significant stride toward achieving a level of theory that is not only accurate but applicable to a wide range of systems. Q-Chem also will soon have the capability to perform perturbative SCF approaches as outlined in the work of Gill et al.[61] These new computational tools will open up larger systems of interest to *ab initio* techniques while introducing errors which are negigible in most applications.

Acknowledgment. This material is based upon work supported by the National Science Foundation (Grant No. CHE-1011360). The computer resources of the Center for Computational Molecular Science and Technology are funded through an NSF CRIF Award (CHE-0946869).

References

(1) Lee, T. J.; Scuseria, G. E. Achieving Chemical Accuracy with Coupled-Cluster Theory. In *Quantum Mechanical Electronic Structure Calculations with Chemical Accuracy*; Langhoff, S. R., Ed.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995; pp 47−108.

(2) Polyansky, O. L.; Zobov, N. F.; Viti, S.; Tennyson, J.; Bernath, P. F.; Wallace, L. *Science* **1997**, *277*, 346.

(3) Polyansky, O. L.; Császár, A. G.; Shirin, S. V.; Zobov, N. F. *Science* **2003**, *299*, 539–542.

(4) Martin, J. M. L. *Chem. Phys. Lett.* **1998**, *292*, 411–420.

(5) Temelso, B.; Valeev, E. F.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 3068–3075.

(6) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 104–112.

(7) Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. *J. Am. Chem. Soc.* **2002**, *124*, 10887–10893.

(8) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200–10207.

(9) Lee, E. C.; Kim, D.; Jurečka, P.; Tarakeshwar, P.; Hobza, P.; Kim, K. S. *J. Phys. Chem. A* **2007**, *111*, 3446–3457.

(10) Pitoňák, M.; Riley, K. E.; Neogrády, P.; Hobza, P. *ChemPhysChem* **2008**, *9*, 1636–1644.

(11) Sherrill, C. D.; Sumpter, B. G.; Sinnokrot, M. O.; Marshall, M. S.; Hohenstein, E. G.; Walker, R. C.; Gould, I. R. *J. Comput. Chem.* **2009**, *30*, 2187–2193.

(12) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910–1918.

(13) Scuseria, G. E.; Scheiner, A. C.; Lee, T. J.; Rice, J. E.; Schaefer, H. F. *J. Chem. Phys.* **1987**, *86*, 2881.

(14) Scuseria, G. E.; Janssen, C. L.; Schaefer, H. F. *J. Chem. Phys.* **1988**, *89*, 7382.

(15) Wu, Q.; Yang, W. *J. Chem. Phys.* **2002**, *116*, 515–524.

(16) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.

(17) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.

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

(19) Gerenkamp, M.; Grimme, S. *Chem. Phys. Lett.* **2004**, *392*, 229–235.

(20) Hill, J. G.; Platts, J. A. *J. Chem. Theory Comput.* **2007**, *3*, 80–85.

(21) Antony, J.; Grimme, S. *J. Phys. Chem. A* **2007**, *111*, 4862–4868.

(22) Takatani, T.; Sherrill, C. D. *Phys. Chem. Chem. Phys.* **2007**, *9*, 6106–6114.

(23) Jung, Y.; Lochan, R. C.; Dutoi, A. D.; Head-Gordon, M. *J. Chem. Phys.* **2004**, *121*, 9793–9802.

(24) Takatani, T.; Hohenstein, E. G.; Sherrill, C. D. *J. Chem. Phys.* **2008**, *128*, 124111.

(25) Whitten, J. L. *J. Chem. Phys.* **1973**, *58*, 4496–4501.

(26) Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. *Int. J. Quantum Chem. Symp.* **1977**, *11*, 81.

(27) Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. *J. Chem. Phys.* **1979**, *71*, 3396–3402.

(28) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. *Chem. Phys. Lett.* **1993**, *208*, 359–363.

(29) Bernholdt, D. E.; Harrison, R. J. *Chem. Phys. Lett.* **1996**, *250*, 477–484.

(30) Kendall, R. A.; Fruchtl, H. A. *Theor. Chem. Acc.* **1997**, *97*, 158–163.

(31) Weigend, F. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285–4291.

(32) Werner, H.-J.; Manby, F. R.; Knowles, P. J. *J. Chem. Phys.* **2003**, *118*, 8149–8160.

(33) Werner, H.; Manby, F. *J. Chem. Phys.* **2006**, *124*, 054114.

(34) Beebe, N. H. F.; Linderberg, J. *Int. J. Quantum Chem.* **1977**, *12*, 683–705.

(35) Roeggen, I.; Wisloff-Nilssen, E. *Chem. Phys. Lett.* **1986**, *132*, 154–160.

(36) Koch, H.; de Meras, A. S.; Pedersen, T. B. *J. Chem. Phys.* **2003**, *118*, 9481–9484.

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

(38) Boström, J.; Aquilante, F.; Pedersen, T. B.; Lindh, R. *J. Chem. Theory Comput.* **2009**, *5*, 1545–1553.

(39) Weigend, F.; Kattannek, M.; Ahlrichs, R. *J. Chem. Phys.* **2009**, *130*, 164106.

(40) Aquilante, F.; Gagliardi, L.; Pedersen, T. B.; Lindh, R. *J. Chem. Phys.* **2009**, *130*, 154107.

(41) Zienau, J.; Clin, L.; Doser, B.; Ochsenfeld, C. *J. Chem. Phys.* **2009**, *130*, 204112.

(42) Boström, J.; Delcey, M. G.; Aquilante, F.; Serrano-Andrés, L.; Pedersen, T. B.; Lindh, R. *J. Chem. Theory Comput.* **2010**, *6*, 747–754.

(43) Chwee, T. S.; Carter, E. A. *J. Chem. Phys.* **2010**, *132*, 074104.

(44) Martinez, T. J.; Mehta, A.; Carter, E. A. *J. Chem. Phys.* **1992**, *97*, 1876–1880.

(45) Martinez, T. J.; Carter, E. A. Pseudospectral Methods Applied to the Electron Correlation Problem. In *Modern Electronic Structure Theory*; Yarkony, D. R., Ed.; World Scientific: Singapore, 1995; Vol. 2, pp 1132−1165.

(46) Friesner, R. A.; Murphy, R. B.; Beachy, M. D.; Ringnalda, M. N.; Pollard, W. T.; Dunietz, B. D.; Cao, Y. *J. Phys. Chem. A* **1999**, *103*, 1913–1928.

(47) Sherrill, C. D.; Takatani, T.; Hohenstein, E. G. *J. Phys. Chem. A* **2009**, *113*, 10146–10159.

(48) Pulay, P. *J. Phys. Lett.* **1980**, *73*, 323–398.

(49) Pulay, P. *J. Comput. Chem.* **1982**, *3*, 556–560.

(50) White, C.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chem. Phys. Lett.* **1996**, *253*, 268–278.

(51) Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Science* **1996**, *271*, 51–53.

(52) Burant, J. C.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1996**, *105*, 8969–8972.

(53) Schwegler, E.; Challacombe, M. *J. Chem. Phys.* **1996**, *105*, 2726–2734.

(54) Ochsenfeld, C.; White, C. A.; Head-Gordon, M. *J. Chem. Phys.* **1998**, *109*, 1663–1669.

(55) Wolinski, K.; Pulay, P. *J. Chem. Phys.* **2003**, *118*, 9497–9503.

(56) Steele, R. P.; DiStasio, R. A.; Shao, Y.; Kong, J.; Head-Gordon, M. *J. Chem. Phys.* **2006**, *125*, 074108.

(57) Steele, R.; Shao, Y.; DiStasio, R.; Head-Gordon, M. *J. Chem. Phys. A* **2006**, *110*, 13915–13922.

(58) Distasio, R.; Steele, R.; Head-Gordon, M. *Mol. Phys.* **2007**, *105*, 2731–2742.

(59) Steele, R.; Head-Gordon, M. *Mol. Phys.* **2007**, *105*, 2455–2473.

(60) Steele, R. P.; DiStasio, R. A.; Head-Gordon, M. *J. Chem. Theory Comput.* **2009**, *5*, 1560–1572.

(61) Deng, J.; Gilbert, A. T. B.; Gill, P. M. W. *J. Chem. Phys.* **2009**, *130*, 231101.

(62) Shao, Y.; et al. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172–3191.

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

(64) Sodt, A.; Beran, G. J. O.; Jung, Y. S.; Austin, B.; Head-Gordon, M. *J. Chem. Theory Comput.* **2006**, *2*, 300–305.

(65) Sodt, A.; Subotnik, J. E.; Head-Gordon, M. *J. Chem. Phys.* **2006**, *125*, 194109.

(66) Aquilante, F.; Pedersen, T. B. *Chem. Phys. Lett.* **2007**, *449*, 354.

(67) Zienau, J.; Clin, L.; Doser, B.; Ochsenfeld, C. *J. Chem. Phys.* **2009**, *130*, 204112.

(68) Doser, B.; Lambrecht, D. S.; Kussmann, J.; Ochsenfeld, C. *J. Chem. Phys.* **2009**, *130*, 064107.

(69) Weigend, F.; Köhn, A.; Hättig, C. *J. Chem. Phys.* **2002**, *116*, 3175–3183.

(70) Hättig, C. *Phys. Chem. Chem. Phys.* **2005**, *7*, 59–66.

(71) Takatani, T.; Hohenstein, E. G.; Malagoli, M.; Marshall, M. S.; Sherrill, C. D. *J. Chem. Phys.* **2010**, *132*, 144104.

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

(73) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243–252.

CT100468F

JCTC Journal of Chemical Theory and Computation

Global Hybrid Functionals: A Look at the Engine under the Hood

Gábor I. Csonka,*[,†,‡] John P. Perdew,[‡] and Adrienn Ruzsinszky[‡]

Department of Inorganic and Analytical Chemistry, Budapest University of Technology and Economics, H-1521 Budapest, Hungary and Department of Physics and Quantum Theory Group, Tulane University, New Orleans, Louisiana 70118, United States

Received August 26, 2010

Abstract: Global hybrids, which add a typically modest fraction of the exact exchange energy to a complement of semilocal exchange−correlation energy, are among the most widely used density functionals in chemistry and condensed matter physics. Here we briefly review the formal and practical advantages and disadvantages of global hybrids. We point out that empiricism seems unavoidable in the construction of global hybrids, as it is not for most other kinds of density functional. Then we use one to three parameters to hybridize many semilocal functionals (including recently developed nonempirical generalized gradient approximations or GGA's and meta-GGA's). We study the performance of these global hybrids for many properties of sp-bonded molecules composed from the lighter atoms of the periodic table: atomization energies, barrier heights, reaction energies, enthalpies of formation, total energies, ionization potentials, electron affinities, proton affinities, and equilibrium bond lengths. We find several new global hybrids that perform better in these tests than standard ones, and we correct some errors in literature assessments. We also discuss the representativity of small fitting sets and the adequacy of various Gaussian basis sets.

1. Introduction to Semilocal and Global Hybrid Functionals

For the description of electronic states in atoms, molecules, and solids, correlated wave function methods can be accurate but very expensive. Kohn and Sham[1] showed that, given the exact density functional for the exchange−correlation energy, the ground-state energy and density of interacting electrons in a multiplicative external potential could be found exactly and efficiently from a single Slater determinant of self-consistent orbitals that see a multiplicative effective potential. The exchange−correlation energy arises from the tendency of electrons to avoid one another due to fermion antisymmetry and Coulomb repulsion. Without it, chemical binding would be weak or absent.[2]

The exact exchange−correlation energy can be expressed as a coupling constant integration or adiabatic connection formula:[3−45]

$$E_{xc} = (1/2)\int d^3r \int d^3r'\, n(\vec{r}) n_{xc}(\vec{r},\vec{r}')/|\vec{r}'-\vec{r}| \quad (1)$$

Here $n_{xc}(\vec{r},\vec{r}') = n_x(\vec{r},\vec{r}') + n_c(\vec{r},\vec{r}')$ is the density at position $\vec{r}'$ of the exchange−correlation hole around an electron at $\vec{r}$. Note that

$$n_{xc}(\vec{r},\vec{r}') = \int_0^1 d\lambda [n_x(\vec{r},\vec{r}') + n_{c,\lambda}(\vec{r},\vec{r}')] \quad (2)$$

is averaged over the strength λ of the Coulomb repulsion $\lambda/|\vec{r}' - \vec{r}|$ at fixed electron density $n(\vec{r})$. The holes satisfy important sign and sum rules:[3−5]

$$n_x(\vec{r},\vec{r}') \le 0, \quad \int d^3r' n_x(\vec{r},\vec{r}') = -1, \quad \int d^3r' n_c(\vec{r},\vec{r}') = 0 \quad (3)$$

* Address correspondence to csonkagi@gmail.com.
† Budapest University of Technology and Economics.
‡ Tulane University.

10.1021/ct100488v
Published on Web 11/18/2010

The first approximation for the spin density functional for exchange and correlation, the local spin density approximation[1,6,7] (LSDA), was constructed to be exact for a uniform electron gas and subsequently shown[3−5] to satisfy the hole constraints of eq 3. Later semilocal functionals[8−15] were constructed nonempirically to satisfy these and many other physical or exact constraints. Although recent nonempirical functionals are not constructed from hole models, constraint-satisfying hole models[16−18] have been reverse engineered from them. All nonempirical semilocal functionals, including LSDA, are by construction exact in the uniform density limit. Some semilocal functionals (e.g., refs 19 and 20) have also been constructed empirically, by fitting to chemical data.

A semilocal (sl) spin density functional:

$$E_{\mathrm{xc}}^{\mathrm{sl}}[n_\uparrow, n_\downarrow] = \int \mathrm{d}^3 r n(\vec{r}) \varepsilon_{\mathrm{xc}}^{\mathrm{sl}}(n_\uparrow(\vec{r}), n_\downarrow(\vec{r}), \nabla n_\uparrow(\vec{r}), \nabla n_\downarrow(\vec{r}), \tau_\uparrow(\vec{r}), \tau_\downarrow(\vec{r})) \quad (4)$$

can be evaluated as a computationally efficient single integral, using ingredients available at position $\vec{r}$, such as the local spin densities $n_\sigma(\vec{r})$ with $\sigma = \uparrow$ or $\downarrow$ (as in LSDA) or additionally their gradients $\nabla n_\sigma(\vec{r})$ as in a generalized gradient approximation (GGA)[8−12,14,19,20] or further the spin-resolved positive orbital kinetic energy densities $\tau_\sigma(\vec{r})$ (as in a meta-GGA).[12,15,21] Fully nonlocal approximations,[22] which require a double integral as in eq 1, can be considerably more costly. For this reason, an efficient method[23] was developed to deal with the full nonlocality of an early van der Waals density functional.[24] Semilocal approximations can be expected to work only when the exact exchange−correlation hole density is well-localized around its electron,[25] as it is in atoms and in many molecules and solids near equilibrium, and even when (at both the interacting and noninteracting levels, the latter possibly with the help of symmetry breaking)[26] electrons are not shared between separated subsystems in the dissociation limit.[27] However, stretched bonds that arise in transition states of chemical reactions or in the dissociation limits of some radical or heteronuclear molecules, etc., require full nonlocality.[25,28−30] Long-range van der Waals interactions also require full nonlocality.

In a typical atom, exchange and correlation can be separately described by good semilocal functionals. In the valence region of a typical molecule near equilibrium, both the exact exchange energy and the exact correlation energy are fully nonlocal, but their full nonlocalities tend to cancel. Thus semilocal approximations for the exchange−correlation energy can work even there. This error cancellation arises because the exact n_{xc} tends to be deeper and more short-ranged in $|\vec{r}' - \vec{r}|$ (and thus more semilocal) than either the exact $n_{\mathrm{x}}(\vec{r},\vec{r}')$ or the exact $n_{\mathrm{c}}(\vec{r},\vec{r}')$ alone. While the dynamic correlation energy of the molecule is approximated by $E_{\mathrm{c}}^{\mathrm{sl}}$, its static or left-right[31] fully nonlocal correlation energy can be estimated from the negative quantity $(E_{\mathrm{x}}^{\mathrm{sl}} - E_{\mathrm{x}}^{\mathrm{exact}}) - \sum(\mathrm{atoms})(E_{\mathrm{x}}^{\mathrm{sl}} - E_{\mathrm{x}}^{\mathrm{exact}})$, where the last sum is over the constituent free atoms.

The global hybrid (gh) idea, due to Becke,[32,33] introduces some full nonlocality into the calculation but only at the level of $E_{\mathrm{x}}^{\mathrm{exact}}$, which can be evaluated semianalytically from the Kohn−Sham orbitals in some computer codes. In its simplest (one parameter) version:[34]

$$E_{\mathrm{xc}}^{\mathrm{gh}} = aE_{\mathrm{x}}^{\mathrm{exact}} + (1 - a)E_{\mathrm{x}}^{\mathrm{sl}} + E_{\mathrm{c}}^{\mathrm{sl}} \quad (5)$$

where the exact-exchange mixing parameter a takes an empirical value in the range $0 \le a \le 1$. For a in this range, both a and $1 - a$ are positive, so all the hole constraints of eq 3 are preserved, including the sign constraint on n_{x}. Because the hybrid functional statistically improves the atomization energies over its semilocal parent, a better estimate of the static correlation energy in the molecule is thus $(1 - a)\{(E_{\mathrm{x}}^{\mathrm{sl}} - E_{\mathrm{x}}^{\mathrm{exact}}) - \sum(\mathrm{atoms})(E_{\mathrm{x}}^{\mathrm{sl}} - E_{\mathrm{x}}^{\mathrm{exact}})\}$, which is more nearly independent of the choice of $E_{\mathrm{x}}^{\mathrm{sl}}$ when a is fitted to the atomization energies for that choice. Because eq 5 is linear in $E_{\mathrm{x}}^{\mathrm{exact}}$, it is properly size extensive: It makes the energy of a system of well-separated subsystems equal to the sum of the energies of the subsystems. Nonlinear mixing should only (and cautiously) be done at deeper levels of the exact-exchange ingredient, such as the exact-exchange energy density $n(\vec{r})\varepsilon_{\mathrm{x}}^{\mathrm{exact}}(\vec{r})$ (as in the local hybrids[25] discussed around eq 6) or even the exact-exchange hole density $n_{\mathrm{x}}^{\mathrm{exact}}(\vec{r},\vec{r}')$ (as in range-separated hybrids, e.g., ref 35). Note that the range-separated or screened exchange hybrid of Heyd, Scuseria, and Ernzerhof (HSE)[36] is designed to imitate the performance of the standard Perdew−Burke−Ernzerhof (PBE) global hybrid PBE0, while eliminating some of the practical problems of exact exchange in metals.

Perhaps the most convincing argument for eq 5 is the least sophisticated one: Semilocal exchange−correlation typically overestimates atomization energies and underestimates energy barriers to chemical reactions, while exact exchange without correlation makes errors of the opposite sign, so that mixing of the two will yield better atomization energies and barriers than either alone.

Becke[32] also justified eq 5 on the basis of the adiabatic connection in eqs 1 and 2: The errors of semilocal exchange and correlation are expected to cancel substantially at $\lambda = 1$ but cannot cancel at $\lambda = 0$ where only the exchange hole survives (since $n_{\mathrm{x}} \sim \lambda^0$, while $n_{c,\lambda} \sim \lambda^1$), so some exact exchange should be mixed with semilocal exchange−correlation. This argument has been quantified[34] to predict a mixing parameter $a = 0.25$, which is the basis for the standard PBE hybrid PBE0.[37,38] This prediction for a uses a model for the λ dependence of $\Delta E_{\mathrm{c},\lambda}$ and must be understood for what it is. First, it applies only to atomization energies, not to other properties. Second, it relies on the qualitative empirical observation that fourth-order perturbation theory in the electron−electron interaction is fairly accurate for atomization energies. Third, even for atomization energies (or for formation enthalpies based upon calculated atomization energies), it is only a rough estimate, reliable within perhaps a factor of 2. As we will see below, values of a fitted to formation enthalpies vary from from 0.60 for PBEsol[14] GGA to 0.32 for the PBE[11] GGA to 0.1 for the Tao−Perdew−Staroverov−Scuseria (TPSS)[13] and revised TPSS (revTPSS)[15] meta-GGA's. Improving the underlying semilocal functional reduces the value of a needed to fit the atomization energies or formation enthalpies and so worsens the barrier heights. The converse is that semilocal functionals

with highly overestimated atomization energies, like the PBE variant for solids (PBEsol),[14] can hybridize to produce excellent barrier heights, as we will also show. Some global hybrids[39,40] thus employ highly fitted semilocal parts intended not to be accurate by themselves but to work well with a fraction of exact exchange. Starting from a standalone semilocal functional, some fraction of exact exchange can be expected to improve most calculated properties, but there is no reason to expect that a single fraction will be optimal for most or all.

Global hybrids with $a < 1$ typically satisfy most of the exact constraints satisfied by the underlying semilocal functional (and all of those for TPSS and revTPSS) but cannot satisfy any additional exact constraints and thus are unavoidably empirical. To satisfy additional constraints and to make optimum use of sophisticated nonempirical meta-GGA's, one can go to the local hybrid[25,41,42] (lh) level in which the mixing fraction $a(\vec{r})$ varies with position:

$$E_{xc}^{lh} = \int d^3r\, n(\vec{r})[a(\vec{r})\varepsilon_x^{exact}(\vec{r}) + \{1 - a(\vec{r})\}\varepsilon_x^{sl}(\vec{r})] + E_c^{sl} \quad (6)$$

where $\varepsilon_x(\vec{r})$ is the exchange energy per electron at $\vec{r}$. But local hybrids seem to require even more empirical parameters than global hybrids do, because the additional exact constraints are only limits: Parameters are needed to control how these limits are approached.[25] Local hybrids also have a potential problem which global ones do not: the nonuniqueness of energy densities like $n(\vec{r})\varepsilon_x(\vec{r})$.[43,44]

Global hybrids often achieve useful accuracy, even in cases where their success is unexplained. They improve atomization energies and related formation enthalpies of molecules as well as energy barriers for reasons we have already discussed. But they also improve equilibrium bond lengths, and they can slightly improve or worsen total energies, ionization energies, electron affinities, and proton affinities,[45] as we shall see below. In solid-state physics, they can often improve the description of strongly correlated solids.[46] Improvements to fundamental band gaps[47] and point defect energies[48] have also been reported. It should be remembered, however, that the improved orbital-energy gaps arise from the use of a fraction of the nonmultiplicative Hartree−Fock exchange potential. Using instead an optimized effective Kohn−Sham multiplicative potential for the fraction of exact exchange would reduce the orbital-energy gaps back to the semilocal range, without much affecting ground-state energies and energy differences.[49,50] Thus explanations of the solid-state successes of global hybrid functionals, based upon their improved orbital-energy gaps, are unconvincing or at least incomplete. There are also documented cases where standard global hybrids are consistently less accurate or as inaccurate as semilocal functionals, including properties of transition-metal compounds[51] and the adsorption energies of CO on transition-metal surfaces.[52]

While the exact total energy of a separated open system varies linearly as a function of average electron number between adjacent integers,[53] the energy predicted by a semilocal density functional approximation is concave upward, and the exact-exchange or Hartree−Fock energy is concave downward.[54] As a result, semilocal functionals can fail[53] for separated open systems of fluctuating electron number, such as the fragments of stretched molecular AB[28,55] or A_2^+.[29] Global hybrid functionals with sufficient exact exchange can clearly fix these problems,[55] in part. For example, we have computed the many-electron self-interaction error[29] $\Delta = E(He_2^+, R = 200\ Å) - E(He) - E(He^+)$, for which the exact value is 0 kcal/mol. We find that Δ is determined mainly by the value of a, and not by the choice of semilocal functional, and that Δ is almost linear in a. With the very different semilocal functionals BPW91, PBE,[11] or PBEsol,[14] we find $\Delta = -96$ kcal/mol for $a = 0$ and $\Delta = -31$ kcal/mol for $a = 0.60$. We observed a similar tendency for dissociating Li_2^+ and F_2^+.

In the next section, we will discuss three-parameter global hybrids as originally proposed by Becke,[32] which not only fit a fraction of exact exchange but also scale the gradient or inhomogeneity corrections to LSDA in the underlying semilocal functional. Our plausibility arguments for this are that the optimal inhomogeneity corrections are less well-known (and indeed less well-defined for a given set of arguments in eq 4) than is the uniform-density limit and that the optimal inhomogeneity corrections change when some exact exchange is included. Then we will construct one- or three-parameter global hybrids for many semilocal functionals, including the recently developed PBEsol[14] GGA and revTPSS[15] meta-GGA that have not been hybridized before, and assess them for a wide range of molecular properties.

2. B3PW91 and B3LYP Functionals

In 1993 Becke suggested the B3PW91 functional in the following form:[32]

$$E_{xc}^{B3PW91} = a \cdot E_x^{exact} + (1 - a) \cdot E_x^{LSDA} + b \cdot \Delta E_x^{B88} + E_c^{LSDA} + c \cdot \Delta E_c^{PW91} \quad (7)$$

where b and c are multiplying factors that modify the exchange and correlation gradient corrections, respectively. Note that choice $b = 1 - a$ and $c = 1$ leads to eq 5 when the semilocal functional is BPW91 (B88 GGA exchange[19] plus PW91 GGA correlation).[10] Because the b and c parameters provide more freedom than the single-parameter hybrid of eq 5, the three-parameter form of eq 7 can give somewhat better results if all three parameters are fitted to experimental data. (We denote our new one parameter hybrids with a subscript a, as in BPW91h$_a$, and our new three parameter hybrids with a subscript abc, as in BPW91h$_{abc}$). Later the B3LYP empirical hybrid functional[56] was constructed with exactly the same $a = 0.2$, $b = 0.72$, and $c = 0.81$ parameters proposed by Becke but with PW91 correlation swapped out. Instead the linear combination $(1 - c) \cdot E_c^{SVWN3} + c \cdot E_c^{LYP}$ was used for the correlation energy in eq 7, where E_c^{SVWN3} is the LSDA correlation energy in the random phase approximation (which does not yield the correct PW92[7] or SVWN5[57] homogeneous electron gas limit). The LYP functional is self-correlation error free, but it does not give the correct uniform gas limit and so produces serious errors for metallic solids.[58] However, despite these construction problems, the B3LYP functional delivered good

results for small molecules, and it was preferred in many areas of chemistry over the B3PW91 and other functionals. The evidence for this is strong. Each of refs 20 and 32 has been cited about 30 000 times, according to the Web of Science. Recent results show several serious failures of the B3LYP functional that may arise from its design problems.[59−64] B3PW91 gives consistently better results for large organic molecules.[45,60] This better performance of B3PW91 can be attributed to its correct behavior for the homogeneous electron gas. (A similar attribution was made for B3PW91 versus B3LYP atomization energies of metals).[63] Note that part of the self-interaction error of BPW91 is removed in B3PW91 and B3LYP.

3. Test Sets

We shall use the following test sets:

In their early works Adamo and Barone use zero-point energy (ZPE) corrected atomization energies, $\sum D_0$, of the small G2-32 test set to measure the performance of the one-parameter hybrids (cf. eq 5).[38,65] The G2-32 test set was derived from the G2/97 test set[66] by selecting the first 32 compounds that contain first-row elements (Z < 10).

The AE6 test set of Lynch and Truhlar[67] provides a quick and supposedly representative evaluation for the ZPE-subtracted atomization energy, $\sum D_e$, of diverse molecular systems. The set includes six molecules: SiH_4, S_2, SiO, C_3H_4, $C_2H_2O_2$, and C_4H_8.

The Kinetics9 database[68,69] contains three forward barrier heights, three reverse barrier heights, and three energies of reaction for the three reactions in the BH6[67] database. We use the optimized QCISD/MG3 geometries and the recent spin−orbit corrected reference energies for AE6 and BH6.[70]

The G3/99 test set[71] includes 223 standard enthalpies of formation (without deleting the COF_2 molecule, contrary to the recommendation in ref 71), 88 ionization potentials, 58 electron affinities, and 8 proton affinities for compounds that contain first- and second-row elements (Z < 18). For the development of the first hybrid functionals, a smaller G2/97 test set was used (148 standard enthalpies of formation). The G3-3 test set includes larger organic molecules and several problematic inorganic compounds. The G3-3 and the G2/97 test sets together form the G3/99 test set.

The IP86 test set for gas-phase ionization potentials was defined in ref 45. As the pure $a = 0$ functionals have convergence problems for H_2S^+ and N_2^+, these were left out from the 88 species of the original G3/99 test set.

The EA58 test set for gas-phase electron affinities and the PA8 test set for gas phase proton affinities were taken from the G3/99 test set.[71]

The T-96R contains equilibrium internuclear distances (r_e) of the 86 neutral diatomic molecules and 10 diatomic molecular cations.[45]

In this paper we shall show how one-parameter global hybrids can be constructed from the BLYP, BPW91, PBE, PBEsol, TPSS, and revTPSS functionals. The PBEsol and revTPSS functionals were programmed by us. As we show later, the selection of the PBEsol functional is particularly interesting as it overbinds the most among the functionals in this study. Thus it requires the largest exact-exchange fraction for optimal performance on atomization energy test sets. The results obtained by the new functionals will be compared to the results obtained by the original three-parameter B3PW91 and B3LYP functionals. We will discuss:

- The dependence of the results on the test sets and the representativity of the G2-32 and AE6 test sets compared to the G3/99 test set.
- The role of the basis sets.
- Is the original B3PW91 parametrization optimal for the G3/99 test set?
- How do parameter values influence the performance of the functionals?
- Is it possible to find better hybrid parameters for eqs 5 and 7 than the currently used values?

4. Thermochemical Properties

4.1. Results for the G2-32 Test Set. First we tried to reproduce the earlier[38,65] ZPE-corrected atomization energies, $\sum D_0$, for the hybrids B1LYP, B1PW91, and PBE0 constructed from the BLYP, BPW91, and PBE functionals for the G2-32 test set with $a = 0.25$. We use the Gaussian 03 program for all calculations,[72] the B3LYP/6-31G(2df, p) optimized geometries, the ultra-fine grid, and the 6-311++G(3df, 3pd) basis set, as in most of the original publications. The amount of 25% exact exchange was set via Iop(3/76 = 0750002500) of the GAUSSIAN 03 program. While we were able to reproduce PBE0 results,[38] our B1LYP and B1PW91 results deviate considerably from the earlier results[65] (cf. Supporting Information). According to our results the PBE0, B1LYP, and B1PW91 mean absolute errors (MAE) are 2.8, 4.8, and 5.0 kcal/mol, respectively, compared to the published 2.6, 3.1, and 5.4 kcal/mol.[38,65] Thus the good PBE0 results were reproduced, but the performance of B1LYP is not significantly better than that of B1PW91, in disagreement with the results of Adamo et al.[65] The differences between the published B1LYP and our results are the following (cf. Supporting Information): (1) The calculated $\sum D_0$ energies do not agree for BeH (ours vs Adamo et al.: 54.4 vs 48.5 kcal/mol), ethane (656.0 vs 654.1 kcal/mol), and hydrazine (402.1 vs 401.2 kcal/mol). For 22 compounds our results agree precisely, and for the other 7 compounds the difference is less than 0.6 kcal/mol. (2) Using the calculated $\sum D_0$ energies by Adamo et al., we obtain MAE = 4.8 kcal/mol, in agreement with our current results. Consequently the published value for the MAE of B1LYP (3.1 kcal/mol)[65] should be corrected. (See the detailed energies in the Supporting Information). Below we show that the BLYP functional is not suitable for the construction of a one parameter hybrid functional. The B1LYP functional yields worse thermochemistry than the BLYP functional for more adequate AE6 and G3/99 test sets. The origin of the 'success' of the B1LYP functional for the G2-32 test set is the inappropriate choice of the test set.

4.2. Results for the AE6, BH6, and K9 Test Sets. Next we used small but so-called representative test sets. Analysis of mean error (ME) and MAE of the thermochemical results obtained for AE6, BH6, and Kinetics9 shows the following

Table 1. Summary of Deviations (ME and MAE) from Experiment of the Atomization Energies in the AE6 set and the Reaction Energy Barriers in the BH6 set As Well As the RMSE for the Kinetics9 (K9) Test Set Calculated with the 6-311+G(3df, 3pd) Basis Set[a]

				AE6		BH6		K9
functional	*a*	*b*	*c*	ME	MAE	ME	MAE	RMSE
BLYP	0.00	1.00	1.00	−1.7	6.6	−8.1	8.1	7.1
BPW91	0.00	1.00	1.00	2.4	6.9	−7.7	7.7	6.9
PW86PBE	0.00	1.00	1.00	2.1	7.8	−8.1	8.1	7.4
PBE	0.00	1.00	1.00	12.3	15.3	−9.6	9.6	8.8
PBEsol	0.00	1.00	1.00	35.8	35.8	−13.0	13.0	12.0
TPSS	0.00	1.00	1.00	3.9	5.6	−8.6	8.6	7.4
revTPSS	0.00	1.00	1.00	3.1	5.9	−7.6	7.6	7.6
M05-2X	0.56	n.a.	n.a.	0.2	2.6	−0.5	**1.4**	**1.4**
M06-2X	0.54	n.a.	n.a.	−0.2	**1.2**	−0.7	**1.2**	**1.1**
B3LYP	0.20	0.72	0.81	−2.2	2.6	−5.1	5.1	4.5
B3PW91	0.20	0.72	0.81	−0.3	4.0	−4.7	**4.7**	**4.0**
BPW91h$_a$	0.10	0.90	1.00	−1.6	5.4	−5.9	5.9	5.2
BPW91h$_{abc}$	**0.15**	**0.75**	**0.35**	−0.5	**2.1**	−4.2	**4.2**	**4.0**
BPW91h$_{abc}$	0.15	0.79	0.75	−0.6	3.1	−5.0	**5.0**	**4.4**
BPW91h$_{abc}$	0.15	0.80	0.85	−0.6	3.7	−5.2	5.2	4.6
BPW91h$_{abc}$	0.15	0.81	0.95	−0.6	4.3	−5.4	5.4	4.8
BPW91h$_a$	0.15	0.85	1.00	−3.5	5.2	−5.1	5.1	4.5
PW86PBEh$_a$	0.15	0.85	1.00	−3.7	5.5	−5.4	5.4	4.9
BPW91h$_{abc}$	**0.20**	**0.70**	**0.57**	−0.6	**2.1**	−5.0	**5.0**	**4.4**
BPW91h$_{abc}$	0.20	0.72	0.77	−0.6	3.4	−4.6	**4.6**	**3.9**
BPW91h$_a$	0.20	0.80	1.00	−5.4	6.2	−4.2	**4.2**	**3.7**
BPW91h$_{abc}$	**0.21**	**0.70**	**0.57**	−1.8	**2.1**	−4.3	**4.3**	**3.7**
PBE0	0.25	0.75	1.00	0.3	6.2	−4.9	**4.9**	4.5
PBEh$_a$	**0.32**	**0.68**	**1.00**	−2.8	5.3	−3.7	**3.7**	**3.5**
PBEsolh$_a$	0.50	0.50	1.00	3.2	10.0	−2.6	**2.6**	**2.6**
PBEsolh$_a$	**0.60**	**0.40**	**1.00**	−2.9	11.1	−0.9	**1.4**	**1.8**
TPSSh	0.10	0.90	1.00	0.6	6.1	−7.0	7.0	6.3
revTPSSh$_a$	0.10	0.90	1.00	0.1	7.8	−6.2	6.2	6.7

[a] The weight of exact exchange (*a*) and the *b* and *c* parameters are also shown for the hybrid functionals (cf. eq 7). Note that one-parameter hybrids have $a > 0$, $b = 1 - a$, and $c = 1$. In our notation, h$_a$ and h$_{abc}$ denote global hybrids with new or refitted empirical parameters *a* and *a*, *b*, and *c*. Note also that Becke's original B3PW91 has $a = 0.20$, $b = 0.72$, and $c = 0.81$. All values are in kcal/mol. Error = theory − experiment. The results that are better than the B3LYP results are shown bold. The mean experimental atomization energy for AE6 is 517.8 kcal/mol, and the mean barrier height for BH6 is 11.7 kcal/mol.

(cf. Table 1): The ZPE-subtracted atomization energy, $\sum D_e$, results in Table 1 show that most of the GGA functionals yield positive ME for the AE6 test set (error = theory − experiment) and thus partly conserve the overbinding tendency of the LDA. PBEsol, which uses the exact second-order gradient coefficient for exchange valid for the slowly varying limit over a wide range of *s*, shows the largest overbinding, while PBE uses a coefficient about twice as big and shows a considerably reduced overbinding, about one-third of that for PBEsol. The BPW91, PW86PBE, TPSS, and revTPSS functionals give much better results, small but positive MEs (2−4 kcal/mol) and relatively small MAEs (6−8 kcal/mol). The TPSS and revTPSS functionals give the best results. The BLYP functional differs qualitatively from these functionals, as it underbinds slightly (ME = −1.7 kcal/mol).

Increasing the weight of the exact exchange according to eq 5 decreases the ME values (increases the underbinding). This leads to improvements of the MAE of the overbinding functionals for the AE6 test set (cf. BPW91h$_a$ and PBEh$_a$ in Figure 1). However, as the BLYP functional underbinds for the AE6 test set, no improvement is possible via exact exchange mixing alone (eq 5). The good B3LYP results for the AE6 test set mainly come from the reduced gradient contribution to the exchange and from mixing the LDA and LYP correlation functionals (cf. eq 7, *b* and *c* parameters and the results in the Table 1). Note that the BLYP functional overbinds for the G2-32 test set (ME = 4.3 kcal/mol, cf. Table S1, Supporting Information) and can be improved by the exact-exchange mixing according to eq 5. In order to resolve this contradiction between AE6 and G2-32 test set, we shall check the validity of these results on the large G3/99 test set (vide infra).

The BPW91, PW86PBE, PBE, and PBEsol overbinding for the AE6 test set is readily decreased by the simple inclusion of exact exchange (eq 5). We observed for BPW91h$_a$, PBEh$_a$, and PBEsolh$_a$ hybrids that the ME shows a quasilinear dependence on the value of the parameter *a*. Close to the optimal values of *a*, the slopes are −38 (BPW91h) and −45 kcal/mol (PBEh). Figure 1 shows how the larger positive ME is compensated by a larger amount of exact exchange for the BPW91 and PBE functionals. The optimal values for *a* are 0.15, 0.15, 0.32, and 0.5 for BPW91h$_a$, PW86PBEh$_a$, PBEh$_a$, PBEsolh$_a$, respectively (cf. Table 1). Notice that meta-GGAs behave differently. Inclusion of 10% of exact exchange worsens the MAEs for the AE6 test set despite some improvements for the MEs.

Applying $a = 0.2$ for the BPW91h$_{abc}$ hybrid as proposed by Becke gives too strong underbinding (cf. eq 5 and Table 1). However, this is compensated by the reduced exchange and correlation gradient contributions via *b* and *c* parameters in B3PW91 (cf. eq 7). The *b* and *c* parameters compensate each other's effect on the ME. A similar compensation effect was observed earlier for H_2 bond distance, total energy, and electron density.[73] We observed that, at fixed *a*, a 0.01 increase of the value of *b* is compensated by a 0.1 increase of the value of *c*. For the BPW91h hybrid functional with $a = 0.15$, we can obtain ME = −0.6 kcal/mol with several different combinations *b* and *c*. Table 1 shows that the same ME can be obtained with $b = 0.81$ and $c = 0.95$, $b = 0.80$ and $c = 0.85$, $b = 0.79$ and $c = 0.75$, or even $b = 0.75$ and $c = 0.35$. This latter combination provides also the smallest (2.1 kcal/mol) MAE, and this can be reproduced with $a = 0.20$, $b = 0.70$, and $c = 0.57$ values in eq 7. These BPW91h$_{abc}$ hybrids are considerably better than the B3LYP or B3PW91 functionals for the AE6 test set (cf. MAE = 2.6 and 4.0 kcal/mol, respectively, in Table 1). Note the excellent performances of the M05-[74] and M06-2X[75] functionals for these test sets (c.f. Table 1). The AE6 and BH6 species were included in the extensive (more than 400 species) fitting sets of these hybrid meta-GGA functionals. The M06-2X functional performs the best and it applies more than 30 fitted parameters. The M05- and M06-2X calculations were performed with the Gaussian 09 program.[76]

Table 1 shows that, among the semilocal functionals applied to the BH6 and K9 test sets, the revTPSS functional gives the best (but poor) results. The BPW91, BLYP, and PW86PBE results are only slightly worse. The PBE and PBEsol functionals give the worst results. The hybrids give

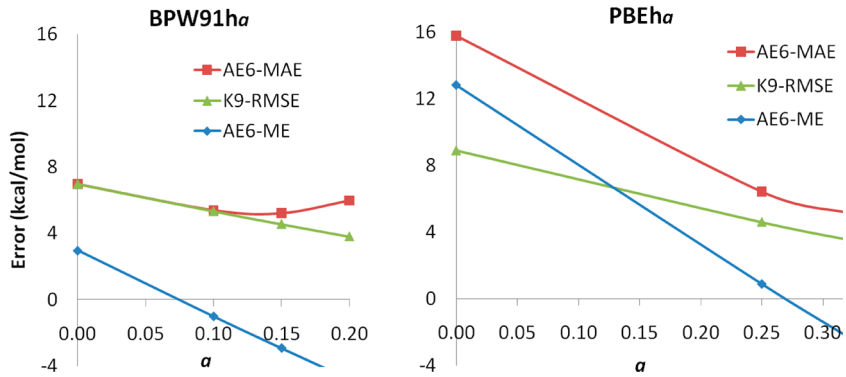


Figure 1. Quasilinear dependence of the ME for the AE6 test set as a function of the weight of exact exchange (*a*) for the BPW91 and PBE hybrids (cf. eq 5). The MAE for the AE6 test set and the RMSE for Kinetics9 (K9) test set are also shown.

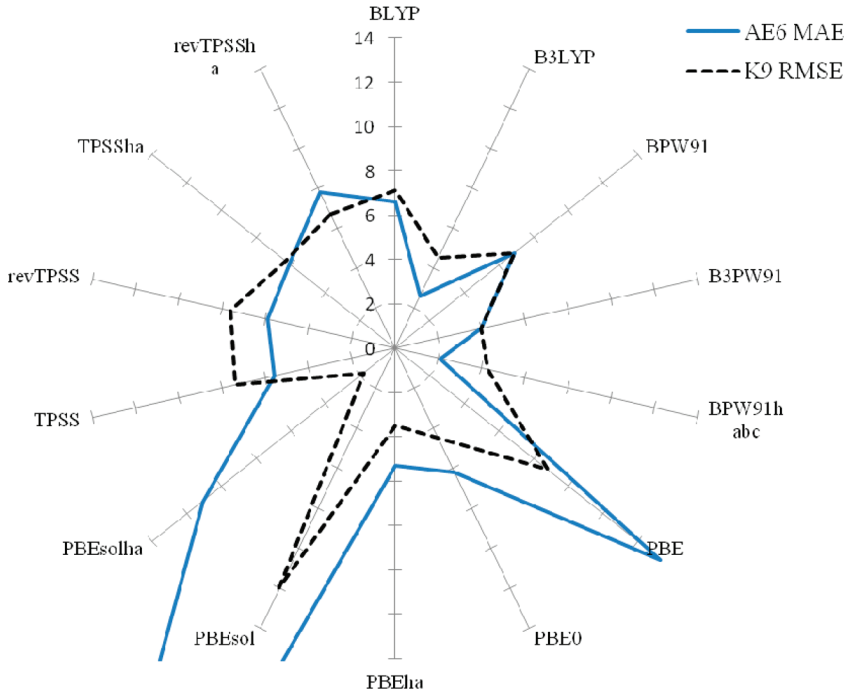


Figure 2. Radar or polar chart summary of the MAE's (kcal/mol) for the AE6 test set and the RMSEs (kcal/mol) for the Kinetics9 (K9) test set for several density functionals discussed in this paper. The connecting lines are simply for guiding the eye. Smaller value means better performance.

improved results for the energy barriers and for kinetics too. The larger weight of the exact exchange considerably decreases the root-mean-square errors (RMSE) for the K9 test sets, and $a = 0.6$ in eq 5 leads to the excellent performance of the PBEsol hybrid (RMSE = 1.8 vs 12.0 kcal/mol for PBEsol). Here again the M06-2X functional shows the best performance ($a = 0.54$, RMSE = 1.1 kcal/mol).

The statistical results for several nonempirical or at most three-parameter functionals are visualized in Figure 2. The radar chart of the MAEs (kcal/mol) for the AE6 test set and the RMSEs (kcal/mol) for the Kinetics9 (K9) test set shows the improvements due to the hybrid functional. Notice that B3PW91 improves the AE6 and K9 results about the same extent but B3LYP does not. BPW91h$_{abc}$ improves only the AE6 results compared to B3PW91. The new PBEh$_a$ ($a = 0.32$) results are uniformly better than those of the PBE0 ($a = 0.25$). Notice that the PBEsolh$_a$ functional performs poorly for the atomization energies (MAE = 11.1 kcal/mol). However, after removing the errors arising from the inconsistencies in the energies of the free atoms and molecules, as suggested in ref 60, we obtain MAE = 2.6 kcal/mol.

Figure 3 shows how the errors of individual PBEsol hybrid atomization energies for the molecules of the AE6 test set depend on the weight of the exact exchange. Increasing the weight of the exact exchange changes the atomization energies in the direction of underbinding for five molecules (see negative slopes in Figure 3), and the effect is the strongest for $C_2H_2O_2$. For SiH_4 the opposite effect can be observed: a very slight overbinding effect occurs. (See the positive slope in Figure 3.) The figure shows that the optimal value of a depends on the composition of the test set. Consequently it is advisible to use larger test sets for parameter optimizations (vide infra).

4.3. Results for the G3/99 Test Set. The results for the enthalpies of formation $\Delta_f H^0_{298}$ of the large G3/99 test set (223 enthalpies of formation) will be discussed here. In this work we apply the procedure of the G3X theory, which uses

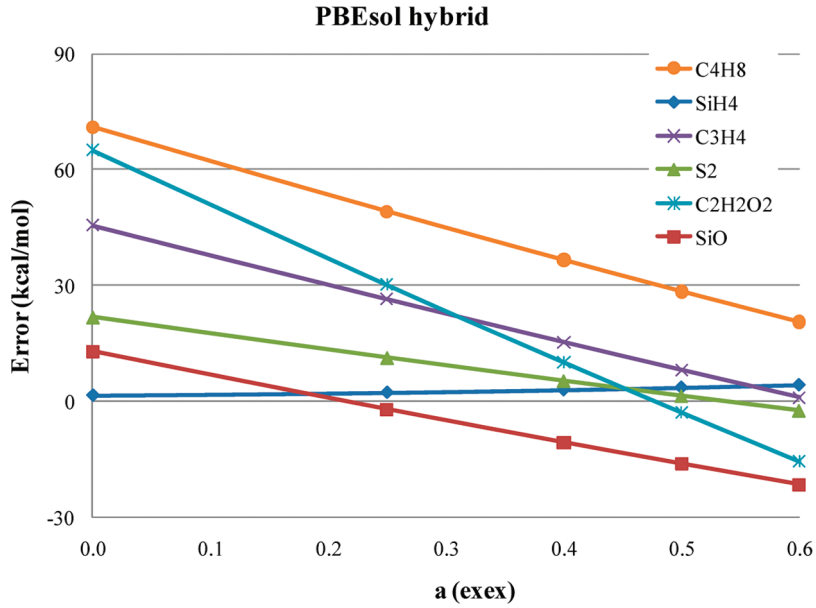


Figure 3. Quasilinear dependence of the errors of the atomization energies of the components of the AE6 test set as functions of the weight of exact exchange (*a*) for the PBEsol hybrid (cf. eq 5).

the equilibrium B3LYP/6-31G(2df, p) geometries in combination with the B3LYP/6-31G(2df, p) zero-point vibration energies and thermal corrections obtained with a frequency scale factor of 0.9854.[77] Note that the $\Delta_f H^0_{298}$ values are calculated from the negatives of the $\sum D_e$ values, atomic enthalpies, and thermal corrections; see eq 1 of ref 60. Thus a more negative ME of the calculated $\Delta_f H^0_{298}$ values means a change in the overbinding direction. The results in Table 2 show that BLYP underbinds (ME = 3.81 kcal/mol) in the same way as for the AE6 test set. Thus the overbinding experienced for the G2-32 test set is seemingly accidental, due to the nonrepresentativity of the latter test set. The other GGAs overbind the same way as for the AE6 test set (cf. ME = −5.6 kcal/mol for BPW91 or ME = −21.69 kcal/mol for PBE in Table 2). Here we include the combination of the PW86 exchange and the PBE correlation functional, since this functional is free of almost all the attractive dispersion forces for noble gas dimers, and thus it is particularly suitable for an additive dispersion correction.[78] We also include TPSS and revTPSS functionals that perform particularly well for thermochemistry even without the inclusion of the exact exchange. We shall present here the first hybrid constructed from revTPSS, the revTPSSh$_a$.[45]

Inspection of Table 2 shows that many of the new hybrid functionals perform considerably better than B3LYP or B3PW91 (cf. Table 2). The best performer is the BPW91h$_{abc}$ with $a = 0.20$, $b = 0.70$, and $c = 0.57$ (MAE = 3.1 kcal/mol). This result is better than the best results published in ref 45 for VSXC (MAE = 3.5 kcal/mol), but M06-2X remains the best performer with MAE = 2.6 kcal/mol (cf. Table 2). The BPW91h$_{abc}$ with $a = 0.21$, $b = 0.70$, and $c = 0.57$ (MAE = 3.3 kcal/mol) and with $a = 0.15$, $b = 0.75$, and $c = 0.35$ (MAE = 4.3 kcal/mol) performs worse. The next best performers are TPSSh and revTPSSh$_a$ with $a = 0.10$ (eq 5) (MAE = 3.9 and 4.3 kcal/mol, respectively). Table 2 also shows that the new PBEh$_a$ with $a = 0.32$ (eq 5) performs better than B3LYP (MAE = 4.7 vs 4.9 kcal/mol) or PBE0[37,38] with $a = 0.25$ in eq 5 (MAE = 6.7 kcal/mol). The revTPSS functional performs slightly better than TPSS (MAE = 5.1 vs 5.8 kcal/mol).

Comparison of the results obtained for the G3/99 and AE6 test sets shows that AE6 is suitable for rough parameter optimization; however, optimization on it does not yield precisely the optimal parameters for the G3/99 test set. For example we observed a considerable difference for the PBEsol hybrid. Optimization for the AE6 test gives $a = 0.5$ (cf. ref 79 and Table 1), while our current results for G3/99 test set give $a = 0.6$ (cf. Table 2). Also the BPW91h$_{abc}$, TPSSh, and revTPSSh$_a$ functionals perform differently on the two test sets.

4.4. Basis Set Effects. Our results show that simplification of the 6-311++G(3df, 3pd) to the 6-311+G(3df, 3pd) basis set leads to a slight overbinding that is proportional to the number of the H atoms in the molecule (cf. Table 2). The origin of this difference is simply the less negative H atom energy calculated with the + basis set compared to the diffuse ++ basis set. The molecular energies practically do not change. To demonstrate this, we have performed PBEh calculations with the 6-311++G(3df, 3pd) basis set and with a new basis set defined here (cf. Table 2). In the new 6-311+′G(3df, 3pd) basis set, the H atom is calculated with the 6-311++G(3df, 3pd) basis set, and all the molecules are calculated with the 6-311+G(3df, 3pd) basis set. The results agree very well as shown in Table 2. Comparison of the PBEh/6-311++G(3df, 3pd) and PBEh/6-311+G(3df, 3pd) results shows the slight relative overbinding of the latter smaller basis set (cf. ME = −1.51 and −1.74 kcal/mol and the −27.7 to −28.2 kcal/mol error for naphthalene in Table 2). Note that the difference is proportional to the number of H atoms. The smaller + basis set is considerably less expensive than the ++ basis set for H atom containing molecules, so for extensive comparisons the use of the smaller basis set is advantageous in computer time, without altering the conclusions.

Table 2. Summary of Deviations (ME, MAE) from Experiment of Standard Enthalpies of Formation, $\Delta_f H^o_{298}$, for the 223 Compounds of the G3/99 Test Set Computed with Various Methods Using Various Basis Sets[a]

functional	basis	*a*	*b*	*c*	ME	MAE	max	(+)	min	(-)	ref
BLYP	6-311++G(3df, 3pd)	0.00	1.00	1.00	3.81	9.49	41.0	(n-octane)	−28.1	(NO_2)	45
BPW91	6-311+G(3df, 3pd)	0.00	1.00	1.00	−5.58	9.03	21.4	($Si(CH_3)_4$)	−32.4	(NO_2)	this work
BPW91	6-311+G(2df, d)	0.00	1.00	1.00	−2.85	8.60	26.3	($Si(CH_3)_4$)	−31.0	(NO_2)	this work
PW86PBE	6-311+G(3df, 3pd)	0.00	1.00	1.00	−5.60	9.76	24.2	($Si(CH_3)_4$)	−36.1	(NO_2)	this work
PBE	6-311++G(3df, 3pd)	0.00	1.00	1.00	−21.69	22.22	8.7	(SiH_4)	−79.7	(azulene)	45
TPSS	6-311++G(3df, 3pd)	0.00	1.00	1.00	−5.20	5.81	16.2	(SiF_4)	−22.90	(ClF_3)	45
revTPSS	6-311+G(3df, 3pd)	0.00	1.00	1.00	−4.04	5.09	16.4	(SiF_4)	−25.70	(ClF_3)	this work
revTPSS	cc-pVTZ	**0.00**	**1.00**	**1.00**	−1.72	**4.54**	21.5	(SiF_4)	−22.90	(ClF_3)	this work
M06-2X[b]	6-311+G(3df, 3pd)	0.54	n.a.	n.a.	−1.33	**2.63**	15.6	(O_3)	−17.9	(P_4)	this work
B3LYP	6-311++G(3df, 3pd)	**0.20**	**0.72**	**0.81**	3.51	**4.93**	20.8	(SF_6)	−8.1	(BeH)	45
B3PW91	6-311++G(3df, 3pd)	**0.20**	**0.72**	**0.81**	−1.80	**3.90**	21.6	(SiF_4)	−17.0	(naphtalene)	45
BPW91h$_{abc}$	6-311+G(3df, 3pd)	**0.15**	**0.75**	**0.35**	1.57	**4.29**	26.9	(SF_6)	−9.3	(BeH)	this work
BPW91h$_{abc}$	6-311+G(3df, 3pd)	**0.15**	**0.79**	**0.75**	−0.70	**3.81**	22.4	(SiF_4)	−12.8	(pyrimidine)	this work
BPW91h$_a$	6-311+G(3df, 3pd)	0.15	0.85	1.00	2.59	5.37	24.7	(SiF_4)	−12.5	($F_2C{=}CF_2$)	this work
PW86PBEh$_a$	6-311+G(3df, 3pd)	0.15	0.85	1.00	2.50	6.19	27.8	(SiF_4)	−16.9	($C_4H_4N_2$)	this work
BPW91h$_{abc}$	6-311+G(3df, 3pd)	**0.20**	**0.70**	**0.57**	−0.04	**3.11**	23.7	(SiF_4)	−8.4	(BeH)	this work
BPW91h$_{abc}$	6-311+G(3df, 3pd)	**0.21**	**0.70**	**0.57**	1.84	**3.32**	25.3	(SiF_4)	−8.4	(BeH)	this work
PBE0	6-311++G(3df, 3pd)	0.25	0.75	1.00	−4.73	6.66	21.3	(SiF_4)	−35.6	(naphtalene)	45
PBEh$_a$	6-311++G(3df, 3pd)	**0.30**	**0.70**	**1.00**	−1.51	**4.89**	24.3	(SiF_4)	−27.7	(naphtalene)	this work
PBEh$_a$	6-311+'G(3df, 3pd)	**0.30**	**0.70**	**1.00**	−1.49	**4.89**	24.3	(SiF_4)	−27.7	(naphtalene)	this work
PBEh$_a$	6-311+G(3df, 3pd)	0.30	0.70	1.00	−1.74	5.02	24.3	(SiF_4)	−28.2	(naphtalene)	this work
PBEh$_a$	6-311+G(2df, d)	**0.30**	**0.70**	**1.00**	0.84	**4.91**	29.8	(SiF_4)	−24.7	(naphtalene)	this work
PBEh$_a$	6-311++G(3df, 3pd)	**0.32**	**0.68**	**1.00**	−0.23	**4.73**	25.5	(SiF_4)	−24.6	(naphtalene)	this work
PBEsolh$_a$	6-311+G(3df, 3pd)	0.20	0.80	1.00	−38.98	39.07	3.9	(Li_2)	−128.2	(naphtalene)	this work
PBEsolh$_a$	6-311+G(3df, 3pd)	0.25	0.75	1.00	−34.07	34.21	4.1	(Li_2)	−115.3	(naphtalene)	this work
PBEsolh$_a$	6-311+G(3df, 3pd)	0.50	0.50	1.00	−10.39	15.06	37.5	(O_3)	−56.2	(n-octane)	this work
PBEsolh$_a$	6-311+G(3df, 3pd)	0.60	0.40	1.00	−1.30	11.98	56.3	(O_3)	−41.1	(n-octane)	this work
TPSSh	6-311+G(3df, 3pd)	0.10	0.90	1.00	−0.86	**3.90**	22.0	(SiF_4)	−18.00	(Si_2H_6)	45
revTPSSh$_a$	6-311+G(3df, 3pd)	**0.10**	**0.90**	**1.00**	−0.06	**4.32**	22.2	(SiF_4)	−25.00	(ClF_3)	this work
revTPSSh$_a$	6-311++G(3df, 3pd)	0.20	0.80	1.00	3.91	6.60	27.7	(SiF_4)	−24.80	(Si_2H_6)	this work

[a] The geometries and zero-point energies were obtained at the B3LYP/6-31G(2df, p) level using a frequency scale factor of 0.9854. See also the footnote of Table 1. All values are in kcal/mol. Error = theory − experiment. 6-311+'G(3df, 3pd) basis set: all calculations were performed with 6-311+G(3df, 3pd), but the free H atom energy was calculated with the 6-311++G(3df, 3pd) basis set. [b] It was not possible to obtain self-consistent energy for Li atom. We calculated the energy using the PBE electron density. Less serious SCF convergence problems were observed for Na and several other atoms.

Further simplification of the 6-311+G(3df,3pd) basis sets by removing polarization functions leaves the atomic energies practically unchanged (especially for spherical atoms such as Li, Be, and N) but makes the molecular energies less negative and leads to an underbinding effect. Comparison of the BPW91/6-311+G(3df, 3pd) and BPW91/6-311+G(2df, p) results in Table 2 clearly shows this tendency. The general overbinding of BPW91 is partly compensated by the underbinding effect of the 6-311+G(2df, p) basis set compared to the 6-311+G(3df, 3pd) basis set (cf. ME = −5.6 vs −2.9 kcal/mol in Table 2). Even the MAE is slightly improved. Similar improvement can be observed for revTPSS/cc-pVTZ results in Table 2, in agreement with the observation that TPSS/6-311G(d, p) results[60] agree better with the experiment than the TPSS/6-311++G(3df, 3pd) results published in ref 45. Our results for the AE6 test set show that the PBE/cc-pVQZ and PBE/6-311+G(3df, 3pd) atomization energies are similar: The difference of the MAEs is less than 0.4 kcal/mol.

Table 3. Relative Percentage Errors, MPE, and MAPE of Calculated PBEsol and PBEsolh$_a$/6-311+G(3df, 3pd) Total Energies Compared to Best Estimates of Nonrelativistic Total Energies vs Weight of the Exact Exchange, *a*, of Eq 5 for Atoms and Molecules Included in AE6 Test Set[a]

		a (cf. eq 5)				
species	best energy hartree	0.00	0.25	0.40	0.50	0.60
H	−0.5000	−2.33	−1.38	−0.80	−0.42	−0.03
C	−37.8450	−0.59	−0.43	−0.34	−0.27	−0.21
O	−75.0674	−0.42	−0.31	−0.25	−0.20	−0.15
Si	−289.3600	−0.23	−0.17	−0.13	−0.11	−0.08
S	−398.1110	−0.20	−0.15	−0.12	−0.09	−0.07
SiH_4	−291.8738	−0.25	−0.18	−0.13	−0.11	−0.08
SiO	−364.7335	−0.27	−0.20	−0.16	−0.13	−0.11
S_2	−796.3840	−0.20	−0.15	−0.11	−0.09	−0.07
C_3H_4	−116.6582	−0.55	−0.41	−0.32	−0.26	−0.20
$C_2H_2O_2$	−227.8341	−0.44	−0.33	−0.27	−0.23	−0.18
C_4H_8	−157.2111	−0.55	−0.40	−0.31	−0.24	−0.18
MPE%		−0.55	−0.37	−0.27	−0.20	−0.12
MAPE%		0.55	0.37	0.27	0.20	0.12

[a] Nonrelativistic total energies are calculated from atomic energies in ref 80 and atomization energies in ref 70.

4.5. About the Origin of the Improvements. Table 3 shows the atoms and the molecules of the AE6 test set, the best estimates of the relevant total energies, the relative percentage errors (calculated/best estimate − 1) 100%, the mean percent errors (MPE%), and the mean absolute percent errors (MAPE%) of the calculated PBEsol/6-311+G(3df, 3pd) total energies as functions of the weight of the exact exchange, *a*, of eq 5. The best estimated energies for the molecules of the AE6 test are derived from the known atomic energies[80] and the reference atomization energies. The results in Table 3 show that the PBEsol functional ($a = 0.0$) gives

Table 4. Relative Percentage Errors, MPE, and MAPE of the Calculated 6-311+G(3df, 3pd) Total Energies Compared to Best Estimates of Nonrelativistic Total Energies vs Weight of Exact Exchange, *a*, of Eq 5 for Atoms and Molecules Included in AE6 Test Set[a]

		PBE	PBE0	PBEh$_a$	PW86PBE	BPW91	BPW91h$_a$	B3PW91	B3LYP
species	best energy hartree	$a = 0$	$a = 0.25$	$a = 0.32$	$a = 0$	$a = 0$	$a = 0.2$	$a = 0.2$	$a = 0.2$
H	−0.5000	−0.076	0.208	0.290	1.040	0.785	0.869	0.796	0.431
C	−37.8450	−0.135	−0.110	−0.103	0.076	−0.003	−0.001	−0.025	0.033
O	−75.0674	−0.084	−0.073	−0.069	0.073	0.014	0.010	−0.010	0.031
Si	−289.3600	−0.047	−0.038	−0.035	0.032	0.007	0.007	−0.007	0.012
S	−398.1110	−0.043	−0.034	−0.032	0.023	0.004	0.004	−0.008	0.006
SiH_4	−291.8590	−0.046	−0.035	−0.031	0.037	0.012	0.013	0.002	0.021
SiO	−364.7391	−0.054	−0.051	−0.049	0.039	0.007	0.001	−0.012	0.012
S_2	−796.4029	−0.043	−0.036	−0.033	0.022	0.004	0.003	−0.009	0.004
C_3H_4	−116.6851	−0.133	−0.122	−0.119	0.076	−0.007	−0.016	−0.031	0.014
$C_2H_2O_2$	−227.8833	−0.099	−0.102	−0.102	0.069	0.002	−0.014	−0.031	0.012
C_4H_8	−157.2426	−0.132	−0.112	−0.106	0.076	−0.004	−0.006	−0.020	0.017
MPE%		−0.081	−0.046	−0.035	0.142	0.075	0.079	0.059	0.054
MAPE%		0.081	0.084	0.088	0.142	0.077	0.086	0.087	0.054

[a] Nonrelativistic total energies are calculated from atomic energies in ref 80 and atomization energies in ref 70. Here B3PW91 and B3LYP are the original parametrizations of refs 32 and 56, respectively.

about −2.3% error for the H atom and −0.6% and −0.4% for the C and O atoms, respectively. It can be observed that relative percentage errors decrease with the increase of the atomic number (cf. −0.2% error for S atom). For molecules, the errors of the constituent atoms dominate in the error (cf. −0.2% error for S_2 molecule). Increasing the weight of the exact exchange consistently improves the performance of PBEsol for the total energies for the atoms and the molecules in the AE6 test set. The inclusion of exact exchange improves the PBEsol hybrid results for the right reason by improving the individual energies. We show the individual errors (calculated − best estimate) in the Supporting Information.

In Table 4 we present the same error analysis for the PBE, PBEh$_a$, PW86PBE, BPW91, BPW91h$_a$, B3PW91, and B3LYP functionals with the 6-311+G(3df, 3pd) basis set. The PBE energies are quite good and consistently more positive than the best-estimate energies (MPE% = −0.08%). Inclusion of exact exchange improves the MPE% via error compensation for PBEh$_a$. (The H atom energy is too negative by 0.9 kcal/mol or about 0.29% percent error, and all the other energies are too positive, cf. Supporting Information). But inclusion of exact exchange does not improve the MAPE%. The individual PBEh$_a$ errors do decrease compared to the PBE errors (cf. Supporting Information). A similar but larger (up to 3 kcal/mol, cf. Supporting Information) error occurs for the H atoms for all the other functionals shown in Table 4. Notice that PW86PBE, BPW91h$_a$, or B3PW91 give worse results than BPW91. B3LYP energies are all too negative, and B3LYP gives the smallest MPE% and MAPE%.

5. Ionization Potentials

Table 5 presents the ME and MAE for ionization potentials (IP) of the IP86[45] test set. IPs were calculated as the difference in total energies at 0 K of the cation and the corresponding neutral, at their corresponding B3LYP/6-31G(2df, p) geometries using scaled B3LYP/6-31G(2df, p) ZPEs. Calculations use the 6-311++G(3df, 3pd) and 6-311+G(2df, p) basis sets. The results show that these basis sets are practically equivalent for these calculations. Our results agree with the results in ref 45.

The best MAEs are slightly below 0.2 eV. The order of performance is B3LYP, PBEsolh, and BPW91h, with almost negligible differences between the functionals (set bold in Table 5). Generally a small underestimation (ME = −0.1, −0.2 eV) of IPs can be observed for nonhybrid functionals, and this underestimation is reduced by the inclusion of the exact exchange. PBEsol with $a = 0.6$ slightly overestimates the IPs (ME = 0.036 eV in Table 5). The ME shows again a quasilinear dependence on the value of *a* in eq 5. In general the hybrid functionals perform better than their nonhybrid counterparts, in agreement with ref 45.

The poorest ionization energy is predicted for the CN molecule because the open-shell $^1\Sigma^+$ singlet state of the CN^+ ion is not correctly described by the functionals used in this study.[66] The incorrect B3LYP geometries of the CH_4^+, BCl_3^+, $B_2F_4^+$, and BF_3^+ cations[81] also lead to large errors. (See the large deviations for BF_3^+ in Table 5.)

6. Electron Affinities

Anions are difficult test cases for the GGA, meta-GGA, and global hybrid functionals, as the self-interaction error can spoil the results. The results are also sensitive to the basis-set, which must be diffuse enough to describe anions but not more diffuse than that. We have calculated the electron affinity (EA) at 0 K as the difference between the total energies of the anion and the corresponding neutral species. We use the B3LYP/6-31G(2df, p) geometries.

Semilocal functionals lead to unstable negative atomic ions in the complete basis set limit, in which a fraction of an electron escapes. Nevertheless, realistic electron affinities can be computed using Gaussian basis sets in which the diffuse basis functions do not have characteristic decay lengths larger than those of the exact Kohn−Sham orbitals. In most cases, there is a plateau on which the electron affinity remains stable as the basis set is expanded within this range.[82]

Table 5. Summary of Deviations from Experiment of IPs of the G3/99 Test Set (86 species) Computed Using the 6-311++G(3df, 3pd) Basis Set[a]

method	basis	*a*	*b*	*c*	ME	MAE	max	(+)	max	(−)
BLYP	6-311++G(3df, 3pd)	0.00	1.00	1.00	−0.191	0.286	1.03	(CN)	−1.06	(BF_3)
BPW91	6-311++G(3df, 3pd)	0.00	1.00	1.00	−0.105	0.241	1.14	(CN)	−0.99	(BF_3)
BPW91	6-311+G(2df, p)	0.00	1.00	1.00	−0.105	0.238	1.15	(CN)	−0.96	(BF_3)
PBE	6-311+G(2df, p)	0.00	1.00	1.00	−0.105	0.233	1.12	(CN)	−0.98	(BF_3)
PBEsol	6-311+G(2df, p)	0.00	1.00	1.00	−0.146	0.237	1.06	(CN)	−0.96	(BF_3)
TPSS	6-311++G(3df, 3pd)	0.00	1.00	1.00	−0.134	0.242	1.22	(CN)	−1.02	(BF_3)
revTPSS	6-311++G(3df, 3pd)	0.00	1.00	1.00	−0.156	0.247	1.20	(CN)	−1.07	(BF_3)
B3LYP	6-311++G(3df, 3pd)	**0.20**	**0.72**	**0.81**	0.009	**0.184**	1.57	(CN)	−0.57	(B_2F_4)
B3PW91	6-311++G(3df, 3pd)	**0.20**	**0.72**	**0.81**	−0.012	**0.190**	1.58	(CN)	−0.65	(B_2F_4)
BPW91h$_a$	6-311+G(2df, p)	0.10	0.90	1.00	−0.084	0.218	1.35	(CN)	−0.77	(B_2F_4)
BPW91h$_a$	6-311+G(2df, p)	0.15	0.85	1.00	−0.074	0.210	1.45	(CN)	−0.72	(B_2F_4)
BPW91h$_a$	6-311+G(2df, p)	0.20	0.80	1.00	−0.064	0.204	1.55	(CN)	−0.67	(B_2F_4)
BPW91h$_{abc}$	6-311+G(2df, p)	**0.20**	**0.70**	**0.57**	0.079	**0.188**	1.67	(CN)	−0.51	(B_2F_4)
PBE0	6-311++G(3df, 3pd)	**0.25**	**0.75**	**1.00**	−0.064	**0.199**	1.61	(CN)	−0.67	(B_2F_4)
PBE0	6-311+G(2df, p)	**0.25**	**0.75**	**1.00**	−0.062	**0.198**	1.62	(CN)	−0.63	(B_2F_4)
PBEh$_a$	6-311+G(2df, p)	**0.32**	**0.68**	**1.00**	−0.051	**0.196**	1.76	(CN)	−0.57	(B_2F_4)
PBEsolh$_a$	6-311+G(2df, p)	**0.25**	**0.75**	**1.00**	−0.064	**0.186**	1.60	(CN)	−0.61	(B_2F_4)
PBEsolh$_a$	6-311+G(2df, p)	**0.60**	**0.40**	**1.00**	0.036	**0.198**	2.37	(CN)	−0.35	(Be)
TPSSh	6-311++G(3df, 3pd)	0.10	0.90	1.00	−0.113	0.229	1.41	(CN)	−0.79	(BF_3)
revTPSSh$_a$	6-311++G(3df, 3pd)	0.10	0.90	1.00	−0.129	0.230	1.40	(CN)	−0.80	(BF_3)
revTPSSh$_a$	6-311++G(3df, 3pd)	0.20	0.80	1.00	−0.102	0.219	1.59	(CN)	−0.62	(B_2F_4)

[a] The geometries and zero-point energies were obtained at the B3LYP/6-31G(2df, p) level using a frequency scale factor of 0.9854. All values are in eV. Error = calculated − experiment. The mean experimental IP is 10.89 eV. See also the footnote of Table 1. The best results (MAE < 0.2 eV) are in bold.

Table 6. Summary of Deviations from Experiment of EAs of the G3/99 Test Set[a]

method	basis	*a*	*b*	*c*	ME	MAE	max		min	
BLYP	6-311++G(3df, 3pd)	**0.00**	**1.00**	**1.00**	0.008	**0.115**	0.70	(C_2)	−0.26	(NCO)
BPW91	6-311++G(3df, 3pd)	**0.00**	**1.00**	**1.00**	0.035	**0.119**	0.78	(C_2)	−0.31	(NO_2)
BPW91	6-311+G(3df, 3pd)	**0.00**	**1.00**	**1.00**	0.030	**0.122**	0.78	(C_2)	−0.31	(NO_2)
BPW91	6-311+G(2df, p)	0.00	1.00	1.00	0.032	0.125	0.78	(C_2)	−0.29	(NO_2)
PBE	6-311+G(2df, p)	**0.00**	**1.00**	**1.00**	0.057	**0.121**	0.78	(C_2)	−0.28	(NO_2)
PBEsol	6-311++G(3df, 3pd)	**0.00**	**1.00**	**1.00**	0.039	**0.113**	0.74	(C_2)	−0.36	(NO_2)
TPSS	6-311++G(3df, 3pd)	0.00	1.00	1.00	−0.020	0.137	0.82	(C_2)	−0.32	(NO_2)
revTPSS	6-311++G(3df, 3pd)	0.00	1.00	1.00	−0.041	0.137	0.82	(C_2)	−0.33	(NO_2)
B3LYP	6-311++G(3df, 3pd)	**0.20**	**0.72**	**0.81**	0.088	**0.124**	1.10	(C_2)	−0.09	(HOO)
B3PW91	6-311++G(3df, 3pd)	0.20	0.72	0.81	0.031	0.137	1.08	(C_2)	−0.26	(HOO)
BPW91h$_a$	6-311+G(2df, p)	0.10	0.90	1.00	0.002	0.138	0.91	(C_2)	−0.28	(HOO)
BPW91h$_a$	6-311+G(2df, p)	0.15	0.85	1.00	−0.012	0.152	0.98	(C_2)	−0.3	(HOO)
BPW91h$_a$	6-311+G(2df, p)	0.20	0.80	1.00	−0.026	0.168	1.04	(C_2)	−0.33	(HOO)
BPW91h$_{abc}$	6-311+G(2df, p)	0.20	0.70	0.57	0.115	0.147	1.16	(C_2)	−0.13	(HOO)
PBE0	6-311+G(2df, p)	0.25	0.75	1.00	−0.028	0.172	1.10	(C_2)	−0.38	(HOO)
PBEh$_a$	6-311+G(2df, p)	0.32	0.68	1.00	−0.050	0.197	1.18	(C_2)	−0.42	(HOO)
PBEsolh$_a$	6-311+G(2df, p)	0.25	0.75	1.00	−0.016	0.150	1.10	(C_2)	−0.40	(HOO)
PBEsolh$_a$	6-311+G(2df, p)	0.60	0.40	1.00	−0.072	0.262	1.58	(C_2)	−0.61	(OH)
TPSSh	6-311++G(3df, 3pd)	0.10	0.90	1.00	−0.046	0.164	0.95	(C_2)	−0.33	(HOO)
revTPSSh$_a$	6-311++G(3df, 3pd)	0.10	0.90	1.00	−0.084	0.188	0.70	(C_2)	−0.41	(OH)
revTPSSh$_a$	6-311++G(3df, 3pd)	0.20	0.80	1.00	−0.061	0.170	0.95	(C_2)	−0.36	(HOO)

[a] The G3/99 has 58 species. Geometries and zero-point energies were obtained at the B3LYP/6-31G(2df, p) level using a frequency scale factor of 0.9854. All values are in eV. Error = theory − experiment. The mean experimental EA is 1.41 eV. See also the footnote of Table 1.

Table 6 presents the ME and MAE for electron affinities of the G3/99 test set (58 species, the EA58 test set is taken from ref 45). The open-shell singlet C_2 is isoelectronic with CN^+, and here again this electron structure causes a problem for single determinant GGAs, leading to extremely large errors in agreement with ref 45. The BPW91 results in the Table 6 show that the 6-311++G(3df, 3pd), 6-311+G(3df, 3pd), and 6-311+G(2df, p) basis sets are practically equivalent. The errors of the functionals are considerably larger than these basis set errors. Our study (not presented here) shows that omitting diffuse functions makes the errors very large.

The results in Table 6 show good agreement with the results in ref 45. GGAs slightly overbind the extra electron. The TPSS meta-GGA, however, shows a slight underbinding of the electron, and this is somewhat more pronounced for revTPSS. The BLYP and the PBEsol functionals show the best performance (MAE = 0.11 eV).

Table 7. Summary of Deviations from Experiment of PAs for the Eight Molecules of the G3/99 Test Set[a]

method	basis	*a*	*b*	*c*	ME	MAE	max	(+)	min	(−)
BLYP	6-311++G(3df, 3pd)	0.00	1.00	1.00	−1.46	1.57	0.4	(C_2H_2)	−3.9	(H_2O)
BPW91	6-311++G(3df, 3pd)	0.00	1.00	1.00	0.86	1.45	3.8	(C_2H_2)	−1.3	(PH_3)
BPW91	6-311+G(3df, 3pd)	0.00	1.00	1.00	0.86	1.47	3.8	(C_2H_2)	−1.3	(PH_3)
BPW91	6-311+G(2df, p)	0.00	1.00	1.00	0.68	1.27	3.4	(C_2H_2)	−1.4	(PH_3)
PBE	6-311++G(3df, 3pd)	0.00	1.00	1.00	−0.82	1.60	2.4	(C_2H_2)	−3.6	(PH_3)
PBEsol	6-311++G(3df, 3pd)	0.00	1.00	1.00	−2.69	2.88	0.7	(C_2H_2)	−6.8	(PH_3)
TPSS	6-311++G(3df, 3pd)	0.00	1.00	1.00	1.68	1.82	4.4	(C_2H_2)	−0.5	(H_2O)
revTPSS	6-311++G(3df, 3pd)	0.00	1.00	1.00	1.77	2.06	4.8	(C_2H_2)	−1.2	(H_2O)
B3LYP	6-311++G(3df, 3pd)	**0.20**	**0.72**	**0.81**	−0.77	**1.16**	1.6	(C_2H_2)	−2.3	(H_2)
B3PW91	6-311++G(3df, 3pd)	**0.20**	**0.72**	**0.81**	0.97	**1.07**	4.2	(C_2H_2)	−0.3	(SiH_4)
BPW91h$_a$	6-311+G(3df, 3pd)	0.10	0.90	1.00	1.08	1.32	4.3	(C_2H_2)	−0.5	(PH_3)
BPW91h$_a$	6-311+G(3df, 3pd)	0.15	0.85	1.00	1.20	1.26	4.5	(C_2H_2)	−0.1	(PH_3)
BPW91h$_a$	6-311+G(3df, 3pd)	0.20	0.80	1.00	1.31	1.33	4.7	(C_2H_2)	−0.1	(SiH_4)
BPW91h$_{abc}$	6-311+G(3df, 3pd)	0.20	0.70	0.57	1.22	1.27	4.1	(C_2H_2)	−0.1	(H_2O)
PBE0	6-311+G(3df, 3pd)	**0.25**	**0.75**	**1.00**	0.18	**1.14**	3.9	(C_2H_2)	−1.7	(SiH_4)
PBEh$_a$	6-311+G(3df, 3pd)	**0.32**	**0.68**	**1.00**	0.46	**1.10**	4.3	(C_2H_2)	−1.8	(SiH_4)
PBEsolh$_a$	6-311++G(3df, 3pd)	0.25	0.75	1.00	−1.25	1.88	2.5	(C_2H_2)	−4.2	(SiH_4)
PBEsolh$_a$	6-311++G(3df, 3pd)	0.50	0.50	1.00	0.22	1.66	4.4	(C_2H_2)	−3.7	(SiH_4)
PBEsolh$_a$	6-311++G(3df, 3pd)	0.60	0.40	1.00	0.82	2.04	5.1	(C_2H_2)	−3.5	(SiH_4)
TPSSh	6-311++G(3df, 3pd)	0.10	0.90	1.00	1.77	1.77	4.8	(C_2H_2)		
revTPSSh$_a$	6-311++G(3df, 3pd)	0.10	0.90	1.00	1.84	1.99	5.2	(C_2H_2)	−0.6	(H_2O)

[a] The geometries and zero-point energies were obtained at the B3LYP/6-31G(2df, p) level using a frequency scale factor of 0.9854. All values are in kcal/mol. Error = theory − experiment. The mean experimental PA is 158.0 kcal/mol. See also the footnote of Table 1.

Mixing exact exchange with the GGA or meta-GGA functionals shifts the ME to the electron-underbinding direction, and again we observed a quasilinear relationship between the value of a in eq 5 and the ME for BPW91h$_a$, PBEh$_a$, and PBEsolh$_a$. The example of BPW91h$_a$ shows that even $a = 0.1$ counterbalances the overbinding, but this does not improve the MAE value (cf. Table 6). This shows that for EA the inclusion of exact exchange deteriorates the results. The largest deterioration can be observed for functionals that underbind at the semilocal level (like TPSS or revTPSS) or for hybrids functionals with a large weight of exact exchange (cf. PBEsolh$_a$ in Table 6). This clearly shows a weakness of the global hybrids.

7. Proton Affinities

The proton affinity (PA) is calculated as the energy difference between the neutral and the protonated molecule M: PA(M) $= E_0(\mathrm{M}) - E_0(\mathrm{MH}^+)$. Adding a proton to a neutral molecule does not change the number of the electrons but alters the geometry and the distribution of the electron density. This usually leads to a more negative exchange and correlation energy.

We selected the eight PAs included in the G3/99 test set as published in ref 45 (the PA8 test set). We used the geometries and the scaled frequencies obtained at the B3LYP/6-31G(2df, p) level. Earlier results show that the LSDA seriously underbinds (ME = −5.9 kcal/mol), while the Hartree−Fock method slightly overbinds (ME = 1.8 kcal/mol), and these model chemistries give a poor MAE for the PA8 test set (MAE = 5.9 and 3.1 kcal/mol, respectively).

Table 6 presents the ME and MAE in proton affinities of the PA8 test set as in ref 45. Calculations use the 6-311++G(3df, 3pd), 6-311+G(3df, 3pd) and 6-311+G(2df, p) basis sets. The results show that the 6-311+G(2df, p) basis set error is large for the PA8 test set (cf. Table 7 BPW91 results), and thus we use only the 6-311+G(3df, 3pd) and 6-311++G(3df, 3pd) basis sets. The GGA functionals qualitatively differ from each other. The order of the GGA functionals from the most underbinding to the overbinding direction is PBEsol (closest to LSDA), BLYP, PBE, BPW91, TPSS, and revTPSS (ME = −2.7, −1.5, −0.8, +0.9, 1.7, and 1.8 kcal/mol, respectively, cf. Table 7). The hybrid functionals describe more correctly the protonation of PA8 and decrease the MAE from 1.5 to around 1 kcal/mol. The largest positive error was observed for C_2H_2 for all functionals.

As the BPW91 GGA overbinding is further aggravated by the overbinding effect of the exact exchange mixing, BPW91h$_a$ does not give good results (cf. Table 7). However, the reduced gradient contributions of exchange and correlation lead to some improvement [cf. MAE = 1.07 and 1.33 kcal/mol for B3PW91 and BPW91h$_{abc}$ ($a = 0.2$, $b = 0.8$, and $c = 1.0$) in Table 7]. This improvement is conserved in B3LYP results too (MAE = 1.16 kcal/mol). The underbinding of PBE is efficiently compensated by $a = 0.32$, and this hybrid is among the best functionals with MAE = 1.10 kcal/mol (cf. Table 7). The large proportion of exact exchange in PBEsolh$_a$ compensates the strong underbinding of PBEsol, but the MAE remains large (1.7−2.0 kcal/mol, Table 7). The TPSSh and revTPSSh$_a$ results conserve the overbinding errors of the parent functionals, that is slightly aggravated by the small portion of exact exchange ($a = 0.1$).

8. Bond Lengths and Visual Summary

Table 8 presents the ME and the MAE for equilibrium internuclear distances (r_e) of the T-96R test set as defined in ref 45. The geometry optimizations were carried out using the 6-311++G(3df, 3pd), 6-311G(2df, p), cc-pVTZ, and aug-cc-pVTZ basis sets with Opt = Tight and Int(Grid = Ultrafine) keywords. Comparison of the basis set dependence of the geometries shows that the 6-311G(2df, p) basis set gives similar, but slightly longer (by 0.010 Å), r_e values than the 6-311++G(3df, 3pd) basis set. This small deviation

Table 8. Summary of Deviations from Experiment of Equilibrium Bond Lengths (r_e) for the T96R Test Set[a]

method	basis	*a*	*b*	*c*	ME	MAE	max	(+)	max	(−)
BLYP	6-311++G(3df, 3pd)	0.00	1.00	1.00	0.0212	0.0223	0.055	(Al_2)	−0.032	(Na_2)
BPW91	6-311++G(3df, 3pd)	0.00	1.00	1.00	0.0166	0.0168	0.070	(Li_2)	−0.007	(F_2^+)
BPW91	6-311G(2df, p)	0.00	1.00	1.00	0.0178	0.0180	0.066	(Li_2)	−0.005	(F_2^+)
PBE	6-311G(2df, p)	0.00	1.00	1.00	0.0164	0.0170	0.052	(Li_2)	−0.008	(F_2^+)
PBEsol	6-311++G(3df,3pd)	0.00	1.00	1.00	0.0104	0.0127	0.067	(Li_2)	−0.021	(F_2^+)
TPSS	6-311++G(3df, 3pd)	0.00	1.00	1.00	0.0138	0.0142	0.078	(Li_2)	−0.008	(P_4)
TPSS	cc-pVTZ	0.00	1.00	1.00	0.0176	0.0180	0.062	(Li_2)	−0.013	(Be_2)
TPSS	aug-cc-pVTZ	0.00	1.00	1.00	0.0178	0.0182	0.062	(Li_2)	−0.014	(Be_2)
revTPSS	6-311++G(3df, 3pd)	0.00	1.00	1.00	0.0137	0.0141	0.081	(Li_2)	−0.010	(F_2^+)
B3LYP	6-311++G(3df, 3pd)	**0.20**	**0.72**	**0.81**	0.0050	**0.0104**	0.041	(Be_2)	−0.040	(Na_2)
B3PW91	6-311++G(3df, 3pd)	**0.20**	**0.72**	**0.81**	0.0026	**0.0093**	0.060	(Li_2)	−0.042	(F_2^+)
BPW91h$_a$	6-311G(2df, p)	0.10	0.90	1.00	0.0108	0.0127	0.064	(Li_2)	−0.024	(F_2^+)
BPW91h$_a$	6-311G(2df, p)	0.15	0.85	1.00	0.0076	0.0112	0.063	(Li_2)	−0.033	(F_2^+)
BPW91h$_a$	6-311G(2df, p)	0.20	0.80	1.00	0.0045	**0.0102**	0.065	(Be_2)	−0.041	(F_2^+)
BPW91h$_{abc}$	6-311G(2df, p)	**0.20**	**0.70**	**0.57**	0.0047	**0.0102**	0.059	(Be_2)	−0.038	(F_2^+)
PBE0	6-311G(2df, p)	**0.25**	**0.75**	**1.00**	0.0005	**0.0098**	0.069	(Be_2)	−0.050	(F_2^+)
PBEh$_a$	6-311G(2df, p)	0.32	0.68	1.00	−0.0033	0.0110	0.096	(Be_2)	−0.061	(F_2^+)
PBEsolh$_a$	6-311G(2df, p)	0.60	0.40	1.00	−0.0179	0.0249	0.231	(Be_2)	−0.098	(F_2^+)
TPSSh	6-311++G(3df, 3pd)	0.10	0.90	1.00	0.0068	**0.0096**	0.062	(Li_2)	−0.024	(F_2^+)
revTPSSh$_a$	6-311++G(3df, 3pd)	**0.10**	**0.90**	**1.00**	0.0070	**0.0100**	0.078	(Li_2)	−0.028	(F_2^+)
revTPSSh$_a$	6-311++G(3df, 3pd)	**0.20**	**0.80**	**1.00**	0.0020	**0.0091**	0.102	(Be_2)	−0.044	(F_2^+)

[a] All values are in Å. Error = calculated − experiment. The mean experimental r_e is 1.565 Å. See also the footnote of Table 1.

usually does not influence the total energies as the energy surfaces around the equilibria are flat. This is in agreement with the earlier observation that the 6-31G(2df, p) basis set gives reasonable geometries.[77] This is a useful observation to speed up considerably the calculations, as geometry optimizations with large diffuse basis sets are very slow and time consuming. The cc-pVTZ and aug-cc-pVTZ basis sets both give consistently longer (by 0.04 Å) r_e values than the 6-311++G(3df, 3pd) basis set as shown for TPSS (cf. Table 8). We have observed a similar 0.05 Å basis set effect for the average bond lengths calculated with the PBEsolh$_a$ functional (not shown in Table 8). This discrepancy shows that even large, polarized, and diffuse triple-ζ basis sets might introduce a considerable systematic error into the calculated equilibrium bond lengths.

It was observed for the T-96R test set that LSDA gives particularly good r_e values, and the Hartree−Fock method systematically underestimates the values of r_e (ME = −0.01 Å in Table 6 of ref 45, Be_2 excluded). We identified several problematic compounds in the T-96R: Be_2, Li_2, and Na_2. Be_2 has a large equilibrium distance and a very flat potential energy curve and is bound by a longer range dispersion interaction that is not described well by GGA or meta-GGA. Be_2 is unbound by the Hartree−Fock method. Although TPSS, PBE, and PBEsol bind Be_2,[83] inclusion of exact exchange makes r_e very large and leads to errors up to 0.231 Å (see PBEsolh (a = 0.6) in Table 8). We note that inclusion of the a posteriori dispersion correction at no cost considerably improves such results without deteriorating the covalent results.[84] Li_2 is problematic with all the functionals (too long r_e), but Na_2 is considerably better described with the hybrid functionals. The GGA and the hybrid results are spoiled by the self-interaction error for F_2^+. These results show that the weakly bound molecules are not described well with the functionals in this study.

While LSDA error compensation gives a reasonable prediction for the r_e values of the T-96R test set, the introduction of gradient correction (GGA and meta-GGA) gives too long r_e values, and that is efficiently compensated by the exact exchange. Again a quasilinear dependence of the ME on the value of exact exchange mixing (a in eq 5) was observed for BPW91h$_a$, PBEh$_a$, and PBEsolh$_a$ functionals. Notice that the ME value for BPW91h$_{abc}$ depends mainly on the value of a and shows much less dependence on b and c parameters of eq 7 (cf. Table 8). The best functionals are revTPSSh$_a$, B3PW91, TPSSh, PBEh$_a$, BPW91h$_a$, and B3LYP. All these functionals show MAEs about 0.01 Å in Table 8.

Visual Summary over the Test Sets. The summary of the statistics is visualized in Figure 4. The radar chart of the relative MAEs (kcal/mol) compared to B3LYP MAE for K9, G3/99, IP86, EA58, PA8, and T96R test sets shows, for example, that B3PW91 performs better or the same as B3LYP except for the EA58 test set, where hybrid functionals fail in general compared to the parent GGA. (Notice the better performance of BPW91 on the figure.)

9. Conclusions

Our formal discussion in Sections 1and 2 presented an explanation and a critique of the concept of global hybrid functionals.

Our numerical studies have introduced global hybrids starting from the previously unhybridized semilocal functionals PBEsol and revTPSS. Interestingly, PSEsol requires a large fraction (60%) of exact exchange to correct its strong overestimation of molecular atomization energies, and its hybrid then predicts accurate energy barriers and improved but still inaccurate atomization energies. But revTPSS requires only a small fraction (10%) of exact exchange to correct its slight overestimation of atomization energies, so its hybridization only slightly improves its strong underestimation of barriers.

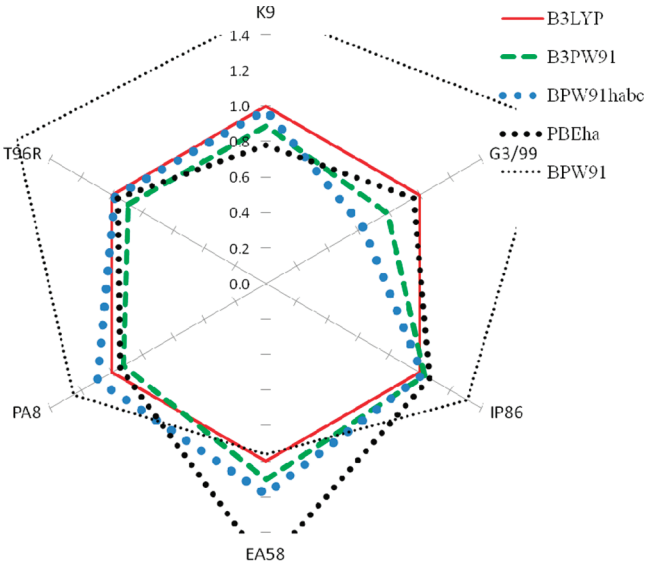


Figure 4. Radar or polar chart summary of the relative MAEs compared to B3LYP MAE (MAE(functional)/MAE(B3LYP)) for the K9, G3/99, IP86, EA58, PA8, and T96R test sets. The connecting lines are simply for guiding the eye. Smaller value means better performance.

We have also constructed optimized empirical parameters which surpass the standard ones for many popular hybrids, except the M06-2X hybrid meta-GGA that contains more than 30 empirical parameters. Perhaps the best overall performance is achieved by our refitted BPW91h$_{abc}$ (e.g., $a = 0.20$, $b = 0.70$, and $c = 0.57$, although different parameter combinations can produce similar results). Its performance seems better than that of the standard B3PW91, B3LYP, and PBE0. Our PBEh$_a$ with $a = 0.32$ also seems to perform better than the standard PBE0 with $a = 0.25$ for thermochemistry.

The molecular properties we have considered are those of the following test sets: G2-32 (32 ZPE-corrected atomization energies), AE6 (6 atomization energies), G3/99 (223 enthalpies of formation), BH6 (6 energy barriers), and K9 (kinetics) for thermochemistry as well as IP86 (86 ionizations), EA58 (58 electron affinities), PA8 (8 proton affinities), and T-96R (96 bond lengths). We tested B3LYP, B3PW91, BPW91, PBE, PW86PBE, PBEsol, TPSS, and revTPSS functionals and their global hybrids, using large triple-ζ basis sets. We also investigated the effects of fitting and basis sets. Here we will summarize the conclusions from our numerical studies in greater detail:

9.1. Test Sets. Our results show that the G2-32 test set is not suitable for testing the performance of any method for thermochemistry; the test set is not representative. The AE6 test set performs considerably better, but we obtained somewhat different results on the large G3/99 test set. The AE6 and the G3/99 test sets give different optimal weight of the exact exchange for the PBEsol functional. Moreover the BPW91h$_{abc}$, TPSSh, and revTPSSh$_a$ functionals perform differently on the two test sets.

The atomization energies for the molecules of the AE6 test set depend almost linearly on the weight of exact exchange, but this dependence varies. For five out of six molecules, the binding is reduced as the weight of exact exchange increases from zero, but SiH_4 shows the opposite behavior. Consequently the optimal weight of the exact exchange depends on the composition of the test set. Larger test sets are needed for reliable parameter optimizations.

9.2. Why B3LYP Works. The overbinding of the BPW91, PW86PBE, PBE, TPSS, revTPSS, and PBEsol functionals is efficiently compensated by a variable amount of exact exchange (with a larger overbinding requiring a larger portion of exact exchange). In contrast the BLYP functional typically underbinds. Since inclusion of exact exchange worsens the underbinding, it leads to worse results for BLYPh$_a$ on the AE6 and G3/99 test sets. The origin of the good results for B3LYP is the reduced gradient correction for the B88 exchange and the mixing of SVWN3 and LYP correlation. (In the first paragraph of Section 4, we corrected a numerical error in the literature that had overly favored B1LYP over B1PW91 on the nonrepresentative G2-32 test set.)

9.3. PBEsol Total Energies Are Improved Consistently by Hybridization. We observed for PBEsolh$_a$ on the AE6 test set that all atomic and molecular energies improve as the weight of the exact exchange increases to 60%. The total energies of other semilocal functionals can be overcorrected by hybridization. For example, PBEh$_a$ and other hybrid functionals give too negative H atom energies, and B3LYP gives too negative atomic and molecular energies for all components of the AE6 test set.

9.4. Hybrid Parameters *a*, *b*, and *c* Are Not Independent. In the three-parameter hybrids, the b and c parameters compensate each other's effects on the ME and partly on MAE. This shows that many different parametrizations might give practically the same results, as was demonstrated for several BPW91h$_{abc}$ hybrids.

9.5. Performance of Refitted Hybrid Functionals for Thermochemistry. We have constructed new BPW91h$_{abc}$, PBEh$_a$, PBEsolh$_a$, and revTPSSh$_a$ functionals. The best performer for the G3/99 test set is the BPW91h$_{abc}$ hybrid with $a = 0.20$, $b = 0.70$, and $c = 0.57$ (MAE = 3.1 compared to the MAE = 4.9 and 3.9 kcal/mol for the B3LYP and B3PW91 functionals). This functional performs particularly well for the AE6 test set (MAE = 2.1 compared to the MAE = 2.6 and 4.0 kcal/mol for the B3LYP and B3PW91 functionals). The new PBEh ($a = 0.32$) also performs better for the G3/99 test set than the original PBE0 ($a = 0.25$) (MAE = 4.7 vs 6.7 kcal/mol, respectively), and it performs better for the AE6 test set and for reaction kinetics (RMSE = 3.5 kcal/mol for the K9 test set). The PBEsolh$_a$ ($a = 0.60$) gives particularly accurate results for reaction kinetics (RMSE = 1.8 kcal/mol for the K9 test set), but it does not perform well for the G3/99 test set (MAE = 12.0 kcal/mol). The revTPSSh$_a$ ($a = 0.10$) shows good performance for the G3/99 test set (MAE = 4.3 kcal/mol), a small improvement over the revTPSS results (MAE = 5.0 kcal/mol). TPSSh performs slightly better (MAE = 3.9 kcal/mol) than revTPSSh$_a$.

9.6. Basis Sets. Simplification of the 6-311++G(3df, 3pd) to 6-311+G(3df, 3pd) basis sets leads to a slight overbinding that is proportional to the number of the H atoms in the molecule. The use of the smaller basis set is advantageous in computer time, without altering the conclusions. Further

simplification of the 6-311+G(3df, 3pd) basis set by removing polarization functions leads to underbinding. A general overbinding of a functional can be partly compensated by the underbinding effect of the 6-311+G(2df, p), 6-311G(d, p), or cc-pVTZ basis-sets. Comparison of the cc-pVQZ and 6-311+G(3df, 3pd) basis sets shows that these basis sets are close to the basis set limit for density functional theory (DFT) calculations.

9.8. Performance for Ionization Energies and Electron and Proton Affinities. For ionization potentials of the IP86 test set, the best MAEs are slightly below 0.2 eV. The order of performance is B3LYP, PBEsolh$_a$, and BPW91h$_a$, with negligible differences between the functionals. The ME shows again a quasilinear dependence on the value of a in eq 5. In general the hybrid functionals perform better than their nonhybrid counterparts, in agreement with ref 45. For several elements of the test set, the errors are large, as the open-shell $^1\Sigma^+$ singlet state of the CN^+ ion is not correctly described by the functionals and the B3LYP geometries of the CH_4^+, BCl_3^+,$B_2F_4^+$ and BF_3^+ cations used for calculations are incorrect.

For electron affinities (EA58 test set), the GGAs slightly overbind the extra electron. The TPSS meta-GGA however shows a slight underbinding of the electron, and this is somewhat more pronounced for revTPSS. The BLYP and the PBEsol functionals show the best performance (MAE = 0.11 eV). Mixing exact exchange with GGA or meta-GGA functionals shifts the ME to the electron-underbinding direction, and again we observed a quasilinear relationship between the value of a in eq 5 and the ME for BPW91h$_a$, PBEh$_a$, and PBEsolh$_a$. For EA58, inclusion of exact exchange deteriorates the results. The largest deterioration can be observed for functionals that underbind at the semilocal level (like TPSS or revTPSS) or for hybrids functionals with large weight of exact exchange (PBEsolh$_a$).

For proton affinities (PA8 set), the GGA functionals qualitatively differ from each other, and the order of the functionals from the most underbinding to the most overbinding is PBEsol, BLYP, PBE, BPW91, TPSS, and revTPSS (ME = −2.7, −1.5, −0.8, +0.9, 1.7, and 1.8 kcal/mol). The hybrid functionals describe more correctly the protonation and decrease the MAE from 1.5 to around 1 kcal/mol. The largest positive error was observed for C_2H_2 for all functionals.

9.7. Performance for Bond Lengths. While the LSDA error compensation gives a reasonable prediction for the r_e values of the T-96R test set, the introduction of gradient correction (GGA and meta-GGA) gives too long r_e values and is efficiently compensated by the exact exchange. Again a quasilinear dependence of the ME on the value of exact exchange mixing (a in eq 5) was observed for the BPW91h$_a$, PBEh$_a$, and PBEsolh$_a$ functionals. Notice that the ME values for BPW91h$_{abc}$ depend only on the value of a and show much less dependence on the b and c parameters of eq 7 (cf. Table 8). The best functionals are revTPSSh$_a$, B3PW91, TPSSh, PBEh, BPW91h$_a$, and B3LYP. All these functionals show MAEs about 0.01 Å. We identified several problematic compounds in the T-96R molecular geometry test set: Be_2, Li_2, and Na_2. Be_2 is bound by dispersion interaction that is not described well by GGA or meta-GGA. (It is unbound in the Hartree−Fock description.) Although TPSS, PBE, and PBEsol bind Be_2, inclusion of exact exchange makes r_e very large (with an error up to 0.231 Å). The dispersion-bound complexes cannot be described correctly by GGA, metaGGA, or global hybrid functionals. An a posteriori dispersion correction might remedy such errors at no cost by adding the missing C_6, C_8, and C_{10} terms.

Acknowledgment. This work was supported in part by the U.S. National Science Foundation under grant no. DMR-0854769. Computational resources were provided by the Center for Computational Science at Tulane University. This work is connected to the scientific program of the "Development of quality-oriented and harmonized R+D+I strategy and functional model at BME" project, supported by the New Hungary Development Plan (Project ID: TÁMOP-4.2.1/B-09/1/KMR-2010-0002).

Supporting Information Available: Zero-point energy corrected calculated atomization energies and statistical data for the molecules included in the G2-32 test set. Errors for the atoms and molecules included in the AE6 test set. This material is provided free of charge via the Internet at http://pubs.acs.org.

References

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

(2) Kurth, S.; Perdew, J. P. *Int. J. Quantum Chem.* **2000**, *77*, 814.

(3) Langreth, D. C.; Perdew, J. P. *Solid State Commun.* **1976**, *18*, 85.

(4) Gunnarsson, O.; Lundqvist, B. I. *Phys. Rev. B: Solid State* **1976**, *13*, 4274.

(5) Langreth, D. C.; Perdew, J. P. *Phys. Rev. B: Solid State* **1977**, *15*, 2884.

(6) von Barth, U.; Hedin, L. *J. Phys. C: Solid State Phys.* **1972**, *5*, 1629.

(7) Perdew, J. P.; Wang, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *45*, 13244.

(8) Langreth, D. C.; Mehl, M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1983**, *28*, 1809.

(9) Perdew, J. P.; Wang, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *33*, 8800. Perdew, J. P.; Wang, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *40*, 3399 (E).

(10) Perdew, J. P.; Chevary, S. H.; Vosko, S. H.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *46*, 6671. Perdew, J. P.; Chevary, S. H.; Vosko, S. H.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1993**, *48*, 4978 (E).

(11) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865. Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396 (E).

(12) Perdew, J. P.; Burke, K.; Wang, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *54*, 16533. Perdew, J. P.; Burke, K.; Wang, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *57*, 14999 (E).

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

(14) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L.; Zhou, X.; Burke, K. *Phys.*

Rev. Lett. **2008**, *100*, 136406. Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L.; Zhou, X.; Burke, K. *Phys. Rev. Lett.* **2009**, *102*, 039902 (E).

(15) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Constantin, L. A.; Sun, J. *Phys. Rev. Lett.* **2009**, *103*, 026403.

(16) Ernzerhof, M.; Perdew, J. P. *J. Chem. Phys.* **1998**, *109*, 3313.

(17) Constantin, L. A.; Perdew, J. P.; Tao, J. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2006**, *73*, 205104.

(18) Constantin, L. A.; Perdew, J. P.; Pitarke, J. M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2009**, *79*, 075126.

(19) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.

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

(21) Becke, A. D.; Roussel, M. R. *Phys. Rev. A: At., Mol., Opt. Phys.* **1989**, *39*, 3761.

(22) Becke, A. D. *J. Chem. Phys.* **2003**, *119*, 2972.

(23) Roman-Perez, G.; Soler, J. M. *Phys. Rev. Lett.* **2009**, *103*, 096102.

(24) Dion, M.; Rydberg, H.; Schroeder, E.; Lundqvist, B. I. *Phys. Rev. Lett.* **2004**, *92*, 246401.

(25) Perdew, J. P.; Staroverov, V. N.; Tao, J.; Scuseria, G. E. *Phys. Rev. A: At., Mol., Opt. Phys.* **2008**, *78*, 052513.

(26) Perdew, J. P.; Savin, A.; Burke, K. *Phys. Rev. A: At., Mol., Opt. Phys.* **1995**, *51*, 4531.

(27) Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I. *J. Phys. Chem. A* **2005**, *109*, 11006.

(28) Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2006**, *125*, 194112.

(29) Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E. *J. Chem. Phys.* **2007**, *126*, 104102.

(30) Cohen, A. J.; Mori-Sanchez, P.; Yang, W. *Science* **2008**, *321*, 5890.

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

(32) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 1372.

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

(34) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982.

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

(36) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2003**, *118*, 8207. Heyd, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2006**, *124*, 219906 (E).

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

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

(39) Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554.

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

(41) Jimenez-Hoyos, C. A.; Janesko, B. G.; Scuseria, G. E.; Staroverov, V. N.; Perdew, J. P. *J. Chem. Phys.* **2009**, *107*, 1077.

(42) Haunschild, R.; Janesko, B. G.; Scuseria, G. E. *J. Chem. Phys.* **2009**, *131*, 154112.

(43) Burke, K.; Cruz, F. G.; Lam, K.-C. *J. Chem. Phys.* **1998**, *109*, 8161.

(44) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. A: At., Mol., Opt. Phys.* **2008**, *77*, 012509.

(45) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2003**, *119*, 12129. Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. *J. Chem. Phys.* **2004**, *121*, 11507 (E).

(46) Kudin, K. N.; Scuseria, G. E.; Martin, R. L. *Phys. Rev. Lett.* **2002**, *89*, 266402.

(47) Brothers, E. N.; Izmaylov, A. F.; Normand, J. O.; Barone, V.; Scuseria, G. E. *J. Chem. Phys.* **2008**, *129*, 011102.

(48) Batista, E. R.; Heyd, J.; Hennig, R. G.; Uberagua, B. P.; Martin, R. L.; Scuseria, G. E.; Umrigar, C. J.; Wilkins, J. W. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2006**, *74*, 121102.

(49) Perdew, J. P.; Levy, M. *Phys. Rev. Lett.* **1983**, *51*, 1884.

(50) Gruening, M.; Marini, A.; Rubio, A. *J. Chem. Phys.* **2006**, *124*, 154108.

(51) Furche, F.; Perdew, J. P. *J. Chem. Phys.* **2006**, *124*, 044103.

(52) Stroppa, A.; Kresse, G. *New J. Phys.* **2008**, *10*, 063020.

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

(54) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Scuseria, G. E.; Staroverov, V. N.; Tao, J. *Phys. Rev. A: At., Mol., Opt. Phys.* **2007**, *76*, 040501.

(55) Dutoi, A. D.; Head-Gordon, M. *Chem. Phys. Lett.* **2006**, *422*, 230.

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

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

(58) Kurth, S.; Perdew, J. P.; Blaha, P. *Int. J. Quantum Chem.* **1999**, *75*, 889.

(59) Redfern, P. C.; Zapol, P.; Curtiss, L. A.; Raghavachari, K. *J. Phys. Chem. A* **2000**, *104*, 5850.

(60) Csonka, G. I.; Ruzsinszky, A.; Tao, J.; Perdew, J. P. *Int. J. Quantum Chem.* **2005**, *101*, 506.

(61) Check, C. E.; Gilbert, T. M. *J. Org. Chem.* **2005**, *70*, 9828.

(62) Grimme, S. *Angew. Chem., Int. Ed.* **2006**, *45*, 4460.

(63) Paier, J.; Marsman, M.; Kresse, G. *J. Chem. Phys.* **2007**, *127*, 024103.

(64) Wodrich, M. D.; Corminboeuf, C.; Schleyer, P. v. R. *Org. Lett.* **2006**, *8*, 3631.

(65) Adamo, C.; Barone, V. *Chem. Phys. Lett.* **1997**, *274*, 242.

(66) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 42.

(67) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 8996. Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 1460 (E).

(68) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715. Boese, A. D.; Martin, J. M. L. *J. Chem. Phys.* **2004**, *121*, 3405.

(69) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656.

(70) Lynch, B. J.; Truhlar, D. G. Minnesota Database Collection, 2006; http://t1.chem.umn.edu/misc/database_group/database_therm_bh/ae6bh6.pl. Accessed October 10, 2010.

(71) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374.

(72) 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.02; Gaussian, Inc.: Wallingford, CT, 2004.

(73) Csonka, G. I.; Nguyen, N. A.; Kolossváry, I. *J. Comput. Chem.* **1997**, *18*, 1534.

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

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

(76) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski , Fox, D. J. *Gaussian 09*, revision A.2; Gaussian, Inc.: Wallingford, CT, 2009.

(77) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **2001**, *114*, 108.

(78) Kannemann, F. O.; Becke, A. D. *J. Chem. Theory Comput.* **2009**, *5*, 719.

(79) Ruzsinszky, A.; Csonka, G. I.; Scuseria, G. E. *J. Chem. Theory Comput.* **2009**, *5*, 902.

(80) Chakravorty, S. J.; Gwaltney, S. R.; Davidson, E. R.; Parpia, F. A.; Fischer, C. F. *Phys. Rev. A: At., Mol., Opt. Phys.* **1993**, *47*, 3649.

(81) Baboul, A. G.; Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **1999**, *110*, 7650.

(82) Lee, D.; Furche, F.; Burke, K. *J. Phys. Chem. Lett.* **2010**, *1*, 2124, and references therein.

(83) Ruzsinszky, A.; Csonka, G. I.; Perdew, J. P. *J. Phys. Chem. A.* **2005**, *109*, 11015.

(84) Steinmann, S. N.; Csonka, G. I.; Corminboeuf, C. *J. Chem. Theory Comput.* **2009**, *5*, 2950, and references cited therein.

CT100488V

JCTC Journal of Chemical Theory and Computation

Optoelectronic and Excitonic Properties of Oligoacenes: Substantial Improvements from Range-Separated Time-Dependent Density Functional Theory

Bryan M. Wong*,† and Timothy H. Hsieh‡

Materials Chemistry Department, Sandia National Laboratories, Livermore, California 94551, United States, and Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

Received September 16, 2010

Abstract: The optoelectronic and excitonic properties in a series of linear acenes (naphthalene up to heptacene) are investigated using range-separated methods within time-dependent density functional theory (TDDFT). In these rather simple systems, it is well-known that TDDFT methods using conventional hybrid functionals surprisingly fail in describing the low-lying L_a and L_b valence states, resulting in large, growing errors for the L_a state and an incorrect energetic ordering as a function of molecular size. In this work, we demonstrate that the range-separated formalism largely eliminates both of these errors and also provides a consistent description of excitonic properties in these systems. We further demonstrate that reoptimizing the percentage of Hartree−Fock exchange in conventional hybrids to match wave function-based benchmark calculations still yields serious errors, and a full 100% Hartree−Fock range separation is essential for simultaneously describing both of the L_a and L_b transitions. From an analysis of electron−hole transition density matrices, we finally show that conventional hybrid functionals over-delocalize excitons and underestimate quasiparticle energy gaps in the acene systems. The results of our present study emphasize the importance of both a range-separated and asymptotically correct contribution of exchange in TDDFT for investigating optoelectronic and excitonic properties, even for these simple valence excitations.

1. Introduction

Conjugated organic structures have attracted significant recent attention due to their potential applications in single-molecule transistors and organic photovoltaics. In the quest for smaller and more efficient electronics, organic semiconductors serve as a promising alternative to their silicon counterparts because of their increased electronic efficiency[1−5] and ease of chemical functionalization.[6−10] In this context, oligoacenes which are composed of linearly fused benzene rings (Figure 1) have high application potential since they possess large charge-carrier mobilities and tunable electronic band gaps. Most notably, pentacene is already utilized as an organic field-effect transistor due to its large hole mobility (5.5 cm^2/V s) which exceeds that of amorphous silicon.[11−13] In general, the linear acenes are especially important since they form the basic fundamental units of armchair graphene nanoribbons, which continue to garner enormous interest as novel nanoscale materials.[14−19]

In addition to their promising photovoltaic applications, the oligoacenes are also noteworthy as a unique system in which the successes and failures of time-dependent density functional theory (TDDFT) can be assessed and addressed. In 2003, Grimme and Parac noted a dramatic failure (>0.5 eV error in excitation energies) of TDDFT using standard hybrid functionals for the lowest-lying $\pi \rightarrow \pi^*$ states of the oligoacenes.[20] Their findings were particularly unusual since these types of valence excitations are typically well described (within 0.1 eV) by hybrid TDDFT calculations. While it is well-known that long-range charge-transfer and Rydberg excitations provide a significant challenge for TDDFT,[21−32]

* Corrsponding author e-mail: bmwong@sandia.gov.
† Sandia National Laboratories.
‡ Massachusetts Institute of Technology.

10.1021/ct100529s
Published on Web 11/11/2010

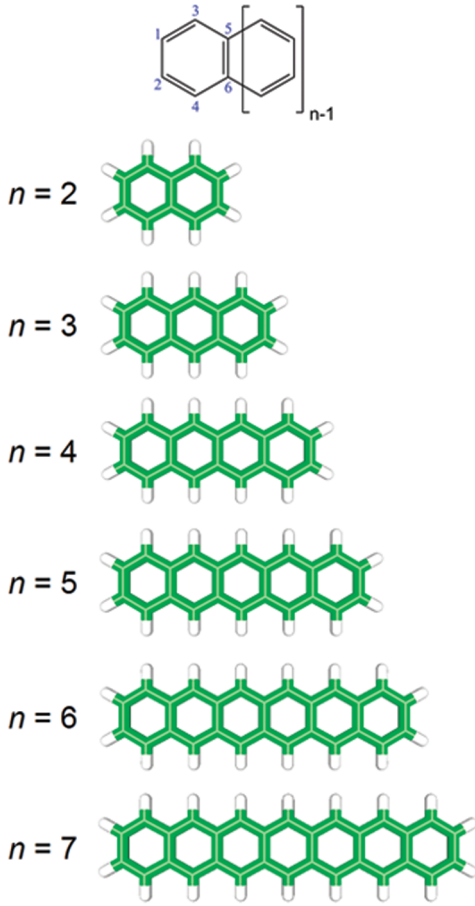


Figure 1. Molecular structure and atom labels for the linear acenes. The specific atom numbers depicted in this figure define an ordered basis for generating the transition density matrices discussed in section 3.

these effects are not present in the acene systems since none of the valence excitations possess Rydberg character or involve any long-range charge transfer (both the ground- and excited-state dipole moments are exactly zero by molecular symmetry). As a result, the unexpected failure of TDDFT in these simple valence excitations is most unusual and somewhat surprising.

The present study has two aims. First, we show that certain range-separated functionals,[33−46] which incorporate *both* a position-dependent admixture and an asymptotically correct contribution of Hartree−Fock (HF) exchange, yield substantial improvements over conventional hybrids for the various oligoacene excitations. Numerical optimization of parameters in the range-separated and hybrid functionals is carried out to understand their effect on excitation energies and their overall trends. Following the two-dimensional real-space analysis approach of Tretiak et al.,[47−50] we then examine excitonic effects for the various excitations and TDDFT methods. The transition densities and electron difference density maps enable us to understand why conventional hybrids fail and how range-separated functionals accurately reproduce excitation energies and quasiparticle energy gaps for each of the different transitions. We begin by briefly reviewing these two different formalisms and then compare their accuracy in predicting oligoacene excitation properties.

2. Theory and Methodology

2.1. Global Hybrid Functionals. Recall that DFT is an exact theory in which the only inaccuracies encountered in practice arise from approximations to the (still unknown) exchange-correlation functional. One of the most widely used DFT schemes for the exchange-correlation energy is Becke's three-parameter B3LYP method,[51] which has a relatively simple formulation given by

$$E_{xc}^{\text{global}} = a_0 E_{x,\text{HF}} + (1 - a_0)E_{x,\text{Slater}} + a_x \Delta E_{x,\text{Becke88}} + (1 - a_c)E_{c,\text{VWN}} + a_c \Delta E_{c,\text{LYP}} \quad (1)$$

In this expression, $E_{x,\text{HF}}$ is the HF exchange energy based on Kohn−Sham orbitals, $E_{x,\text{Slater}}$ is the uniform electron gas exchange-correlation energy,[52] $\Delta E_{x,\text{Becke88}}$ is Becke's 1998 generalized gradient approximation (GGA) for exchange,[53] $E_{c,\text{VWN}}$ is the Vosko−Wilk−Nusair 1980 correlation functional,[54] and $\Delta E_{c,\text{LYP}}$ is the Lee−Yang−Parr correlation functional.[55] Depending on the choice of GGA, there are numerous other hybrid functionals in the literature which combine different GGA treatments of exchange and correlation with varying coefficients. In these "global hybrid" functionals, the fraction of nonlocal HF exchange, a_0, is held constant in space and fixed to a GGA-specific value (the B3LYP functional, for example, is parametrized with $a_0 = 0.20$).

2.2. Range-Separated Functionals. In contrast to conventional hybrids which incorporate a constant fraction of HF exchange, the long-range-corrected[35,37,40,41] (abbreviated as LC or LRC in the literature) formalism mixes HF exchange densities nonuniformly by partitioning the electron repulsion operator as

$$\frac{1}{r_{12}} = \frac{1 - \text{erf}(\mu r_{12})}{r_{12}} + \frac{\text{erf}(\mu r_{12})}{r_{12}} \quad (2)$$

The "erf" term denotes the standard error function. $r_{12} = |\mathbf{r}_1 - \mathbf{r}_2|$ is the interelectronic distance between electrons at coordinates $\mathbf{r}_1$ and $\mathbf{r}_2$, and μ is the range-separation parameter in units of Bohr^{-1}. The first term in eq 2 is a short-range interaction which decays rapidly on a length scale of $\sim 2/\mu$, and the second term is the long-range part of the Coulomb potential. For a general GGA or hybrid functional, the corresponding exchange-correlation energy according to the LC formalism is

$$E_{xc}^{\text{LC}} = E_{c,\text{DFT}} + (1 - a_{\text{HF}})E_{x,\text{DFT}}^{\text{SR}} + a_{\text{HF}}E_{x,\text{HF}}^{\text{SR}} + E_{x,\text{HF}}^{\text{LR}} \quad (3)$$

In this expression, $E_{c,\text{DFT}}$ is the original, unmodified DFT correlation contribution, $E_{x,\text{DFT}}^{\text{SR}}$ and $E_{x,\text{HF}}^{\text{SR}}$ are the respective DFT and HF contributions computed with the short-range part of the Coulomb operator (first term in eq 2), and $E_{x,\text{HF}}^{\text{LR}}$ is the HF exchange contribution evaluated using the long-range part of the Coulomb potential[56] (second term in eq 2). The a_{HF} parameter is the coefficient of HF exchange present in the original hybrid functional ($a_{\text{HF}} = 0$ if the original functional is a pure density functional, i.e., BLYP, BOP, or PBE).

It is important to mention at this point that there are also several other range-separation techniques and functionals in

the literature, and that the prescription given in eqs 2 and 3 is only one of many LC forms. For example, the range-separation technique has been further modified by Handy et al.[39,42] with their CAM-B3LYP (Coulomb-attenuating method−B3LYP) methods. Similarly, the Scuseria group has also developed several new range-separated functionals based on a semilocal exchange-hole approach.[43−46] These exchange-hole models have been further refined by the Herbert group to design new functionals which accurately describe both ground and excited states.[25,28] In terms of chemical applications, Jacquemin et al. have also presented benchmarks for several families of excitations including the electronic spectra of anthroquinone dyes,[57] $n \rightarrow \pi^*$ transitions in nitroso and thiocarbonyl dyes,[58] and $\pi \rightarrow \pi^*$ excitations in organic chromophores.[59] Very recently, there has also been pioneering work by the Baer group and the Kronik group in constructing range-separation functionals tuned entirely from first principles.[29,30] The key to their success is the choice of a range-separation parameter, μ, which minimizes the difference between the ionization energy (IE) and the negative of the highest-occupied molecular orbital (HOMO) energy, $-E_{\mathrm{HOMO}}$, of the molecule. Since the ionization energy is rigorously equal to $-E_{\mathrm{HOMO}}$ for an "exact functional," the formalism by Baer and co-workers is entirely self-consistent and does not require any experimental input or high-level benchmark calculations.

In all of these various range-separated methods, the key improvement in their accuracy is the smooth separation of DFT and nonlocal HF exchange interactions through the parameter μ. Specifically, the exchange-correlation potentials of conventional density functionals exhibit the wrong asymptotic behavior, but the LC scheme ensures that the exchange potential smoothly recovers the exact $-1/r$ dependence at large interelectronic distances. It is important to point out that the length-scale partitioning in the LC formalism is essential for obtaining accurate TDDFT results. More precisely, a 100% global HF exchange fraction without range separation can corrupt the delicate balance between exchange and correlation contributions, resulting in large errors in excitation energies. For extended charge-transfer processes, the long-range exchange corrections are also particularly vital since these types of excitations are especially sensitive to the asymptotic part of the nonlocal exchange-correlation potential.

2.3. Computational Details. For the linear acenes in this work, we compared the performance of global hybrid functionals against range-separated and wave function-based calculations. In order to investigate the role of different HF exchange schemes in the various TDDFT methods, we explored the effect of changing the HF exchange fraction, a_0, in the global hybrid model and the result of varying the range-separation parameter μ within the LC formalism. For the parametric study on global hybrids, we kept the same functional form in Becke's three-parameter model (eq 1) and computed vertical singlet excitation energies as a function of a_0 ranging from 0.0 to 1.0 in increments of 0.05. In these calculations, we fixed $a_x = 1 - a_0$ in eq 1 but kept the correlation contribution with $a_c = 0.81$ unchanged. The $a_x = 1 - a_0$ convention is a common choice used in many hybrid functionals[60−63] such as Becke's B1 convention[61] (in a previous study on large oligothiophenes,[31] we had carried out calculations with a_x fixed to the original 0.72 value recommended by Becke and found that all of the excitation energies were nearly identical compared to the $a_x = 1 - a_0$ convention).

To explore the effect of range-separated exchange on the optoelectronic properties of the acenes, we computed vertical singlet excitation energies as a function of μ ranging from 0 to 0.90 Bohr^{-1} (in increments of 0.05 Bohr^{-1}) while keeping the correlation contribution $E_{c,\mathrm{DFT}}$ in the LC-BLYP functional unchanged. In our study, we utilized several range-separated functionals including CAM-B3LYP, LC-BOP, LC-PBE, LC-ωPBE, and LC-BLYP but found that all of the full-HF-exchange LC functionals gave similar results for the linear acenes. It is very important to note that the original CAM-B3LYP functional is defined[39] with a coulomb-attenuating parameter of $\alpha + \beta = 0.65$ and, therefore, exhibits a $-0.65/r$ dependence for the exchange potential. As a result, the CAM-B3LYP functional is particularly different than the other LC functionals considered in this work since it does not incorporate a full 100% HF exchange at large interelectronic distances. The very similar results obtained from the other full-exchange LC functionals imply that the excitation energies are not very sensitive to the specific DFT correlation contribution used, and that the systematic error observed previously by Grimme and Parac[20] is largely due to the HF exchange component for the acene systems. In light of these similarities, much of our parametric study focuses on the LC-BLYP results since the other full-HF-exchange LC methods give very similar energies as a function of μ. We should also note that a direct comparison between LC-BLYP, CAM-B3LYP, and the global hybrid model in eq 1 allows a very fair and consistent evaluation since all of these methods have similar correlation contributions.

As benchmarks for assessing the quality of the various TDDFT methods, we calculated CC2/cc-pVTZ excitation energies for the linear acenes ranging from $n = 2$ to 7 benzene rings (we stop at $n = 7$ since our CC2/cc-pVTZ calculations indicate a very abrupt and large multireference/diradical character for $n = 8$). We use the CC2 excitation energies as reference values since EOM-CCSD and CASPT2 calculations with the cc-PVTZ basis set were out of reach for larger acenes containing five or more benzene rings. Furthermore, we consider the CC2 results as reliable reference values since they accurately reproduce solvent-corrected experimental excitation energies[20] (see Table 1) and are close to CC3 benchmark calculations for the smaller acenes.[64] As an additional check on the quality of the CC2 calculations, we found that none of the acene systems required a multireference treatment of electron correlation (D1 diagnostic values were in the 0.04−0.06 range), and contributions from single excitations were always greater than 90%.

In order to maintain a consistent comparison across the B3LYP, CAM-B3LYP, LC-BOP, LC-PBE, LC-ωPBE, LC-BLYP, and CC2 levels of theory, identical molecular geometries were used for each of these methods. These reference geometries were optimized at the B3LYP/cc-PVTZ level of theory and are available in the Supporting Informa-

Table 1. Comparison of TDDFT, CC2, and Experimental Excitation Energies [eV] (Wavelengths [nm] Are in Parentheses) for the L_a and L_b States in the Linear Acenes[a]

number of rings	B3LYP ($a_0 = 0.20$)	B3LYP$_{opt}$ ($a_0 = 0.50$)	CAM-B3LYP ($\alpha + \beta = 0.65$)	LC-BOP ($\mu = 0.29$)	LC-PBE ($\mu = 0.29$)	LC-ωPBE ($\mu = 0.29$)	LC-BLYP ($\mu = 0.29$)	CC2	experiment[20]
				L_a state					
2	4.39 (282)	4.69 (264)	4.68 (265)	4.76 (260)	5.05 (246)	4.80 (258)	4.76 (260)	4.89 (254)	4.66 (266)
3	3.22 (385)	3.54 (350)	3.54 (350)	3.64 (341)	3.66 (339)	3.67 (338)	3.63 (342)	3.70 (335)	3.60 (344)
4	2.44 (508)	2.75 (451)	2.77 (448)	2.89 (429)	2.89 (429)	2.91 (426)	2.88 (431)	2.90 (428)	2.88 (431)
5	1.89 (656)	2.19 (566)	2.22 (558)	2.36 (525)	2.36 (525)	2.38 (521)	2.35 (528)	2.35 (528)	2.37 (523)
6	1.49 (832)	1.77 (700)	1.83 (438)	1.97 (629)	1.98 (626)	1.99 (623)	1.96 (633)	1.95 (636)	2.02 (614)
7	1.18 (1051)	1.46 (849)	1.53 (810)	1.68 (738)	1.69 (734)	1.71 (725)	1.68 (738)	1.66 (747)	
MAE (eV)	0.42	0.13	0.11	0.04	0.10	0.06	0.04	0.09	
				L_b state					
2	4.48 (277)	4.75 (261)	4.63 (268)	4.59 (270)	4.62 (268)	4.61 (269)	4.59 (270)	4.47 (277)	4.13 (300)
3	3.87 (320)	4.14 (299)	4.04 (307)	4.02 (308)	4.04 (307)	4.03 (308)	4.02 (308)	3.90 (318)	3.64 (341)
4	3.48 (357)	3.73 (332)	3.66 (339)	3.65 (340)	3.67 (338)	3.66 (339)	3.65 (340)	3.52 (352)	3.39 (366)
5	3.21 (386)	3.46 (358)	3.40 (365)	3.40 (365)	3.41 (364)	3.41 (364)	3.39 (366)	3.27 (379)	3.12 (397)
6	3.02 (411)	3.26 (380)	3.21 (386)	3.22 (385)	3.23 (384)	3.23 (384)	3.22 (385)	3.09 (401)	2.87 (432)
7	2.88 (431)	3.11 (399)	3.08 (403)	3.09 (401)	3.10 (400)	3.10 (400)	3.08 (403)	2.97 (417)	-
MAE (eV)	0.18	0.44	0.36	0.35	0.36	0.36	0.34	0.22	-

[a] The mean absolute errors (MAE) relative to solvent-corrected experimental results are listed below each of the various methods. Excitation energies were computed with the cc-pVTZ basis with the same reference geometry for all of the different methods.

tion. For all of the TDDFT excitation energies, we used a cc-PVTZ basis set and a high-accuracy Lebedev grid consisting of 96 radial and 302 angular quadrature points. All TDDFT calculations were performed with a locally modified version of GAMESS,[65] and the CC2 calculations were carried out with the TURBOMOLE package.[66]

3. Results and Discussion

We focus on two different valence excitations in the linear acenes, commonly labeled in the literature[20,67] as L_a (lowest excited state of B_{2u} symmetry) and L_b (B_{3u} symmetry). The L_a excited state results from a HOMO → LUMO transition with polarization along the molecular short axis, and the L_b state is characterized by a nearly equal mixture of HOMO−1 → LUMO and HOMO → LUMO+1 excitations with a total polarization along the long axis.[68] Using the CC2 excitations as reference values, we performed a total root-mean-square error (RMSE) analysis for all 12 energies (six L_a and six L_b transitions) as a function of μ and a_0. As seen in Figure 2a, the RMSE curve for LC-BLYP has a minimum at $\mu = 0.29$ Bohr^{-1} with a RMS error of 0.10 eV. Perhaps, surprisingly, this RMSE-optimized value of μ is nearly identical to the 0.31 Bohr^{-1} value recommended for simultaneously describing excitation and fluorescence energies in large oligothiophenes.[31] The RMSE in Figure 2b for the B3LYP-like global hybrid functional has a minimum at $a_0 = 0.50$, with a larger error of 0.20 eV. It is worth noting that our RMSE-minimization with $a_0 = 0.50$ and $a_x = 1 - a_0$ yields a functional very similar to the BHHLYP functional (originally defined with $a_c = 0$) with the exception that our choice has an extra correlation contribution due to the $\Delta E_{c,LYP}$ term in eq 1. We denote this reoptimized hybrid functional with $a_0 = 0.50$ as B3LYP$_{opt}$ in the remainder of this work. Unless otherwise noted, all further LC-TDDFT calculations indicate a range-separation parameter of $\mu = 0.29$ Bohr^{-1}.

Table 1 compares the L_a and L_b excited-state energies between B3LYP, B3LYP$_{opt}$, CAM-B3LYP, LC-BOP, LC-PBE, LC-ωPBE, LC-BLYP, and CC2, and Figure 3a,b depicts in more detail the general trends in transition energies (expressed in wavelength units) between the various TDDFT and CC2 results. It is most important to note in these figures that the energetic ordering of the two electronic states is different, depending on the size of the acene. Specifically, both CC2 and experimental studies[69] indicate that a curve crossing between the L_a and L_b states occurs slightly before $n = 3$ benzene rings (anthracene). For all of the other larger acenes, the L_a state lies energetically below the L_b state. Examining Table 1 and Figure 3b, we find that the full-

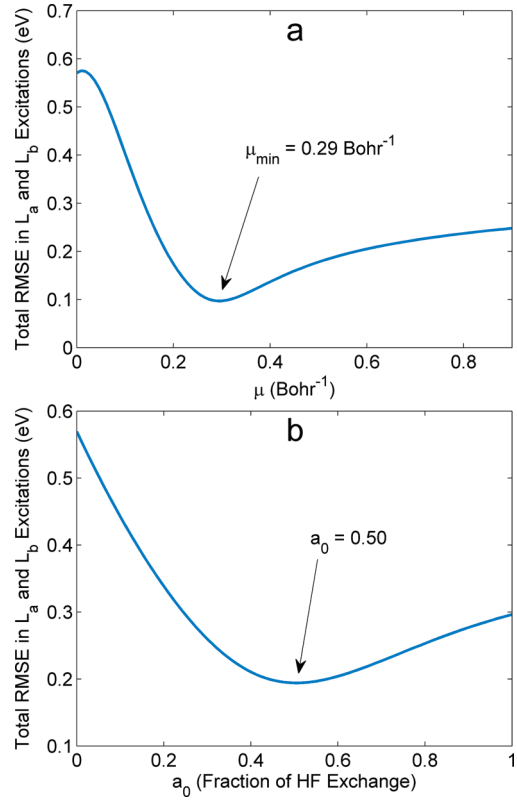


Figure 2. Total root-mean-square errors (RMSE) as a function of (a) the range-separation parameter μ in the LC-BLYP functional and (b) the HF exchange fraction a_0 in a B3LYP-like hybrid functional. Part a shows the RMSE curve having a minimum at $\mu = 0.29$ Bohr^{-1}, and part b shows the RMSE curve having a minimum at $a_0 = 0.50$.

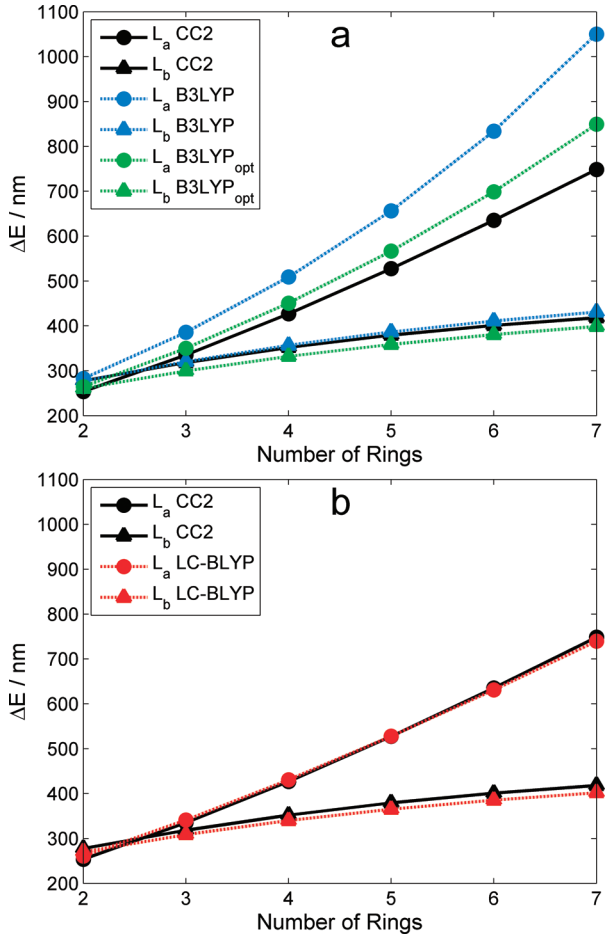


Figure 3. Comparison between TDDFT and CC2 excitation energies (in wavelength units) for (a) conventional global hybrid and (b) range-separated LC-BLYP functionals. The B3LYP$_{opt}$ functional denotes a modified B3LYP functional with a RMSE-optimized exchange fraction of $a_0 = 0.50$, as discussed in the main text.

exchange LC-TDDFT calculations are unique in that they show excellent agreement with CC2 energies for both the L_a and L_b excitations. Moreover, all of the LC-TDDFT methods preserve the correct ordering of electronic states between $n = 2$ and $n = 3$ benzene rings. In the case of the CAM-B3LYP functional though (which only has 65% HF exchange at long-range), there are still some systematic discrepancies for the L_a excitations which are still somewhat underestimated. Although the energetic ordering of the L_a and L_b states is correctly predicted by CAM-B3LYP, the energy differences are almost negligible, with only a 0.05 eV difference separating the L_a and L_b states of naphthalene (compared to a ~0.2 eV difference with the full-exchange LC functionals). These observations strongly indicate that a range-separated partitioning of exchange alone, without 100% asymptotic HF exchange, is not sufficient, and a full asymptotic contribution of exchange is essential for accurately describing both the L_a and L_b excitations. Turning now to the global hybrids, Figure 3a shows that the B3LYP functional severely underestimates excitation energies (i.e., overestimates absorption wavelengths) for the L_a electronic state. The situation is somewhat improved upon using the RMSE-optimized $a_0 = 0.50$ value in B3LYP$_{opt}$; however,

Table 2. Comparison of the TDDFT Λ-Overlap Diagnostic for the L_a and L_b Excited States in the Linear Acenes

number of rings	B3LYP ($a_0 = 0.20$)	LC-BLYP ($\mu = 0.29$)
	Λ-overlap for L_a states	
2	0.89	0.89
3	0.88	0.88
4	0.88	0.88
5	0.89	0.89
6	0.89	0.89
7	0.90	0.90
	Λ-overlap for L_b states	
2	0.65	0.64
3	0.65	0.65
4	0.63	0.63
5	0.62	0.62
6	0.60	0.61
7	0.59	0.60

this procedure results in L_b excitations which are now *overestimated* and L_a excitations which are still quite underestimated. Most importantly, both B3LYP and B3LYP$_{opt}$ give an incorrect ordering of electronic states—the crossing between L_a and L_b curves occurs much too early in both functionals, and the electronic symmetries in naphthalene have the wrong order. In general, the accuracy in excitation energies and trends is significantly improved with the LC scheme, while conventional hybrids are unable to reproduce the qualitative behavior in excitations *even if the fraction of HF exchange is optimized*.

From these results, it is interesting to note that long-range charge transfer is not responsible for the unexpected failure of B3LYP in these highly symmetrical systems. In a recent benchmark study, Peach et al.[26] introduced a diagnostic test which quantifies the spatial overlap, Λ, between the occupied and virtual orbitals involved in an excitation. This diagnostic metric is typically used to postprocess a converged TDDFT calculation and has an intuitive form given by

$$\Lambda = \frac{\sum_{i,a} (X_{ia} + Y_{ia})^2 O_{ia}}{\sum_{i,a} (X_{ia} + Y_{ia})^2} \tag{4}$$

In this expression, X_{ia} and Y_{ia} are the virtual−occupied and occupied−virtual transition amplitudes, respectively, and O_{ia} is the spatial overlap integral of the moduli of the two orbitals, $O_{ia} = \int |\phi_i(\mathbf{r})||\phi_a(\mathbf{r})|\, d\mathbf{r}$. By construction, the diagnostic metric Λ is bounded between 0 and 1, with small values signifying a long-range excitation and large values indicating a localized, short-range transition. On the basis of their extensive benchmarks, if Λ is less than 0.3, indicating little overlap and significant long-range charge transfer character, hybrid functionals are predicted to yield inaccurate results. In Table 2, we computed the Λ diagnostic for both the L_a and L_b states and found that all values were well above the 0.3 threshold (some of them even approaching 0.9), indicating a substantial overlap and no long-range charge transfer in these systems (it is rather interesting though that the diagnostic incorrectly predicts the L_a excited state to be more accurately described than the L_b state in both the B3LYP and LC-BLYP functionals). Thus, instead of long-

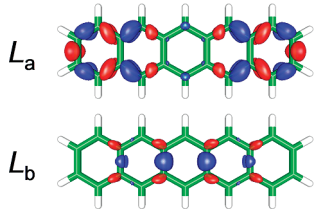


Figure 4. Electron density difference maps ($\rho_{excited} - \rho_{ground}$) for the L_a and L_b excited states of anthracene computed at the CC2 level of theory. Red regions denote a positive density difference (accumulation of density upon electronic excitation), and blue regions represent a negative density difference (depletion of density upon excitation). Both densities are plotted using the same isosurface contour value.

range charge transfer from one end of the molecule to the other, we do find that the L_a excitation involves a sizable local rearrangement of electron density. In support of this assertion, Figure 4 depicts the electron density difference map ($\rho_{excited} - \rho_{ground}$) for the L_a and L_b excited states in pentacene computed at the CC2 level of theory (difference maps for the other acenes can be found in the Supporting Information). The electron density difference map gives a dynamic visualization of electronic rearrangement for a transition, with red regions (positively valued) denoting an accumulation of density and blue regions (negatively valued) representing a depletion of density upon excitation. As depicted in Figure 4, the L_a state involves significantly more local charge redistribution than the higher-energy L_b state. In contrast, the L_b excited-state density is very similar to the ground state, as evidenced by the very small and sparsely distributed isosurface regions. These CC2 difference densities confirm the long-held valence-bond viewpoint[70−72] that the L_a state possesses an "ionic" character whereas the L_b transition is primarily covalent in nature. Notice also that the length scale of charge redistribution is on the order of the carbon−carbon bond length (∼1.4 Å), which is comparable to the length scale at which LC-BLYP predicts long-range HF exchange to dominate short-range DFT correlation ($1/\mu \sim 1.8$ Å). Even though none of these transitions have long-range charge transfer character, our findings do support the physical interpretation that a range-separated contribution of full HF exchange on the length scale of the molecule is still necessary for accurately describing these local charge rearrangements.

In order to provide further insight into these optoelectronic trends, we carried out an investigation of excitonic effects by analyzing electron−hole transition density matrices for the various excitations and TDDFT methods. Following the two-dimensional real-space analysis approach of Tretiak et al.,[47−50] one can construct coordinate $\mathbf{Q}_v$ and momentum $\mathbf{P}_v$ matrices with elements given by

$$(Q_v)_{mn} = \langle\psi_v|c_m^\dagger c_n|\psi_g\rangle + \langle\psi_g|c_m^\dagger c_n|\psi_v\rangle \quad (5)$$

$$(P_v)_{mn} = \langle\psi_v|c_m^\dagger c_n|\psi_g\rangle - \langle\psi_g|c_m^\dagger c_n|\psi_v\rangle \quad (6)$$

where ψ_g and ψ_v are ground and excited states, respectively. The Fermi operators $c_i^\dagger$ and c_i represent the creation and annihilation of an electron in the ith basis set orbital in ψ.

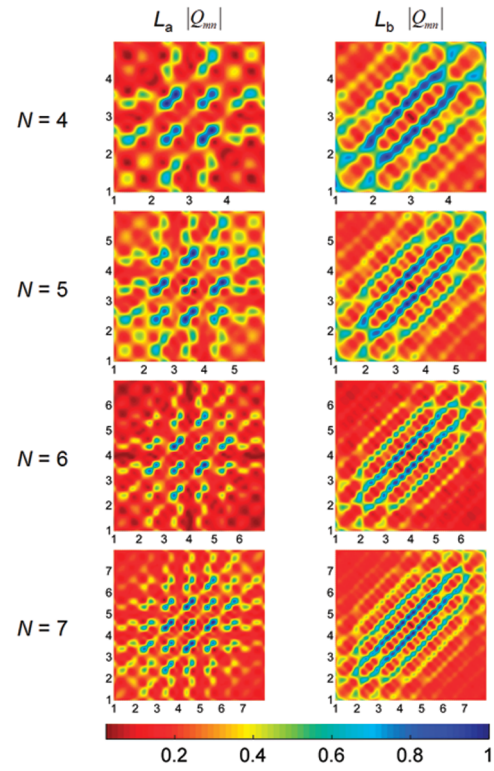


Figure 5. Contour plots of coordinate density matrices (**Q**) for the L_a and L_b excited states computed at the LC-BLYP level of theory. The *x*- and *y*-axis labels represent the number of benzene repeat units in the molecule. The elements of the coordinate matrix, Q_{mn}, give a measure of exciton delocalization between sites *m* (*x* axis) and *n* (*y* axis). The color scale is given at the bottom.

For the acene systems analyzed in this work, the $\mathbf{Q}_v$ and $\mathbf{P}_v$ matrices each form a two-dimensional *xy* grid over all of the carbon sites along the *x* and *y* axes. The specific ordering of the carbon sites used in this work is shown in Figure 1. The $(Q_v)_{mn}$ coordinate matrix gives a measure of exciton delocalization between sites *m* and *n*, and the $(P_v)_{mn}$ momentum matrix represents the probability amplitude of an electron−hole pair oscillation between carbon sites *m* and *n*, respectively. Each of these matrices provides a complementary view of exciton delocalization and electron−hole coherence for optical transitions within the acene systems.

Figure 5 displays the absolute value of the coordinate density matrix elements, $|(Q_v)_{mn}|$, for the L_a and L_b excitation energies computed at the LC-BLYP level of theory. The *x* and *y* axes in this figure represent the benzene repeat units in the molecule, and the individual matrix elements are depicted by the various colors. On the basis of its construction, off-diagonal elements with large intensities represent widely separated electron−hole pairs between different atoms. As shown in Figure 5, the L_a density matrix has more off-diagonal elements than the corresponding L_b excitation, whose matrix elements are primarily confined along the diagonal. These figures reflect the more delocalized nature of the L_a state, in agreement with the electron density difference maps discussed previously. It is also important to note that all the transition density plots are symmetric along the counterdiagonal, verifying that no long-range charge transfer occurs in these systems (an asymmetric

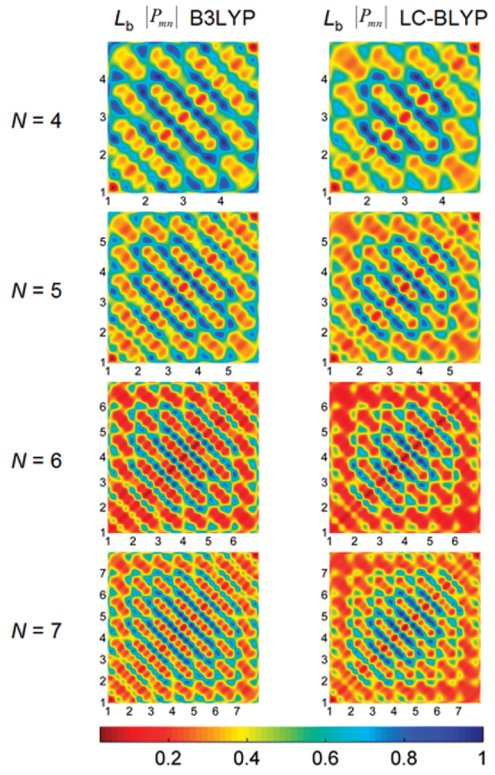


Figure 6. Contour plots of momentum density matrices (**P**) for the L_b excited state computed at the B3LYP and LC-BLYP levels of theory. The *x*- and *y*-axis labels represent the number of benzene repeat units in the molecule. The elements of the momentum matrix, P_{mn}, represent the probability amplitude of an electron−hole pair oscillation between sites *m* (*x* axis) and *n* (*y* axis). The color scale is given at the bottom.

transition density along the counterdiagonal implies more electrons than holes are localized on one side of the molecule).

The coherence size, which characterizes the distance between an electron and a hole, is given by the width of the momentum density matrix, $\mathbf{P_v}$. To compare excitonic effects between global and range-separated hybrids, we plot the absolute value of the momentum density matrix elements, $|(P_v)_{mn}|$, for both the B3LYP and LC-BLYP functionals in Figure 6 (transition density plots for all of the different functionals and excited states can be found in the Supporting Information). These figures show that the B3LYP functional gives a more delocalized density-matrix pattern and a larger coherence size compared to the LC-BLYP functional. Furthermore, the coherence size as predicted by the B3LYP functional is larger by nearly one repeat unit in comparison to the LC-BLYP results. These findings are consistent with the B3LYP formalism which only incorporates a global fraction of 20% HF exchange and, therefore, exhibits a $-0.2/r$ dependence for the exchange potential. As a result, the asymptotically incorrect B3LYP exchange potential is not attractive enough, leading to an over-delocalized electron−hole pair and, therefore, an overestimated coherence size in the acene systems.

Finally, it is interesting to compare quasiparticle energy gaps predicted by both global hybrid and range-separated functionals in the acene systems. Within Kohn−Sham theory,[73] the quasiparticle gap can be approximated by the difference between the lowest unoccupied and highest unoccupied molecular orbital energies, $E_{LUMO} - E_{HOMO}$. Table 3 compares $-E_{LUMO}$, $-E_{HOMO}$, and the experimental ionization energies (IE) for the linear acenes computed at the B3LYP, CAM-B3LYP, and LC-BLYP levels of theory. From Kohn−Sham theory, it is well-known that an "exact functional" (if one had access to such a functional), would yield an ionization energy exactly equal to $-E_{HOMO}$. For the pentacene molecule, as a specific example, the B3LYP functional provides a $-E_{HOMO}$ value of 4.78 eV which significantly underestimates the experimental ionization energy[74] of 6.61 eV. The $-E_{HOMO}$ values predicted by CAM-B3LYP are an improvement over the B3LYP energies, but the average deviation of −0.70 eV from the experimental IEs is still quite large. In contrast, the LC formalism, which incorporates a correct asymptotic behavior of the exchange potential by construction, gives $-E_{HOMO}$ values in exceptional agreement with all of the experimental IEs, resulting in an impressive average deviation of 0.07 eV. These results complement our previous discussion of L_a and L_b excitation energies by further demonstrating that a full 100% asymptotic contribution of HF exchange is necessary to provide a consistent description of electronic properties in these systems. Furthermore, these findings demonstrate that the range-separated formalism with full asymptotic HF exchange is very self-consistent—*both* the excitation energies and quasiparticle properties in these systems are predicted accurately while simultaneously satisfying the energy constraints as required by Kohn−Sham theory.

4. Conclusion

In conclusion, the present study clearly indicates that both a range-separated partitioning as well as an asymptotically

Table 3. Comparison of $-E_{LUMO}$, $-E_{HOMO}$, and Experimental Ionization Energies (IE) for the Linear Acenes Computed at the B3LYP, CAM-B3LYP, and LC-BLYP Levels of Theory[a]

	B3LYP ($a_0 = 0.20$)		CAM-B3LYP ($\alpha + \beta = 0.65$)		LC-BLYP ($\mu = 0.29$)		
number of rings	$-E_{LUMO}$ (eV)	$-E_{HOMO}$ (eV)	$-E_{LUMO}$ (eV)	$-E_{HOMO}$ (eV)	$-E_{LUMO}$ (eV)	$-E_{HOMO}$ (eV)	exp.[74] IE (eV)
2	1.21	6.00	0.10	7.40	−0.60	8.21	8.14
3	1.85	5.43	0.84	6.72	0.17	7.51	7.44
4	2.29	5.05	1.34	6.27	0.69	7.03	6.97
5	2.59	4.78	1.71	5.95	1.07	6.69	6.63
6	2.81	4.59	1.98	5.71	1.36	6.44	6.36
7	2.98	4.45	2.18	5.53	1.57	6.25	–
$\langle -E_{HOMO} - IE \rangle$	–	−1.94	–	−0.70	–	0.07	–

[a] The average deviation of $-E_{HOMO}$ relative to the experimental IE is listed below each of the various methods.

correct contribution of exchange play a vital role in predicting optoelectronic properties in the linear acenes. Even though none of the excitations involve extended long-range charge transfer, we find that a range-separated contribution of full exchange is still necessary to accurately describe both the valence excitation energies and the $L_a \rightarrow L_b$ curve crossing in these simple systems. The results of our observations also strongly indicate that a range-separated partitioning of exchange by itself, without 100% asymptotic HF exchange (i.e., CAM-B3LYP), is not sufficient to accurately describe both the L_a and L_b states. Conversely, reoptimization of functional parameters toward 100% full exchange without range separation in a global hybrid does not improve the situation either; in fact, this reparameterization results in a corruption between exchange and correlation errors with trends in L_a and L_b excitations being even more poorly described. In particular, we find that global hybrid functionals overdelocalize excitons, underestimate quasiparticle energies, and are unable to reproduce general trends in both L_a and L_b, even if the fraction of HF exchange is optimized. The most important results of our observations indicate that a simultaneous use of range-separated partitioning as well as a full contribution of exchange at large interelectronic distances is essential for accurately describing both the L_a and L_b states in these systems.

As acenes form the basis of nanoribbons and other polycyclic aromatic hydrocarbons,[75] this study serves an important role in determining which TDDFT methods are most appropriate for these systems, especially since wave function-based calculations on carbon nanostructures are still prohibitively demanding. Looking forward, it would be extremely interesting to see if the range-separated formalism also provides a similar accuracy for describing triplet states in acenes and other chromophores. While this study focused on only singlet excitations, further work is still needed to understand triplet excitations since exciton fission to low-lying triplet states ultimately control the electronic efficiencies in photovoltaic systems.[76] We are currently investigating these triplet states, with further calculations on extended organic light-harvesting systems,[9] to help predict the efficiencies of these materials. With this in mind, we anticipate that the LC-TDDFT technique will play a significant role in understanding and accurately predicting the optoelectronic properties in these novel nanostructures.

Acknowledgment. This research was supported in part by the National Science Foundation through TeraGrid resources (Grant No. TG-CHE1000066N) provided by the National Center for Supercomputing Applications. Funding for this effort was provided by the Laboratory Directed Research and Development (LDRD) program at Sandia National Laboratories, a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

Supporting Information Available: Electron density difference maps, contour plots of coordinate (**Q**) and momentum (**P**) density matrices, and Cartesian coordinates of all of the optimized structures. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Lane, P. A.; Kafafi, Z. H. Solid-state organic photovoltaics: a review of molecular and polymeric devices. In *Organic Photovoltaics: Mechanisms, Materials, and Devices*; Sun, S., Sariciftci, N. S., Eds.; CRC Press: Boca Raton, FL, 2005; pp 49−104.

(2) Picciolo, L. C.; Murata, H.; Kafafi, Z. H. *Appl. Phys. Lett.* **2001**, *78*, 2378.

(3) Kelley, T. W.; Baude, P. F.; Gerlach, C.; Ender, D. E.; Muyres, D.; Haase, M. A.; Vogel, D. E.; Theiss, S. D. *Chem. Mater.* **2004**, *16*, 4413.

(4) Wong, B. M.; Morales, A. M. *J. Phys. D: Appl. Phys.* **2009**, *42*, 055111.

(5) Zade, S. S.; Bendikov, M. *Angew. Chem., Int. Ed.* **2010**, *49*, 4012.

(6) Dimitrakopoulos, C. D.; Malenfant, P. R. L. *Adv. Mater.* **2002**, *14*, 99.

(7) Odom, S. A.; Parkin, S. R.; Anthony, J. E. *Org. Lett.* **2003**, *5*, 4245.

(8) Anthony, J. E. *Angew. Chem., Int. Ed.* **2008**, *47*, 452.

(9) Zhou, X.; Zifer, T.; Wong, B. M.; Krafcik, K. L.; Léonard, F.; Vance, A. L. *Nano Lett.* **2009**, *9*, 1028.

(10) Kaur, I.; Jazdzyk, M.; Stein, N. N.; Prusevich, P.; Miller, G. P. *J. Am. Chem. Soc.* **2010**, *132*, 1261.

(11) Koch, N. *ChemPhysChem* **2007**, *8*, 1438.

(12) Yamashita, Y. *Sci. Technol. Adv. Mater.* **2009**, *10*, 024313.

(13) Hasegawa, T.; Takeya, J. *Sci. Technol. Adv. Mater.* **2009**, *10*, 024314.

(14) Wakabayashi, K.; Fujita, M.; Ajiki, H.; Sigrist, M. *Phys. Rev. B* **1999**, *59*, 8271.

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

(16) Son, Y.-W.; Cohen, M. L.; Souie, S. G. *Phys. Rev. Lett.* **2006**, *97*, 216803.

(17) White, C. T.; Li, J.; Gunlycke, D.; Mintimire, J. W. *Nano Lett.* **2007**, *7*, 825.

(18) Raza, H.; Kan, E. C. *Phys. Rev. B* **2008**, *77*, 245434.

(19) López-Bezanilla, A.; Triozon, F.; Roche, S. *Nano Lett.* **2009**, *9*, 2537.

(20) Grimme, S.; Parac, M. *ChemPhysChem* **2003**, *3*, 292.

(21) Dreuw, A.; Weisman, J. L.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 2943.

(22) Tozer, D. J. *J. Chem. Phys.* **2003**, *119*, 12697.

(23) Dreuw, A.; Head-Gordon, M. *J. Am. Chem. Soc.* **2004**, *126*, 4007.

(24) Day, P. N.; Nguyen, K. A.; Pachter, R. *J. Chem. Theory Comput.* **2008**, *4*, 1094.

(25) Lange, A. W.; Rohrdanz, M. A.; Herbert, J. M. *J. Phys. Chem. B* **2008**, *112*, 6304.

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

(27) Wong, B. M.; Cordaro, J. G. *J. Chem. Phys.* **2008**, *129*, 214703.

(28) Rohrdanz, M. A.; Herbert, J. M. *J. Chem. Phys.* **2009**, *130*, 054112.

(29) Stein, T.; Kronik, L.; Baer, R. *J. Am. Chem. Soc.* **2009**, *131*, 2818.

(30) Stein, T.; Kronik, L.; Baer, R. *J. Chem. Phys.* **2009**, *131*, 244119.

(31) Wong, B. M.; Piacenza, M.; Della Sala, F. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4498.

(32) Day, P. N.; Nguyen, K. A.; Pachter, R. *J. Chem. Theory Comput.* **2010**, *6*, 2809.

(33) Gill, P. M. W.; Adamson, R. D.; Pople, J. A. *Mol. Phys.* **1996**, *88*, 1005.

(34) Leininger, T.; Stoll, H.; Werner, H. J.; Savin, A. *Chem. Phys. Lett.* **1997**, *275*, 151.

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

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

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

(38) Toulouse, J.; Colonna, F.; Savin, A. *Phys. Rev. A* **2004**, *70*, 062505.

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

(40) Kamiya, M.; Sekino, H.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2005**, *122*, 234111.

(41) Sato, T.; Tsuneda, T.; Hirao, K. *Mol. Phys.* **2005**, *103*, 1151.

(42) Yanai, T.; Harrison, R. J.; Handy, N. C. *Mol. Phys.* **2005**, *103*, 413.

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

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

(45) Henderson, T. M.; Izmaylov, A. F.; Scuseria, G. E.; Savin, A. *J. Chem. Phys.* **2007**, *127*, 221103.

(46) Henderson, T. M.; Janesko, B. G.; Scuseria, G. E. *J. Chem. Phys.* **2008**, *128*, 194105.

(47) Tretiak, S.; Chernyak, V.; Mukamel, S. *Chem. Phys. Lett.* **1996**, *259*, 55.

(48) Mukamel, S.; Tretiak, S.; Wagersreiter, T.; Chernyak, V. *Science* **1997**, *277*, 781.

(49) Tretiak, S.; Igumenshchev, K.; Chernyak, V. *Phys. Rev. B* **2005**, *71*, 033201.

(50) Wong, B. M. *J. Phys. Chem. C* **2009**, *113*, 21921.

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

(52) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.

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

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

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

(56) Adamson, R. D.; Dombroski, J. P.; Gill, P. M. W. *J. Comput. Chem.* **1999**, *20*, 921.

(57) Jacquemin, D.; Perpète, E. A.; Scalmani, G.; Frisch, M. J.; Kobayashi, R.; Adamo, C. *J. Chem. Phys.* **2007**, *126*, 144105.

(58) Jacquemin, D.; Perpète, E. A.; Vydrov, O. A.; Scuseria, G. E.; Adamo, C. *J. Chem. Phys.* **2007**, *127*, 094102.

(59) Jacquemin, D.; Perpète, E. A.; Scuseria, G. E.; Ciofini, I.; Adamo, C. *J. Chem. Theory Comput.* **2008**, *4*, 123.

(60) Burke, K.; Ernzerhof, M.; Perdew, J. P. *Chem. Phys. Lett.* **1997**, *265*, 115.

(61) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040.

(62) Poater, J.; Duran, M.; Solà, M. *J. Comput. Chem.* **2001**, *22*, 1666.

(63) Güell, M.; Luis, J. M.; Rodríguez-Santiago, L.; Sodupe, M.; Solà, M. *J. Phys. Chem. A* **2009**, *113*, 1308.

(64) Kadantsev, E. S.; Stott, M. J.; Rubio, A. *J. Chem. Phys.* **2006**, *124*, 134901.

(65) 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.

(66) Hättig, C.; Weigend, F. *J. Chem. Phys.* **2000**, *113*, 5154.

(67) Platt, J. R. *J. Chem. Phys.* **1949**, *17*, 484.

(68) Handa, T. *B. Chem. Soc. Jpn.* **1963**, *36*, 235.

(69) Scher, H.; Montroll, E. W. *Phys. Rev. B* **1975**, *12*, 2455.

(70) Roos, B. O.; Andersson, K.; Fülscher, M. P. *Chem. Phys. Lett.* **1992**, *192*, 5.

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

(72) Hirao, K.; Nakano, H.; Nakayama, K. *J. Chem. Phys.* **1997**, *107*, 9966.

(73) Parr, R. G.; Yang, W. The Kohn-Sham method: basic principles. In *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989; pp 142−168.

(74) NIST Chemistry WebBook. NIST Standard Reference Database Number 69. http://webbook.nist.gov/chemistry (accessed Sep. 16, 2010).

(75) Parac, M.; Grimme, S. *Chem. Phys.* **2003**, *292*, 11.

(76) Zimmerman, P. M.; Zhang, Z.; Musgrave, C. B. *Nat. Chem.* **2010**, *2*, 648.

CT100529S

JCTC Journal of Chemical Theory and Computation

g_wham—A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates

Jochen S. Hub,*,† Bert L. de Groot,‡ and David van der Spoel†

Department of Cell and Molecular Biology, Uppsala University, Box 596, 75124 Uppsala, Sweden, and Computational Biomolecular Dynamics Group, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

Received August 30, 2010

Abstract: The Weighted Histogram Analysis Method (WHAM) is a standard technique used to compute potentials of mean force (PMFs) from a set of umbrella sampling simulations. Here, we present a new WHAM implementation, termed g_wham, which is distributed freely with the GROMACS molecular simulation suite. g_wham estimates statistical errors using the technique of bootstrap analysis. Three bootstrap methods are supported: (i) bootstrapping new trajectories based on the umbrella histograms, (ii) bootstrapping of complete histograms, and (iii) Bayesian bootstrapping of complete histograms, that is, bootstrapping via the assignment of random weights to the histograms. Because methods ii and iii consider only complete histograms as independent data points, these methods do not require the accurate calculation of autocorrelation times. We demonstrate that, given sufficient sampling, bootstrapping new trajectories allows for an accurate error estimate. In the presence of long autocorrelations, however, (Bayesian) bootstrapping of complete histograms yields a more reliable error estimate, whereas bootstrapping of new trajectories may underestimate the error. In addition, we emphasize that the incorporation of autocorrelations into WHAM reduces the bias from limited sampling, in particular, when computing periodic PMFs in inhomogeneous systems such as solvated lipid membranes or protein channels.

Introduction

The concept of potentials of mean force (PMFs), originally introduced by Kirkwood,[1] is frequently used to characterize the energetics of transitions in solid, fluid, and biomolecular systems. A routinely used technique to compute the PMF along a given reaction coordinate ξ is umbrella sampling. That technique aims to overcome limited sampling at energetically unfavorable configurations by restraining the simulation system with an additional (typically harmonic) potential.[2] Accordingly, a set of N_w separate umbrella simulations are carried out, with an umbrella potential

$$w_i(\xi) = K_i/2(\xi - \xi_i^c)^2 \quad (1)$$

which restrains the system at the position ξ_i^c ($i = 1, ..., N_w$) with a force constant K_i. From each of the N_w umbrella simulations (sometimes referred to as "umbrella windows"), an umbrella histogram $h_i(\xi)$ is recorded, representing the probability distribution $P_i^b(\xi)$ along the reaction coordinate biased by the umbrella potential $w_i(\xi)$. The probably most widely used technique to compute the PMF from histograms, that is, to unbias the distributions $P_i^b(\xi)$, is the weighted histogram analysis method (WHAM).[3]

On the basis of the histogram method of Ferrenberg and Swendsen,[4] the idea of WHAM is to estimate the statistical

* Author to whom correspondence should be addressed. Tel.: +46-(0)18-4715056. Fax: +46-(0)18-511755. E-mail: jochen@xray.bmc.uu.se.

† Uppsala University.

‡ Max-Planck-Institute for Biophysical Chemistry.

10.1021/ct100494z
Published on Web 11/16/2010

uncertainty of the unbiased probability distribution given the umbrella histograms, and subsequently to compute the PMF that corresponds to the smallest uncertainty. For a derivation of the equations, we refer to the original publication by Kumar et al.[3] An excellent (and less technical) review on umbrella simulations and the WHAM procedure has been presented by Roux.[5] The WHAM equations read[3]

$$P(\xi) = \frac{\sum_{i=1}^{N_w} g_i^{-1} h_i(\xi)}{\sum_{j=1}^{N_w} n_j g_j^{-1} \exp[-\beta(w_j(\xi) - f_j)]} \quad (2)$$

and

$$\exp(-\beta f_j) = \int \mathrm{d}\xi \exp[-\beta w_j(\xi)]\, P(\xi) \quad (3)$$

Here, β denotes the inverse temperature $1/k_B T$, with the Boltzmann constant k_B and the temperature T, and n_j is the total number of data points in histogram h_j. The statistical inefficiency g_i is given by $g_i = 1 + 2\tau_i$, with the integrated autocorrelation time τ_i of umbrella window i (in units of the simulation frame time step.) Note that the g_i's cancel from the WHAM equations if (and only if) the autocorrelation times in all umbrella windows equal. In contrast, if the g_i's differ between different histograms, the factors g_i^{-1} assign lower weights to histograms with longer autocorrelations. $P(\xi)$ denotes the unbiased probability distribution that is related to the PMF via $\mathscr{W}(\xi) = -\beta^{-1} \ln[P(\xi)/P(\xi_0)]$. Here, ξ_0 is an arbitrary reference point where the PMF $\mathscr{W}(\xi_0)$ is defined to zero. The WHAM equations contain two unknown quantities, that is, the free energy constants f_j and the unbiased distribution $P(\xi)$, and must therefore be solved iteratively. Depending on the number of histograms and the height of the barriers in the PMF, the WHAM equations typically converge within tens of iterations and up to tens of thousands of iterations.

Alternative approaches to derive the PMF and the uncertainty from a set of umbrella simulations have been proposed,[6−8] as well as several extensions to the umbrella sampling technique.[9,10]

Despite the fact that WHAM has been widely used to derive PMFs from biomolecular simulations, a standard protocol to compute the statistical errors for the derived PMF has not yet evolved. Therefore, we here present a new WHAM implementation, termed g_wham, that allows one to compute robust error estimates using different bootstrap techniques. We apply the techniques on two test systems to demonstrate the potential and the limitations of the bootstrap methods. Besides the ability to estimate the statistical error, g_wham supports a number of features that are expected to be useful to the community. To compute PMFs along periodic reaction coordinates such as dihedral angles or coordinates in a simulation box with periodic boundary conditions, a periodic WHAM is implemented. Nonharmonic umbrella potentials can be provided as tabulated potentials. g_wham allows for the estimatation of autocorrelation times and the incorporation of these into WHAM. As shown in the Results, this procedure may yield more realistic PMF estimates in the presence of long autocorrelations.

The software is freely distributed with the GROMACS simulation suite.[11] If the umbrella simulations were carried out using the GROMACS pull options, g_wham conveniently reads the GROMACS output files. In the case of more complex reaction coordinates, or if the simulations were not carried out using GROMACS, the user may provide g_wham input files in text format. A detailed description of g_wham, including all options, is provided in the Appendix and is available with the command line g_wham -h.

Methods

Error Estimates from Bootstrap Analysis. g_wham estimates the statistical uncertainty of the PMF using bootstrap analysis.[12] Bootstrapping is a resampling technique that can be applied to estimate the uncertainty of a quantity $A(a_1, ..., a_n)$ which is computed from a large set of n observations a_l ($l = 1, ..., n$). To calculate the uncertainty in A, one could redo the n observations multiple times, yielding several independent estimates for A and hence the uncertainty in A. That procedure would require many more observations and is therefore often not tractable.

The observations a_l are typically drawn from an unknown underlying probability distribution $P(a)$. The idea of bootstrapping is to estimate $P(a)$ using the n observations and subsequently generate new random sets of n hypothetical observations, based on the estimated distribution. Each of the sets of n hypothetical observations is used to calculate a hypothetical value for A. The uncertainty in A is then given by the standard deviation of the hypothetical values for A. For a detailed introduction into the bootstrap technique, we refer to the monograph by Chernick.[13]

Bootstrapping Trajectories Based on Umbrella Histograms. The WHAM procedure computes the PMF based on the N_w trajectories $\xi_i(t)$ along the reaction coordinate, each taken from one of the umbrella windows ($i = 1, ..., N_w$). All positions ξ_i during the N_w simulations may thus be considered as the large set of observations, which we referred to as a_l in the previous paragraph.[14] Alternatively, complete umbrella histograms may be considered as the individual observations (see next section).[15] Note that the probability distributions of ξ_i are already available as the umbrella histograms. Thus, we can generate new hypothetical observations, that is, a "bootstrapped" trajectory $\xi_{b,i}(t)$ for each umbrella histogram $h_i(\xi)$, such that $\xi_{b,i}(t)$ is distributed according to the respective histogram. Each bootstrapped trajectory $\xi_{b,i}(t)$ yields a new histogram $h_{b,i}(\xi)$. The new set of N_w histograms $h_{b,i}$ is subsequently applied in WHAM to compute a bootstrapped PMF $\mathscr{W}_b(\xi)$. The whole procedure is repeated N_b times (e.g., $N_b = 200$), yielding a large set of N_b bootstrapped PMFs $\mathscr{W}_{b,k}(\xi)$ ($k = 1, ..., N_b$). The uncertainty of the PMF is then given by the standard deviation as calculated by the N_b bootstrapped PMFs, that is via

$$\sigma_{\mathrm{PMF}}(\xi) = [(N_b - 1)^{-1} \sum_{k=1}^{N_b} (\mathscr{W}_{b,k}(\xi) - \langle \mathscr{W}_b(\xi) \rangle)^2]^{1/2} \quad (4)$$

Here, $\langle \mathscr{W}_b(\xi)\rangle = N_b^{-1}\sum_{i=k}^{N_b} \mathscr{W}_{b,k}(\xi)$ denotes the average of the bootstrapped PMFs at position ξ. One could also calculate the uncertainty via the standard deviation of the respective probabilities $\propto \exp(-\beta \mathscr{W}_{b,k}(\xi))$, which could subsequently be translated into the uncertainty of the PMF. We found that that this procedure yields similar error estimates compared to the definition in eq 4 applied here.

Any property generated from MD simulations has a natural time correlation. In order for the bootstrapping procedure to generate correct error estimates, that autocorrelation must be taken into account explicitly. Here, we chose the following procedure to generate autocorrelated bootstrapped trajectories $\xi_b(t)$ with a given integrated autocorrelation time (IACT) τ, and distributed according to a histogram $h(\xi)$. (Here, $h(\xi)$ may denote any of the given histograms, and the procedure is repeated for each histogram.) First, given a normally distributed random variable of zero mean and unit variance $R_t \sim \mathscr{N}(0,1)$, we generate a time series $x(t)$ via

$$x(0) = R_0 \tag{5}$$

$$x(t+1) = ax(t) + \sqrt{1-a^2}R_{t+1} \tag{6}$$

where $a = \exp(-1/\tau)$. Then, $x(t) \sim \mathscr{N}(0,1)$ and the IACT of $x(t)$ equals τ. The normally distributed $x(t)$ is translated into an evenly distributed series on [0,1) using the error function via $x'(t) = (1 + \mathrm{erf}[x(t)\sqrt{2}])/2$. Eventually, we solve the equation

$$x'(t) = C_h(\xi_b(t)) \equiv \int_{-\infty}^{\xi_b(t)} h(\xi')\,\mathrm{d}\xi' \tag{7}$$

for $\xi_b(t)$, where $C_h(\xi_b(t))$ denotes the cumulative distribution function of the (normalized) histogram. Then, $\xi_b(t)$ will be distributed according to $h(\xi)$, with an approximate IACT of τ.

Bootstrapping Complete Histograms. The conformational sampling of macromolecules during MD simulations is frequently affected by long autocorrelations, with autocorrelation times ranging from pico- to microseconds or even longer. A complete sampling of all coordinates perpendicular to the reaction coordinate is therefore often intractable, in particular during a typically short umbrella simulation. In such situations, the individual umbrella histograms do not represent all accessible areas of phase space. Bootstrapped trajectories based on such nonconverged histograms, following the procedure in the previous paragraph, would also not represent all accessible areas of phase space. In addition, note that bootstrapping trajectories from given histograms require at least approximate knowledge of the IACT. Given only incomplete sampling, however, the IACT may be severely underestimated because slow transitions may not occur during the short umbrella simulations. Bootstrapping trajectories based on incomplete histograms in combination with underestimated IACTs would severely underestimate the uncertainty.

If the simulations are affected by such long autocorrelations, we suggest carrying out the simulation of each umbrella window multiple times from independent initial frames. Then, we consider complete histograms as individual observations and randomly select a new set of N_w histograms from the given set of N_w histograms, allowing one to multiply select a specific histogram (sampling with replacement).[15] Hence, in contrast to the bootstrapping of trajectories based on umbrella histograms (see previous paragraph), we do *not* generate new trajectories and histograms. To ensure that the bootstrapped histograms span the whole reaction coordinate, that is, that no gaps between the bootstrapped histograms are generated, the histograms can be grouped along the reaction coordinate, and histograms can be bootstrapped within each group separately. We show that, given limited sampling, bootstrapping of complete histograms allows for a more accurate estimation of the uncertainty (see Results).

Bayesian Bootstrapping of Complete Histograms. As pointed out in the previous paragraph, introducing groups of histograms (and subsequent bootstrapping only within each group) avoids gaps along the reaction coordinate between bootstrapped histograms, but an appropriate choice for the number of histograms per group may be unclear. Therefore, we propose a method related to the so-called *Bayesian bootstrap* that avoids the introduction of groups of histograms by instead assigning random weights to all histograms within each bootstrap.

When applying the usual bootstrap on individual observations, n observations are selected with replacement from the given n observations a_i ($i = 1, ..., n$), where the probability of selecting any of the specific observations equals $1/n$. Hence, all observations a_i are selected with equal probability. Rubin proposed an alternative procedure, known as the Bayesian bootstrap, that instead assigns random weights ω_i to each observation.[16] Then, each observation a_i is selected with probability ω_i (instead of $1/n$), or alternatively, the weights ω_i are assigned to the observations when computing the observable $A(a_1, ..., a_n)$ from the observations. According to the Bayesian bootstrap, the weights ω_i are generated as follows: draw $n-1$ uniform random variables between 0 and 1, and let $u_{(1)}, u_{(2)}, ..., u_{(n-1)}$ denote their values in increasing order. In addition, let $u_{(0)} = 0$ and $u_{(n)} = 1$. The random weights are then given by the gaps between two consecutive random numbers, i.e., $\omega_i = u_{(i)} - u_{(i-1)}$, where $i = 1, ..., n$. For each bootstrap turn, new random weights are generated.

Note that the bootstrapping of complete histograms (compare previous section) is equivalent to the assignment of random weights to the histograms, if these random weights are an integer multiple of $1/N_w$. Here, we suggest the assignment of continuous random weights to the histograms, and selection of the weights according to the Bayesian bootstrap. That procedure resembles the bootstrapping of compete histograms in the sense that it considers only complete histograms as independent data points and thus is expected to yield realistic error estimates in the presence of long autocorrelations. However, because the continuous weights ω_i are (almost) never exactly zero, it excludes the possibility of generating gaps along the reaction coordinate in the bootstrapped histogram set. The WHAM procedure with weighted histograms was implemented by multiplying the inverse statistical inefficiencies g_i^{-1} in eq 2 by ω_i.

Autocorrelations. The normalized autocorrelation function of umbrella window i is given by

$$R_i(\Delta t) = \frac{\langle(\xi_i(t) - \langle\xi_i\rangle)(\xi_i(t + \Delta t) - \langle\xi_i\rangle)\rangle}{\sigma^2_{\xi,i}} \quad (8)$$

where $\xi_i(t)$ denotes the reaction coordinate during simulation i, $\sigma^2_{\xi,i} = \langle(\xi_i(t) - \langle\xi_i\rangle)^2\rangle$ is the respective variance, and $\langle ... \rangle$ represents the average over the simulation frames. Following the nomenclature in Kumar et al.,[3] the integrated autocorrelation time (IACT) of window i is defined by

$$\tau_{i,\mathrm{int}} = \sum_{\Delta t=1}^{\infty} R_i(\Delta t) \quad (9)$$

The autocorrelation function derived from a short umbrella simulation is typically very noisy. Sophisticated methods to compute the IACT such as fitting of a single or double exponential to $R_i(\Delta t)$ or any kind of binning analysis turned out to be too unstable for the present purpose. Note that the IACT should be computed automatically for hundreds (or thousands) of possibly poorly converged $R_i(\Delta t)$'s. Therefore, we chose to compute $\tau_{i,\mathrm{int}}$ directly via eq 9 but carried out the summation only until $R_i(\Delta t)$ dropped under a predefined threshold of 0.05.

Simulation Details. Test simulations were carried out using the GROMACS simulation suite.[11] As a test system, we have computed the PMF along the distance between two methanol molecules in a vacuum. These simulations were set up by placing one methanol molecule in the origin and placing the second molecule at the distance ξ_i^c of the corresponding umbrella window. The molecules were randomly rotationally oriented. In addition, the initial distance between the two molecules was varied randomly by $\pm\sigma_u$, where $\sigma_u = \sqrt{(k_B T/K)}$ denotes the width of the umbrella histogram (assuming a flat underlying PMF). Here, $K = 800$ kJ/mol/nm^2 is the umbrella force constant, k_B is the Boltzmann factor, and T is the temperature. The sampling was carried out using a stochastic dynamics integrator ($\tau = 0.07$ ps, $T = 300$ K), with an independent random seed for each simulation. Lennard-Jones and electrostatic interactions were computed in direct space without a cutoff. Bonds were constrained using LINCS,[17] allowing a time step of 2 fs. The umbrella positions were recorded every 10th step during simulations with a total simulation time of 50 ps and were recorded at every step during simulations with a simulation time of 4 ps. Methanol parameters were taken from the GROMOS96 force field.[18] The methods applied to compute the PMF for the Rhesus channel Rh50 and for the lipid membrane have been published elsewhere.[19]

Results and Discussion

Error Analysis for a Model System and a Lipid Membrane PMF. As a test system, we compute the PMF of the center-of-mass distance between two methanol molecules in a vacuum. Such a simple system allows us to carry out the complete set of umbrella simulations many times and hence to accurately compute the "true" statistical uncertainty of the PMF. Subsequently, we test whether the bootstrapping procedures are able to estimate the "true" uncertainty using only the data from one set of umbrella simulations.

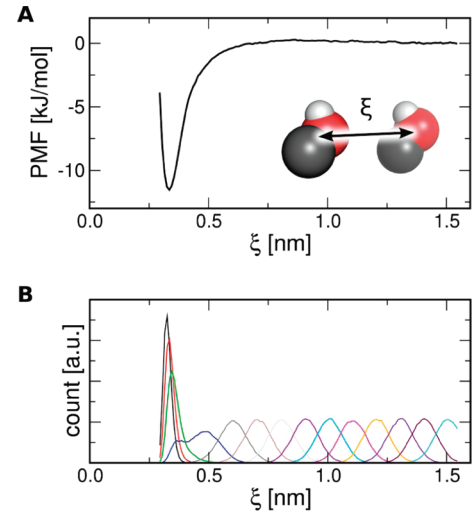


Figure 1. (A) Converged PMF (black curve) of the center of mass distance between two methanol molecules in vacuum. (B) Converged umbrella histograms, each derived from a 3-ns simulation.

As a reference for the following discussion, Figure 1A presents the converged PMF and Figure 1B, the respective umbrella histograms. Here, each of the 14 umbrella windows was simulated for 3 ns, yielding well-converged statistics as visible from the Gaussian histograms at a great distance ξ. The PMF at $\xi = 1.5$ nm was chosen as a reference point and defined to zero. To arrive at a flat PMF at a great distance, the PMF was corrected by $k_B T \ln(4\pi\xi^2)$, which removes the entropic decrease in the PMF because of the increase in the number of configurations on a sphere of radius ξ.[20]

To assess whether the bootstrapping procedure provides a reliable error estimate, we have repeatedly computed the same PMF using limited statistics, with each umbrella window simulated for 50 ps. The nonconverged histograms of one set of these umbrella simulations is shown in Figure 2A. The complete set of umbrella simulations was carried out 50 times with different initial random seeds for the stochastic forces and different initial orientations and velocities of the methanol molecules, yielding 50 independent estimates for the PMF (Figure 2B). The uncertainty (67% confidence interval) for a single set of umbrella simulations as derived from these 50 PMFs is plotted as a green curve in Figure 2D. Note that the error at $z = 1.5$ nm equals zero since all PMFs were defined to zero at that point. Next, an autocorrelated bootstrapped trajectory was generated for each of the histograms plotted in Figure 2A using eqs 5, 6, and 7, allowing one to compute a new "hypothetical" estimate for the PMF based on the umbrella histograms. That bootstrapping procedure was repeated 200 times, yielding 200 bootstrapped PMFs (Figure 2C, colored curves). As expected, the bootstrapped PMFs substantially differ, in line with the 50 PMFs calculated from independent simulations (Figure 2B). The standard deviation computed from the bootstrapped PMFs (Figure 2D, black) is in good agreement with the uncertainty calculated from the 50 independent

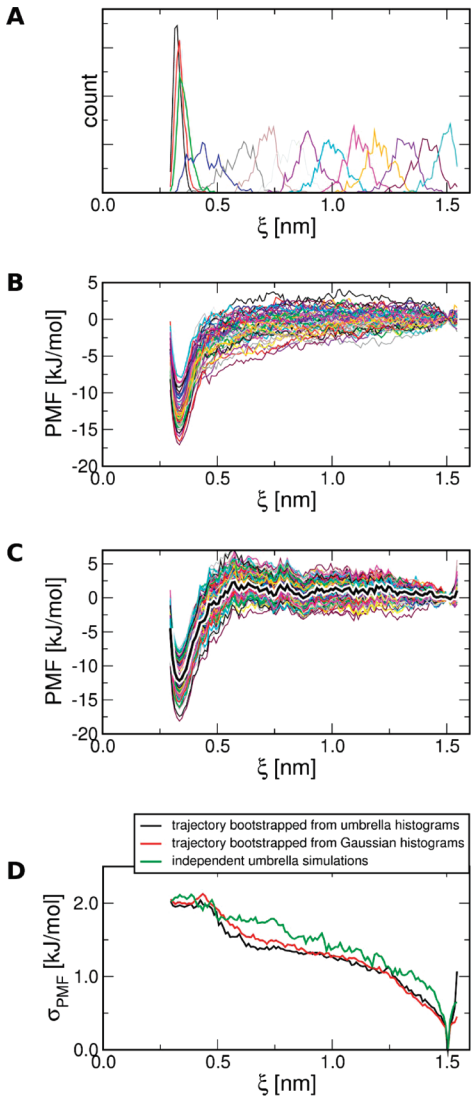


Figure 2. (A) Nonconverged histograms, each derived from 50 ps simulations. (B) 50 PMFs derived from 50 fully independent sets of umbrella simulations. (C) PMF (black curve) derived from the set of nonconverged histograms (A). Autocorrelated trajectories were bootstrapped from the histograms shown in A 200 times, yielding 200 bootstrapped PMFs (colored curves in C). (D) Statistical uncertainty calculated from the 50 independent simulations (green) shown in B and from the 200 bootstrapped PMFs (black) shown in C. Alternatively, the uncertainty was estimated from trajectories that were bootstrapped from Gaussian distributions of the average and σ taken from the umbrella histograms (red).

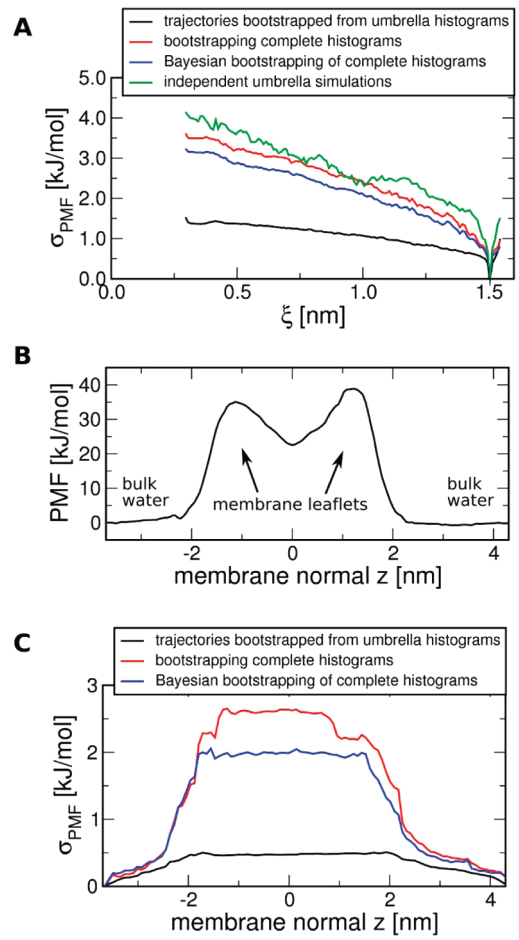


Figure 3. Estimating uncertainties in the presence long autocorrelations. (A) The PMF along the methanol−methanol distance (not shown) was computed from 140 umbrella histograms, each derived from a 4 ps simulation. As a reference, the uncertainty σ_{PMF} was computed from 100 independent sets of umbrella simulations (green curve). Generating bootstrapped trajectories for each umbrella histogram leads to an underestimated uncertainty (black curve). Estimating the uncertainty by bootstrapping complete histograms (red curve) or using the Bayesian bootstrap on complete histograms (blue curve) yields more accurate error estimates. (B) PMF for ammonia permeation across a lipid membrane containing 40 mol % cholesterol. (C) Statistical uncertainty of the ammonia PMF computed by bootstrapping trajectories for each umbrella histogram (black curve) and by (Bayesian) bootstrapping of complete histograms (red and blue curves).

simulations (Figure 2D, green), demonstrating that the bootstrapping procedure provides a reliable error estimate without the requirement to carry out new independent simulations. Alternatively, the uncertainty was estimated from trajectories that were bootstrapped from Gaussian distributions with the average and width taken from the respective umbrella histogram (Figure 2D, red), yielding almost identical and hence equally accurate error estimates.

Biomolecular simulations naturally contain long autocorrelations. The histograms based on short umbrella simulations may therefore not represent all parts of phase space. In addition, the IACTs may be severely underestimated since slow transitions do not occur during the short simulations. Consequently, bootstrapping trajectories based on these histograms (in combination with underestimated IACTs) will underestimate the uncertainty. This fact is demonstrated in Figure 3A. To emulate umbrella sampling of a biomolecular system with long autocorrelations, we computed the PMF of the methanol distance based on 4 ps simulations (using the first 0.5 ps for equilibration), resulting in highly nonconverged histograms. Ten independent umbrella simulations were carried out for each of the 14 umbrella window positions, yielding 140 histograms. The whole set of umbrella simulations was carried out 100 times, allowing one to compute the true uncertainty (as one standard deviation) in the PMF (Figure 3A, green curve). Figure 3A compares the

true uncertainty to the estimated uncertainty derived from three different bootstrapping methods. Because the estimated uncertainties vary slightly between the different sets of independent umbrella simulations, Figure 3A plots estimated uncertainties averaged from 15 (of the 100) sets of umbrella simulations. The uncertainty computed by bootstrapping trajectories is shown as a black curve, demonstrating that this procedure greatly underestimates the uncertainty in that case. The red curve in Figure 3A presents the uncertainty estimated by bootstrapping complete histograms. Here, the histograms were grouped into 14 sets of 10 histograms, with each group containing the 10 histograms at the same umbrella position. Consequently, 10 histograms were bootstrapped from each of the 14 sets, and the PMF was computed from the 140 bootstrapped histograms using WHAM. The whole procedure was repeated 200 times, providing 200 hypothetical estimates for the PMF (not shown) and allowing one to compute the uncertainty using eq 4. As visible from Figure 3A, bootstrapping complete histograms yields a more accurate estimate of the uncertainty, despite the poor sampling within each umbrella window. The blue curve in Figure 3A presents the uncertainty estimated using Bayesian bootstrapping of complete histograms, that is, by assigning random weights to the individual histograms (see Methods). The Bayesian bootstrap also yields a reasonable error estimate because the method considers only complete histograms as independent data points, similar to the bootstrapping of complete histograms.

For a second comparison between the different bootstrapping methods, Figure 3B presents the PMF for ammonia permeation across a biological membrane composed of the lipids POPE and POPC plus 40 mol % cholesterol. The flat regions at small and large z correspond to the ammonia molecule in the two bulk water regions above and below the membrane, whereas the two maxima in the PMF correspond to the hydrophobic regions of the two membrane leaflets. The PMF has been computed from 656 histograms (not shown), each taken from 1 ns of simulation, where the first 50 ps were removed for equilibration. The initial frames for the umbrella simulations at a specific z coordinate were generated by inserting ammonia at various randomly chosen positions in the membrane plane, justifying the assumption that the histograms are independent. Figure 3C shows the estimated uncertainty computed via (i) bootstrapped trajectories (black), (ii) bootstrapping of complete histograms with 12 histograms within each group (red), and (iii) Bayesian bootstrapping of complete histograms (blue). Method i yields a very small uncertainty of only 0.5 kJ/mol, whereas methods ii and iii yield an uncertainty of ~2 kJ/mol at the main barriers in the PMF. Because considerable computational effort is required to compute the PMF in Figure 3B, we cannot compute the uncertainty from independent sets of umbrella simulations for this example. However, Figure 3C suggests that the individual histograms do not represent all accessible areas of phase space, leading to an underestimated uncertainty as computed from method i. Presumably, slow transitions on a multi-nanosecond time scale may affect the sampling in this case, whereas the autocorrelation analysis based on the shorter simulations yields spuriously short IACTs. In contrast to method i, methods ii and iii do not depend on the accurate computation of the IACTs but only require the histograms to be independent. Therefore, methods ii and iii are expected to yield a reliable error estimate in this case.

To estimate uncertainties in the presence of long (possibly unknown) autocorrelations, we therefore suggest carrying out many short umbrella simulations instead of a few long umbrella simulations, such that each position along the reaction coordinate is covered by at least several independent histograms. Given sufficiently many independent histograms, the error can be estimated using bootstrapping of complete histograms or using the Bayesian bootstrap of complete histograms.

Effect of Autocorrelations. As visible from the WHAM equations, eqs 2 and 3, the IACTs cancel if (and only if) the IACTs are equal in all umbrella windows. In nonhomogeneous systems, however, that assumption may not hold. An example would be umbrella simulations for solute permeation across a lipid membrane or across a protein channel surrounded by bulk water. Here, the IACTs of windows in the bulk are typically lower than the IACTs of windows inside the lipid membrane or inside the protein channel. We found that neglecting the IACTs may lead to artifacts in particular when computing the PMF along a periodic reaction coordinate. As an example, Figure 4A presents a nonconverged PMF for ammonia permeation across the Rhesus protein channel Rh50 from *N. europaea* (Figure 4C). The PMF was derived from 365 400-ps histograms, taken from 500 ps simulations, using the first 100 ps for equilibration. The simulations were carried out with periodic boundary conditions, implying that a PMF for solute permeation should yield the same free energy in the two bulk-water regions below and above the channel. The black curve in Figure 4A was computed by a nonperiodic WHAM. The PMF is not converged, as apparent from the substantial offset of ~15 kJ/mol between the two bulk-water regions. To account for the periodicity of the system, a periodic WHAM assuming equal IACTs of all umbrella windows could be carried out (red curve). However, with equal IACTs, the WHAM procedure assigns equal weights to all histograms and, hence, equally distributes the offset of 15 kJ/mol along the reaction coordinate to enforce a periodic PMF. As a consequence, an unphysical slope is induced in the bulk-water regions of the PMF ($|z| > 2$ nm). A more realistic procedure is therefore to compute the IACTs for each umbrella window and to apply them within WHAM. The IACT derived by direct integration of the autocorrelation function for the displacement for each umbrella window is plotted in Figure 4B as black dots. Because the IACTs cannot be accurately computed from the limited sampling in the umbrella windows, we suggest smoothing the IACT along the reaction coordinate yielding a semiquantitative autocorrelation measure (Figure 4B, red curve). Whereas the IACTs are small in bulk water, substantial autocorrelations limit the sampling within the channel, suggesting that the 15 kJ/mol is a consequence of slow sampling within the channel. The PMF computed by a periodic WHAM that takes IACTs into account is shown in Figure 4A as a blue curve. As expected, the PMF is flat in the bulk-water regions (in agreement with the nonperiodic WHAM result, black curve), whereas corrections

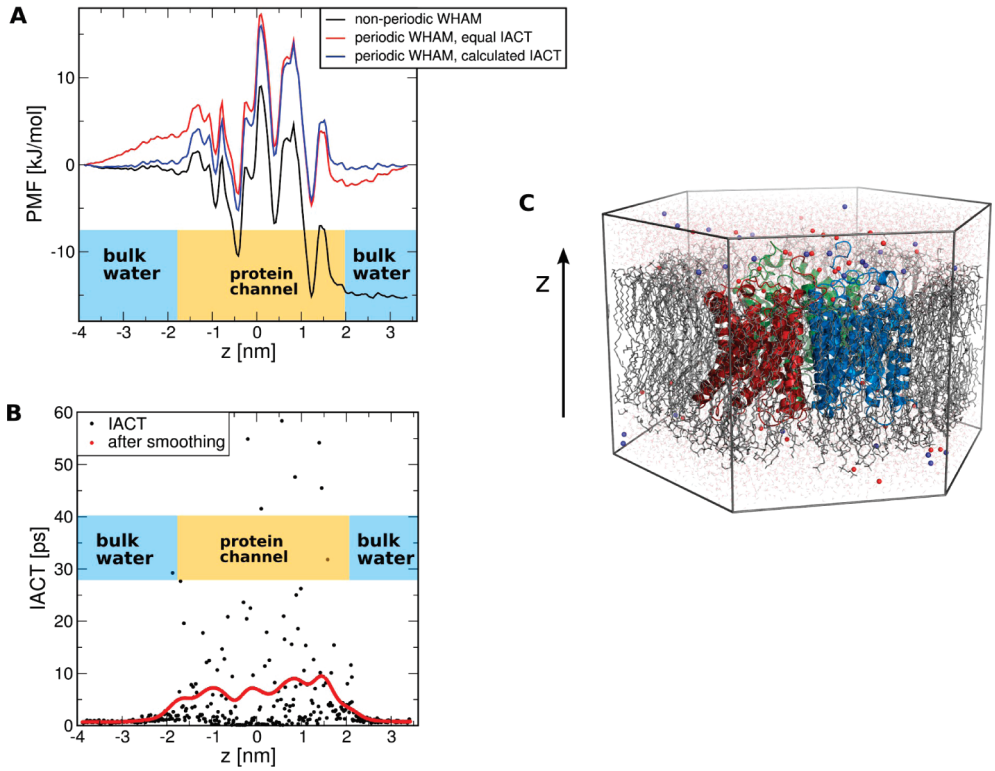


Figure 4. Effect of autocorrelations in a periodic WHAM. (A) Nonconverged PMF of ammonia permeation across the Rhesus protein channel Rh50 (black). The limited sampling accounts for a substantial offset of ~15 kJ/mol between the two end points of the PMF corresponding to the two bulk water regions. A periodic WHAM assuming equal integrated autocorrelation times (IACTs) accounts for the periodicity of the system (red curve) but induces approximately a linear slope in the complete PMF, including the well-sampled bulk water regions. Blue curve: PMF derived from periodic WHAM incorporating the calculated IACTs. The PMFs in the bulk-water regions are almost flat, in accordance with the bulk-water regions in the nonperiodic PMF (black). (B) IACTs calculated by direct integration of the autocorrelation functions (black dots), and by subsequent smoothing with a Gaussian filter (red curve). (C) Simulation box of an Rh50 trimer embedded in a lipid membrane and solvated in water and 150 mM electrolyte.

were introduced in the less sampled channel region to yield a periodic PMF.

Converged PMFs for ammonia permeation across the Rh50 channel as well as the biological implications have been published elsewhere.[19]

Conclusions

We have presented a new WHAM implementation, termed g_wham, that is freely distributed with the GROMACS simulation suite. The g_wham software is easy to use, flexible, and efficiently implemented. Statistical uncertainties are quantified using different bootstrap analysis methods: (i) bootstrapping of hypothetical trajectories based on the umbrella histograms together with the respective autocorrelation time, (ii) by bootstrapping complete histograms, or (iii) by using the Bayesian bootstrap of complete histograms, that is, by assigning random weights to the histograms. We have shown that method i provides an accurate error estimate if (and only if) the histograms are sufficiently converged. If the histograms are affected by long autocorrelations, as frequently occurrs in simulations of large biomolecules, methods ii and iii provide a more accurate error estimate. In nonhomogeneous systems such as a protein channel or a lipid membrane surrounded by bulk water, the autocorrelation times may substantially vary along the reaction coordinate and thus not cancel from the WHAM equations. Consistent application of the autocorrelations has here been shown to yield a more accurate estimate for the PMF in such systems, in particular when computing a periodic PMF.

Acknowledgment. This study was supported by a Marie Curie Intra-European Fellowship within the 7th European Community Framework Programme, by the Max-Planck-Society, and by the Deutsche Forschungsgemeinschaft (SFB:803).

Appendix

g_wham Input Modes. A help file, including all command line options, is provided by the g_wham tool via the command g_wham -h. g_wham supports three input modes. In modes 1 and 2, g_wham reads specific GROMACS files, These modes are thus convenient for GROMACS users. In mode 3, g_wham reads only text files and is therefore suitable for non-GROMACS users as well.

1. With option -it, the user provides a file which contains the *file names* of the umbrella simulation run-input files (GROMACS tpr-files). In addition, with option -ix, the user provides a file which contains the file names of the pull position output files (pullx.xvg etc.) written by the GROMACS mdrun program.

2. This mode is the same as mode 1, except that the user provides with option -if a file which contains the file names of the pull force output files (pullf.xvg etc.) written by the GROMACS mdrun program.
3. With option -ip, the user provides a file which contains the file names of the pull output files written by GROMACS 3 (pdo files). pdo files are text files and can be generated by non-GROMACS users. Each pdo file contains a header with the umbrella positions and force constants, and the body contains the simulation time versus the displacement of the system with respect to the umbrella center. The pdo file format (with a typical header) is explained with the g_wham help file provided with g_wham -h.

WHAM Options. Default values for the following options are listed in square brackets:

-min, -max: boundaries of the profile [0,0]

-auto: determine boundaries automatically [yes]

-bins: number of bins used [200]

-temp: temperature in Kelvin [298.15]

-tol: tolerance. The WHAM iterations stop when the probabilities change less than the tolerance. [10^{-6}]

-b, -e, -dt: specify simulation times in picoseconds (begin, end, time step) that are used in WHAM [50, infinity, 0]

-cycl: periodic (or cyclic) WHAM [no]

-tab: file name with tabulated potential in the case of nonharmonic umbrella potentials

Output Control

-o: file name of PMF output file

-hist: file name of histogram output file

-histonly: write histograms and exit [no]

-boundsonly: determine boundaries automatically and exit [no]

-log: write negative logarithm of the probabilities; that is, enable output in energy units; otherwise, write probabilities [yes]

-unit: define energy unit (kJ/mol, kcal/mol, k_BT) [kJ/mol]

-zprof0: set profile to zero at this position [0]

-sym: symmetrize profile around $\xi = 0$ (useful for membranes, for instance) [no]

-v: verbose mode [no]

Autocorrelation Handling

-ac: calculate integrated autocorrelation times (IACTs) using eqs 8 and 9 and use in WHAM [no]

-acsig: smooth IACTs along reaction coordinate using a Gaussian filter of width defined here [0]

-ac-trestart: when computing the autocorrelation functions for $\xi_i(t)$, restart the calculation after the time delay defined here [1 ps]

-oiact: (smoothed) IACT output file name

-iiact: IACT input file name. If the user prefers to calculate the IACTs *not* using g_wham, the IACTs can be provided to g_wham using this option.

Bootstrapping Control

-bsprof: output file name of all bootstrapped profiles

-bsres: output file name with average and standard deviation of bootstrapped profiles (that is, the uncertainty of the PMF)

-nBootstrap: number of bootstraps carried out to estimate the uncertainty (use, e.g., 100) [0]

-bs-method: bootstrap method applied ('b-hist', 'hist', 'traj', or 'traj-gauss'); Bayesian bootstrapping of complete histograms, bootstrap complete histograms, bootstrap new trajectories from the umbrella histograms, or bootstrap new trajectories from Gaussian distributions with average and width taken from the respective histogram [b-hist]

-bs-tau: specify integrated autocorrelation time used for all histograms with bootstrap methods 'traj' or 'traj-gauss'; if not provided (default), use calculated IACTs (options -ac and -acsig)

-histbs-block: number of histograms in one group with bootstrap method 'hist'; histograms will be bootstrapped only within each group separately; that procedure avoids gaps without any histogram data along the reaction coordinate.

-bs-seed: random seed for bootstrapping (−1 generates a seed) [−1]

-vbs: verbose bootstrapping (output cumulative distribution functions for each histogram and a histogram file for each bootstrapped PMF) [no].

References

(1) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300–313.

(2) Torrie, G. M.; Valleau, J. P. *Chem. Phys. Lett.* **1974**, *28*, 578–581.

(3) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(4) Ferrenber g, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195–1198.

(5) Roux, B. *Comput. Phys. Commun.* **1995**, *91*, 275–282.

(6) Kästner, J.; Thiel, W. *J. Chem. Phys.* **2005**, *123*, 144104.

(7) Kästner, J.; Thiel, W. *J. Chem. Phys.* **2006**, *124*, 234106.

(8) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105.

(9) Bartels, C.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 865–880.

(10) Souaille, M.; Roux, B. *Comput. Phys. Commun.* **2001**, *135*, 40–57.

(11) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(12) Efron, B. *Ann. Stat.* **1979**, *7*, 1–26.

(13) Chernick, M. R. *Bootstrap Methods: A Guide for Practitioners and Researchers*, 2nd ed.; Wiley-Interscience: New York, 2007.

(14) Grossfield, A. An implementation of WHAM: the weighted histogram analysis method. http://membrane.urmc.rochester.edu/Software/WHAM/WHAM.html (accessed October 6, 2010).

(15) Hub, J. S.; de Groot, B. L. *Biophys. J.* **2006**, *91* (3), 842–848.

(16) Rubin, D. B. *Ann. Stat.* **1981**, *9*, 130–134.

(17) Hess, B. *J. Chem. Theory Comput.* **2008**, *4*, 116–122.

(18) Van Gunsteren, W. F.; Berendsen, H. J. C. *Gromos Manual*; BIOMOS, Biomolecular Software, Laboratory of Physical Chemistry, University of Groningen: Groningen, The Netherlands, 1987.

(19) Hub, J. S.; Winkler, F. K.; Merrick, M.; de Groot, B. L. *J. Am. Chem. Soc.* **2010**, *132*, 13251–13263.

(20) Neumann, R. *Am. J. Phys.* **1980**, *48*, 354–357.

CT100494Z

JCTC Journal of Chemical Theory and Computation

Excited States in Solution through Polarizable Embedding

Jógvan Magnus Olsen,[†] Kęstutis Aidas,[‡] and Jacob Kongsted*,[†]

Department of Physics and Chemistry, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark, and Department of Chemistry, H. C. Ørsted Institute, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark

Received July 8, 2010

Abstract: We present theory and implementation of an advanced quantum mechanics/molecular mechanics (QM/MM) approach using a fully self-consistent polarizable embedding (PE) scheme. It is a polarizable layered model designed for effective yet accurate inclusion of an anisotropic medium in a quantum mechanical calculation. The polarizable embedding potential is described by an atomistic representation including terms up to localized octupoles and anisotropic polarizabilities. It is generally applicable to any quantum chemical description but is here implemented for the case of Kohn−Sham density functional theory which we denote the PE-DFT method. It has been implemented in combination with time-dependent quantum mechanical linear and nonlinear response techniques, thus allowing for assessment of electronic excitation processes and dynamic ground- and excited-state molecular properties using a nonequilibrium formulation of the environmental response. In our formulation of polarizable embedding we explicitly take into account the full self-consistent many-body environmental response from both ground and excited states. The PE-DFT method can be applied to any molecular system, e.g., proteins, nanoparticles and solute−solvent systems. Here, we present numerical examples of solvent shifts and excited-state properties related to a set of organic molecules in aqueous solution.

1. Introduction

Accurate modeling of excited states and molecular properties of large molecular samples represents one of the greatest challenges to modern quantum chemistry. The description of excited states requires the use of quantum mechanics. However, in many cases it is not necessary to use a full quantum mechanical description of the total system. This is, for example, the case when dealing with a solute−solvent system or in more general terms a molecule subjected to a structured environment. In these cases the part of the system not directly involved in the electronic processes can be described effectively using, e.g., classical mechanics. Even though linear scaling techniques are becoming more advanced and may be used to describe larger and more complex systems,[1] effects due to conformational sampling still persist and may become more important as the size of the molecular system is increased. In fact, in many cases it is mandatory to include effects of nuclear dynamics in combination with the electronic structure in order to pursue a direct comparison with experimental data. Thereby, formulation of accurate effective Hamiltonian methods becomes of crucial importance. This should particularly be seen in the light of recent trends aiming at a complete quantitative description of biological functions with the necessary step of bringing quantum chemistry into the life sciences.[2]

With the aim of describing large molecular systems we present in this paper a focused model based on the quantum mechanics/molecular mechanics (QM/MM) approach using a fully self-consistent polarizable embedding scheme which we denote the polarizable embedding (PE) model. The electrostatic embedding potential, i.e., the permanent charge distribution of the environment, is represented by a multi-

* Corresponding author e-mail: kongsted@ifk.sdu.dk.
[†] University of Southern Denmark.
[‡] University of Copenhagen.

10.1021/ct1003803
Published on Web 10/26/2010

center multipole expansion. The expansion centers are defined to be located either at the atomic nuclei of the molecules defining the environment or at the atomic nuclei and bond midpoints. The latter description would, in principle, lead to a more accurate representation of the embedding potential with an improved radius of convergence for the multipole model. The electrostatic embedding potential only accounts for the permanent charge distribution of the environment, and in order to account for many-body induction effects, i.e., the polarization of the environment both internally and by the quantum mechanical core, we assign a set of localized anisotropic dipole polarizability tensors at the expansion centers, giving rise to an induced charge distribution in the environment. The latter is represented in terms of induced dipoles which are determined on the basis of classical response theoretical methods.[3] The localized multipoles and polarizabilities are determined using quantum mechanical methods. In this paper we show that a careful representation of the permanent and induced potentials by multipoles and polarizabilities, respectively, leads to a very accurate mid- and long-range electrostatic potential as compared to quantum mechanical data. The PE model is generally applicable but here it is implemented for the case of Hartree−Fock and Kohn−Sham density functional theory, which we denote the PE-HF and PE-DFT methods. The functional form of the polarizable embedding potential resembles that of the EFP method by Gordon et al.;[4−6] however, the strength of our model is the ability to describe excited states on the same footing as ground states. This is achieved through a formulation of the PE model within the context of time-dependent quantum mechanical response theory. Pertinent to our formulation of polarizable embedding within response theory is the self-consistent many-body environmental response. Here it is important to emphasize that the response of the environment due to the differential change between the ground- and excited-state electron density is fully self-consistent, whereas this is approximated in other similar implementations.[7−10] The consequences of typical approximations as compared to the inclusion of the fully self-consistent environmental response is investigated with numerical examples. The PE model is presented and implemented up to and including quadratic response, with the possibility of straightforwardly extending it to higher order response. This allows for evaluation of vertical electronic excitation energies and the related one- and two-photon transition moments. Furthermore, electronic second- and third-order ground-state molecular properties, such as static and dynamic (hyper)polarizabilities, are available, as are excited-state first-order molecular properties. In addition, magnetic properties, such as magnetizabilities, nuclear shielding constants, and spin−spin coupling constants, may also be computed, using gauge invariant atomic orbitals (GIAOs) when needed.

Nuclear dynamics is in the present method considered by performing classical molecular dynamics (MD) simulations. Here we proceed in a sequential manner; i.e., we first perform the MD simulations and then, using an appropriate number of configurations extracted from the MD simulations, simulate the electronic structure. In this respect we neglect the effect of the electronic structure on the configurations, and the accuracy of our approach relies first of all on the use of an accurate classical potential to be used for the MD simulations.

Inclusion of explicit polarization into force field methods have in recent years received much attention.[11] The current status is that polarization may contribute significantly and specifically to specific solvation process. For example, polarization causes a significant increase in the dipole moment of a water molecule in the liquid state and may in addition constitute as much as 50% of the total interaction energy.[12] An important point here is to be able to calculate all properties characterizing the intermolecular interactions by quantum mechanical methods.

In the present paper the focus is on the effects from a water solvent on the excitation energies of a set of organic molecules, i.e., the solvent shifts. We emphasize that our computational method is not restricted to consideration of solute−solvent systems. However, predictions and rationalizations related to solvent shifts have, for a long time, been a very active and important research area in chemistry[13−17] and serve here as a valuable benchmark for this newly developed computational method. Furthermore, solvent shifts are also highly relevant to, e.g., the studies of biological samples. In fact, certain organic molecules are frequently used as molecular chameleons in order to characterize the degree of polarity of an environment. This is possible since a change in the environmental polarity will lead to a differential stabilization of the ground and excited states of the probe molecule, and thereby to a change in the energy difference between these two states. Consequently, variations in the intensity and especially the position of the absorption or emission spectra becomes a direct measure of the polarity and related specific interactions between the probe and the environment. The key to an accurate rationalization and modeling of, e.g., such chameleons is a flexible computational model formulated toward excited states of large molecular samples.

2. Theory

Below we detail the theoretical aspects of the PE model and its formulation within Kohn−Sham density functional theory and time-dependent response theory.

2.1. Ground-State Polarizable Embedding. The PE model presented in this work uses the QM/MM approach[18−22] to describe the interactions between the environment and the central molecular system. We use an advanced force field representation of the environment which is derived by quantum mechanical calculations. Thus, we assign a multicenter multipole expansion to each molecule in the surrounding environment to represent the electrostatic embedding potential. Furthermore, we place localized anisotropic dipole−dipole polarizability tensors on all expansion centers to allow polarization of the electrostatic embedding potential. This enables us to formulate the PE model where the ground-state electron density of the molecular core is optimized while simultaneously taking into account the explicit electrostatic interactions and many-body induction effects of the surrounding environment in a self-consistent manner. All other

interactions, mainly short-range repulsion and dispersion, can be modeled with a standard 6−12 Lennard-Jones (LJ) potential. The LJ potential does not depend on electronic coordinates and will therefore not affect the electron density of the molecular core.

We begin by describing the general PE model and will later apply it to density functional theory (DFT). Our model focuses on a central molecular system which we will refer to as the QM core. The effects from interactions with the environment are described through the PE potential. The energy of the QM core can thus be separated into two terms

$$E_{\text{PE-QM}} = E_{\text{QM}} + E_{\text{PE}} \tag{1}$$

where E_{QM} is the energy of the isolated QM core and E_{PE} is the energy due to the interactions with the PE potential. The interaction energy, E_{PE}, is given by

$$E_{\text{PE}} = E_{\text{PE}}^{\text{es}} + E_{\text{PE}}^{\text{ind}} + E_{\text{PE}}^{\text{LJ}} \tag{2}$$

where $E_{\text{PE}}^{\text{es}}$ is the electrostatic interaction energy, $E_{\text{PE}}^{\text{ind}}$ is the induction energy, and $E_{\text{PE}}^{\text{LJ}}$ is the energy due to the LJ interactions. The electrostatic contribution is composed of interactions between the permanent multipoles in the environment and the nuclei and electrons in the QM core; i.e.,

$$E_{\text{PE}}^{\text{es}} = \sum_{s=1}^{S}\sum_{k=0}^{K}\frac{(-1)^k}{k!}\left(\sum_{m=1}^{M} Z_m \mathbf{T}_{ms}^{(k)} - \sum_{i=1}^{N}\mathbf{T}_{is}^{(k)}\right)\mathbf{Q}_s^{(k)} \tag{3}$$

Here S is the total number of sites in the environment and K is the maximum order of the multipole expansion assigned to the molecules in the environment. The quantities M and N are the numbers of nuclei and electrons, respectively, in the QM core. The $\mathbf{T}^{(k)}$ factors are the interaction tensors, defined as $\mathbf{T}_{ab}^{(k)} = \nabla^k[1/(|\mathbf{r}_b - \mathbf{r}_a|)]$, and $\mathbf{Q}_s^{(k)}$ is the kth order multipole moment assigned to the sth site in the environment; e.g., $\mathbf{Q}_s^{(0)} = q_s$, $\mathbf{Q}_s^{(1)} = \boldsymbol{\mu}_s$, and so on.

The induction energy due to the polarization of the environment both internally and by the QM core is given by

$$E_{\text{PE}}^{\text{ind}} = -\frac{1}{2}\boldsymbol{\mu}^{\text{ind}} \cdot (\mathbf{F}^{\text{nuc}} + \mathbf{F}^{\text{elec}} + \mathbf{F}^{\text{es}}) \tag{4}$$

where $\boldsymbol{\mu}^{\text{ind}}$ is a $3S$-dimensional vector which contains the full set of induced dipole moments, i.e., $\boldsymbol{\mu}^{\text{ind}} = [\boldsymbol{\mu}_1^{\text{ind}}, \boldsymbol{\mu}_2^{\text{ind}}, ..., \boldsymbol{\mu}_S^{\text{ind}}]^{\text{T}}$, and $\mathbf{F}^{\text{nuc}}$, $\mathbf{F}^{\text{elec}}$, and $\mathbf{F}^{\text{es}}$ are the corresponding electric field vectors, which contain the electric fields from the nuclei and electrons in the QM core and the permanent multipole moments in the environment at the positions of the induced dipoles. An induced dipole moment is determined by the total electric field which is the sum of the fields from the nuclei and electrons in the QM core and the permanent multipoles and all the other induced dipoles in the environment. The set of induced dipoles can be conveniently determined as a simple matrix−vector multiplication[3]

$$\boldsymbol{\mu}^{\text{ind}} = \mathbf{B}(\mathbf{F}^{\text{nuc}} + \mathbf{F}^{\text{elec}} + \mathbf{F}^{\text{es}}) = \mathbf{BF} \tag{5}$$

where $\mathbf{B}$ is the symmetric $(3S \times 3S)$-dimensional classical response matrix connecting the electric fields and the set of induced dipoles. The response matrix is defined as

$$\mathbf{B} = \begin{pmatrix} \boldsymbol{\alpha}_1^{-1} & \mathbf{T}_{12}^{(2)} & \cdots & \mathbf{T}_{1S}^{(2)} \\ \mathbf{T}_{21}^{(2)} & \boldsymbol{\alpha}_2^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{T}_{(S-1)S}^{(2)} \\ \mathbf{T}_{S1}^{(2)} & \cdots & \mathbf{T}_{S(S-1)}^{(2)} & \boldsymbol{\alpha}_S^{-1} \end{pmatrix}^{-1} \tag{6}$$

where the polarizability tensors are along the diagonal and the off-diagonal elements are the dipole−dipole interaction tensors.

The PE model is applied to DFT by constructing an effective Kohn−Sham (KS) operator; i.e.,

$$\hat{f}_{\text{eff}} = \hat{f}_{\text{KS}} + \hat{v}_{\text{PE}} \tag{7}$$

where $\hat{f}_{\text{KS}}$ is the ordinary vacuum Kohn−Sham operator and $\hat{v}_{\text{PE}}$ is the PE potential operator. The contribution to the effective KS operator due to the polarizable environment, i.e., the PE potential operator $\hat{v}_{\text{PE}}$, is determined by minimization of the total energy functional with respect to the electron density. Therefore, we only need to consider terms that depend on the electron density, i.e., the last term in eq 3, where the electron charge is replaced by an integral over the density, and eq 4 which has a density dependence through the electric field from the electrons both explicitly and through the induced dipole moments (eq 5). Thus, the PE contribution in second quantized (SQ) form is found to be

$$\hat{v}_{\text{PE}} = \sum_{s=1}^{S}\sum_{k=0}^{K}\frac{(-1)^{(k+1)}}{k!}\mathbf{Q}_s^{(k)}\sum_{pq}\mathbf{T}_{s,pq}^{(k)}\hat{E}_{pq} - \sum_{s=1}^{S}\boldsymbol{\mu}_s^{\text{ind}}(\mathbf{F}[\rho])\sum_{pq}\mathbf{T}_{s,pq}^{(1)}\hat{E}_{pq} \tag{8}$$

where the subscripts pq indicate a matrix element of the corresponding operator in the KS orbital basis. The excitation operator $\hat{E}_{pq}$ is expressed in terms of creation and annihilation operators, i.e., $\hat{E}_{pq} = a_{p\alpha}^{\dagger}a_{q\alpha} + a_{p\beta}^{\dagger}a_{q\beta}$. The first term in eq 8 contains the electrostatic embedding potential introduced here in terms of a set of localized multipole moments, while the second term accounts for polarization of the environment by the electron density. The induced dipole moments are updated in each self-consistent field (SCF) iteration, thus leading to a fully self-consistent treatment of the polarization. The derivation of the PE-HF method proceeds in a similar manner with the construction of an effective Fock operator.

2.2. Polarizable Embedding for Excited States. In this work we will only give an overview of the derivation of the linear and quadratic quantum mechanical response functions with emphasis on the contributions that are due to the PE potential. For a detailed discussion of the implementation of linear and quadratic response theory in vacuum, we refer the reader to the work by Sałek et al.[23]

The starting point for the derivation of the response functions is to consider the time dependence of an expectation value of a time-independent operator $\hat{A}$. It can be expanded in orders of a time-dependent perturbation

$$\langle t|\hat{A}|t\rangle = \langle t|\hat{A}|t\rangle^{(0)} + \langle t|\hat{A}|t\rangle^{(1)} + \langle t|\hat{A}|t\rangle^{(2)} + ... \tag{9}$$

where the first term on the right-hand side is the time-independent expectation value and the second and third terms

describe the linear and quadratic response to the perturbation, respectively. The Fourier representation of the linear and quadratic response is given by

$$\langle t|\hat{A}|t\rangle^{(1)} = \int \langle\langle \hat{A};\hat{V}^{\omega}\rangle\rangle_{\omega} \exp(-i\omega t)\, d\omega \tag{10}$$

$$\langle t|\hat{A}|t\rangle^{(2)} = \frac{1}{2}\int\int \langle\langle \hat{A};\hat{V}^{\omega_1},\hat{V}^{\omega_2}\rangle\rangle_{\omega_1,\omega_2} \exp(-i(\omega_1 + \omega_2)t)\, d\omega_1\, d\omega_2 \tag{11}$$

where $\langle\langle \hat{A};\hat{V}^{\omega}\rangle\rangle_{\omega}$ and $\langle\langle \hat{A};\hat{V}^{\omega_1},\hat{V}^{\omega_2}\rangle\rangle_{\omega_1,\omega_2}$ are the linear and quadratic response functions, respectively, and $\hat{V}^{\omega}$ is a Fourier transformed perturbation operator; i.e., $\hat{V}(t) = \int \hat{V}^{\omega} \exp(-i\omega t)\, d\omega$. We use an exponential parametrization of the time evolution of the reference Kohn−Sham determinant; i.e., $|t\rangle = \exp(-\hat{\kappa}(t))|0\rangle$, where $|0\rangle$ is the unperturbed Kohn−Sham determinant and $\hat{\kappa}(t)$ is the anti-Hermitian one-electron operator defined as $\hat{\kappa}(t) = \sum_{rs} \kappa_{rs}(t)\hat{E}_{rs}$.

The response functions are derived using the Ehrenfest theorem, which can be written as[24]

$$\langle 0|[\hat{Q}, \exp(\hat{\kappa}(t))\left(\hat{H}(t) + \hat{V}(t) - i\frac{d}{dt}\right)\exp(-\hat{\kappa}(t))]|0\rangle = 0 \tag{12}$$

where $\hat{Q}$ is a general one-electron time-independent operator which we will define as the vector $\hat{\mathbf{q}}$ containing the excitation operators $\hat{E}_{pq}$. Equation 12 is then expanded in a Baker−Campbell−Hausdorff (BCH) expansion, and perturbation expansions of the time evolution operator $\hat{\kappa}(t)$ and the KS Hamiltonian $\hat{H}(t)$ are inserted. Terms that are first and second order in the perturbation are collected and used to derive the linear and quadratic response functions, respectively, in the frequency domain. The SQ form of the expanded time-dependent Kohn−Sham Hamiltonian is

$$\hat{H}(t) = \sum_n \hat{H}^{(n)} = \sum_n \sum_{pq} f_{pq}^{(n)}\hat{E}_{pq} = \sum_n \sum_{pq} (\delta_{0n}h_{pq} + j_{pq}^{(n)} + \nu_{xc,pq}^{(n)} + \nu_{PE,pq}^{(n)})\hat{E}_{pq} \tag{13}$$

where h_{pq} is an integral over the kinetic energy and nuclear-attraction operators, $j_{pq}^{(n)}$ is an nth-order Coulomb integral and $\nu_{xc,pq}^{(n)}$ is an nth-order integral over the exchange-correlation potential. Finally, the integral $\nu_{PE,pq}^{(n)}$ is an nth-order integral over the PE potential which gives the contribution from the polarizable environment.

2.2.1. Linear Response. The linear response function $\langle\langle \hat{A};\hat{V}^{\omega}\rangle\rangle_{\omega}$ for the property $\hat{A}$ perturbed by a periodic perturbation $\hat{V}^{\omega}$ with associated frequency ω is given by

$$\langle\langle \hat{A};\hat{V}^{\omega}\rangle\rangle_{\omega} = -\mathbf{A}^{\dagger}\boldsymbol{\kappa}^{\omega} \tag{14}$$

where $\mathbf{A} = \langle 0|[\hat{\mathbf{q}},\hat{A}]|0\rangle$ and the time evolution parameters are collected in the vector $\boldsymbol{\kappa}^{\omega}$. The time evolution parameters are determined from the matrix equation

$$(\mathbf{E} - \omega\mathbf{S})\boldsymbol{\kappa}^{\omega} = \mathbf{V}^{\omega} \tag{15}$$

which is derived from the Ehrenfest theorem in eq 12. Here it has been used that $\hat{\kappa}^{\omega} = \hat{\mathbf{q}}^{\dagger}\boldsymbol{\kappa}^{\omega}$. The $\mathbf{E}$ matrix is defined through

$$\mathbf{E}\boldsymbol{\kappa}^{\omega} = -\langle 0|[\hat{\mathbf{q}}, [\hat{\kappa}^{\omega}, \hat{H}^{0}] + \hat{H}^{\omega}]|0\rangle \tag{16}$$

The generalized overlap matrix is defined as

$$\mathbf{S} = \langle 0|[\hat{\mathbf{q}}, \hat{\mathbf{q}}^{\dagger}]|0\rangle \tag{17}$$

and the perturbation vector is given by

$$\mathbf{V}^{\omega} = \langle 0|[\hat{\mathbf{q}}, \hat{V}^{\omega}]|0\rangle \tag{18}$$

We observe that explicit contributions from the PE potential only enter the linear response function through the $\mathbf{E}$ matrix as

$$\mathbf{E}_{PE}\boldsymbol{\kappa}^{\omega} = -\langle 0|[\hat{\mathbf{q}}, [\hat{\kappa}^{\omega}, \hat{v}_{PE}^{0}] + \hat{v}_{PE}^{\omega}]|0\rangle \tag{19}$$

Using one-index transformed integrals,[25] we define a new set of operators

$$\hat{Q}_1^{\omega} = [\hat{\kappa}^{\omega}, \hat{v}_{PE}^{0}] = \hat{v}_{PE}^{0}(\boldsymbol{\kappa}^{\omega}) \tag{20}$$

$$\hat{Q}_2^{\omega} = \hat{v}_{PE}^{\omega} = -\sum_{s=1}^{S} \boldsymbol{\mu}_s^{ind}(\tilde{\mathbf{F}}^{\omega})\hat{\mathbf{T}}_s^{(1)} \tag{21}$$

where the induced dipole moments are determined from eq 5 and the transformed electric field is evaluated according to

$$\tilde{\mathbf{F}}^{\omega} = \langle 0|[\hat{\kappa}^{\omega}, \hat{\mathbf{T}}_s^{(1)}]|0\rangle = \langle 0|\hat{\mathbf{T}}_s^{(1)}(\boldsymbol{\kappa}^{\omega})|0\rangle \tag{22}$$

Finally, we can express the PE contribution to the linearly transformed $\mathbf{E}$ matrix as

$$\mathbf{E}_{PE}\boldsymbol{\kappa}^{\omega} = -\langle 0|[\hat{\mathbf{q}}, \hat{Q}_1^{\omega} + \hat{Q}_2^{\omega}]|0\rangle \tag{23}$$

The $\hat{Q}_1^{\omega}$ operator gives the zero-order PE contribution to the linear response which corresponds to a static environment; i.e., the environment does not respond to the applied perturbation. The $\hat{Q}_2^{\omega}$ operator, on the other hand, describes the dynamical response of the environment due to the perturbation. It is important to note that this is the fully self-consistent many-body environmental response without approximations as opposed to other similar implementations[7,8] where the dynamical response is approximated by using a block-diagonal classical response matrix (eq 6) in the response calculations. The approximated block-diagonal response matrix includes the polarizability tensors but neglects the off-diagonal interaction tensors, whereas we include the full response matrix. The consequences of the approximation is investigated in section 5.2.

2.2.2. Quadratic Response. The quadratic response function $\langle\langle \hat{A};\hat{V}^{\omega_1},\hat{V}^{\omega_2}\rangle\rangle_{\omega_1,\omega_2}$ for the property $\hat{A}$ perturbed by two periodic perturbations $\hat{V}^{\omega_1}$ and $\hat{V}^{\omega_2}$ with associated frequencies ω_1 and ω_2, respectively, is given by

$$\langle\langle \hat{A};\hat{V}^{\omega_1},\hat{V}^{\omega_2}\rangle\rangle_{\omega_1,\omega_2} = \boldsymbol{\kappa}^{\mathbf{A}^{\dagger}}\mathbf{V}^{\omega_1,\omega_2} + \hat{P}_{12}\langle 0|[\hat{\kappa}^{\omega_1}, [\hat{\kappa}^{\omega_2}, \hat{A}]]|0\rangle \tag{24}$$

where the perturbation vector is given by

$$\mathbf{V}^{\omega_1,\omega_2} = \hat{P}_{12}(\langle 0|[\hat{\mathbf{q}},[\hat{\kappa}^{\omega_1},[\hat{\kappa}^{\omega_2},\hat{H}^0]]]|0\rangle + \omega_2\langle 0|[\hat{\mathbf{q}},[\hat{\kappa}^{\omega_1},\hat{\kappa}^{\omega_2}]]|0\rangle + 2\langle 0|[\hat{\mathbf{q}},[\hat{\kappa}^{\omega_1},\hat{H}^{\omega_2}+\hat{V}^{\omega_2}]]|0\rangle + \langle 0|[\hat{\mathbf{q}},\hat{\bar{H}}^{\omega_1,\omega_2}]|0\rangle) \quad (25)$$

Here $\hat{P}_{12}$ is the idempotent symmetrizer defined through $\hat{P}_{12}A(\omega_1,\omega_2) = (1/2)[A(\omega_1,\omega_2) + A(\omega_2,\omega_1)]$. Furthermore, we have used that the second-order density matrix elements can be separated into components due to either first- or second-order parameters such that the Hamiltonian can be similarly separated; i.e., $\hat{H}^{\omega_1,\omega_2} = \hat{\bar{H}}^{\omega_1,\omega_2} + \hat{\bar{\bar{H}}}^{\omega_1,\omega_2}$. The time evolution parameters are determined by solving three linear response equations

$$\kappa^{\mathbf{A}^\dagger}(\mathbf{E} - (\omega_1+\omega_2)\mathbf{S}) = \mathbf{A}^\dagger \quad (26)$$

$$(\mathbf{E} - \omega_1\mathbf{S})\kappa^{\omega_1} = \mathbf{V}^{\omega_1} \quad (27)$$

$$(\mathbf{E} - \omega_2\mathbf{S})\kappa^{\omega_2} = \mathbf{V}^{\omega_2} \quad (28)$$

The explicit PE contributions to the quadratic response function enter the $\mathbf{E}$ matrix and the $\mathbf{V}^{\omega_1,\omega_2}$ vector. Contributions that appear in the $\mathbf{E}$ matrix are analogous to the linear response case; i.e.,

$$\kappa^{\mathbf{A}^\dagger}\mathbf{E}_{\mathrm{PE}} = -\langle 0|[\hat{\mathbf{q}},\hat{Q}_1^{\omega_1,\omega_2} + \hat{Q}_2^{\omega_1,\omega_2}]|0\rangle \quad (29)$$

where

$$\hat{Q}_1^{\omega_1,\omega_2} = [\hat{\kappa}^{\mathbf{A}^\dagger},\hat{v}_{\mathrm{PE}}^0] = \hat{v}_{\mathrm{PE}}^0(\kappa^{\mathbf{A}^\dagger}) \quad (30)$$

$$\hat{Q}_2^{\omega_1,\omega_2} = \hat{\bar{v}}_{\mathrm{PE}}^{\omega_1,\omega_2} = -\sum_{s=1}^{S}\boldsymbol{\mu}_s^{\mathrm{ind}}(\tilde{\mathbf{F}}^{\omega_1,\omega_2})\hat{\mathbf{T}}_s^{(1)} \quad (31)$$

The induced dipoles are calculated using eq 5 and the transformed electric field is given by

$$\tilde{\mathbf{F}}^{\omega_1,\omega_2} = \langle 0|[\hat{\kappa}^{\mathbf{A}^\dagger},\hat{\mathbf{T}}_s^{(1)}]|0\rangle = \langle 0|\hat{\mathbf{T}}_s^{(1)}(\kappa^{\mathbf{A}^\dagger})|0\rangle \quad (32)$$

Contributions to the perturbation vector are obtained from eq 25; i.e.,

$$\mathbf{V}_{\mathrm{PE}}^{\omega_1,\omega_2} = \hat{P}_{12}(\langle 0|[\hat{\mathbf{q}},[\hat{\kappa}^{\omega_1},[\hat{\kappa}^{\omega_2},\hat{v}_{\mathrm{PE}}^0]]]|0\rangle + 2\langle 0|[\hat{\mathbf{q}},[\hat{\kappa}^{\omega_1},\hat{v}_{\mathrm{PE}}^{\omega_2}]]|0\rangle + \langle 0|[\hat{\mathbf{q}},\hat{\bar{v}}_{\mathrm{PE}}^{\omega_1,\omega_2}]|0\rangle) \quad (33)$$

which we can rewrite to a more convenient form by defining the following set of operators:

$$\hat{Q}_3^{\omega_1,\omega_2} = \hat{P}_{12}[\hat{\kappa}^{\omega_1},[\hat{\kappa}^{\omega_2},\hat{v}_{\mathrm{PE}}^0]] = \hat{P}_{12}\hat{v}_{\mathrm{PE}}^0(\kappa^{\omega_2},\kappa^{\omega_1}) \quad (34)$$

$$\begin{aligned}\hat{Q}_4^{\omega_1,\omega_2} &= 2\hat{P}_{12}[\hat{\kappa}^{\omega_1},\hat{v}_{\mathrm{PE}}^{\omega_2}] \\ &= -2\hat{P}_{12}\sum_{s=1}^{S}\boldsymbol{\mu}_s^{\mathrm{ind}}(\tilde{\mathbf{F}}^{\omega_2})[\hat{\kappa}^{\omega_1},\hat{\mathbf{T}}_s^{(1)}] \\ &= -2\hat{P}_{12}\sum_{s=1}^{S}\boldsymbol{\mu}_s^{\mathrm{ind}}(\tilde{\mathbf{F}}^{\omega_2})\hat{\mathbf{T}}_s^{(1)},(\kappa^{\omega_1})\end{aligned} \quad (35)$$

$$\begin{aligned}\hat{Q}_5^{\omega_1,\omega_2} &= \hat{P}_{12}\hat{\bar{v}}_{\mathrm{PE}}^{\omega_1,\omega_2} \\ &= -\hat{P}_{12}\sum_{s=1}^{S}\boldsymbol{\mu}_s^{\mathrm{ind}}(\tilde{\mathbf{F}}^{\omega_1,\omega_2})\hat{\mathbf{T}}_s^{(1)}\end{aligned} \quad (36)$$

Here the perturbed electric field $\tilde{\mathbf{F}}^{\omega_2}$ is defined in eq 22 and $\tilde{\mathbf{F}}^{\omega_1,\omega_2}$ is defined as

$$\tilde{\mathbf{F}}^{\omega_1,\omega_2} = \langle 0|[\hat{\kappa}^{\omega_1},[\hat{\kappa}^{\omega_2},\hat{\mathbf{T}}_s^{(1)}]]|0\rangle = \hat{\mathbf{T}}_s^{(1)}(\kappa^{\omega_2},\kappa^{\omega_1}) \quad (37)$$

The induced dipole moments are determined from eq 5 as before. Using the newly defined operators, we obtain the PE contribution to the perturbation vector as

$$\mathbf{V}_{\mathrm{PE}}^{\omega_1,\omega_2} = \langle 0|[\hat{\mathbf{q}},\hat{Q}_3^{\omega_1,\omega_2} + \hat{Q}_4^{\omega_1,\omega_2} + \hat{Q}_5^{\omega_1,\omega_2}]|0\rangle \quad (38)$$

Just as in the linear response case there are terms, here it is $\hat{Q}_1^{\omega_1,\omega_2}$ and $\hat{Q}_3^{\omega_1,\omega_2}$, that contain the zero-order PE potential operator which also here give the contributions to the response function that arise from a static environment. All the other contributions, i.e., $\hat{Q}_2^{\omega_1,\omega_2}$, $\hat{Q}_4^{\omega_1,\omega_2}$, and $\hat{Q}_5^{\omega_1,\omega_2}$, account for the dynamical response of the environment due to the periodic perturbations.

3. Implementation

The presented PE-DFT method has been implemented in a developmental version of the Dalton program.[26] The implementation also trivially includes the PE-HF method due to the nature of the DFT implementation in the Dalton program. The electrostatic part is currently able to use permanent multipoles up to and including octupoles. The implementation includes the use of anisotropic dipole−dipole polarizability tensors leading to induced dipole moments which are calculated using either a direct or an iterative approach. In the direct approach, which is the default, we calculate the classical response matrix (eq 6) explicitly and store it on disk. The induced dipoles are subsequently calculated using a simple matrix−vector multiplication. This is the most efficient and fastest method; however, for calculations on very large molecular systems and/or on computers with low memory, where it is not possible to form the response matrix explicitly, the iterative approach becomes useful because of very low memory requirements. The iterative approach per default uses the Jacobi method to calculate the induced dipole moments. To avoid the so-called "polarizability catastrophe", we have added the option to use modified dipole interactions according to the model by Thole.[27,28]

The implementation of the contributions to the linear and quadratic response functions is based on the work by Sałek et al.,[23] who implemented DFT and DFT response functions in the Dalton program. Thus, our approach was to add the relevant contributions due to the PE potential to the existing DFT response code. The contributions we considered are

presented in eqs 23, 29, and 38. However, terms which contain zero-order PE contributions $\hat{v}_{PE}^{0}$, i.e., the $\hat{Q}_1^{\omega}$ operator (eq 20) and the $\hat{Q}_1^{\omega_1,\omega_2}$ and $\hat{Q}_3^{\omega_1,\omega_2}$ operators (eqs 30 and 34), are accounted for via the effective zero-order Hamiltonian. In the case of a nonpolarizable environment these are the only terms that contribute since the other terms account for the induced polarization in the environment due to the applied perturbation. The dynamic response of the environment in the linear response part is accounted for by the $\hat{Q}_2^{\omega}$ operator (eq 21) which is formed by first calculating the transformed electric field according to eq 22 and subsequently updating the induced dipoles via eq 5 using the classical response matrix from eq 6 which has been stored on disk during the ground-state optimization. The $\hat{Q}_4^{\omega_1,\omega_2}$ and $\hat{Q}_5^{\omega_1,\omega_2}$ operators (eqs 35 and 36) in the case of quadratic response are constructed in a similar manner using eqs 32 and 37 for the transformed electric fields.

4. Computational Details

To illustrate the capabilities of the implemented PE-DFT method, we here consider the UV/vis vertical absorption energies of a range of organic compounds in aqueous solution. In particular, we have computed the lowest $n \rightarrow \pi^*$ excitation energy of acetone and acrolein as well as the lowest $\pi \rightarrow \pi^*$ excitation energy of acrolein, pyridine, uracil, coumarin 151, and coumarin 153. To model nuclear dynamical effects, we performed classical MD simulations, using a polarizable force field, of each solute in an aqueous environment in order to extract a number of statistically uncorrelated solute−solvent configurations. These configurations were then subjected to the PE-DFT calculations where the solute is treated using DFT and the solvent molecules are represented by a PE potential. The excitation energy in solution is evaluated as the statistical average over the molecular configurations. The solvent shift of the excitation energy is defined as the difference between the excitation energy in solution and in vacuum.

4.1. Molecular Structures, Force Fields, and MD Simulations. In this work, we used the molecular configurations of acetone, *s-trans*-acrolein, and uracil in aqueous solution derived in our previous studies.[17,29,30] However, the computational procedure to obtain the solute−solvent configurations for the rest of the considered molecular probes follows the same strategy. The solvated geometries of pyridine and coumarin 151 were obtained from geometry optimizations using the B3LYP exchange-correlation functional[31] and the aug-cc-pVTZ basis set[32] along with the integral equation formalism PCM model[33] to account for bulk solvent effects. For coumarin 153 we used the same method but the aug-cc-pVDZ basis[32] due to cost-effectiveness, and we only considered the lowest energy conformation as obtained from our PCM based test calculations. The same methods were utilized to derive the molecular geometries in vacuum. The Gaussian 03 program[34] was used for all geometry optimizations.

The force fields used in the MD simulations consist of partial point charges, isotropic polarizabilities, and LJ parameters. The charges were calculated by fitting to the quantum-mechanical electrostatic potential according to the CHelpG algorithm[35] at the B3LYP/aug-cc-pVTZ level for pyridine and coumarin 151, and the B3LYP/aug-cc-pVDZ level for coumarin 153 in vacuum using the Gaussian 03 program.[34] To model induction interactions, we assigned isotropic polarizabilities to the atomic sites of the solutes. The distributed polarizabilities were computed using the LoProp method[36] available in the Molcas program[37] at the B3LYP/aug-cc-pVTZ level for pyridine and coumarin 151 and the B3LYP/aug-cc-pVDZ level for coumarin 153. The Dunning basis sets were recontracted so as to be of atomic natural orbital type as required by the LoProp method. The LJ parameters for pyridine were taken from ref 38. For both coumarin molecules we used the LJ parameters from the optimized potential for liquid simulations (OPLS) force field given in ref 39, except for the amino group in coumarin 151 where we used the OPLS parameters from ref 40. The water molecules were modeled using the polarizable force field of Ahlström et al.,[41] which represents a water molecule by three atomic point charges and an isotropic molecular polarizability located at the center of mass. The internal geometry of the water molecules was fixed to $R_{OH} = 0.9572$ Å and $\angle HOH = 104.49°$. The geometries of the solvated molecules and force field parameters for pyridine, uracil, coumarin 151, and coumarin 153 as used in the MD simulation are available as Supporting Information.

The MD simulations of a rigid solute molecule, i.e., either pyridine, coumarin 151, or coumarin 153, and 511 rigid water molecules were performed within the NVT ensemble at the temperature of 298.15 K. The cubic box length was always set so as to reproduce the experimental density of liquid water—24.91, 25.05, and 25.12 Å for pyridine, coumarin 151 and coumarin 153, respectively. The velocity Verlet integration algorithm was employed with the time step of 2 fs along with periodic boundary conditions. The electrostatic and LJ interactions were truncated at half of the box length, and the reaction field correction was applied beyond this cutoff. The induced dipole moments were recalculated every third time step with the relative tolerance of 10^{-7}. In addition, the linear damping of the dipole−dipole interactions was employed.[28] Lorentz−Berthelot rules were applied for the LJ interactions of unlike atoms.[42] The system was equilibrated for 200 ps, and the molecular configurations were recorded every 10th ps during the production run of 1.2 ns. We thus have 120 molecular snapshots from each of the MD runs to use in the PE-DFT calculations. All MD simulations were performed using the Molsim software.[43]

4.2. Electronic Structure Calculations. We used the CAM-B3LYP hybrid exchange-correlation functional[44] to compute the excitation energies. This functional exhibits improved long-range behavior which is due to the splitting of the $1/r$ operator into short- and long-range contributions in the exact HF exchange term. The CAM-B3LYP functional has been shown to provide improved results for long-range properties including excitation energies.[44−46] Furthermore, we recently demonstrated that the CAM-B3LYP based solvent shifts of the $\pi \rightarrow \pi^*$ type excitation energies are more reliable as compared to the corresponding B3LYP results.[17,30] The parametrization of the CAM-B3LYP functional as proposed in the original work[44] was used. For all molecules,

except coumarin 153, the aug-cc-pVDZ basis set was used in the calculations of excitation energies, which is adequate for local n → π^* and $\pi \rightarrow \pi^*$ excitation energies.[17,30,47] A smaller 6-31++G* basis[48] was used in the calculations on coumarin 153.

To model the solvent molecules, we used PE potentials based on a hierarchy of force fields computed using the LoProp approach.[36] The force field parameters, i.e., multipole moments and polarizabilities, were derived at the B3LYP/aug-cc-pVTZ level of theory in vacuum using the Molcas program. In this work we use the designation M*X*P*Y* for the force fields, where *X* denotes the highest order of the multipole moments and *Y* indicates whether it includes isotropic ($Y = 1$) or anisotropic ($Y = 2$) polarizabilities. All these parameters are attributed to the atomic sites of the water molecules. In addition to the M*X*P*Y* force fields we also used a force field denoted M2P2BM, which includes multipoles up to quadrupoles and anisotropic polarizabilities assigned to both atomic sites and bond midpoints. Furthermore, we derived a series of force fields using the PCM model which we denote M*X*PCM. In this case the induction effects are incorporated implicitly into the multipole moments. The parameters for these force fields were calculated at the B3LYP/aug-cc-pVTZ/PCM level. Finally, we used the Ahlström force field presented in the previous section and the standard nonpolarizable TIP3P force field due to Jorgensen.[49] The LoProp based force fields are provided in the Supporting Information.

The PE-DFT results for the excitation energies are the statistical averages over 120 molecular configurations. A spherical cutoff radius equal to 12 Å based on the distance between the center of masses of the solute and solvent molecules was used for every configuration. We have previously shown that this cutoff radius provides converged results in terms of electrostatics and that 120 molecular configurations used in the averaging of the liquid-phase results represents statistically converged properties (see, for example, refs 17 and 30). The standard error of the mean is evaluated as $s/\sqrt{N}$, where s is the sample standard deviation and N is the number of samples.

4.3. Benchmarks of the Force Fields. In this work we assess the quality of the force fields used for the water molecules by comparing the molecular electrostatic potentials in the vicinity of the molecule on the basis of the force field and quantum chemical reference calculations. The electrostatic potential is the most suitable observable for such a comparison since it enters directly in the PE potential operator in eq 8. The electrostatic potential was probed at a number of points forming the grid around the water molecule. The grid is formed between two concentric van der Waals surfaces of the molecule. The inner boundary of the grid is a conventional van der Waals surface of water molecule constructed from atom-centered interlocking spheres with the van der Waals radii of 1.55 Å for oxygen[50] and 1.20 Å for hydrogen[51] atoms. The outer boundary is obtained as the van der Waals surface formed using van der Waals radii scaled by a factor of 4. The grid points are then homogeneously distributed between the two van der Waals surfaces with a separation between two neighboring grid points of 0.2 Å in all three directions. The resulting grid is thus composed of 122 263 points in total. The electrostatic potential due to the multipoles at the ath grid point was calculated according to

$$\varphi_a = \sum_{s=1}^{S} \sum_{k=0}^{K} \frac{(-1)^k}{k!} \mathbf{T}_{as}^{(k)} \mathbf{Q}_s^{(k)} \tag{39}$$

where the summations are over all multipole expansion centers in the molecule and all multipoles. The QM electrostatic potential at the ath grid point is the expectation value of the $1/(|\mathbf{r}_a - \mathbf{r}|)$ operator plus the nuclear contribution. The B3LYP exchange-correlation functional and the aug-cc-pVTZ basis set, i.e., the same method used to derive the LoProp force fields, was used to evaluate the QM reference electrostatic potential. The analysis is then performed in terms of the root-mean-square deviation (rmsd)

$$\mathrm{rmsd} = \sqrt{\frac{1}{N} \sum_a (\varphi_a - \varphi_a^{\mathrm{QM}})^2} \tag{40}$$

where N is the number of grid points. In particular, we performed the analysis on the subsets of the grid points between two close-lying van der Waals surfaces so as to investigate the behavior of the rmsd with respect to the distance from the molecule.

The quality of the polarizabilities can be assessed by applying an external homogeneous electric field in the calculation of the electrostatic potential. This field will give rise to induced dipoles which in turn creates an electrostatic potential around the molecule. Two calculations are then performed at the B3LYP/aug-cc-pVTZ level—one in vacuum and another in the external electric field. The QM reference is then obtained by subtracting the electrostatic potential in vacuum from that in the external field at every grid point. In this work we applied an electric field with the magnitude of 0.01 a.u. and all Cartesian components positive and equal. We used the Dalton program for the QM calculations, whereas the construction of the grid, calculations of the electrostatic potential and the analysis are performed using the Whirlpool program.[52]

5. Results and Discussion

Below we detail the results from the calculations performed using the PE-DFT approach with regards to the assessment of the force fields.

5.1. Quality of the Force Fields. We inspect the quality of the force fields used for a water molecule by comparing the electrostatic potentials due to classical and quantum mechanical representations of the molecule. This is important in the present context as it is desirable to use, for example, permanent multipole expansions truncated at the lowest possible order and still obtain converged electrostatic interactions. Similarly, it is of interest to investigate if the induction effects are sufficiently accurately described by using the simpler isotropic form of the polarizabilities or if anisotropic polarizabilities have to be used. The distribution of the force field parameters over the molecule is also an open question.

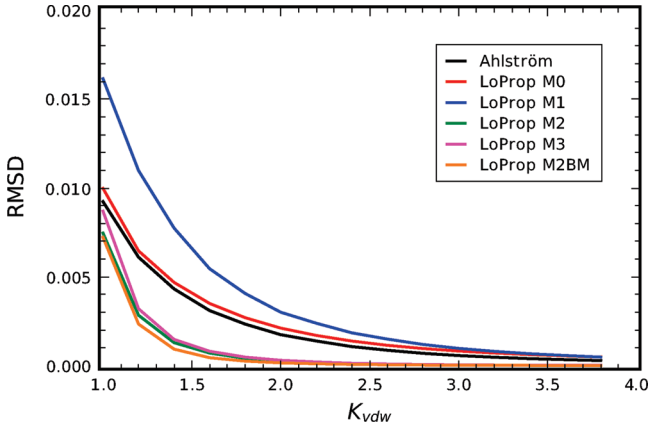


Figure 1. rmsd of the molecular electrostatic potential due to the multipoles of a water molecule as a function of the distance from the molecular van der Waals surface. The distance from the surface is given as the factor scaling the van der Waals radii. rmsd is in a.u.

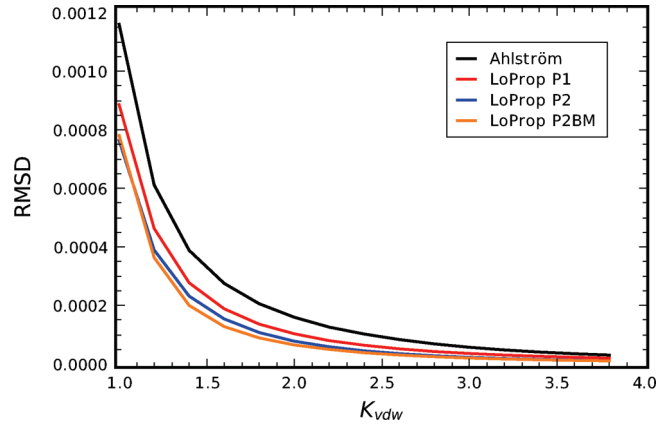


Figure 2. rmsd of the molecular electrostatic potential due to the induced dipole moments of a water molecule as a function of the distance from the molecular van der Waals surface. The distance from the surface is given as the factor scaling the van der Waals radii. rmsd is in a.u.

In Figure 1 we show the rmsd obtained by comparing the molecular electrostatic potential due to multipole moments taken from the considered force field and the quantum mechanically computed potential for different distances from the molecular van der Waals surface. In particular, we consider the three atomic point charges from the Ahlström force field and multipoles up to octupoles taken from the LoProp force fields. Figure 1 clearly illustrates that the multipole expansion is appropriate at large distances from the molecule and that higher order multipoles are mandatory to consider when the molecular potential close to the molecule is probed. We observe that the electrostatic potential due to the LoProp force field including atomic charges and dipole moments (M1) is poorer recovered than using the atomic point charges (M0) only. Quadrupole moments have a very pronounced effect and improve the molecular electrostatic potential considerably, as it was also found in ref 53. The octupole moments contribute little to the electrostatic potential. The atomic point charges in the Ahlström force field are constructed so as to implicitly include higher order multipoles. However, it is evident from Figure 1 that the improvement is negligible compared to the M0 force field and cannot match the performance of the M2 force field. We see that the LoProp force field which includes multipoles at bond midpoints (M2BM) offers minor improvement as well. We also inspected the M3BM force field, and octupoles were found to provide virtually no improvement. To conclude, we find that the LoProp force field with multipoles up to quadrupole moments assigned to the atomic sites of the water molecule provides apparently converged electrostatic interactions in terms of the multipole expansion.

In Figure 2 we compare the induced changes in the electrostatic potentials due to an external electric field. Here the electrostatic potential is due to the dipole moments induced by the external electric field. The QM reference electrostatic potential is the difference between the potential with and without the external field. We observe that the distributed isotropic polarizabilities in the LoProp force field lead to a more accurate account of the polarization of the electrostatic potential as compared to the single molecular isotropic polarizability assigned to the oxygen site of the water molecule in the Ahlström force field. Further improvement, though not that pronounced, is achieved by using the anisotropic polarizabilities. However, we would expect a larger difference in other molecules with a higher degree of anisotropy than a water molecule. We note that the rmsd in the case of induced dipoles is smaller by at least an order of magnitude than that due to the multipoles. This indicates that the specific description of the polarization is not as important as an accurate account of the electrostatics. However, as detailed later, explicit inclusion of polarization is generally found to be important for solvent induced shifts of excitation energies.

5.2. Excitation Energies. We examine the effects from the water solvent on the excitation energies of the solute molecules using our PE-DFT method. More precisely, we will look at the behavior of the solvent induced shifts of the excitation energies as we vary the complexity of the force field used in the PE-DFT calculations. We only include a solute molecule in the QM core; thereby only electrostatic and induction interactions are considered even though other interactions are known to be important, especially for $\pi \rightarrow \pi^*$ transitions.[17,30] This allows us to systematically investigate the effects from the electrostatic and induction interactions on the solvent shifts.

The computed excitation energies used as gas-phase references for the solvent shifts are shown in Table 1 together with the corresponding experimental data.[54−58] The computed excitation energies are generally in good agreement with the experimental values. For the $n \rightarrow \pi^*$ transitions the calculated values are off by 0.01 and 0.1 eV for acetone and acrolein, respectively. In the case of $\pi \rightarrow \pi^*$ transitions the deviation from experiment ranges from 0.02 eV in acrolein to 0.6 eV in pyridine. For uracil and the coumarins the computed excitation energies are overestimated by about 0.3 eV. It should be noted that the presented experimental value for coumarin 151 is our estimate of the excitation energy at the absorption maximum in vapor phase. It is based on values given by Ernsting et al.[58] In that work the authors measured the wavelengths of the absorption maxima of several

Table 1. Calculated and Experimental Gas-Phase Reference Vertical Excitation Energies of the Lowest $n \rightarrow \pi^*$ and/or $\pi \rightarrow \pi^*$ Transitions[a]

	ΔE_{vac} (eV)	
	calcd	exptl
n $\rightarrow \pi^*$		
acetone	4.473[b]	4.46[e]
acrolein	3.783[c]	3.69[f]
$\pi \rightarrow \pi^*$		
acrolein	6.405[c]	6.42[f]
pyridine	5.595	4.99[g]
uracil	5.384[d]	5.08[h]
coumarin 151	4.020	3.81 ± 0.06[i]
coumarin 153	3.675	3.37[j]

[a] The calculations were performed at the CAM-B3LYP/aug-cc-pVDZ level. [b] Reference 29. [c] Reference 17. [d] Reference 30. [e] Reference 54. [f] Reference 55. [g] Reference 56. [h] Reference 57. [i] Estimated value based on experimental data in ref 58 (see section 5.2 for more details). [j] Reference 58.

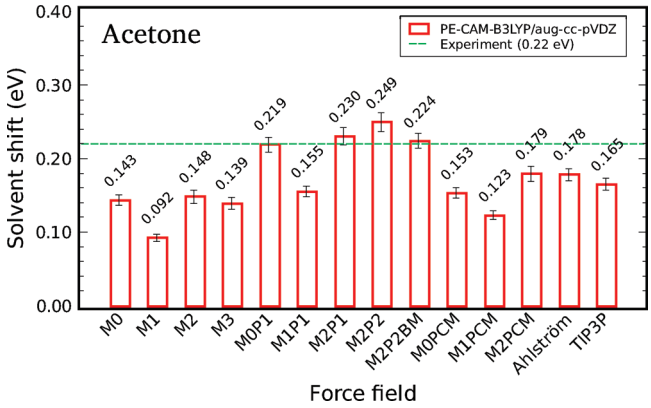


Figure 3. Gas-to-aqueous solvent shift of the lowest n $\rightarrow \pi^*$ excitation energy in acetone.

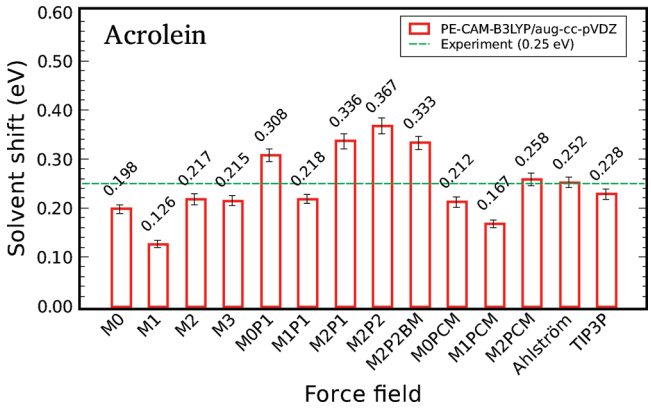


Figure 4. Gas-to-aqueous solvent shift of the lowest n $\rightarrow \pi^*$ excitation energy in acrolein.

coumarins and also the first strong vibronic band observed in a supersonic jet. The difference between the wavelengths of the absorption maxima and the vibronic bands are between about 18 and 28 nm. Subtracting the differences from the measured vibronic band of coumarin 151 leads to our estimate of the excitation energy at the absorption maximum in vapor phase given in the table.

The calculated solvent shifts are shown in Figures 3−9 together with the experimental solvent shift.[54−61] The first columns, i.e., M0, M1, M2 and in the cases of acetone,

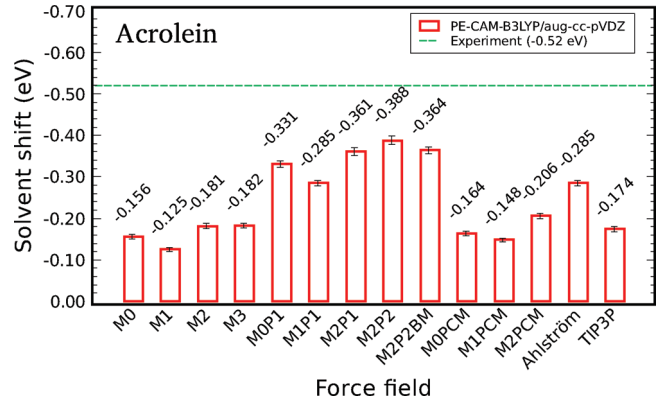


Figure 5. Gas-to-aqueous solvent shift of the lowest $\pi \rightarrow \pi^*$ excitation energy in acrolein.

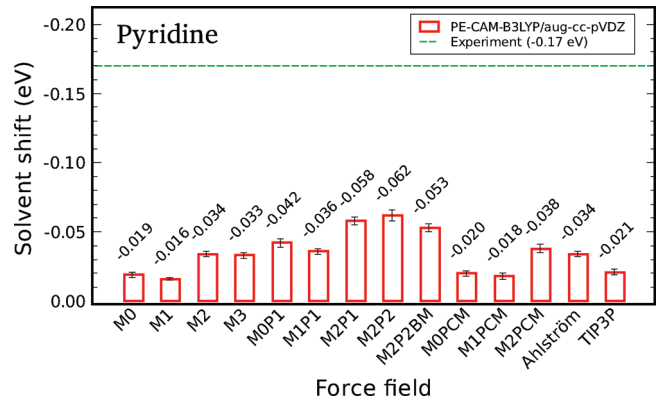


Figure 6. Gas-to-aqueous solvent shift of the lowest $\pi \rightarrow \pi^*$ excitation energy in pyridine.

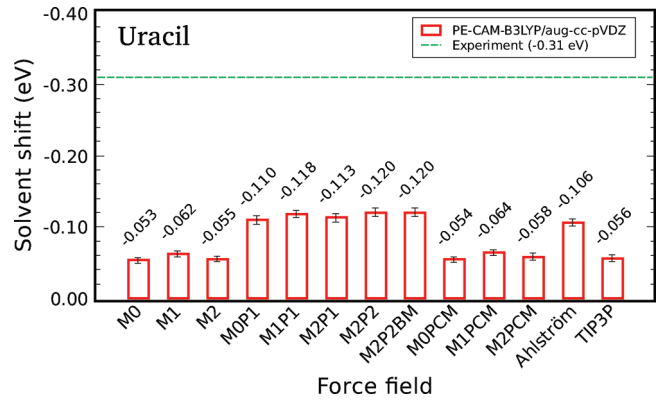


Figure 7. Gas-to-aqueous solvent shift of the lowest $\pi \rightarrow \pi^*$ excitation energy in uracil.

acrolein, and pyridine also M3, in each figure show the trends of the shifts with increasing order of the multipole expansion. These force fields do not model solvent polarization and are only used to investigate the effects of the higher order multipoles on the excitation energies. In all cases we observe a large effect from both the dipole and quadrupole moments. Adding octupoles only leads to very small changes that are within the statistical errors, indicating that we are converged at the quadrupole level with respect to the order of the permanent multipole moments. We thus observe a clear correlation between the trends here and the behavior of the force fields in terms of how well they reproduce the QM electrostatic potential (see Figure 1 and discussion in section

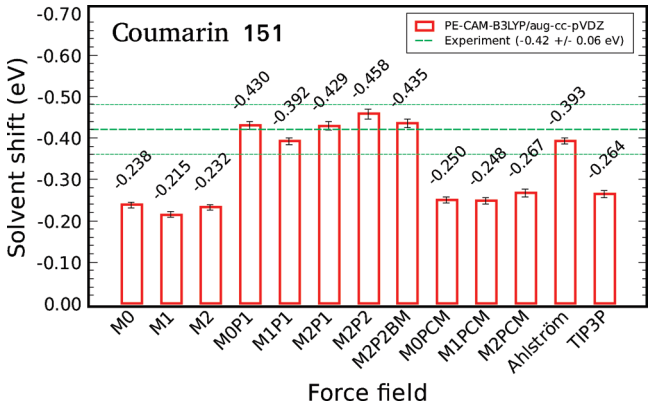


Figure 8. Gas-to-aqueous solvent shift of the lowest $\pi \rightarrow \pi^*$ excitation energy in coumarin 151.

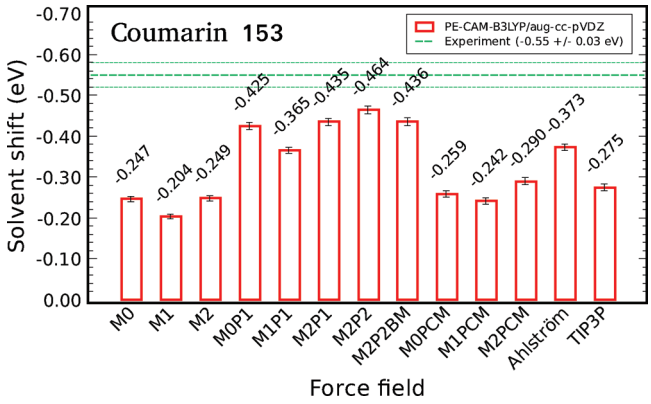


Figure 9. Gas-to-aqueous solvent shift of the lowest $\pi \rightarrow \pi^*$ excitation energy in coumarin 153.

5.1). Note that the shifts at the M0 and M2 (and M3) levels in general are very similar due to the fact that the effects from the dipole and quadrupole moments tend to cancel each other. This is not always the case, however, as we clearly see for pyridine (Figure 6) where the shifts at the M2 (and M3) level are about 75% larger than at the M0 level and, less pronounced, in acrolein (Figures 4 and 5) where the same difference is about 10 and 15%, respectively. Finally, we note that the same tendencies are observed in the M0P1−M2P1 and M0PCM−M2PCM series of force fields. Therefore, we find that it is, in general, necessary to include permanent multipoles up to quadrupoles in the LoProp force fields to get a converged description of the electrostatic interactions. However, using only point charges can, in some cases, lead to satisfying results as well.

Introducing distributed isotropic polarizabilities, in addition to the permanent multipoles, to the LoProp force fields, i.e., M0P1, M1P1, and M2P1, leads to a substantial increase of the solvent shifts. Comparing the shifts calculated at the M2P1 level to the M2 level shows that the solvent shifts are increased from about 50% in the case of $n \rightarrow \pi^*$ transitions in acetone (Figure 3) and acrolein (Figure 4) to more than 100% for the $\pi \rightarrow \pi^*$ transition in uracil (Figure 6). This clearly shows that induction effects have a significant impact on the solvent shifts and therefore must be taken into account. Furthermore, we observe that induction effects are particularly important for $\pi \rightarrow \pi^*$ transitions. Using distributed anisotropic polarizabilities, i.e., the M2P2 force field, gives further 6−10% increase in the solvent shifts as compared to the M2P1 level, which in the case of pyridine and uracil is about the same size as the statistical errors. Thus, the use of distributed anisotropic polarizabilities only gives small improvements as compared to the distributed isotropic polarizabilities. This can be explained by a rather small degree of anisotropy of a water molecule. Using the most sophisticated LoProp force field, M2P2BM, decreases the solvent shifts by a small amount as compared to the M2P2 force field which is of comparable magnitude but opposite sign as the difference between the M2P1 and M2P2 force fields. This indicates that the M2P2 force field has a tendency to overestimate the induction effects. Furthermore, it shows that it can be sufficient to use the M2P1 force field; however, this is only true for solvent molecules with low anisotropy such as water. The improved results at the M2P2BM level can mainly be contributed to an improved description of the induction interactions since we expect that the electrostatic interactions are converged at the quadrupole level. We expect that the M2P2BM force field provides the best model of the environment through an elaborate description of both the electrostatic and induction effects. This is achieved by using a converged distributed multipole expansion and distributed anisotropic polarizabilities with all properties localized on atoms and bond midpoints. It is interesting that the M0P1 force field performs rather well in most cases compared to the full M2P2BM force field. Therefore, we find that an appropriate approximation of the PE potential would be to use the M0P1 force field which captures the main parts of the electrostatics and induction effects due to a water solvent on the vertical excitation energies.

Recognizing that it is necessary to include the induction effects, it is an open question whether explicit inclusion of polarization is mandatory or if it is sufficient to have implicit polarization by using enhanced permanent multipoles. We used the PCM method in combination with the LoProp method to model the bulk solvent effects on the permanent multipoles. Using the M0PCM, M1PCM, and M2PCM force fields, we obtained larger solvent shifts as compared to the M0, M1, and M2 force fields. However, at the point charge level this increase is comparable to the statistical errors. Therefore, to obtain an improved description of the environment using the LoProp method in combination with PCM, it is necessary to include multipoles up to quadrupoles, i.e., the M2PCM force field. Here it appears that the largest effect is achieved for $n \rightarrow \pi^*$ transitions, which indicates that explicit modeling of induction effects are important for $\pi \rightarrow \pi^*$ transitions. The results for the solvent shifts of the $n \rightarrow \pi^*$ transitions calculated using the M2PCM force field are about 20% smaller than the shifts obtained using the M2P2BM force field. For the $\pi \rightarrow \pi^*$ transitions this difference is even larger and ranges from roughly 30 to 50%. Therefore, we find that the nonpolarizable M2PCM force field can be a reasonable representation of water molecules when $n \rightarrow \pi^*$ transitions are considered depending on the desired accuracy. However, explicit treatment of polarization is essential for $\pi \rightarrow \pi^*$ transitions.

It is interesting to compare the results obtained using the elaborate LoProp force field, M2P2BM, with other com-

Table 2. Comparison of Solvent Shifts Where the Environmental Response Due to the Differential Change between the Ground- and Excited-State Electron Density Is Approximated (See Section 5.2 for Details)[a]

solute	Q_1	Q_2^*	Q_2
acetone ($n \rightarrow \pi^*$)	0.226	0.224	0.224
coumarin 151 ($\pi \rightarrow \pi^*$)	−0.353	−0.451	−0.435

[a] The calculations were performed at the PE(M2P2BM)-CAM-B3LYP/aug-cc-pVDZ level.

Table 3. Molecular Dipole Moments of Solvated Acetone and Coumarin 151 in the Ground State and Excited States Corresponding to the Lowest $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ Transitions in Acetone and Coumarin 151, Respectively[a]

solute	μ^{gs} (D)	μ^{ex} (D)
acetone	5.0	3.7
coumarin 151	11.0	15.8

[a] The calculations were performed at the PE(M2P2BM)-CAM-B3LYP/aug-cc-pVDZ level.

monly used force fields. For this purpose we chose the standard nonpolarizable TIP3P force field and the polarizable Ahlström force field. The solvent shifts obtained using the TIP3P force field are all significantly smaller than those obtained using the M2P2BM force field. As expected we also observe substantially larger deviations for the $\pi \rightarrow \pi^*$ transitions since these require explicit modeling of the induction effects. The shifts of the $n \rightarrow \pi^*$ transitions are about 20% smaller compared to the results obtained using the M2P2BM force field, whereas the shifts of the $\pi \rightarrow \pi^*$ transitions vary between roughly 40 and 60%. The Ahlström force field, on the other hand, provides results that are in much better agreement with the results computed using the M2P2BM force field. Using this force field results in shifts that are about 10−35% smaller as compared to the M2P2BM level.

It is common to neglect the dynamical response of the environment that is due to the differential change between the ground-state and excited-state electron density or to include it in an approximate form.[7,8] We will denote the first case as the Q_1 approximation because it corresponds to the neglect of the $\hat{Q}_2^{\omega}$ operator (eq 21) in the case of linear response. The second case we will denote as the Q_2^* approximation, where the $\hat{Q}_2^{\omega}$ operator is included in an approximated form which corresponds to a block-diagonal classical response matrix (eq 6) with the polarizability tensors along the diagonal; i.e., the interaction tensors are omitted. In the Q_1 approximation the induced dipoles from the optimization of the ground-state wave function are also used in the response calculations. This can be a good approximation if the electronic density of the solute does not change significantly upon excitation. The Q_2^* approximation partly captures the dynamical environmental response, and the size of the effect is also connected to the difference between the electronic densities of the ground and excited states. To examine the effects of both approximations, we made additional calculations on acetone and coumarin 151. The results can be found in Table 2, where the Q_1 column presents the solvent shifts using the Q_1 approximation and the Q_2^* column contains the results where the Q_2^* approximation is used and in the final Q_2 column are the shifts where the full dynamical response of the environment is included. First of all we observe that both approximations have negligible or no effect on the $n \rightarrow \pi^*$ transition in acetone compared to the full inclusion of dynamical response. However, for the $\pi \rightarrow \pi^*$ transition in coumarin 151 we observe substantial effects from the Q_1 approximation and to a lesser degree from the Q_2^* approximation. The Q_1 approximation results in a solvent shift that is roughly 20% smaller than the shift calculated without approximations, while the Q_2^* approximation overshoots by roughly 4%. As a measure of the difference between the electron density of the ground and excited states we also calculated the difference between the dipole moment of the ground and excited states of acetone and coumarin 151 in aqueous solution at the PE(M2P2BM)-CAM-B3LYP/aug-cc-pVDZ level. This quantity is conveniently obtained as the residue of a quadratic response function. The calculated dipole moments are shown in Table 3. Here we see that the magnitude of the dipole moment of acetone decreases by 1.3 D upon excitation, whereas in coumarin 151 it goes up by 4.8 D thus explaining the much larger effects due to the Q_1 and Q_2^* approximations on the $\pi \rightarrow \pi^*$ transition in coumarin 151. These results clearly show that completely neglecting the dynamical response of the environment when computing excitation energies is a severe approximation if there is a significant change in the electron density upon excitation. The results also indicate that the approximate inclusion of dynamical response is a much better approximation; however, the error is still significant compared to the statistical errors. Moreover, we would expect the error to become larger in molecular systems where the difference of the ground- and excited-state electron density is even larger.

The $n \rightarrow \pi^*$ electronic absorption energies of acetone and acrolein have been extensively studied using different theoretical solvation models, and we refer to refs 17, 47, and 62 and references therein for discussion of some previously reported results. Very recent experimental measurements by Renge[54] estimate the gas-to-aqueous solvent shift of the $n \rightarrow \pi^*$ transition in acetone at 0.22 eV. The experimental estimate of the corresponding solvent shift in acrolein of 0.25 eV[55] has recently been confirmed.[17] Yoo et al.[8] used DFT in combination with the EFP method to compute the solvent shift of acetone, and the resulting shift of 0.21 eV compares very well to experimental result. Very recently, Kaminski et al.[63] have used the orbital-free embedding potential due to the statistically averaged solvent density through three-dimensional reference interaction site model (OFE/RISM) to study the lowest excitation energies of several organic probes in solution. The computed gas-to-aqueous solvent shift of the $n \rightarrow \pi^*$ transition in acetone is 0.19 eV, while the corresponding shift in acrolein of 0.33 eV is found to be somewhat overestimated compared to experimental data. Recently, several studies have elucidated the $n \rightarrow \pi^*$ transition in acetone and/or acrolein in water solution using electronic structure approaches rooted in coupled cluster theory. Caricato et al.[64] have evaluated this solvent shift in acrolein using EOM-CCSD/PCM and obtained 0.23 eV. A solvent shift of 0.18 eV in acetone was predicted on the basis of CCSD/MM calculations using a

Table 4. Calculated and Experimental Vertical Excitation Energies of the Lowest $n \rightarrow \pi^*$ and/or $\pi \rightarrow \pi^*$ Transitions in the Solvated Molecules[a]

	ΔE_{aq} (eV)	
	calcd	exptl
$n \rightarrow \pi^*$		
acetone	4.697 ± 0.010	4.68[b]
acrolein	4.116 ± 0.014	3.94[c]
$\pi \rightarrow \pi^*$		
acrolein	6.041 ± 0.009	5.90[c]
pyridine	5.542 ± 0.003	4.82[d]
uracil	5.264 ± 0.006	4.77[e]
coumarin 151	3.585 ± 0.011	3.39[f]
coumarin 153	3.239 ± 0.010	2.82 ± 0.03[g]

[a] The calculations were performed at the PE(M2P2BM)-CAM-B3LYP/aug-cc-pVDZ level. [b] Reference 54. [c] Reference 55. [d] Reference 59. [e] Reference 57. [f] Reference 60. [g] Value measured on a spectrum of coumarin 153 in aqueous 1-propanol solution ($X_{PrOH} = 0.05$) provided in ref 61.

nonpolarizable water potential.[65] Mata[66] has investigated the many-body effects on the solvent shifts and obtained 0.12 and 0.24 eV for the solvent shifts of acetone and acrolein, respectively, using a reduced two-body expansion with the EOM-CCSD method. Sneskov et al.[67] have elucidated the importance of triples excitations in the coupled cluster expansion on the solvent shift of the $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ excitation energies of acrolein and have found generally small effects as compared to CCSD. This short overview of the most recent results demonstrates the capability of different theoretical methods to describe the solvent effects on the $n \rightarrow \pi^*$ transition of acrolein and acetone. In this work we have observed a satisfactory agreement between theoretical and experimental results for the solvent shifts of the $n \rightarrow \pi^*$ transition of acetone. However, the solvent shift of 0.33 eV for acrolein is overestimated as compared to the experimentally measured 0.25 eV. The experimental data[54,55,57,59−61] for the solutes in aqueous solution are provided in Table 4 together with the corresponding theoretical values obtained using the M2P2BM force field.

Nonelectrostatic interactions have been found to substantially contribute to the solvent shift of the $\pi \rightarrow \pi^*$ transitions.[17,30,66] In this study we have neglected all effects other than electrostatic and induction interactions, and therefore the computed solvent shifts of the $\pi \rightarrow \pi^*$ excitation energies are in general considerably underestimated against experimental data, as illustrated in Figure 5 to Figure 9. Calculations by Caricato et al.[64] using EOM-CCSD and PCM gave a solvent shift of −0.38 eV of the $\pi \rightarrow \pi^*$ transition in acrolein, which is very similar to the results of the present work and also to the results based on CCSDR(3)/MM calculations in ref 67. In ref 66 the EOM-CCSD method coupled to the TIP3P force field for the water molecules gave a solvent shift of −0.27 eV in acrolein. Furthermore, extensive two-body expansions were found to be mandatory to obtain good agreement with experimental data. High-level coupled-cluster calculations have estimated a blue-shift of the $\pi \rightarrow \pi^*$ transition in uracil.[68] In that work, a gas-phase result which is in much better agreement with experiment, was obtained; however, the effects from the solvent are not well described which emphasizes the importance of a good embedding potential, especially when considering $\pi \rightarrow \pi^*$ transitions. Very recent CC2/MM calculations using the Ahlström force field predicted the solvent shift of −0.20 eV.[30] In the present paper a smaller shift of around −0.11 eV was found, indicating that in this case the CC2 model accounts more accurately for the differential effects of dynamical correlation compared to the CAM-B3LYP functional. A solvent shift of −0.25 eV in coumarin 151 was obtained from OFE/RISM calculations,[63] which is somewhat underestimated as compared to our PE-DFT results and experimental estimate. Sulpizi et al.[69] have obtained a solvent shift of −0.33 eV for coumarin 153 from TD-DFT/MM simulations which is somewhat smaller compared to our predictions and experimental findings, likely due to the implicit treatment of intermolecular polarization.

6. Summary and Conclusions

We have presented the theoretical details and implementation of the PE-DFT (and PE-HF) method. This method is a focused model based on a self-consistent polarizable embedding scheme, i.e., the PE model, applied to Kohn−Sham density functional theory. The method includes ground-state density optimization and calculation of molecular properties through time-dependent response theory, wherein the effects from a polarizable atomistic environment are taken into account in a self-consistent manner. The electrostatic interactions are modeled using a multicenter multipole expansion which includes multipoles up to and including octupoles. The polarization of the environment is described by using distributed anisotropic polarizability tensors. The multipoles and polarizability tensors are derived from quantum mechanical calculations and are distributed on the atomic sites or on the atomic sites and bond midpoints.

To evaluate the performance of the method, we benchmarked a series of force fields for a water molecule. The electrostatic potential due to the permanent multipole moments was sampled and compared to the quantum mechanically derived electrostatic potential. We found that the multipole expansion converges at the quadrupole level where it performs well at long and medium distances compared to the reference electrostatic potential. The errors increase at short distances but still show a big improvement compared to force fields with multipole moments truncated at lower order. The electrostatic potential due to induced dipole moments was also benchmarked and showed that the contribution from induction effects to the electrostatic potential is small compared to the permanent multipole moments. The best performance was observed with distributed anisotropic polarizabilities, although the improvement over distributed isotropic polarizabilities was not impressive. This was ascribed to the low degree of anisotropy of a water molecule.

The capability of the implemented method was demonstrated by computing the gas-to-aqueous solvent induced shifts of the lowest $n \rightarrow \pi^*$ vertical excitation energy in acetone and acrolein and the lowest $\pi \rightarrow \pi^*$ vertical excitation energy in acrolein, pyridine, uracil, coumarin 151, and coumarin 153. The solute−solvent dynamics were taken into account through classical molecular dynamics simulations

using polarizable force fields. A systematic investigation of the effects from electrostatic and induction interactions on the solvent shifts showed that the effects on the solvent shifts from electrostatic interactions converge at the quadrupole level, consistent with the benchmarking of the force fields. Furthermore, we found that modeling of the induction interactions is essential for the calculation of accurate solvent shifts, particularly for $\pi \rightarrow \pi^*$ transitions. The best performance was observed with the most detailed force field which includes multipoles up to quadrupoles and anisotropic polarizability tensors distributed on atomic sites and bond midpoints of the water molecules. Including the induction effects implicitly through an enhancement of the permanent multipole moments improved the solvent shifts although it could not match the results where explicit modeling was used. Moreover, for $\pi \rightarrow \pi^*$ transitions the effects were rather small and we therefore found that explicit account of the induction interactions is necessary for excitations of this nature. Finally, we found that an appropriate approximation of a water solvent should as a minimum consist of point charges and isotropic polarizabilities, which capture the main parts of the electrostatics and induction effects on the vertical excitation energies.

Our formulation of polarizable embedding within response theory includes the fully self-consistent many-body response of the environment. We investigated the effects on the solvent shifts when the dynamical response of the environment due to the differential change between the ground- and excited-state electron density is either neglected or approximated. Both are valid approximations if the electron density does not change significantly upon excitation. Calculations on acetone and coumarin 151, which have a small and large difference, respectively, between the dipole moment of the ground and relevant excited state, showed that complete neglect of the dynamical response can introduce significant errors. For coumarin 151 it comprised as much as 20% of the total calculated shift. The error due to an approximate inclusion of the dynamical environmental response was only 4% of the total shift in coumarin 151 and therefore presents a much better approximation. These errors are expected to increase in molecular systems where the electron density difference between the ground and excited states becomes even larger.

Comparisons with experiment showed satisfactory agreement. The sign of the solvent shifts, i.e., blue shift of $n \rightarrow \pi^*$ transitions and red shift of $\pi \rightarrow \pi^*$ transitions, were in all cases correctly predicted. For $n \rightarrow \pi^*$ transitions the solvent shifts tend to be slightly overestimated while the opposite applied to $\pi \rightarrow \pi^*$ transitions. This was ascribed to the neglect of certain intermolecular interactions, e.g., short-range repulsion and dispersion, as well as the limitations inherent in current exchange-correlation functionals.

Acknowledgment. We thank the Danish Center for Scientific Computing for the computational resources. J.K. thanks the Danish Natural Science Research Council/The Danish Councils for Independent Research for financial support.

Supporting Information Available: Tables showing solvated molecular geometries and force field parameters for pyridine, coumarin 151, and coumarin 153, the parameters of the water molecules as used in the MD simulation, LoProp based force fields, and Ahlström and TIP3P force fields used for the water molecules in the PE-DFT calculations. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) van Mourik, T. *Philos. Trans. R. Soc. London, A* **2004**, *362*, 2653.

(2) Ochsenfeld, C.; Kussmann, J.; Lambrecht, D. S. *Rev. Comput. Chem.* **2007**, *23*, 1.

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

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

(5) Gordon, M. S.; Freitag, M. A.; Bandyopadhyay, P.; Jensen, J. H.; Kairys, V.; Stevens, W. J. *J. Phys. Chem. A* **2001**, *105*, 293.

(6) Gordon, M. S.; Slipchenko, L. V.; Li, H.; Jensen, J. H. *Annu. Rep. Comput. Chem.* **2007**, *3*, 177.

(7) Nielsen, C. B.; Christiansen, O.; Mikkelsen, K. V.; Kongsted, J. *J. Chem. Phys.* **2007**, *126*, 154112.

(8) Yoo, S.; Zahariev, F.; Sok, S.; Gordon, M. S. *J. Chem. Phys.* **2008**, *129*, 144112.

(9) Arora, P.; Slipchenko, L. V.; Webb, S. P.; DeFusco, A.; Gordon, M. S. *J. Phys. Chem. A* **2010**, *114*, 6742.

(10) Slipchenko, L. V. *J. Phys. Chem. A* **2010**, *114*, 8824.

(11) Jorgensen, W. L. *J. Chem. Theory Comput.* **2007**, *3*, 1877.

(12) Yu, H.; van Gunsteren, W. F. *Comput. Phys. Commun.* **2005**, *172*, 69.

(13) McRae, E. G. *J. Phys. Chem.* **1957**, *61*, 562.

(14) Kongsted, J.; Mennucci, B. *J. Phys. Chem. A* **2007**, *111*, 9890.

(15) Pavone, M.; Cimino, P.; De Angelis, F.; Barone, V. *J. Am. Chem. Soc.* **2006**, *128*, 4338.

(16) Brancato, G.; Barone, V.; Rega, N. *Theor. Chem. Acc.* **2007**, *117*, 1001.

(17) Aidas, K.; Møgelhøj, A.; Nilsson, E. J. K.; Johnson, M. S.; Mikkelsen, K. V.; Christiansen, O.; Söderhjelm, P.; Kongsted, J. *J. Chem. Phys.* **2008**, *128*, 194503.

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

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

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

(21) Gao, J.; Xia, X. *Science* **1992**, *258*, 631.

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

(23) Sałek, P.; Vahtras, O.; Helgaker, T.; Ågren, H. *J. Chem. Phys.* **2002**, *117*, 9630.

(24) Olsen, J.; Jørgensen, P. *J. Chem. Phys.* **1985**, *82*, 3235.

(25) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic-Structure Theory*; Wiley: New York, 2000.

(26) DALTON, a molecular electronic structure program, Release 2.0, 2005; see: http://www.kjemi.uio.no/software/dalton/dalton.html.

(27) Thole, B. T. *Chem. Phys.* **1981**, *59*, 341.

(28) van Duijnen, P. T.; Swart, M. *J. Phys. Chem. A* **1998**, *102*, 2399.

(29) Aidas, K.; Møgelhøj, A.; Kjær, H.; Nielsen, C. B.; Mikkelsen, K. V.; Ruud, K.; Christiansen, O.; Kongsted, J. *J. Phys. Chem. A* **2007**, *111*, 4199.

(30) Olsen, J. M.; Aidas, K.; Mikkelsen, K. V.; Kongsted, J. *J. Chem. Theory Comput.* **2010**, *6*, 249.

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

(32) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.

(33) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999.

(34) Frisch, M. J. *Gaussian 03*, Revision B.05; Gaussian: Wallingford, CT, 2004.

(35) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361.

(36) Gagliardi, L.; Lindh, R.; Karlström, G. *J. Chem. Phys.* **2004**, *121*, 4494.

(37) 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.

(38) Jorgensen, W. L.; McDonald, N. A. *J. Mol. Struct. (THEOCHEM)* **1998**, *424*, 145.

(39) Cinacchi, G.; Ingrosso, F.; Tani, A. *J. Phys. Chem. B* **2006**, *110*, 13633.

(40) Pranata, J.; Wierschke, S. G.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1991**, *113*, 2810.

(41) Ahlström, P.; Wallqvist, A.; Engström, S.; Jönsson, B. *Mol. Phys.* **1989**, *68*, 563.

(42) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, U.K., 1987.

(43) Linse, P. *MOLSIM 3.3.0*; Lund University: Lund, Sweden, 2001.

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

(45) Peach, M. J. G.; Helgaker, T.; Sałek, P.; Keal, T. W.; Lutnæs, O. B.; Tozer, D. J.; Handy, N. C. *Phys. Chem. Chem. Phys.* **2006**, *8*, 558.

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

(47) Aidas, K.; Kongsted, J.; Osted, A.; Mikkelsen, K. V.; Christiansen, O. *J. Phys. Chem. A* **2005**, *109*, 8001.

(48) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. R. *J. Comput. Chem.* **1983**, *4*, 294.

(49) Jorgensen, W. L. *J. Am. Chem. Soc.* **1981**, *103*, 335.

(50) Batsanov, S. S. *Inorg. Mater.* **2001**, *37*, 871.

(51) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441.

(52) Aidas, K. *Whirlpool, a QM/MM Analysis program*, version 1.00; 2010.

(53) Söderhjelm, P.; Krogh, J. W.; Karlström, G.; Ryde, U.; Lindh, R. *J. Comput. Chem.* **2007**, *28*, 1083.

(54) Renge, I. *J. Phys. Chem. A* **2009**, *113*, 10678.

(55) Moskvin, A. F.; Yablonskii, O. P.; Bondar, L. F. *Theor. Exp. Chem.* **1966**, *2*, 469.

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

(57) Clark, L. B.; Peschel, G. G.; Tinoco, I., Jr. *J. Phys. Chem.* **1965**, *69*, 3615.

(58) Ernsting, N. P.; Asimov, M.; Schäfer, F. P. *Chem. Phys. Lett.* **1982**, *91*, 231.

(59) Andon, R. J. L.; Cox, J. D.; Herington, E. F. G. *Trans. Faraday Soc.* **1954**, *50*, 918.

(60) Gayathri, B. R.; Mannekutla, J. R.; Inamdar, S. R. *J. Mol. Struct.* **2008**, *889*, 383.

(61) Shirota, H.; Castner, E. W., Jr. *J. Chem. Phys.* **2000**, *112*, 2367.

(62) Hoyau, S.; Ben Amor, N.; Borini, S.; Evangelisti, S.; Maynau, D. *Chem. Phys. Lett.* **2008**, *451*, 141.

(63) Kaminski, J. W.; Gusarov, S.; Wesolowski, T. A.; Kovalenko, A. *J. Phys. Chem. A* **2010**, *114*, 6082.

(64) Caricato, M.; Mennucci, B.; Scalmani, G.; Trucks, G. W.; Frisch, M. J. *J. Chem. Phys.* **2010**, *132*, 084102.

(65) Aidas, K.; Mikkelsen, K. V.; Mennucci, B.; Kongsted, J. *Int. J. Quantum Chem.*, in press (DOI: 10.1002/qua.22624).

(66) Mata, R. A. *Mol. Phys.* **2010**, *108*, 381.

(67) Sneskov, K.; Matito, E.; Kongsted, J.; Christiansen, O. *J. Chem. Theory Comput.* **2010**, *6*, 839.

(68) Epifanovsky, E.; Kowalski, K.; Fan, P.-D.; Valiev, M.; Matsika, S.; Krylov, A. I. *J. Phys. Chem. A* **2008**, *112*, 9983.

(69) Sulpizi, M.; Röhrig, U. F.; Hutter, J.; Rothlisberger, U. *Int. J. Quantum Chem.* **2005**, *101*, 671.

CT1003803

JCTC Journal of Chemical Theory and Computation

Prediction of ^{57}Fe Mössbauer Parameters by Density Functional Theory: A Benchmark Study

Arteum D. Bochevarov,[†] Richard A. Friesner,*[,†] and Stephen J. Lippard[‡]

Department of Chemistry, Columbia University, New York, New York 10027, United States and Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

Received July 18, 2010

Abstract: We report the performance of eight density functionals (B3LYP, BPW91, OLYP, O3LYP, M06, M06-2X, PBE, and SVWN5) in two Gaussian basis sets (Wachters and Partridge-1 on iron atoms; cc-pVDZ on the rest of atoms) for prediction of the isomer shift (IS) and quadrupole splitting (QS) parameters of Mössbauer spectroscopy. Two sources of geometry (density functional theory optimized and X-ray) are used. Our data set consists of 31 iron-containing compounds (35 signals), the Mössbauer spectra of which were determined at liquid helium temperature and where the X-ray geometries are known. Our results indicate that the larger and uncontracted Partridge-1 basis set produces slightly more accurate linear correlations of electronic density used for prediction of IS and noticeably more accurate results for the QS parameters. We confirm and discuss the earlier observation of Noodleman and co-workers that different oxidation states of iron produce different IS calibration lines. The B3LYP and O3LYP functionals have the lowest errors for either IS or QS. BPW91, OLYP, PBE, and M06 have mixed success, whereas SVWN5 and M06-2X demonstrate the worst performance. Finally, our calibrations and conclusions regarding the best functional to compute the Mössbauer characteristics are applied to candidate structures for the peroxo and Q intermediates of the enzyme methane monooxygenase hydroxylase (MMOH) and are compared to experimental data in the literature.

Introduction

Mössbauer spectroscopy[1,2] is a valuable experimental technique, especially suitable for the study of iron-containing substances. It is sensitive to the distribution of charge density around the ^{57}Fe nuclei and is used to probe the geometric and electronic structures of molecular systems of all types: from simple inorganic salts to metalloproteins. The investigation of complex catalytic cycles of iron-containing proteins has particularly benefited from Mössbauer spectroscopy.[3–8]

Recently, there has been a surge of applications of density functional theory (DFT) to predict Mössbauer spectra.[9–14,16–20] DFT, combining rigorous physics and numerical parametrization, serves as a de facto standard in theoretical investigations of biological processes that are dependent on quantum effects. The following parameters are most often computed and compared to experiment: the isomer shift (IS) δ, quadrupole splitting (QS) ΔE, and asymmetry parameter η. Several papers report an acceptable agreement between theory and experiment[14,17,20,21] for medium-sized complexes, suggesting that DFT can be regarded as a well-behaved method for predicting Mössbauer characteristics. This situation appears to be in contrast to many other applications where pure DFT suffers from drawbacks such as defective description of dispersion interaction and systematic errors in predicting thermochemistry[22,23] that need to be remedied by separate treatments.[24,25] However, several cases have been described in which DFT completely failed to agree with experiment in accounting for the quadrupole splitting parameter,[26,27] arguably owing to DFT's single-reference character.

Despite extensive applications of DFT to computing Mössbauer parameters there have been very few studies with the goal of assessing the accuracy of various functionals and

* Corresponding author e-mail: rich@chem.columbia.edu.
[†] Columbia University.
[‡] Massachusetts Institute of Technology.

10.1021/ct100398m
Published on Web 11/09/2010

basis sets on a broad set of iron-containing compounds. The complexity of novel DFT functionals and quantity of empirical parameters on which they are based have reached a point at which it is difficult to reliably predict their accuracy, judging only from the physical approximations that went into constructing the functional. In this situation benchmarking using a diverse data set becomes a necessary step. A different view of the problem is provided by the practical difficulty encountered when predicting the geometries of intermediates in catalytic cycles of iron-containing proteins. These intermediates are of often unknown structure, are typically unstable, and cannot be crystallized. When the experimental Mössbauer spectra are available, the computation of IS and QS allows one to choose the most probable candidates from the variety of considered model structures. However, the tendency of different functionals to produce somewhat different IS and especially QS parameters may not only lead to confusion (provided a few functionals have been used in the study)[3] but also may create the possibility of reconciling the experimental values with the theoretical ones in ref 20 or several candidate structures. Determining the 'error bars' and placing the functionals into 'accuracy classes' with respect to their ability to predict Mössbauer characteristics would alleviate the uncertainty of such situations. At the end of this paper we demonstrate the application of this strategy to the enzyme methane monooxygenase hydroxylase.

Several authors have made progress in assessing the quality of the functionals with respect to the accuracy of their Mössbauer parameter predictions. Nemykin and Hadt compared the performance of BPW91 and B3LYP on a broad range of ferrocene derivatives and other compounds.[17] The group of Oldfield tested these two functionals on porphyrin derivatives[12] and some unusual 2- and 3-coordinate Fe(II) complexes.[21] The first of these works concludes that B3LYP performs slightly better than BPW91 (except for ferrocene derivatives, where it seriously overestimates quadrupole splittings), whereas the second found no significant difference between the accuracy of these two functionals. At the same time, the group of Noodleman uses another set of functionals, OLYP, OPBE, and PW91,[4,15,16,28,29] with variable success in the case of PW91. Römelt et al.[19] evaluated five diverse functionals (BP86, TPSS, TPSSh, B3LYP, and B2PLYP) based on their accuracy in predicting isomer shift on a set of 20 compounds.

Even though these efforts set useful guidelines, a more thorough benchmarking is desired in order to improve some of the aspects of these studies. Below we review some of their deficiencies and describe the improvements that we have made to achieve our goal in the present study.

All previous investigations considered the Mössbauer data points collected at different temperatures: from room temperature to 4.2 K. Although this choice may have been driven by the relative paucity of Mössbauer data, such data sets nevertheless seem highly questionable because Mössbauer spectra typically have a strong temperature dependence. The dependence arises from different populations of low-lying excited states as well as the corresponding geometry changes. Theoretical calculations of ground states that do not take thermal smearing into account should be compared to the Mössbauer spectra measured at the lowest practical temperature, 4.2 K. The linear extrapolations of both IS and QS to 4.2 K performed by Noodleman and co-workers[16,28] may improve the quality of the data set, but this operation should be avoided in a benchmark study because nontrivial temperature dependence of Mössbauer parameters or absence thereof is very common (see, for example, refs 30 and 79). The data set employed in the present work is restricted solely to spectra taken at liquid helium temperature. Another important restriction on a data set is constructing it exclusively from the molecular systems for which X-ray and Mössbauer spectra are known and belong to the same compound. Very often simple ions (such as $FeCl_4^{2-}$) included in the data set have multiple Mössbauer spectra available, corresponding to different counterions and/or crystallization conditions.[30] Notable changes in the geometry of these ions, apparent in X-ray results, naturally give rise to slightly different Mössbauer parameters and sometimes lead to the citation of different experimental values for the same ion, comparing them to Mössbauer data obtained from DFT-optimized geometries, as in the case of $FeCl_4^{2-}$ in refs 13 and 29. Whether or not the DFT can accurately predict Mössbauer parameters corresponding to different crystallographic variants of the same ion is a separate question which we leave for another study. The experimentally determined geometries selected for our test set were unimpaired: they did not have any parts of the structure missing and did not require additional reconstruction. One more restriction on the test set entries is that the multiplicity and the spin state of the iron atoms is known for certain, and those displaying spin crossover were excluded.

The issues discussed in the previous paragraph pertain to the stricter selection of the molecular systems for the test set. However, it is possible to also introduce improvements of an 'extensive' nature: a larger set of parameters and conditions under which benchmarking is performed. Here, we study the performance of eight functionals: B3LYP,[31,35,36] BPW91,[32–34] M06,[40] M06-2X,[40] O3LYP,[31,37] OLYP,[31,38] PBE,[39] and SVWN5.[41–43] The first two of these have been widely used for predicting Mössbauer parameters. M06 and M06-2X are two interesting new functionals worth investigating in the light of their superior accuracy in various calculations.[40] M06 is parametrized for organometallic and inorganometallic chemistry, whereas M06-2X should have a well-balanced, universal applicability. However, there is no guarantee that the optimization protocols used to populate over 30 adjustable parameters for these functionals leads to improved performance in predicting Mössbauer spectra that were not included in the training set.

Here, M06 and M06-2X are applied for prediction of the Mössbauer parameters for the first time. O3LYP and OLYP are built upon Handy's OPTX exchange functional and showed superior accuracy in predicting electronic densities among several functionals[44] as well as in other benchmarks.[45] The last two functionals, PBE and the local density approximiation (LDA) SVWN5, are considered more 'physical' rather than 'empirical', and even though they are not functionals of choice in numerous modern applications, at

times they display unexpected prominence. For LDA's impressive result for the dispersion interaction energetics, see ref 46. This set is the largest selection of functionals to be studied systematically in the context of the Mössbauer spectra. Unfortunately, we were unable to evaluate such promising modern functionals as TPSSh[47] and B2PLYP[48] because they were not available in our computational package, Jaguar. B2PLYP was recently shown to yield the best linear fits of the isomer shift.[19] However, the performance of B2LYP was only slightly superior to that of B3LYP, which is included in this study.

All calculations in this work were performed with Gaussian-type (GT) basis sets. The alternative is Slater-type (ST) basis sets, which better approximate orbitals in the proximity of the nuclei. However, GT bases are much more computationally tractable, which makes software capable of handling ST rare. Zhang and Oldfield compared the effect of these two types of basis sets on the predicted Mössbauer parameters and found no apparent advantage of ST versus GT bases.[21] In this work we use two Gaussian bases on iron nuclei: Partridge[49] and Wachters.[50] They both contain very large exponential coefficients on *s*-primitives to facilitate the reproduction of nuclear cusp, but the Partridge basis set is completely uncontracted which makes it significantly larger than Wachters. The groups of Neese[13] and Filatov[20,51] prefer uncontracted bases, but other authors[12,17] suggest that the quality of the Wachters basis set (with contraction and smaller exponential coefficients) is sufficient. Kurian and Filatov,[18] who computed the isomer shift from Filatov's nonempirical approach,[51] also conclude that completely uncontracted bases are not necessary to achieve good accuracy. Comparison of the accuracy of the contracted and uncontracted basis sets on a large data set seems logical.

Another logical comparison for a comprehensive benchmarking study is that between Mössbauer parameters obtained from the X-ray geometries and DFT-optimized geometries. Both sources introduce different types of errors. The coordinates of heavy atoms in X-ray geometries, dependent on resolution and thermal atomic motion, are generally considered highly accurate for most measurements. However, such measurements are not always available. Additionally, the positions of the hydrogen atoms cannot sometimes be extracted from the X-ray data, and such atoms have to be added by means of computational algorithms. DFT-optimized geometries, although readily procurable, are subject to errors inherent in functionals and finite basis sets. Functionals such as B3LYP and O3LYP have established an excellent reputation in optimizing the geometries of organic molecules. Their performance on metal-containing systems is not so well tested and is therefore less reliable. Despite this deficiency, Han and Noodleman argue that it is more fair to use DFT-optimized geometries in DFT Mössbauer calculations because these geometries correspond to energetic minima within the DFT model. It is perhaps assumed that the reported X-ray geometries differ, through either crystallization, distortion, or bad resolution, from the 'relaxed' configurations for which the Mössbauer parameters are measured, and DFT optimization in principle alleviates this proposed effect. It must be noted that DFT optimizations are typically performed at a lower level of theory than that used for subsequent Mössbauer calculations, so that in such cases the single-point electronic density still does not correspond to the optimal geometry. Nevertheless, it is not clear which of these errors dominates. Nemykin and Hadt,[17] who studied the question, observed little variation of results, but their data set was not very diverse and predominantly consisted of high- and medium-temperature experimental data points.

In this work we compute the Mössbauer parameters from both X-ray and DFT-optimized geometries using a typical protocol for the optimization (see the Computational Details section for the details). When we pick a particular optimization protocol out of many possible protocols we do not attempt to conclude after the analysis of the results which source of the geometries (theoretical or experimental) produces better accuracy, except in a few cases where there are large differences in the results as compared to experiment. In comparing results obtained from both types of geometries we are only trying to assess the particular optimization method we have chosen and also understand to what degree the small changes of the positions of atoms influence the Mössbauer results.

To summarize, we investigate the prediction of IS and QS by 8 functionals combined with 2 Gaussian basis sets of different composition and 2 sources of geometries (X-ray and optimized). The experimental data for our test set consisting of 31 compounds and 35 individual Mössbauer signals was obtained at liquid helium temperature. We believe ours to be the most comprehensive such study to date.

Finally, we apply our optimized DFT Mossbauer protocol to computation of Mossbauer spectra for various intermediates in the catalytic cycle of methane monooxygenase hydroxylase (MMOH), a nonheme diiron protein that converts methane to methanol under room temperature conditions.[53,54] Using criteria developed in our benchmark study, we achieve good agreement between calculated and experimental Mossbauer data for a number of MMOH structures. The benchmarking data are essential in assessing the degree to which Mossbauer comparisons can be used to confirm, or rule out, assignment of three-dimensional structures to experimentally observed intermediates.

Computational Details

All DFT calculations were performed using the locally modified Jaguar 7.5 program.[52] Although Jaguar commonly employs the pseudospectral approach,[55–59] speeding up the solution of the Kohn−Sham equations significantly, we computed all the integrals analytically so as not to introduce a potential source of error in the benchmarking study. The unrestricted Kohn−Sham equations converged to 10^{-8} hartree in energy, and fine grids were employed for computation of densities on iron atoms. The medium quality of the wave function (converged to 10^{-6} hartree) and coarse grids were generally sufficient in the case of the Wachters basis but not in the case of the Partridge basis set (see ref 13 for a detailed discussion of this effect).

The uncontracted Partridge basis set had the structure (20s,12p,9d), and the contraction scheme of the Wachters was 62111111/331211/3111. The basis set on all the noniron

Table 1. Test Set of Compounds Used for Benchmarking in This Work[a]

no.	system	code	Fe oxidation	spin	δ, mm/s	$\lvert\Delta E\rvert$, mm/s	J, cm^{-1}	R factor, %	ref
1	$Fe_2(O_2CH)_2(BIPhMe)_2$	SISKOU	+2	0	1.26	2.56	~0	4.1	78
			+2		1.25	3.30			
2	$Fe(HB(mtda^R)_3)_2$	JOHCEP	+2	0	0.49	0.26		4.83	79
3	$Fe_2(OAc)_2(TPA)_2^{2+}$	VUNMIA	+2	0	1.12	3.33	~1	4.7	80
4	$Fe_2(ImH)_2(XDK)(O_2CPh)_2(MeOH)$	YUZKAF10	+2	0	1.35	3.04	−0.51	8	81
			+2		1.12	2.83			
5	$Fe_2(py)_2(O_2CAr^{Mes})_4$	XIGDIA	+2	0	1.14	3.23	30	5.44	82
6	Fe(OEP)CO	YEQPOA	+2	0	0.27	1.84		2.85	83
7	Fe(OEP)	DEDWUE	+2	1	0.63	2.55		7.2	84
8	Fe(OEC)	BUYKUB10	+2	1	0.62	1.71		6	84
9	$Fe_3(SPh)_6(CO)_6$	FATBOR	+2	2	1.00	2.00		3.9	85
			+2		0.10	0.22			
10	$Fe_2(H_2O)(O_2CPh)_4(TMEN)_2$	VUPJUL	+2	2	1.25	3.11	small	4.4	86
			+2		1.26	2.70			
11	$Fe_2(H_2O)(OAc)_4(TMEN)_2$	VUPJOF	+2	2	1.27	2.75	small	5.5	86
12	$Fe(DTSQ)_2^{2-}$	PTSQFE10	+2	2	0.67	4.01		5.5	87
13	$Fe(SPh)_4^{2-}$	PTHPFE10	+2	2	0.66	3.24		4.7	87
14	$Fe_2O(HBpz_3)_2(OAc)_2$	CACZIP10	+3	0	0.52	1.60	−121	4	88
15	$Fe_2(OH)(HBpz_3)_2(OAc)_2$	COCJIN10	+3	0	0.47	0.37	~−17	4.8	93
16	$Fe_2(OH)(O_2P(OPh)_2)_3(HBpz_3)_2^{2+}$	PIMTAG	+3	0	0.44	0.44	~−15	6.9	93
17	$Fe_2O(Piv)_2(Me_3TACN)_2^{2+}$	ZOCPEM	+3	0	0.48	1.54	−111	6.6	89
18	$Fe_2(O)_2(6\text{-}Me_3\text{-}TPA)_2^{2+}$	YOCKAC	+3	0	0.50	1.93	54	6.5	90
19	$Fe_2(NO)_2(Et\text{-}HPTB)(O_2CPh)$	RABHAD	+3[1]	0	0.67	1.44	−23	8.5	91
20	$Fe_2(S\text{-}t\text{-}Bu)_2(NO)_2$	GIDKIN02	+3	0	0.15	0.90		2.11	92
21	$(Fe(Me_3TACN)(TTC))_2O$	YOHMOX	+3	0	0.46	1.41	−90	6.5	94
22	$Fe_2O(TMIP)_2(OAc)_2^{2+}$	JIGNUI	+3	0	0.52	1.61	−120	5.7	95
23	$Fe(OEP)(4\text{-}NMe_2Py)_2^{2+}$	VOFLOR	+3	1/2	0.26	2.15		6.7	96
24	$Fe(S\text{-}t\text{-}Bu)_3NO$	WEDXAF	+3	3/2	0.26	0.46		2.93	92
25	$FeCl_5(H_2O)^{2-}$	VOCBAQ	+3	5/2	0.49	0.56		3.2	97
26	$Fe(SEt)_4^-$	CANDAW10	+3	5/2	0.25	0.62		5	98
27	$Fe(NO)_2(S(p\text{-}Me)Ph)_2^-$	SONMUE	+3	5/2	0.18	0.69		3.55	99
28	$FeCl(MBTHx)_2$	CELVEU	+3	5/2	0.43	0.98		3.8	100
29	$(Fe(TAML)_2)_2O^-$	KAJBIH	+4	0	−0.07	3.30	>100	5.32	101
30	$Fe(PPh_3)_2(\text{“S2”})_2$	SOCVUB	+4	1	0.16	1.52		5.9	102
31	$Fe(PPh_3)(\text{“S2”})_2$	SOCWAI	+4	1	0.12	3.03		4.6	102

[1] Deduced from Mössbauer, magnetic susceptibility, and SCF-Xα data. [a] The six-letter codes refer to the Cambridge Structural Database identifiers. The coupling constant J is given when known. The ligands are encoded as follows: BIPhMe = bis(1-methylamidazol-2-yl)phenylmethoxymethane, DTSQ = bis(dithiodithiosquarato-*S*,*S*′) dianion, Et-HPTB = *N*,*N*,*N*′,*N*′-tetrakis(*N*-ethyl-2-benzimidazolylmethyl)-1,3-diaminopropane), ImH = imidazole, $HB(mtda^R)_3$ = tris(mercaptothiadiazolyl)borate, $HBpz_3$ = hydrotris-1-(pyrazolyl)borate, HO_2CAr^{Mes} = 2,6-bis(mesityl)benzoic acid, MBTHx = bis(*N*-methylbenzothiohydroxamato) anion, Me_3TACN = 1,4,7-trimethyl-1,4,7-triazacyclonane, OEP = dianion of octaethylporphyrin, OEP = dianion of *trans*-7,8-dihydro-octaethylporphyrin, Piv = pivalate, “S2” = 1,2-benzenedithiolato-*S*,*S*′ dianion, TAML = tetra-amido macrocyclic ligand,[101] TMEN = *N*,*N*,*N*′,*N*′-tetramethylethylenediamine, TMIP = tris(methylimidazol-2-yl)phosphine, TPA = tris(2-pyridylmethyl)amine, TTC = tetrachlorocatecholato-*O*,*O*′ dianion, XDK = acid anion of *m*-xylenediamine bis(Kemp's triacid)-imide.

atoms in single-point calculations was cc-pVDZ. These basis sets can be obtained from the EMSL Basis Set Exchange database.[60] The d orbitals carried pure orbital momentum. For geometry optimizations, the pseudospectral approach was used, with a B3LYP functional, the LACVP* basis set[61] on iron atoms, and 6-31G* on all other atoms. Many complexes in our data set contained over 100 atoms (the largest being 134 atoms), so this approximation level was necessary. We realize that the offered optimization model does not serve as the single 'ultimate' representation of all the optimization models, to be evaluated against the X-ray crystallography in its function to furnish accurate geometries; rather, it is a functional method comparable to that which served us reliably in the past.[62]

The models representing the compounds from Table 1 were initially constructed from the experimental crystal structures. Counterions and solvent molecules were not included. In constructing the models, we took care to preserve as much of the original structure as possible. In some rare cases, however, we had to replace extremely large (and seemingly not important for the electric field gradient on the iron atoms) groups in the ligands by smaller ones, to reduce the size of the system. These structures were used as the starting point in the geometry optimizations. All X-ray-based geometries used in our calculations are available in the Supporting Information.

Data Set. In constructing the data set, apart from the criteria for the soundness of the X-ray and Mössbauer experimental data, we included diverse chemical structures. The data set used in this work is presented in Table 1. It comprises compounds with the iron atoms in oxidations states +2, +3, and +4 and in low- as well as high-spin states. The total spin is $n/2$ where n varies from 0 to 5.

The main chemical classes from Table 1 are as follows, where the numbers in parentheses refer to the identity numbers in the table: (i) diiron(II) (1, 3−5, 10, and 11) and diiron(III) (14−18, 22) ligand complexes with bridging carboxylates, which may be regarded as structural models for the active sites of bacterial multicomponent monooxy-

genases,[63] (ii) nitrosyl complexes (19, 20, 24, 27), (iii) porphyrin derivatives (6−8), and (iv) compounds with multiple sulfur atoms in the coordination sphere (9, 12, 13, 26, 30, 31). Complexes 6 and 9 contain CO. With the exception of $[Fe(H_2O)Cl_5]^{2-}$ (entry 25), we did not include simple complexes containing 7−15 atoms due to the absence of a complete package of experimental data. Many Mössbauer studies of such complexes were conducted over four decades ago and are not accompanied by X-ray crystal structures, whereas the recently deciphered crystal structures with counterions such as $Fe(H_2O)_6^{3+}$ and $Fe(CN)_6^{3-}$ offer little interesting in the configuration of these counterions to warrant investigation by Mössbauer spectroscopy.

Some of these complexes have been included in the test sets of others,[15,16,28] but several were investigated theoretically in this work for the first time. No systematic examination of the electronic structure of these compounds was attempted, for to do so would reach beyond the benchmarking objectives of this work. Although we discuss some of the numbers produced for the individual compounds, our focus is on statistics and comparison of the general performance of theoretical methods with respect to the IS and QS properties.

Counterions or crystal inclusions were not included in the systems in our test set for the following reasons. First, the location of counterions was not always available in the crystallographic data. Including counterions in some systems and omitting them in others might introduce a bias that would be undesirable in a benchmark study. Second, some counterions were fairly large organic systems, containing more than a dozen atoms each, such as BPh_4^-, cocrystallized with our VUNMIA system.[80] Their inclusion and treatment within a DFT scheme would make some of our calculations prohibitively expensive. Third, our experimentation with environment showed a marginal influence of counterions and cocrystallized neutral molecules on the Mössbauer parameters of the iron complex under study. This observation is in accord with an earlier study[21] that reports no change in predicted isomer shift and quadrupole splitting upon inclusion of a counterion in the DFT calculations. Although some theoretical studies provide special treatment of environment,[20] many other do not and still report accurate Mössbauer parameter predictions.[12,18,19]

The experimental isomer shifts varied between 1.35 and −0.07 mm/s, a typical experimental range for most iron-containing complexes. FeO_4^{2-}, often placed at the bottom of test sets for its extraordinarily low isomer shift value,[64] did not satisfy a number of the filtering criteria mentioned in the Introduction (for instance, no data at liquid helium temperature). There is a noticeable gap in experimental isomer shifts between 0.67 and 1.00 mm/s in our data set. This region, marking a transition between diiron(III) and diiron(II) formal redox states, is also either very sparsely populated or empty in other researchers' data sets.[15–17,29] The absolute value of the experimental quadrupole splittings ranged from 0.22 to 4.01 mm/s. The slightly unusual maximal value corresponds to distorted complex 12, the DFT-predicted quadropole splitting of which is an outlier in ref 28. Apart from the gap between 1.0 and 1.4 mm/s (which does not seem to be a rare interval), the rest of the quadrupole splittings in Table 1 are distributed more or less uniformly.

Isomer Shift. Multiple studies outline the quantum theory of the Mössbauer isomer shift (see refs 65 and 66 and references therein). The parametrization method, which represents the isomer shift as depending linearly on the electronic density on iron nuclei, $\rho(0)$

$$\delta = \alpha\rho(0) + C \quad (1)$$

where α and C are empirically determined coefficients, has been found to be accurate and robust in a great number of DFT publications.[13,15,17,29] This linear relationship has also been observed in the complete active space self-consistent field (CASSCF) calculations on a small test set.[67] For an alternative, nonempirical computation of the isomer shift, which employs derivatives of the electronic energy with respect to the size of the nonpoint-like nucleus and which is more laborious, see refs 18, 51, 68, and 69. The slope α and intercept C are typically obtained with the least-squares method using a parametrization set of experimental isomer shifts and theoretically computed densities $\rho(0)$. Several high-quality studies that attempt to match experimental and theoretical values of the slope α are available.[19,20,70] Because the densities on the nuclei vary with the basis set and functional used,[44] each such combination necessitates a separate parametrization (eq 1). Since many researchers are likely to prefer different bases and ways of computing the wave function they will have to perform a parametrization of their own, but one way to simplify their task is to make available the atomic coordinates of the entries of a large and reliable test set as we do in the present work.

The mean unsigned errors (MUEs) obtained from the X-ray geometries for 8 functionals and both basis sets are given in Table 2. The results can be broadly classified into two categories. For hybrid functionals such as B3LYP or O3LYP, the data lie on a single straight line, providing a parametrization (converting the calculated IS into the experimental IS) that can be applied to an arbitrary complex. For many of the other functionals, including the LDA, gradient-corrected functionals, and the M06 and M06-2X functionals, the data can be better characterized as lying on two parallel straight lines, with the location of each point depending upon whether the Fe atoms have a formal oxidation state of +2 or +3/+4. Using a dual parametrization for these functionals substantially improves the results, although it can create problems for molecules where the assignment of formal oxidation state is problematic or ambiguous, e.g., Fe−nitrosyl complexes, which are well known to have a complex electronic structure in many cases.[72,73] We report results using both types of parametrizations for all functionals for completeness, although the difference for the hybrid functionals is quite small. The ability to use a single parametrization is an advantage for the hybrid functionals, making them (in our view) the approach of choice for computation of IS values.

The basis set effects are relatively minor, although the larger and more flexible Partridge basis set unsurprisingly yields slightly better linear correlations overall. Either basis set can be judged suitable for practical applications. The

Table 2. Isomer Shift MUE's for the Data Set Presented in Table 1 Computed with the Wachters and Partridge-1 Basis Sets and Tight Convergence Criteria[a]

	Wachters			Partridge-1		
functional	MUE (+2)	MUE (+3, +4)	all	MUE (+2)	MUE (+3, +4)	all
B3LYP	0.0283	0.0285	0.0324	0.0272	0.0238	0.0296
BPW91	0.0358	0.0289	0.0772	0.0349	0.0253	0.0733
M06	0.0334	0.0352	0.0370	0.0524	0.0344	0.0450
M06-2X	0.0380	0.0342	0.0426	0.0581	0.0345	0.0564
OLYP	0.0352	0.0321	0.0756	0.0372	0.0269	0.0703
O3LYP	0.0320	0.0291	0.0461	0.0285	0.0277	0.0420
PBE	0.0355	0.0291	0.0793	0.0350	0.0253	0.0746
SVWN5	0.0355	0.0318	0.0845	0.0383	0.0274	0.0822

[a] The MUE's are given in mm/s. The +2 column includes the signals from iron atoms with the oxidation state +2. Similarly, the +3, +4 column includes the signals from iron atoms with the oxidation state +3 and +4. The 'all' column includes all signals. There are 17 signals from the +2 irons and 18 signals from the +3, +4 ions.

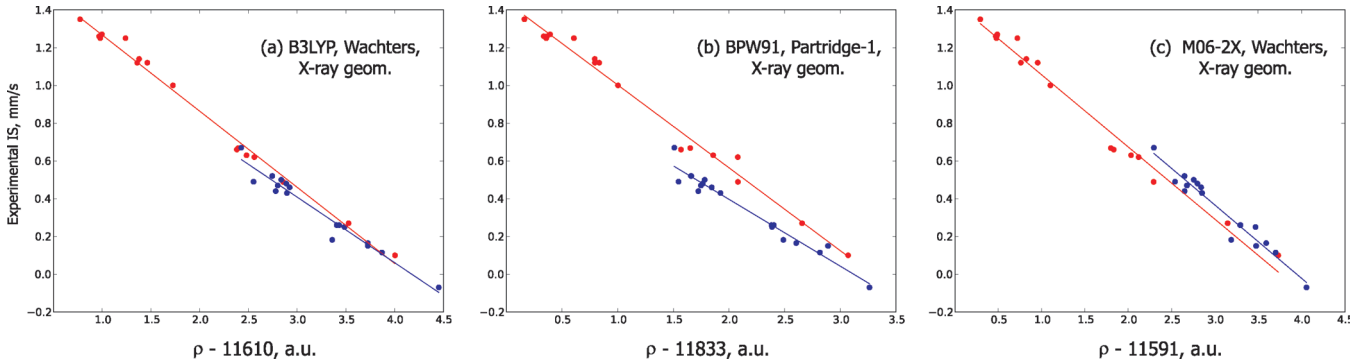


Figure 1. Some linear correlations observed between electronic density on iron and experimental isomer shift for the X-ray-based geometries. The data in red correspond to the oxidation state +2, and the data in blue correspond to oxidation states +3 and +4.

average errors reported in Table 2 represent a best-case scenario; in reality, there will be additional noise due to the use of DFT (as opposed to X-ray) geometries (discussed further below) and also introduction of novel chemistry not covered in this data set. Nevertheless, our results provide a reasonable starting point for estimating an acceptable deviation of theory from experiment, which we develop further in the Discussion section.

We report the slope and intercept constants for all combinations of the functionals and the basis sets in the Supporting Information. Direct comparison of our slope constants with the most accurate theoretical value available to date, -0.31 ± 0.04 a_0^3mm/s,[20] is not possible because our study did not include relativistic treatment. However, our slope and intersect are in excellent agreement with those obtained by others[74] in a nonrelativistic calibration for the B3LYP/Wachters combination. Additionally, if we use a correction factor of 1.30 employed to scale nonrelativistic densities to relativistic ones,[74,75] our best performing functional/basis combinations produce the α constant in the range from -0.31 to -0.32 a_0^3mm/s, in excellent agreement with the best available theoretical value.

The optimized geometries produced mixed-quality isomer shift results (see Figure 3 and compare it to Figure 1a). Most of the +2 oxidation-type densities show a very good linear correlation versus the experimental isomer shifts. However, the +3,+4-type points are visibly quite scattered, showing a poorer linear correlation. The two greatest outliers are the structurally similar nitrosyl complexes $Fe_2(S\text{-}t\text{-}Bu)_2(NO)_2$ (GIDKIN02) and $Fe(NO)_2(S(p\text{-}Me)Ph)_2^-$ (SONMUE), the former being particularly prominent. This discrepancy is all the more remarkable for the fact that the outliers of this magnitude have not been observed for any of the isomer shifts computed from the experimental geometries. Because the isomer shift of GIDKIN02 lies greatly beyond the rest of the data points regardless of the functional used, we need to seek the answer to this behavior in the optimized geometry of the complex. Table 3 compares the experimental and DFT-optimized geometries of the two above-mentioned outliers. Clearly, our DFT optimization protocol does not reproduce some of the geometrical elements of GIDKIN02 correctly. The Fe−N and Fe−S distances are greatly overestimated,

Table 3. Geometrical Parameters of the Complexes $Fe_2(S\text{-}t\text{-}Bu)_2(NO)_2$ (GIDKIN02) and $Fe(NO)_2(S(p\text{-}Me)Ph)_2^-$ (SONMUE) (which are the greatest outliers in Figure 3) Obtained from Two Sources: DFT Optimization and X-ray Crystallography[a]

	GIDKIN02		SONMUE	
parameter	DFT optimization	X-ray	DFT optimization	X-ray
d(Fe−N)	1.85	1.67	1.76	1.71
d(Fe−S)	2.38	2.25	2.38	2.33
d(N−O)	1.17	1.17	1.18	1.18
d(S−C)	1.90	1.87	1.78	1.80
∠(NFeN)	74.8	116.6	120.3	115.7
∠(SFeS)	100.1	106.2	115.0	111.9

[a] The distances are given in Angstroms, and the angles are in degrees.

and the N−Fe−N angle is much smaller than that reported by X-ray crystallography. The rest of the optimized geometry matches the experimental geometry quite accurately (compare, for example, the N−O and S−C distances which almost coincide). Some structural variations are also observed for the optimized structure of the SONMUE complex, but they are not as serious as in the case of GIDKIN02. The DFT optimization of the other complexes did not result in obvious structural incongruities.

Inspection of the isomer shift correlations for the other combinations of the functional and the basis set (not shown) obtained with the DFT-optimized geometries reveals a bigger scatter of points in comparison with the correlations obtained from the X-ray geometries, even when the obvious outliers (for instance, GIDKIN02) are excluded. Apparently, the structural inaccuracies introduced by the optimization method outweigh the inaccuracies of the experimental determination of atomic positions.

Quadrupole Splitting. The quadrupole splitting ΔE_Q in ^{57}Fe is the transition energy between the $I_z = \pm 1/2$ and $I_z = \pm 3/2$ substates of the nuclear excited state with $I = 3/2$. These substates $E_{I, Iz}$ originate from the interaction between the nuclear quadrupole moment and the electric field gradient created at the excited ^{57}Fe nucleus by the surrounding nuclei and electrons, eq 2,

$$E_{I,Iz} = \frac{1}{6}\sum_{\alpha\beta} Q_{\alpha\beta}(I, I_z)V_{\alpha\beta} \tag{2}$$

where α, β are the Cartesian coordinates, $Q_{\alpha\beta}(I, I_z)$ are the components of the nuclear quadrupole tensor, and $V_{\alpha\beta}$ are the derivatives of the environmental electric field potential V with respect to the Cartesian coordinates

$$V_{\alpha\beta} = \frac{\partial^2 V}{\partial\alpha\partial\beta} \tag{3}$$

Substituting the expression for the quadrupole moment components[2] in the nuclear shell approximation[76] into (eq 2) and taking the difference of the states with the appropriate quantum numbers we arrive at the well-known formula used to compute quadrupole splitting in zero magnetic field, eq 4,

$$\Delta E_Q = \frac{1}{2}eQV_3\left[1 + \frac{(V_1 - V_2)^2}{3V_3^2}\right]^{1/2} \tag{4}$$

where e is the absolute value of the electron charge, Q is the nuclear electric quadrupole moment for the $I = 3/2$ state (taken to be 0.16 barn[77]), and V_k are the eigenvalues of the tensor $V_{\alpha\beta}$ with the convention that V_3 has the maximal absolute value.

The sign of ΔE_Q defines the relative position of the $I_z = \pm 1/2$ and $I_z = \pm 3/2$ states but it is usually not reported in experimental studies. It is also not always reliably predicted by theoretical calculations. Because the electric field satisfies the Laplace's equation, the eigenvalues of eq 3 sum up to zero: $V_1 + V_2 + V_3 = 0$. When one of the eigenvalues, for example V_1, is much smaller in absolute value than the other two, V_3 is approximately the negative of V_2. In such a situation a small error in predicting the field V (due to the basis set or the functional) might result in V_3 changing sign, because it is always defined as the largest by absolute value, and flipping the sign of ΔE_Q. For these reasons, we compare only the absolute values of the quadrupole splittings, adopting the approach of Noodleman and co-workers.[28]

Table 4. Quadrupole Splitting MUE's for the Data Set Presented in Table 1 Computed with the Wachters and Partridge-1 Basis Sets and Tight Convergence Criteria[a]

	Wachters		Partridge-1	
functional	<2.0 mm/s	all	<2.0 mm/s	all
B3LYP	0.144	0.352	0.0952	0.337
BPW91	0.172	0.314	0.133	0.284
M06	0.140	0.332	0.114	0.304
M06−2X	0.208	0.496	0.250	0.532
OLYP	0.180	0.319	0.147	0.284
O3LYP	0.147	0.320	0.0937	0.291
PBE	0.175	0.329	0.136	0.283
SVWN5	0.175	0.410	0.149	0.384

[a] The MUE's are given in mm/s. The <2.0 column includes the signals whose experimental quadrupole splittings are less than 2.0 mm/s except the obvious outlier Fe(OEC) or BUYKUB10, which is excluded. The 'all' column includes all the signals.

Let us first discuss the data obtained from the X-ray geometries. The MUE's of the quadrupole splittings are presented in Table 4, while the actual data points for a few representative combinations of the functional and basis set are shown in Figure 2. Figure 2a and 2b demonstrates some of the best performers (O3LYP, Wachters and O3LYP, Partridge-1, respectively), but Figure 2c illustrates the worst combination M06-2X, Partridge-1. Note that the red line in these plots is simply the $y = x$ function and not the best fit. Looking at the first two of these figures we immediately notice that the prediction of small (less than about 2.0 mm/s) quadrupole splittings is significantly more accurate than that of the larger ones. This conclusion is corroborated by the actual numbers in Table 4: the MUE's for the region <2.0 mm/s in experimental values [excluding Fe(OEC), or BUYKUB10, an outlier in almost all cases] are two to three times smaller than the overall MUE's (including all points). Except for the M06-2X functional, the Partridge-1 basis is noticeably more accurate than Wachters. The two best performers for the smaller QS are, just as in the case of the IS, B3LYP and O3LYP. They produce a remarkable accuracy of less than 0.1 mm/s on this region. Because some applications are concerned solely with the absolute QS values <2.0 mm/s,[3,29] using either B3LYP or O3LYP in these studies is highly recommended. The overall best performance for the QS is shared by BPW91, OLYP, and PBE, with O3LYP following closely behind. M06, parametrized on the metal-containing compounds, brings a significant improvement over the related M06-2X functional even though it performs only slightly better than M06-2X for the IS. M06-2X gives some of the worst agreements with the experimental QS (see also Figure 2c). The second worst performer, after M06-2X is, unsurprisingly, the LDA functional SVWN5.

The observation that the QS is predicted more accurately for smaller values has not been reported before, to our knowledge. The QS data published by some other researchers[17,29] does not indicate a much greater prevalence of errors in the region above 2.0 mm/s, although the quadrupole splittings computed by Oldfield and co-workers[12] with the

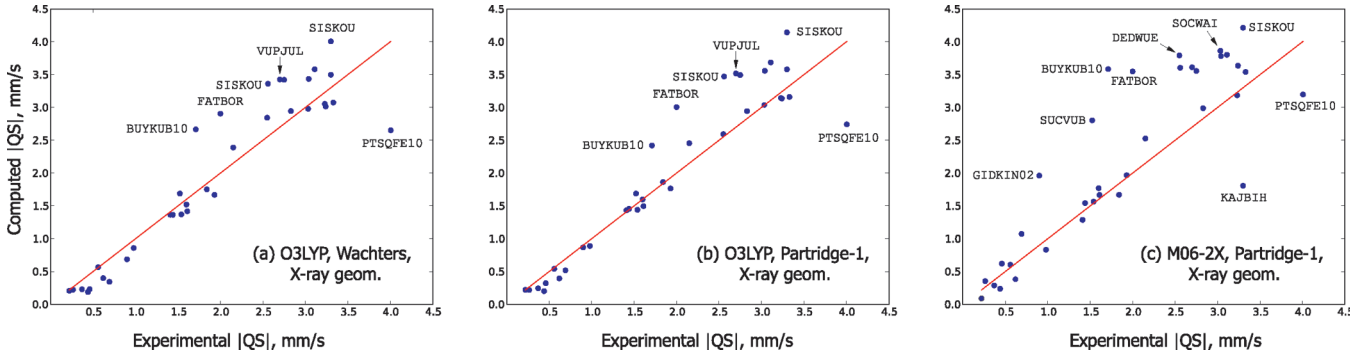


Figure 2. Some comparisons between the experimental and computed absolute values on the quadrupole splittings. The red line is $y = x$: the points lying on it represent perfect agreement between experiment and theory. The obvious outliers are indicated by their Cambridge Structural Database code which can also be found in Table 1.

B3LYP functional do appear to have greater errors for large QS values. Currently, it is not clear if our observation is simply a specious effect related to the accidental presence of several outliers with |QS| > 2.0 mm/s in our data set. For example, removal of the five outliers from Figure 2a and 2b would not result in much better accuracy of QS prediction in the region below 2.0 mm/s. Although we do not read much importance into this curious imbalance between the regions smaller and greater 2.0 mm/s, we think that more accurate experimental data (conforming to our selection criteria from the Introduction) are needed to make a conclusion.

Let us now discuss the QS outliers, paying greater attention to the functionals that perform well (O3LYP, B3LYP). The IS data did not reveal any obvious outliers as prominent as those in Figure 2. There may be several reasons for this observation. First, the determination of the QS from the experiment has inherently a greater incidence of error than the same procedure for the IS. The determination of IS involves finding the middle point between two idealized peaks of finite width. Each of these peaks is recorded with a certain error of position $\pm\Delta$, so that the middle point is determined with a maximal error of Δ (when both peaks have the positioning error of the same sign). The QS is determined as the difference between the positions of these peaks, so that the maximal error becomes 2Δ (when the peak centering errors are of different signs). If peak positioning errors dominate the rest of the errors, this analysis has a consequence that particularly large QS errors would be accompanied by particularly small IS errors and vice versa. Second, the theoretical computation of the IS is more local in nature, depending on the density of *s*-electrons on the iron nuclei, which is mostly determined by the immediate surroundings of the iron atom. The QS, in contrast, is potentially influenced by long-distance effects, namely, the overall symmetry of the electronic density, which is determined by all atoms of the system. In other words, the IS is computed from the values of the *s*-orbitals at a certain point, whereas QS requires integration over global density. Thus, QS is more sensitive to small changes in geometry. Moreover, DFT functionals have the potential to mishandle the global density because they typically incorporate either local or not entirely satisfactory gradient-corrected density formulas. Global errors in density would be less important in the typical optimization of DFT functionals than the local errors. Finally, solvation effects, counterion effects, possible protonation of ligands away from the iron atoms, and the error of placement of hydrogen atoms in X-ray structures all would produce a greater influence on the QS than on the IS.

The greatest QS outlier is a small distorted complex $Fe(DTSQ)_2^{2-}$ (PTSQFE10) with the unusually large experimental QS value (over 4.0 mm/s). This QS is greatly underestimated by almost all DFT functionals studied in this work (except M06-2X) by 1.0 mm/s or more. Interestingly, the same complex also turned out to be the largest outlier among 21 complexes in the recent work by Noodleman and co-workers.[28] The B3LYP geometry optimization starting from the X-ray geometry converged to a symmetric structure, which, however, only increased the discrepancy between experiment and theory. The inclusion of the two bulky PPh_4^{4+} counterions PTSQFE10 cocrystallized with $Fe(DTSQ)_2^{2-}$ typically did not alleviate the problem (only M06-2X produced a slight improvement). There is another compound in our data set taken from the same paper where $Fe(DTSQ)_2^{2-}$ was originally described along with its Mössbauer and X-ray data,[87] $Fe(SPh)_4^{2-}$ (PTHPFE10). Its computed QS is in an excellent agreement with the experimental value. The issue about the greatest QS outlier will perhaps be clarified when the Mössbauer characteristics of $Fe(DTSQ)_2^{2-}$ are computed by some other ab initio approach, preferably including multiple determinants, which would be feasible in view of the small size of the complex.

Among some noticeable outliers in Figure 2a and 2b are a group of diiron, dicarboxylate complexes with small coupling constants *J* (SISKOU, VUPJUL, VUPJOF). The QS of the first one (SISKOU) is regularly overestimated by all the functionals by as much as 1.0 mm/s in the worst cases, whereas the QS of the latter two vary substantially from functional to functional: either seriously overestimated (B3LYP, O3LYP, M06-2X), somewhat underestimated (PBE, OLYP), or seriously underestimated (SVWN5). VUPJUL and VUPJOF contain a protonated oxo moiety in a critical bridging position: an error in the location of the protons might have a large effect on the computed QS.

FATBOR, another (partial) outlier in QS (as predicted by B3LYP, O3LYP, and M06-2X), is a tri-iron carbonyl complex with two equivalent terminal low-spin Fe(II) atoms (having low IS and QS) and one central high-spin Fe(II) atom

(having large IS and QS). The calculations show a much better agreement with the low-spin signals and a significant overestimation of the high-spin quadrupole splittings. In any case, all eight functionals show good qualitative agreement with experiment, predicting high and low IS and QS values for two different types of Fe atoms, as described, whereas BPW91, OLYP, PBE, and SVWN5 are close to the quantitative agreement.

The last minigroup of QS outliers consists of porphyrin complexes Fe(OEC) (BUYKUB10) and Fe(OEP) (DEDWUE). The QS of Fe(OEC) in both bases ranges from 1.0 mm/s and smaller (BPW91) up to almost 4.0 mm/s (M06-2X). It is puzzling that none of the combinations of the functional and the basis comes close to the experimental quadrupole splitting (1.71 mm/s). The closest agreement is registered for O3LYP/Partridge-1 (2.42 mm/s). A related compound Fe(OEP) shows a smaller scattering range of quadrupole splittings than Fe(OEC), but the predicted QS is highly functional dependent. However, here O3LYP yields a very good agreement with its experimental value (2.55 mm/s) in both basis sets, although the other functionals are off by 1.0 mm/s or more, both under- and overestimating. Somewhat surprisingly then, the computed QS of a related porphyrin complex Fe(OEP)CO (YEQPOA) has a quantitative agreement with the experiment in almost all the functionals (except M06 and M06-2X) and bases, with little variation from approach to approach. The work of Oldfield and co-workers[12] reports similar large overestimations of the QS by the B3LYP functional and the Wachters basis set for some porphyrin-based compounds and a more balanced performance of BPW91 on the same compounds. Although we and the cited work studied different porphyrin derivatives and the absolute values of overestimations by B3LYP and underestimations by BPW91 differ, there is an agreement in the trend.

Overall, we find some of the QS outliers (FATBOR, DEDWUE) in qualitative agreement with the experiment, while some others (PTSQFE10, BUYKUB10, SISKOU, VUPJUL, VUPJOF) are probably difficult cases which require further investigation by more reliable theoretical methods than DFT.

Now, let us turn to the quadrupole splitting computed for the DFT-optimized geometries. Figure 4 juxtaposes the quadrupole splittings computed with the B3LYP/Partridge-1 method for the theoretical and experimental geometries. This particular method serves as a typical case, and we believe its results are good for illustrative purposes. The overall pattern of data points in both sides of Figure 4, each side representing one type of geometry, is quite similar: substantially better agreement with the experiment <2.0 mm/s and a bigger scatter and more outliers in the region >2.0 mm/s. However, both visually and quantitatively, the accuracy of prediction in the case of the DFT-optimized geometries is worse. The MUE's in the region <2.0 mm/s are 0.168 and 0.0952 mm/s for the theoretical and experimental geometries, respectively. The MUE's for all 35 points are 0.381 and 0.337 mm/s, respectively. The similar situation is seen for all other combinations of the functional and the basis, again demonstrating, in parallel with the isomer shift data, that the optimization of the geometries has been somewhat detrimental for the Mössbauer calculations.

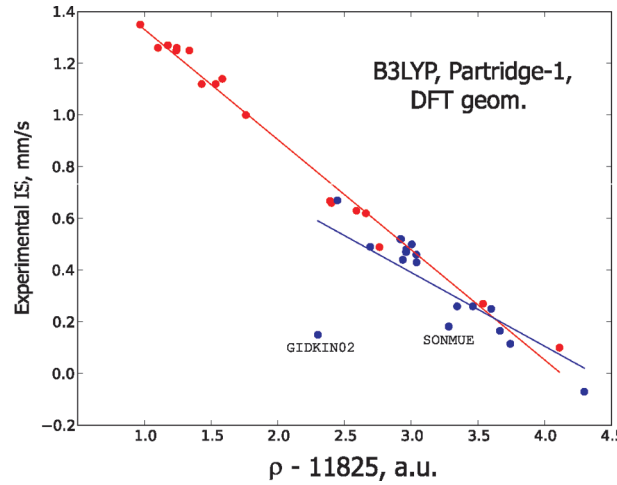


Figure 3. Linear correlations observed between electronic density on iron and experimental isomer shift computed with the B3LYP functional and Partridge-1 basis set for the DFT-based geometries. The data in red correspond to the oxidation state +2, and the data in blue correspond to the oxidation states +3 and +4.

After the optimization, at least three minor outliers (VOCBAQ, GIDKIN02, RABHAD) and one major one (SOCVUB) appeared (compare Figure 4a and 4b) in the region <2.0 mm/s. It is interesting that all these structures are of nonchelate type, being nonrigid and with a geometry that is therefore more likely to be distorted by the optimization method. The experimental SOCVUB structure was already a slight outlier in Figure 4b, and its optimization strongly deteriorated agreement with the experiment. The large outlier BUYKUB10 becomes an even larger one after the optimization. The DFT-optimized GIDKIN02 structure also produced an outlier in the isomer shift calibrations (Figure 3). In the region >2.0 mm/s, regardless of the geometry type, essentially the same outliers are observed (with some alterations of their positions, especially notable for VOFLOR and SISKOU).

Application to Methane Monooxygenase Intermediates. The soluble methane monooxygenase hydroxylase (MMOH) is a well-studied enzyme hosting a diiron active site that catalyzes the oxidation of methane to methanol.[53] Its catalytic cycle involves several intermediates with different oxidation states of iron atoms, and we have been studying these intermediates quite extensively in the last several years, both with DFT[103,104] and QM/MM[62] methods. These intermediates are rather short lived so that their structures cannot be derived from X-ray crystallography. However, their Mössbauer spectra are available, and therefore, the computational Mössbauer approach becomes an important tool for investigating their structures. Han and Noodleman recently computed the Mössbauer parameters of several candidate structures for the key intermediates P and Q.[3,16,29,105] Comparison with the experimental data allowed these authors to select the most probable structures.

Here, as an application to our benchmarking results, we are computing the Mössbauer characteristics of the same

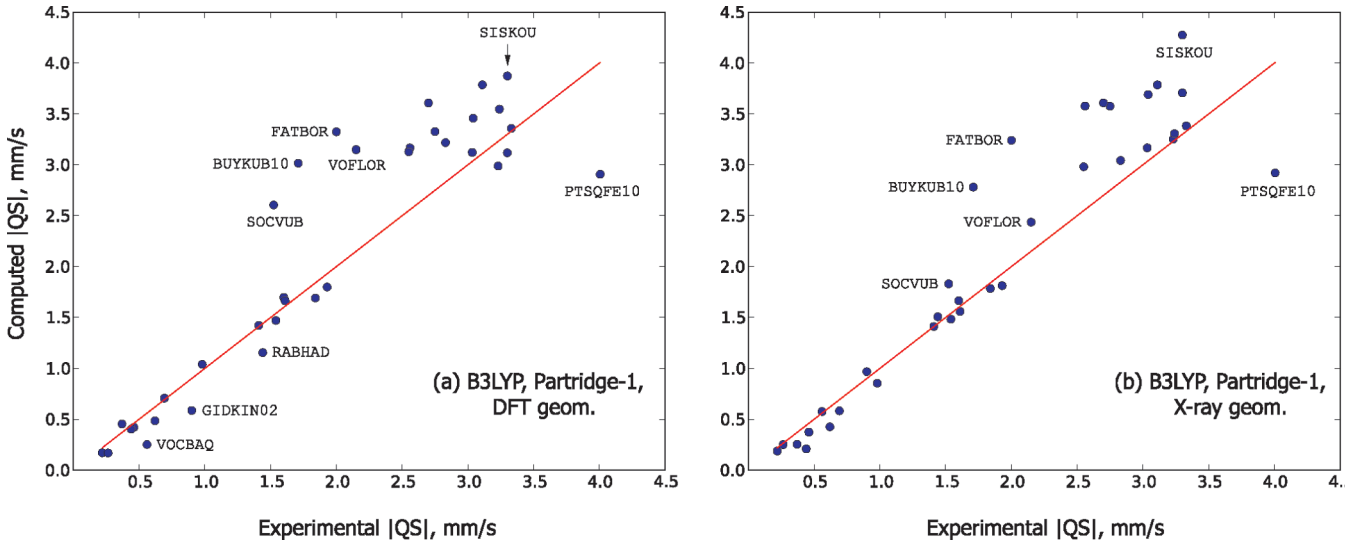


Figure 4. Comparison of the quadrupole splittings computed with the B3LYP functional and Partridge-1 basis set using the DFT (a) and X-ray (b) geometries. The red line is $y = x$: the points lying on it represent perfect agreement between experiment and theory. The obvious outliers are indicated by their Cambridge Structural Database code which can also be found in Table 1.

Table 5. Comparison of the Isomer Shifts and the Quadrupole Splittings of MMOH Reduced (Re) Structure As Well As the Peroxo and Q Intermediates Computed in This Work and in the Works of Han and Noodleman[3,29a]

	this work		refs 3 and 29		experiment	
form	δ	|QS|	δ	|QS|	δ	|QS|
Re	1.27	3.00	1.26	2.87	1.3	2.87; 3.1
	1.16	2.46	1.34	3.00	1.3	2.87; 2.4−3.0
P-1 (μ-η^2, η^1)	0.62	1.03			0.66	1.51
	0.78	0.44			0.66	1.51
P-2 (μ-η^2, η^2)	0.57	0.77	0.60	0.57	0.66	1.51
	0.63	1.15	0.57	0.97	0.66	1.51
P-3 (A-μ-1,2)	0.73	1.60	0.72	1.69	0.66	1.51
	0.70	1.25	0.63	1.12	0.66	1.51
P-4 (S-μ-1,2)	0.57	1.86	0.64	1.81	0.66	1.51
	0.58	1.31	0.61	1.21	0.66	1.51
Q	0.11	0.78	0.18	0.33	0.17, 0.21	0.53, 0.68
	0.38	0.69	0.22	0.33	0.17, 0.14	0.53, 0.55

[a] Our geometries were optimized by QM/MM,[62] whereas those in refs 3 and 29 were produced by the DFT in continuous dielectric medium (the so-called COSMO model). The juxtaposition of the coordinations is approximate since the geometries in the same row do not exactly coincide. For the schematic geometries of the QM/MM-optimized intermediates see Figure 4. The IS and QS values of this work are computed with the methods that give the most accurate statistics: B3LYP/Partridge-1 and O3LYP/Partridge-1, respectively. The Mossbauer characteristics corresponding to the geometries of Han and Noodleman were taken from their works.[3,29]

intermediates but using presumably more accurate geometries, those optimized through the QM/MM approach in our recent work.[62] In contrast, the models of Han and Noodleman were obtained from the conductor-like screening solvation model (COSMO),[107,108] which essentially utilizes a DFT method treating molecular systems embedded in a molecular-shaped cavity and surrounded by a dielectric medium.

As an initial calibration, we compute the Mössbauer parameters for a model of the reduced MMOH enzyme, for which there is a crystal structure.[109] Hence, for this case, uncertainty about the geometry should be relatively unimportant (although errors from the QM/MM geometry optimization will still be present). The results shown in Table 5 indicate good agreement for the IS. For the QS, there are two sets of experiments; our results are more consistent with the second set than with the first and reasonable in either case. These calculations give some indication of what to expect in a realistic, complex application, where there can be noise and uncertainty in both theory and experiment.

In ref 62 we identified four P candidate structures and one Q structure. The schematic diagrams of their active sites are given in Figure 5. In Table 5 we bring together the Mössbauer parameters of these models and the models from the works of Han and Noodleman.[3,29] Since these authors used somewhat different geometries the juxtaposition of our models with theirs is approximate. In placing the two models in the same row for comparison we tried to preserve the type of coordination of the oxygen and the ligands to the iron atoms. Then, because Han and Noodleman used different functionals to compute Mossbauer characteristics and they sometimes produced slightly contradictory results we select those IS and QS values computed by these authors which seem to have the best agreement with the experiment and those structures that are selected by the authors as the most likely.

As our computation methodology in the present study, we take the best combination of the functional and basis to compute the IS (B3LYP/Partridge-1) and equivalently with

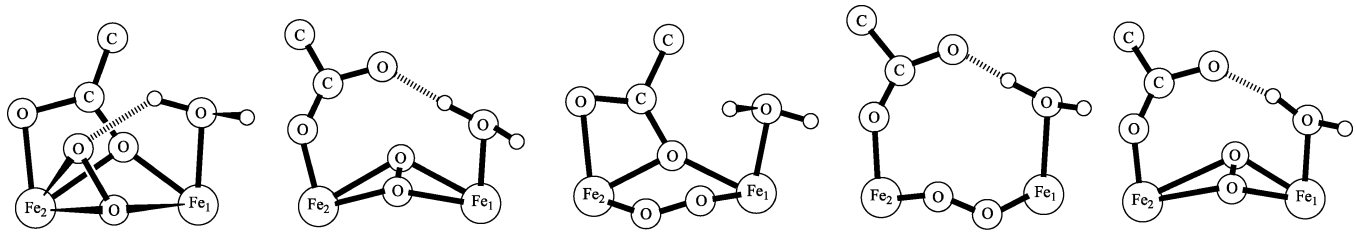


Figure 5. Schematic diagrams of the MMOH intermediates active sites. P stands for Fe^{3+},Fe^{3+} peroxo models and Q represents the Fe^{4+},Fe^{4+} species.

QS (O3LYP/Partridge-1). The corresponding values are given in Table 5. Analyzing the results of our calculations for the four peroxo models, we conclude that the P-3 model (A-μ-1,2) unequivocally has the closest agreement with the experiment. This type of coordination was in fact chosen as most likely by Han and Noodleman,[3] despite some confusion created by multiple energetic and Mössbauer data available from two functionals (OPBE and PW91). This type of structure is also in agreement with the vibrational analysis performed by the groups of Solomon and Yoshizawa.[110,111] All other peroxo structures listed in Table 5 have the Mössbauer characteristics that are substantially farther from experiment than those of P-3. The computed IS of the P-3 model produced by us and the group of Noodleman have about the same proximity to the experimental values. Our QS range is 'narrower' however (1.60 and 1.25 mm/s), bringing the signals from both iron centers within the error bars of the single experimental value (1.51 mm/s). Encouragingly, the other two types of structures in which we and the group of Noodleman overlap (μ-η^2,η^2, S-μ-1,2) are in good qualitative and perhaps even quantitative agreement with one another. Thus, we have to conclude that despite the fact that our previous QM/MM calculations predicted the lowest energy for the P-2 (μ-η^2,η^2) structure, the analysis of the computationally obtained vibrational and Mössbauer spectra converges unambiguously on the single type of structure, P-3 (A-μ-1,2). At the same time this apparent consensus should not categorically rule out the P-2 (μ-η^2,η^2) structure as a participant in the MMOH catalytic cycle. It may be still energetically most stable but for some (for example, kinetic) reason undetectable in the Mössbauer experiment. Additionally, recent careful experimental analysis of the MMOH peroxo intermediates revealed the presence of the second peroxo structure the Mössbauer spectra of which are currently unknown.[112]

Finally, let us discuss the Mössbauer data for the Q intermediate (which is a Fe^{4+}, Fe^{4+} species) given in Table 5. The table indicates that there were two corresponding experimental Mössbauer measurements, which produced somewhat different qualitative and quantitative results. The first experiment showed only one signal, but the second one gave two similar signals in both IS and QS. The central geometry of our Q structure is slightly distorted, with two oxygens that form the 'diamond' core being closer to Fe_2 than Fe_1.[62] Therefore, two types of signals in the Mössbauer experiment would be more expected than one. In this regard, our calculations qualitatively agree with the second experiment. The calculation of Han and Noodleman, however, produced two different IS signals but only one QS aggregate signal. Our isomer shift has worse agreement with the experiment than Han and Noodleman's, but our quadrupole splitting is closer to the experiment than theirs. The Q structure initially proposed by these authors as the most likely model had a somewhat better agreement with the experimental QS (two signals 0.70 and 0.37 mm/s).[29] Unfortunately, this structure involved a suspicious proton transfer from the coordinated water to one of the histidine rings and was later dismissed on energetic and spectroscopic grounds.[3] It looks like our computed IS and QS values lie within the error bars and are in good semiquantitative agreement with the experiment.

The application to the MMOH intermediates serves as a good example of the power and utility of combining energetics calculations with Mössbauer spectral calculations in cases where the structures are not experimentally determined, agreeing with conclusions drawn by the group of Noodleman. By applying the statistically most accurate method(s) to compute the IS and QS, we arrived immediately at very good qualitative (and perhaps even quantitative) agreement with the experiment, helping resolve some of the questions that are still disputed in the MMOH intermediates research.

Discussion

We summarize the conclusions that can be drawn with regard to the accuracy of DFT calculations of Mössbauer parameters, including the dependence upon the geometry and the functional employed. We first discuss the results obtained when the X-ray structures are used for the geometry and then examine the effect of performing geometry optimization. An important goal is to understand to what degree one can rely upon Mössbauer calculations in a system like an iron-containing protein to provide accurate discrimination among alternative structures. If the noise in the calculation is larger than the difference predicted in the parameters for the alternative structures, it would be inappropriate to draw any conclusions one way or another. Hence, understanding the level of noise expected in a realistic application is essential if the method is to be profitably used to help in assigning structures in cases where crystallography cannot be carried out. The discussion below focuses on the optimal functionals and basis sets determined in our benchmarking studies.

As indicated above, we first discuss the errors observed for calculations of the IS and QS using X-ray geometries.

For the IS, errors in the range 0.02−0.04 mm/s are observed. However, since this range is observed after the parametrization on the training set, the error is likely to grow somewhat if the methodology is applied to an independent test set. Nevertheless, because the isomer shift data points form a good linear correlation, the MUE that would be observed on a test set is not expected to be much larger than 0.02−0.04 mm/s. For QS, training set errors in the range of 0.15 for the QS less than 2.0 mm/s and 0.4 for QS greater than 2.0 mm/s are observed.

The effects of geometry optimization are more specific and system dependent. As noted in above, there are specific complexes and chemistries where DFT geometry optimization has significant problems reproducing the experimentally observed geometry. In these cases, substantial errors can be introduced from the geometry optimization. However, for typical cases, where the geometry optimization yields a result that agrees reasonably well with the experiment, the deviation between the Mössbauer parameters computed at the two geometries is small. Hence, the principal concern when geometry optimization is employed is whether one is dealing with a system with a potential energy surface that is poorly modeled by the variant of the DFT being employed. However, this is a serious concern in any case, whether one is computing Mössbauer parameters or just energetics; one has to be suspicious of the quality of all the results if the geometry is in poor agreement with experiment. Repairing the occasional, but sometimes large, errors in DFT geometries, particularly for transition metals, thus should be a high priority in the development of the next generation of functionals but will not be discussed further here.

In considering what constitutes "acceptable" agreement in a Mössbauer calculation for a complex system such as an iron-containing protein, somewhat larger error bars have to be used than the MUE averaged over all of the data set, as the case at hand could lie on the high side of the error distribution. If calculations are being done with a crystal structure, an error on the order of 0.1 mm/s for the IS, 0.2 mm/s for the QS if less than 2.0 mm/s, and 0.4 mm/s if the QS is greater than 2.0 mm/s would appear to be reasonable estimates based on our results. If geometry optimization has to be carried out, one might increase the expected level of error by 20−30%, assuming that the geometry is in fact not seriously deviant from the experiment. By this criterion, the use of the Mössbauer data to choose between the proposed MMOH peroxo structures is clear cut, and the results for the P-3 peroxo model are in reasonable agreement with the experimental data. Additionally, the Mössbauer data reported in Table 5 for the reduced MMOH structure modeled through geometry optimization from an oxidized MMOH structure, the X-ray geometry of which is available, lie, satisfactorily, within the error bars. For Q, the IS from the Noodleman group calculations are in good agreement with experiment, but one of the Fe atoms in our calculations is a little outside the range proposed above. On the other hand, our QS results for Q are in good agreement with the experiment, whereas the Noodleman group results are outside of the suggested error bounds, given that the QS is less than 2.0 mm/s. Since the Q models that we and the Noodleman group are employing are very similar but not identical and we are using different functionals, what these data suggest is that we are both probably quite close to the correct structure but that some detailed refinement of the geometry would be required to better match the IS and QS from experiment. This result is unsurprising in that the conformational potential energy surface of a protein is much more complicated than that of the small molecules examined in our training set, and so we can expect to see greater theory/experiment deviations based on errors in the geometry on the complex surface, which contains a large number of minima that are closely spaced in energy.

In summary, the overall results, both on the training set and for the MMOH calculations, are encouraging with respect to the goal of employing Mössbauer data, in combination with energetic criteria, to identify structures in iron-containing proteins. However, there are clearly some quantitative issues to address related to noise introduced by the complexity of the protein potential energy surface.

Conclusions

This work assembled a high-quality, good-size test set (31 compounds; 35 data points) for the benchmarking of isomer shifts and quadrupole splittings. We selected the entries of the test set on the basis of availability of their Mössbauer spectra at liquid helium temperature as well as X-ray crystallographic data. Eight functionals and two bases (on iron atoms) were applied to determine which combination predicted more accurate results, and two sources of geometries (DFT optimized and crystallographic) were used. The geometries obtained through the optimization (B3LYP functional, pseudospecral approach, LACVP* basis on the irons, and 6-31G* on the other atoms) yielded somewhat worse Mössbauer results (greater scatter of points and more outliers). Nonchelate structures were particularly affected by the optimization, probably due to their nonrigidity. It is worth seeking a method for a more accurate optimization of nonchelate metal geometries in a future publication, but for now the usage of X-ray geometries should not be shunned as they produce quite accurate results. For example, the mean unsigned error of quadrupole splitting is less than 0.1 mm/s in the region <2.0 mm/s (if one obvious outlier is excluded) and below 0.3 mm/s in the whole region (all data points are counted in). The conclusions below are based on the X-ray geometries.

The isomer shift statistics indicate that the best functional to compute the isomer shift is B3LYP with O3LYP a close second. Partridge-1 set on iron atoms shows somewhat better performance than Wachters basis. M06 and especially M06-2X show some of the worst IS performance. The best functional to compute quadrupole splitting is O3LYP in the Partridge-1 basis set with B3LYP being the second best. M06-2X produces a large number of outliers and the worst overall QS performance. The Wachters basis is appreciably smaller, and the quality of the density computed with it using the loose convergence criteria is sufficient to compute good-quality isomer shifts and quadrupole splittings. Thus, the combination of the Wachters basis set and lower-quality wave function may be used for the rough estimate of the

Mössbauer characteristics in situations where many different models have to be filtered against the experimental data. For higher quality computations of the Mössbauer spectra, the well-converged wave function computed with the Partridge-1 basis set should be used. To summarize, O3LYP and B3LYP in combination with the Partridge-1 basis on iron atoms show the best performance and SVWN5 and M06-2X are not recommended for Mössbauer calculations.

The results of our statistical analysis were used to select the best computational methods for predicting the Möbssauer spectra of the P and Q intermediates of the enzyme methane monooxygenase hydroxylase (MMOH). We compare our results to those of Han and Noodleman, who recently applied some of their Möbssauer DFT methods to determine which of the P and Q structural candidates was most likely to represent the experimental structure. For P structures, our results agree comfortably with Han and Noodleman's, whereas for the Q structure we find a significantly better agreement in the quadrupole splitting albeit a slightly worse one in the isomer shift. It is also important to emphasize what we see as a methodological improvement on the work of Han and Noodleman. Whereas their results were obtained with the functional and basis that best match the experimental Mössbauer parameters of the systems of interest, we treated our target systems with the functional/basis set combination that was first found the most accurate on an extensive and independent test set, which is a less biased approach.

Acknowledgment. We thank the group of Noodleman for providing us with the Cartesian coordinates of some of their MMOH models. This work was supported in part by a grant from the NIH to R.A.F. (GM 40526) and S.J.L. (GM 32134).

Supporting Information Available: Atomic coordinates of all models from Table 1, Mössbauer parameters computed for all combinations of the functional, basis, and data set entries, as well as calibration constants for all isomer shift parametrizations. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Mössbauer, R. L. *Z. Phys.* **1958**, *151*, 124–143.

(2) Greenwood, N. N.; Gibb, T. C. *Mössbauer spectroscopy*; Chapman and Hall: London, 1971.

(3) Han, W.-G.; Noodleman, L. *Inorg. Chem.* **2008**, *47*, 2975–2986.

(4) Han, W.-G.; Noodleman, L. *Dalton Trans.* **2009**, *30*, 6045–6057.

(5) Jensen, K. P.; Bell III, C. B.; Clay, M. D.; Solomon, E. I. *J. Am. Chem. Soc.* **2009**, *131*, 12155–12171.

(6) Lai, W.; Chen, H.; Cho, K.-B.; Shaik, S. *J. Phys. Chem. A* **2009**, *113*, 11763–11771.

(7) Murray, L. J.; García-Serres, R.; Naik, S.; Huynh, B. H.; Lippard, S. J. *J. Am. Chem. Soc.* **2006**, *128*, 7458−7459.

(8) Song, W. J.; Behan, R. K.; Naik, S. G.; Huynh, B. H.; Lippard, S. J. *J. Am. Chem. Soc.* **2009**, *131*, 6074−6075.

(9) Havlin, R. H.; Godbout, N.; Salzmann, R.; Wojdelski, M.; Arnold, W.; Schulz, C. E.; Oldfield, E. *J. Am. Chem. Soc.* **1998**, *120*, 3144–3151.

(10) Braden, D. A.; Tyler, D. R. *Organometallics* **2000**, *19*, 1175–1181.

(11) Li, M.; Bonnet, D.; Bill, E.; Neese, F.; Weyhermüller, T.; Blum, N.; Sellmann, D.; Wieghardt, K. *Inorg. Chem.* **2002**, *41*, 3444–3456.

(12) Zhang, Y.; Mao, J.; Godbout, N.; Oldfield, E. *J. Am. Chem. Soc.* **2002**, *124*, 13921–13930.

(13) Neese, F. *Inorg. Chim. Acta* **2002**, *337*, 181–192.

(14) Zhang, Y.; Oldfield, E. *J. Phys. Chem. A* **2003**, *107*, 4147–4150.

(15) Liu, T.; Lovell, T.; Han, W.-G.; Noodleman, L. *Inorg. Chem.* **2003**, *42*, 5244–5251.

(16) Han, W.-G.; Liu, T.; Lovell, T.; Noodleman, L. *J. Comput. Chem.* **2006**, *27*, 1292–1306.

(17) Nemykin, V. N.; Hadt, R. G. *Inorg. Chem.* **2006**, *45*, 8297–8307.

(18) Kurian, R.; Filatov, M. *J. Chem. Theor. Comput.* **2008**, *4*, 278–285.

(19) Römelt, M.; Ye, S.; Neese, F. *Inorg. Chem.* **2009**, *48*, 784–785.

(20) Kurian, R.; Filatov, M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 2758–2762.

(21) Zhang, Y.; Oldfield, E. *J. Phys. Chem. B* **2003**, *107*, 7180–7188.

(22) Hobza, P.; Šponer, J.; Reschel, T. *J. Comput. Chem.* **1995**, *16*, 1315–1325.

(23) Wodrich, M. D.; Corminboeuf, C.; Schleyer, P. v. R. *Org. Lett.* **2006**, *8*, 3631–3634.

(24) Friesner, R. A.; Knoll, E. H.; Cao, Y. *J. Chem. Phys.* **2006**, *125*, 124107.

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

(26) Hendrich, M. P.; Gunderson, W.; Behan, R. K.; Green, M. T.; Mehn, M. P.; Betley, T. A.; Lu, C. L.; Peters, J. C. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17107–17112.

(27) Popescu, C. V.; Mock, M. T.; Stoian, S. A.; Dougherty, W. G.; Yap, G. P. A.; Riordan, C. G. *Inorg. Chem.* **2009**, *48*, 8317–8324.

(28) Hopmann, K. H.; Ghosh, A.; Noodleman, L. *Inorg. Chem.* **2009**, *48*, 9155–9165.

(29) Han, W.-G.; Noodleman, L. *Inorg. Chim. Acta* **2008**, *361*, 973–986.

(30) Edwards, P. R.; Johnson, C. E.; Williams, R. J. P. *J. Chem. Phys.* **1967**, *47*, 2074–2082.

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

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

(33) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244–13249.

(34) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1992**, *46*, 6671–6687.

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

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

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

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

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

(40) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(41) Bloch, F. *Z. Phys.* **1929**, *57*, 545–555.

(42) Dirac, P. A. M. *Proc. Cambridge Philos. Soc.* **1930**, *26*, 376–385.

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

(44) Bochevarov, A. D.; Friesner, R. A. *J. Chem. Phys.* **2008**, *128*, 034102.

(45) Xu, X.; Goddard III, W. A. *J. Phys. Chem. A* **2004**, *108*, 8495–8504.

(46) Zhao, Y.; Truhlar, D. G. *J. Chem. Theor. Comput.* **2007**, *3*, 289–300.

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

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

(49) Partridge, H. *J. Chem. Phys.* **1987**, *87*, 6643–6647.

(50) Wachters, A. J. H. *J. Chem. Phys.* **1970**, *52*, 1033–1036.

(51) Filatov, M. *J. Chem. Phys.* **2007**, *127*, 084101.

(52) *Jaguar 7.5*, Schrödinger, Inc.: Portland, OR, 2009.

(53) Merkx, M.; Kopp, D. A.; Sazinsky, M. H.; Blazyk, J. L.; Müller, J.; Lippard, S. J. *Angew. Chem., Int. Ed.* **2001**, *40*, 2782–2807.

(54) Baik, M.-H.; Newcomb, M.; Friesner, R. A.; Lippard, S. J. *Chem. Rev.* **2003**, *103*, 2385–2420.

(55) Friesner, R. A. *Chem. Phys. Lett.* **1985**, *116*, 39–43.

(56) Friesner, R. A. *J. Chem. Phys.* **1986**, *85*, 1462–1468.

(57) Friesner, R. A. *J. Chem. Phys.* **1987**, *86*, 3522–3521.

(58) Friesner, R. A. *J. Phys. Chem.* **1988**, *92*, 3091–3096.

(59) Martinez, T. J.; Carter, E. A. In *Modern Electronic Structure Theory, Part II*; Yarkony, D. R., Ed.; World Scientific: Singapore, 1995; Vol. 92, p 1132.

(60) Schuchardt, K. L.; Didier, B. T.; Elsethagen, T.; Sun, L.; Gurumoorthi, V.; Chase, J.; Li, J.; Windus, T. L. *J. Chem. Inf. Model.* **2007**, *47*, 1045–1052.

(61) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299–310.

(62) Rinaldo, D.; Philipp, D. M.; Lippard, S. J.; Friesner, R. A. *J. Am. Chem. Soc.* **2007**, *129*, 3135–3147.

(63) Tshuva, E. Y.; Lippard, S. J. *Chem. Rev.* **2004**, *104*, 987–1012.

(64) Wertheim, G. K.; Herber, R. H. *J. Chem. Phys.* **1962**, *36*, 2497–2499.

(65) Shirley, D. A. *Rev. Mod. Phys.* **1964**, *36*, 339–351.

(66) Filatov, M. *Coord. Chem. Rev.* **2008**, *253*, 594–605.

(67) Sadoc, A.; Broer, R.; de Graaf, C. *Chem. Phys. Lett.* **2008**, *454*, 196–200.

(68) Kurian, R.; Filatov, M. *J. Chem. Phys.* **2009**, *130*, 124121.

(69) Zwanziger, J. W. *J. Phys.: Condens. Matter* **2009**, *21*, 195501.

(70) Wdowik, U. D.; Ruebenbauer, K. *Phys. Rev. B* **2007**, *76*, 155118.

(71) Han, W.-G.; Liu, T.; Lovell, T.; Noodleman, L. *J. Am. Chem. Soc.* **2005**, *127*, 15778–15790.

(72) Enemark, J. H.; Westcott, B. L. Transition metal nitrosyls. In *Inorganic electronic structure and spectroscopy*; Solomon, E. I., Lever, A. B. P., Eds., John Wiley and Sons: New York, 1999; Vol. II: Applications and case studies, p 403.

(73) McCleverty, J. A. *Chem. Rev.* **2004**, *104*, 403–418.

(74) Zhang, Y.; Mao, J.; Oldfield, E. *J. Am. Chem. Soc.* **2002**, *124*, 7829–7839.

(75) Trautwein, A.; Harris, F. E.; Freeman, A. J.; Desclaux, J. P. *Phys. Rev. B* **1975**, *11*, 4101–4105.

(76) Landau, L. D.; Lifshitz, L. M. *Quantum Mechanics*; Butterworth-Heinmann: Oxford, 1981; p 482.

(77) Sinnecker, S.; Slep, L. D.; Bill, E.; Neese, F. *Inorg. Chem.* **2005**, *44*, 2245–2254.

(78) Tolman, W. B.; Liu, S.; Bentsen, J. G.; Lippard, S. J. *J. Am. Chem. Soc.* **1991**, *113*, 152–164.

(79) Silva, R. M.; Gwengo, C.; Lindeman, S. V.; Smith, M. D.; Long, G. J.; Grandjean, F.; Gardinier, J. R. *Inorg. Chem.* **2008**, *47*, 7233–7242.

(80) Ménage, S.; Zhang, Y.; Hendrich, M. P.; Que, L., Jr. *J. Am. Chem. Soc.* **1992**, *114*, 7786–7792.

(81) Herold, S.; Lippard, S. J. *J. Am. Chem. Soc.* **1997**, *119*, 145–156.

(82) Chavez, F. A.; Ho, R. Y. N.; Pink, M.; Young, V. G., Jr.; Kryatov, S. V.; Rybak-Akimova, E. V.; Andres, H.; Münck, E.; Que, L., Jr.; Tolman, W. B. *Angew. Chem., Int. Ed.* **2002**, *41*, 149–152.

(83) Silvernail, N. J.; Noll, B. C.; Schulz, C. E.; Scheidt, W. R. *Inorg. Chem.* **2006**, *45*, 7050–7052.

(84) Strauss, S. H.; Silver, M. E.; Long, K. M.; Thompson, R. G.; Hudgens, R. A.; Spartalian, K.; Ibers, J. A. *J. Am. Chem. Soc.* **1985**, *107*, 4207–4215.

(85) Walters, M. A.; Dewan, J. C. *Inorg. Chem.* **1986**, *25*, 4889–4893.

(86) Hagen, K. S.; Lachicotte, R. *J. Am. Chem. Soc.* **1992**, *114*, 8741–8742.

(87) Coucouvanis, D.; Swenson, D.; Baenziger, N. C.; Murphy, C.; Holah, D. G.; Sfarnas, N.; Simopoulos, A.; Kostikas, A. *J. Am. Chem. Soc.* **1981**, *103*, 3350–3362.

(88) Armstrong, W. H.; Spool, A.; Papaefthymiou, G. C.; Frankel, R. B.; Lippard, S. J. *J. Am. Chem. Soc.* **1984**, *106*, 3653–3667.

(89) Bossek, U.; Hummel, H.; Weyhermüller, T.; Bill, E.; Wieghardt, K. *Angew. Chem., Int. Ed.* **1996**, *34*, 2642−2645.

(90) Zang, Y.; Dong, Y.; Que, L., Jr.; Kauffmann, K.; Münck, E. *J. Am. Chem. Soc.* **1995**, *117*, 1169–1170.

(91) Feig, A. L.; Bautista, M. T.; Lippard, S. J. *Inorg. Chem.* **1996**, *35*, 6892–6898.

(92) Harrop, T. C.; Song, D.; Lippard, S. J. *J. Am. Chem. Soc.* **2006**, *128*, 3528–3529.

(93) Turowski, P. N.; Armstrong, W. H.; Liu, S.; Brown, S. N.; Lippard, S. J. *Inorg. Chem.* **1994**, *33*, 636–645.

(94) Jüstel, T.; Weyhermüller, T.; Wieghardt, K.; Bill, E.; Lengen, M.; Trautwein, A. X.; Hildebrandt, P. *Angew. Chem., Int. Ed.* **1995**, *34*, 669–672.

(95) Wu, F.-J.; Kurtz, D. M., Jr.; Hagen, K. S.; Nyman, P. D.; Debrunner, P. G.; Vankai, V. A. *Inorg. Chem.* **1990**, *29*, 5174–5183.

(96) Safo, M. K.; Gupta, G. P.; Walker, F. A.; Scheidt, W. R. *J. Am. Chem. Soc.* **1991**, *113*, 5497–5510.

(97) James, B. D.; Bakalova, M.; Liesegang, J.; Reiff, W. M.; Skelton, B. W.; White, A. H. *Inorg. Chem.* **2001**, *40*, 4617–4622.

(98) Maelia, L. E.; Millar, M.; Koch, S. A. *Inorg. Chem.* **1992**, *31*, 4594–4600.

(99) Harrop, T. C.; Tonzetich, Z. J.; Reisner, E.; Lippard, S. J. *J. Am. Chem. Soc.* **2008**, *130*, 15602–15610.

(100) Berry, K. J.; Clark, P. E.; Murray, K. S.; Raston, C. L.; White, A. H. *Inorg. Chem.* **1983**, *22*, 3928–3934.

(101) Ghosh, A.; de Oliveira, F. T.; Yano, T.; Nishioka, T.; Beach, E. S.; Kinoshita, I.; Münck, E.; Ryabov, A. D.; Horwitz, C. P.; Collins, T. J. *J. Am. Chem. Soc.* **2005**, *127*, 2505–2513.

(102) Sellmann, D.; Geck, M.; Knoch, F.; Ritter, G.; Dengler, J. *J. Am. Chem. Soc.* **1991**, *113*, 3819–3828.

(103) Dunietz, B. D.; Beachy, M. D.; Cao, Y.; Whittington, D. A.; Lippard, S. J.; Friesner, R. A. *J. Am. Chem. Soc.* **2000**, *122*, 2828–2839.

(104) Gherman, B. F.; Baik, M.-H.; Lippard, S. J.; Friesner, R. A. *J. Am. Chem. Soc.* **2004**, *126*, 2978–2990.

(105) Noodleman, L.; Lovell, T.; Han, W.-G.; Li, J.; Himo, F. *Chem. Rev.* **2004**, *104*, 459–508.

(106) Lovell, T.; Han, W.-G.; Liu, T.; Noodleman, L. *J. Am. Chem. Soc.* **2002**, *124*, 5890–5894.

(107) Klamt, A. *J. Phys. Chem.* **1995**, *99*, 2224–2235.

(108) Andzelm, J.; Kölmel, C.; Klamt, A. *J. Chem. Phys.* **1995**, *103*, 9312–9320.

(109) Whittington, D. A.; Lippard, S. J. *J. Am. Chem. Soc.* **2001**, *123*, 827–838.

(110) Brunold, T. C.; Tamura, N.; Kitajima, N.; Moro-oka, Y.; Solomon, E. I. *J. Am. Chem. Soc.* **1998**, *120*, 5674−5690.

(111) Yumura, T.; Yoshizawa, K. *Bull. Chem. Soc. Jpn.* **2004**, *77*, 1305–1311.

(112) Tinberg, C.; Lippard, S. J. *Biochemistry,* **2009**, *48*, 12145–12158.

CT100398M

JCTC Journal of Chemical Theory and Computation

Temperature Dependence of Hydrogen-Bond Stability in β-Hairpin Structures

Qiang Shao and Yi Qin Gao*

College of Chemistry and Molecular Engineering, National Laboratory of Molecular Sciences, Peking University, Beijing, China

Received August 6, 2010

Abstract: Understanding the temperature effect in the folding of multiple β-hairpins with different sequence (although based on an approximate solution model) makes possible quantitative characterization of the different contributing factors that are difficult to be obtained from single temperature studies. The detailed thermodynamics analyses performed in this article provide at least a semiquantitative understanding of how temperature and positions affect the stability of individual backbone hydrogen bonds in β-hairpin structures. These effects are then rationalized, according to the separation into enthalpic and entropic contributions. The formation of backbone hydrogen bonds at the terminal position is favored at low temperatures and those near the turn become more favorable at high temperatures, as a result of the differences in their formation entropy. Regardless of the differences in the turn stability, the side-chain hydrophobicity, and the room temperature folding mechanism of these β-hairpins, there is a shift to the "zip-out" mechanism in the assembling of backbone hydrogen bonds as temperature increases for all polypeptides under study. In addition, it was also observed that although the backbone hydrogen-bond formation shows a strong dependence on temperature, the formation order of the three structural elements of β-hairpin (the turn, hydrophobic core cluster, and hydrogen-bond assembly) along the minimum free energy pathway in the free energy landscapes appears to be only sequence dependent and largely unaffected by the temperature change.

Introduction

As important model systems for protein folding, β-hairpin structured polypeptides have been the subject of a large number of experimental and theoretical studies.[1−15] The popular β-hairpins include not only fragments of natural proteins (e.g., those from the B1 domain of protein G (GB1 peptide),[10,16] ubiquitin,[14] and human chorionic gonadotropin)[17] but also artificially designed peptides (e.g., the tryptophan zipper series, TRPZIP 1−6).[11] Since the folding of β-hairpins is a cooperative process which largely resembles that of complex proteins,[8,10] the understanding of their folding mechanism and kinetics is of great interest. A great number of mechanistic studies have been performed, and different folding mechanisms have been proposed. These mechanisms mainly differ in the arrangement of the important structural elements, such as the hydrophobic core cluster, the β-turn, and backbone hydrogen bonds. For instance, the laser-induced temperature-jump experiment[10] and the lattice Monte Carlo (MC) simulation[18] on the GB1 peptide suggested a "zipping" or "hydrogen-bond centric" mechanism. The folding of a β-hairpin starts from the turn and propagates to the terminus. During this process, the native backbone hydrogen bonds are formed. At the last stage, the hydrophobic core is packed. A modified version of the zipping model in which the hydrogen-bond formation instead of hydrophobic core packing occurs at last was also proposed based on molecular dynamics (MD) simulations on a GB1 peptide.[19] On the other hand, "hydrophobic core centric" mechanism as favored by several theoretical simulations[4,20−24] postulated that the hydrophobic core is packed first, during which a few backbone hydrogen bonds could also form. The

* Corresponding author. E-mail: gaoyq@pku.edu.cn. Telephone: +86 010 62752431.

10.1021/ct100436r
Published on Web 11/29/2010

final stage of folding in this model consists of the formation of the remaining hydrogen bonds as well as the native turn structure.

These folding mechanisms of the β-hairpin are surmised based on the experimental observations and the simulation results. To date, many different all-atom force fields have been used for the computational simulations of β-hairpins, e.g., CHARMM, AMBER, OPLS, and GROMOS96, employed with either explicit or implicit solvent models. Different force fields favor different secondary structures, e.g., of the AMBER all-atom force fields, it is well-known that FF96 favors β-structures, whereas FF94, FF99, and FF03 favor α-helical conformations.[25−28] On the other hand, the explicit solvent model is more desirable for elucidating the details of protein folding pathways at the cost of enormous CPU time, whereas the implicit solvent models are developed to economize the computation time with the loss of simulation precision. The most popular implicit solvent model is the generalized Born/surface area (GB/SA) model.[29] One of the potential sources for the inconsistency among these folding mechanisms mentioned above might be the usage of different force fields and solvent models in these simulations.

The question is: On the premise of saving the computation time, which force field and implicit solvent model should we use to best describe the folding pathway of β-hairpins? To answer this question, Zhou performed the replica exchange molecular dynamics (REMD) simulations on the folding of GB1 peptide with different force fields in combination with different implicit models and compared the results to those from explicit solvent models.[30] Of all implicit solvent models tested, only AMBER FF96/GBSA produced reasonable results comparable to the explicit solvent models.

More recently, Shell et al. tested the stability of several polypeptides (GB1 peptide, TRPZIP2, C peptide, and EK helix) with AMBER force fields and several versions of the GB/SA model.[31] It showed that the combination of FF96 with GB^{OBC} model (igb=5)[32] is the best choice to balance the α-helix and β-hairpin tendencies of the polypeptides tested. We also ran MD simulation on the folding of GB1 peptide using FF96 force field with both GB^{OBC} and GBn (igb = 7)[33] models and compared the results to MD simulation using FF96 and TIP3P explicit solvent model.[34] The results showed that the combination of FF96 with GB^{OBC} model generates the more consistent free energy landscapes compared to those in explicit solvent model and therefore is better at describing the folding of GB1 peptide.[34] Furthermore, Ozkan et al. ran the REMD simulations with the zipping and assembly (ZA) search strategy on the folding of nine proteins, including both α- and β-secondary structures, and observed that FF96 force field, combined with GB/SA model, is best balanced for various secondary structures compared to other force fields.[35] All information mentioned above showed that although the combination of AMBER FF96 and GB^{OBC} model might generate the folding thermodynamics data (such as folding/unfolding energy barriers) with the exact values inconsistent (not far away) with those in the explicit solvent simulation,[30] it is still good at depicting the overall folding scene of β-hairpin. The latter is more interesting for us and is the main issue in the present study.

To distinguish between different folding mechanisms as mentioned above by quantifying the thermodynamics of the folding of β-structured polypeptides and, more importantly, to examine the sequence dependence in their folding mechanisms, in the recent study we ran MD simulations on the folding of a series of polypeptides using the integrated tempering sampling (ITS)[36,37] method and performed detailed folding thermodynamics analyses on these polypeptides at room temperature.[38] All polypeptides are modeled using the AMBER FF96 force field[39] and the GB^{OBC} model,[32] the best force field and implicit solvent model combination. The advantage of this study is that the usage of the enhanced energy sampling method ITS allows the simulation to sample the potential energy surface thoroughly and thus to capture plenty of polypeptide configurations and, to some extent, reduces the influence of the force field and implicit solvent model.

It is worth noting that all theoretical models of β-hairpin folding mentioned above are based on the studies of a single system, GB1 peptide. A systematic study on the sequence influence of the folding mechanism of β-structures has not been carried out previously and therefore becomes very necessary. The systems in our study include peptide 1 (sequence: SESYIN^{D}PDGTWTVTE),[40] GB1 (sequence: GEWTYDDATKTFTVTE, PDB code: 2GB1),[41] TRPZIP2 (sequence: SWTWENGKWTWK, PDB code: 1LE1),[11] and TRPZIP4 (sequence: GEWTWDDATKTWTWTE, PDB code: 1LE3).[11] The native structures of the four polypeptides are shown in Figure 1 (backbone hydrogen bonds (HBs 1−6) are named from the terminus to the turn positions in all polypeptides. For TRPZIP2, HB1 located at the terminus is not accounted in the data analysis due to its high instability). The sequence differences among these polypeptides are in their turn structure (either type I or type I′) and side-chain hydrophobicity. TRPZIP2 and TRPZIP4 were designed based on the wild-type GB1 peptide with modified turn sequences and/or hydrophobic core cluster composition. Both of them exhibit reversible and highly cooperative thermal unfolding transition in solution.[11] The fourth hairpin, peptide 1, possesses a pair of very weak hydrophobic interactions but has a very stable (type I′) turn.[40,42] Consequently, the folded peptide 1 and TRPZIP2 but not GB1 or TRPZIP4 possess stable turn structures. On the other hand, TRPZIP2 and TRPZIP4 possess very strong hydrophobic interactions, which are largely absent for GB1 and peptide 1.

Our folding free energy landscape calculations for the four polypeptides at room temperature showed that the folding mechanism of a β-hairpin is strongly dependent on its turn stability and side-chain hydrophobicity.[38] The stable turns of peptide 1 and TRPZIP2 make the turn formation a barrierless and spontaneous process in comparison with the formation of the hydrophobic core and backbone hydrogen bonds, while the turn formation in the other two polypeptides has to overcome significant free energy barriers and becomes the rate-limiting step in the hairpin folding process. Therefore, the turn stability is the key element in determining the formation order of the structural elements of a β-hairpin.

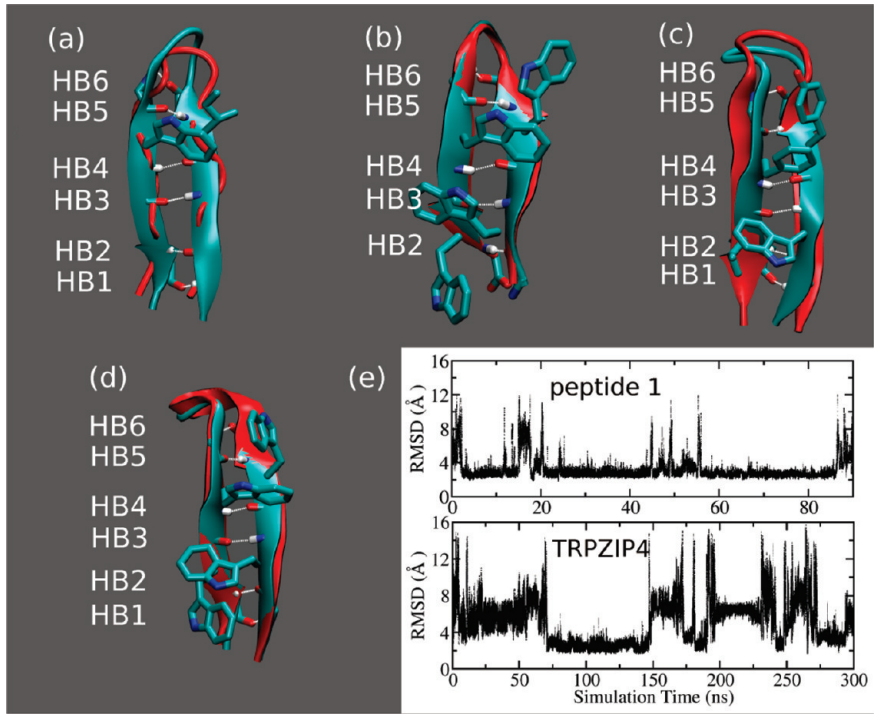


Figure 1. Folded structures of (a) peptide 1, (b) TRPZIP2, (c) GB1, and (d) TRPZIP4 polypeptides obtained in MD simulations (blue color) in comparison to their corresponding native structures (red color). The hydrophobic core is shown in licorice model, and backbone hydrogen bonds are represented by dash lines. (e) Time series of C_α_rmsd value in the typical trajectories of peptide and TRPZIP4.

This is perfectly consistent with the observation in the static IR and CD spectroscopy, and the IR temperature jump experiments by Gai and co-workers, which suggested that the turn plays a key role in the folding of β-hairpin; a strong turn-promoting sequence increases the stability of a β-hairpin by increasing its folding rates.[43,44] More interestingly, both turn structure and side-chain hydrophobicity were observed to strongly affect the backbone hydrogen-bond formation.[38] For instance, the stable turn of TRPZIP2 strongly promotes the formation of hydrogen bonds near the turn, and the hydrogen-bond formation follows a "zip-out" mechanism.[45] In contrast, the unstable turn of TRPZIP4 makes difficult the formation of the inner hydrogen bonds (HB5 and HB6). At the same time the strong hydrophobic interactions among the four tryptophan residues in TRPZIP4 allow the easy formation of the hydrogen bonds in the middle of the strands (H3 and H4), which is then followed by the zipping of the rest of the hydrogen bonds, both near the turn and the terminus.[38]

In this article, based on the folding simulation data obtained earlier[38] we performed detailed analyses on the temperature-dependent folding/unfolding thermodynamics for these polypeptides without explicitly considering the temperature effects in solvation, to understand in more detail the mechanisms of β-hairpin formation. We calculated the unfolding free energy of individual polypeptides and analyzed the stability of individual backbone hydrogen bonds in each polypeptide in a large temperature range (270−380 K). These detailed calculations allow us to determine important thermodynamic parameters, such as the melting temperature and the entropy and enthalpy changes in protein folding and in individual hydrogen-bond formation. As a result of using the approximate solvation model (GB^{OBC} model), as shown later, the calculated thermodynamics parameters, such as the melting temperatures, of all polypeptides are in rough agreement with experiments. Nevertheless, these results should be considered as qualitative; through these analyses, we could quantitatively distinguish the enthalpic and entropic contributions in the formation of native β-hairpin structures as well as individual backbone hydrogen bonds and thus try to understand quantitatively the sequence and the temperature dependence of hairpin formation.

It is worth noting that in a very recent article, Tokmakoff and co-workers reported their experimental observation of the temperature-dependent stability of the backbone hydrogen bonds of TRPZIP2 studied by isotope-edited two-dimensional infrared spectroscopy.[46] It was observed in this experiment that as temperature increases from 298 to 358 K, the turn region and its neighboring backbone hydrogen bond (HB6) become more stable, whereas the hydrogen bond at the terminus (HB2) is easily broken. The hydrogen bonds in the middle of the strands (HB3 and HB4) keep contacting at all temperature, whereas their thermal disorder is increased. This is in nice agreement with our simulation results for TRPZIP2 in the present study, which demonstrates that the methodology used here (AMBER FF96 combined with GB^{OBC} implicit model and ITS sampling method) does provide reasonable qualitative descriptions of β-hairpin folding.

Materials and Methods

All MD simulations were performed using AMBER 9.0 package. In the folding simulations of all polypeptides, GB^{OBC} implicit solvent model[32,34] was used. The polypep-

tides were modeled with AMBER FF96 all-atom force field.[39] In these simulations, the salt concentration is set to 0.2 M, and the default surface tension is 0.005 kcal/mol/Å^2. The SHAKE algorithm[47] with a relative geometric tolerance of 10^{-5} is used to constrain all chemical bonds. No nonbonded cutoff was used in simulations. For each polypeptide, multiple independent trajectories were carried out for several hundred nanoseconds. In each trajectory, the fully extended structure of a polypeptide was first subjected to 2500 steps of minimization, then the temperature of the system was established by velocity rearrangement from a Maxwell−Boltzmann distribution at 300 K. After that the system was maintained at 300 K using the weak-coupling algorithm with a coupling constant of 0.5 ps^{-1}. The ITS method[36,48] was used in the production run of each trajectory to enhance the energy sampling on the potential energy surface.

In the ITS method, a desired potential energy range corresponds to a temperature range in MD simulation. The temperature range could be separated into a series of smaller ranges, each having its own energy distribution. Using a quick and robust method, the ratio among the distributions in all small temperature ranges could be adjusted. As a result, the sampling in the entire energy (temperature) range becomes even.[36,48] Moreover, the energy range sampled in the ITS simulation could be largely extended. In the present study, 50 temperatures, evenly distributed in the range of 240−380 K, were used in the ITS method to ensure the efficient sampling of the desired energy range. In each trajectory, which was run at the constant simulation temperature, a large energy range was covered, and many folding and unfolding transitions were obtained.

Backbone Hydrogen-Bond Definition. A hydrogen bond is considered as formed only if the distance between the carbonyl oxygen and the amide hydrogen [C═O···NH] is less than 3.5 Å and the N−H−O angle is greater than 145°.

Heat capacity *Cp* was calculated by

$$C_P = (\langle E^2 \rangle - \langle E \rangle^2)/kT^2 \quad (1)$$

where *E*is the potential energy (the contribution of kinetic energy is treated as a constant and therefore is not taken into account in the calculation), *k*is the Boltzmann constant, and *T* is the temperature.

Unfolding free energy $\Delta G_U$of a polypeptide was calculated by

$$\Delta G_U = kT \ln \frac{P}{1-P} \quad (2)$$

where P is the formation probability of folded hairpin structures. The folded hairpins are defined as structures with at least three of backbone hydrogen bonds formed. This definition of folded hairpin structures is based on the characters of the folded states in the free energy landscape as a function of the radius of gyration of the hydrophobic core (R_g^{core}) and the number of backbone hydrogen bond formed (N_{HB}) ($R_g^{core} < 5$ Å and $N_{HB} \geq 3$, see Figure 4a in ref 34, Figure 5 in ref 45, and Figure 2 in ref 38). The unfolding entropy and enthalpy were then calculated by $\Delta S_U = -(\partial \Delta G_U/\partial T)_P$ and $\Delta H_U = \Delta G_U + T\Delta S_U$, respectively. After we obtained the figure of ΔG_U as a function of temperature (ΔG_U vs *T*), we assume that ΔS_U (ΔH_U) is uniform in the time range of two neighboring temperature points (T_1 and T_2). Then ΔS_U at that temperature range is equal to the slope of ΔG_U vs *T*, $-(\Delta G_{U1} - \Delta G_{U2})/(T_1 - T_2)$, and ΔH_U is the intersection. For each polypeptide, all calculated trajectories in the simulation are involved in the calculation of C_p, ΔG_U, ΔS_U, and ΔH_U. In the meanwhile, the corresponding thermodynamics parameters calculated from individual trajectories are used to generate error bars.

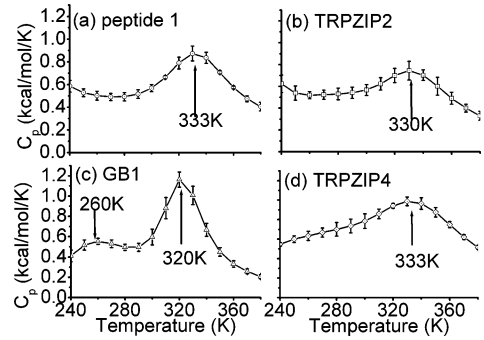


Figure 2. Temperature dependence of the heat capacity for four β-structured polypeptides.

Results and Discussion

Temperature Dependence of the Folded Structure Stability. For each polypeptide under study, more than 10 independent trajectories were run, each starting from the fully extended structure and lasting for several hundred nanoseconds. Consequently, the total simulation time is 1.0 μs for peptide 1, 2.2 μs for GB1, and 2.0 μs for TRPZIP2 and TRPZIP4. Plenty of folding and unfolding events were observed for each polypeptide, e.g., totalling 57 folding events obtained for peptide 1, 19 for GB1, 24 for TRPZIP2, and 26 for TRPZIP4 (see the typical trajectories for peptide 1 and TRPZIP4 as examples in Figure 1). To investigate the conformational transition of the polypeptides, we first calculated their specific heat as a function of temperature in a large temperature range (240−380 K). Using the peak positions in their heat capacity diagrams as shown in Figure 2, we estimated the melting temperatures for the four polypeptides. As seen in this figure, the error bars are small for each polypeptide in the whole temperature range, which means that each trajectory in the simulation is well converged.

Except for the GB1 peptide, the experimentally determined melting temperatures for the four polypeptides are unarguable (e.g., 304.1 ± 0.1 K for peptide 1,[40] 345.0 ± 0.1 K for TRPZIP2,[11] and 343.1 ± 0.1 K for TRPZIP4).[11,43,44] The laser temperature jump experiment by Eaton and co-workers showed that the melting temperature of GB1 peptide is 297.3 K.[10] Nevertheless in the more recent NMR and CD spectroscopy experiments by Anderson and co-workers[49,50] and by Scholtz and co-workers,[51] the GB1 peptide demonstrated less stability in the aqueous solution with the determined melting temperature of 280−285 K or even lower to ∼273 K. The calculated heat capacity diagrams for the four polypeptides except GB1 show a single peak at ∼330 K,

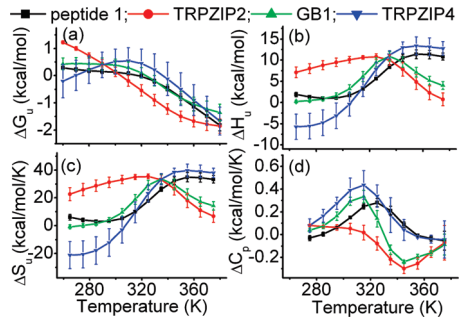


Figure 3. (a) Temperature dependence of the unfolding free energy, (b) enthalpy change, (c) entropy change, and (d) heat capacity change for four polypeptides.

while the diagram of GB1 shows a minor peak at 260 K in addition to the major peak at 320 K. The experimentally determined melting temperature of GB1 is in the middle of the two peaks. For the other three polypeptides, the calculated melting temperatures are about 15 °C too low for TRPZIP2 and TRPZIP4 but are ∼20 °C too high for peptide 1 compared to the experimental data.

The usage of the current force field and the implicit solvent model is considered to be the potential source for the deviation between the calculated and experimental melting temperatures. As mentioned earlier, compared to the other force fields in AMBER, FF96 force field strongly favors the hairpin conformations.[25−28] On the other hand, by comparing REMD simulation results on the folding of polypeptide Ala10 with the GB solvent models to those with explicit TIP3P solvent model, Roe et al. observed that GB models overstabilize α-helical conformations.[52] Moreover, the solvent-accessible surface area (SASA) model, combining with the GB model and accounting for the nonpolar part of the salvation free energy, stabilizes the compact structures and thus artificially increases the transition temperature of the protein.[53] These factors, together, deviate the calculated melting temperatures from the experimentally determined data. Even so, compared to most of the previously reported melting temperatures in various MD simulations on the folding of the above-mentioned hairpins,[4,20,53,54] our results are within reasonable range.

Next we calculated the unfolding free energy as a function of temperature. To fit to the experiment condition, we performed the calculations in the temperature range of 270−380 K. We should note here that the melting temperature calculated using the free energy diagram depends on the definition of the folded structure (see the definition in Materials and Methods Section) and thus can deviate from that obtained from the heat capacity calculations in Figure 2. Figure 3a shows the free energy change for the transition of folded → unfolded state as a function of temperature. As seen from this figure, for peptide 1, TRPZIP2, and GB1, the error bars are rather small in the temperature range under study. For TRPZIP4, the error bars in the low temperature are apparently larger than the other three polypeptides. It is also seen from this figure that all four polypeptides are stable ($\Delta G > 0$) in a rather large temperature range, including room temperature. For all polypeptides except TRPZIP4, only one transition temperature (at which $\Delta G = 0$) is observed in the

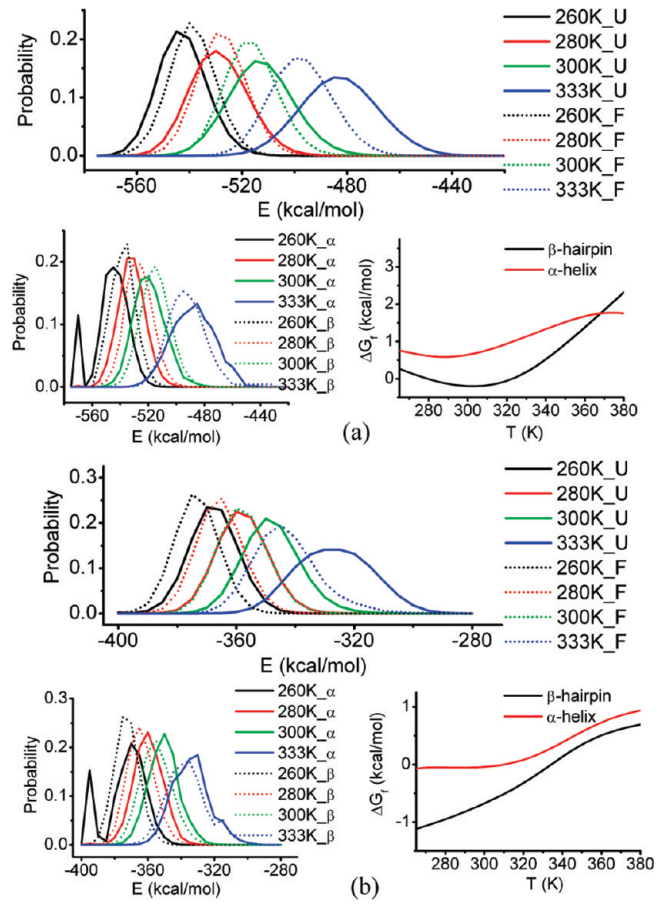


Figure 4. Comparison of unfolded (U) structures to folded (F) structures for (a) TRPZIP4 and (b) TRPZIP2 at different temperatures. Top: the potential energy distribution of the unfolded (solid lines) and folded structures (dash lines); left bottom: the potential energy distribution of α-helix (α, solid lines) and β-hairpin (β, dash lines); and right bottom: the temperature dependence of the unfolding free energy of α-helix and β-hairpin.

entire temperature range studied. The native β-structure of TRPZIP4 becomes unstable at both low and high temperatures, while the other three polypeptides (especially TRPZIP2, the unfolding free energy of which increases with the decreasing temperature in the entire temperature range) have the native state as the stable structure at all low temperatures.

To further understand the structure stability as a function of temperature, we also calculated the distribution of potential energy for both folded and unfolded structures at various temperatures. (As mentioned earlier, the folded structures are defined as the ones with at least three backbone hydrogen bonds formed and the left structures with less than three hydrogen bonds formed refer to the unfolded structures.) Since TRPZIP4 and TRPZIP2 represent two extreme behaviors among the four systems, the results are only shown for these two polypeptides (see Figure 4a and b). It is shown in Figure 4a that at low temperatures, the potential energy of unfolded TRPZIP4 is lower than that of the folded structure. This result shows that the breaking of the TRPZIP4 native structure at low temperatures is mainly due to an enthalpy effect. [A cluster analysis of its non-native structures shows that as temperature decreases, the probability of

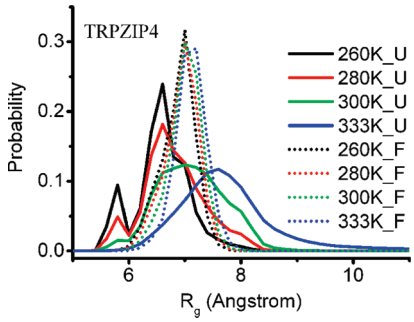


Figure 5. Radius gyration (R_g) distribution of unfolded (U, solid lines) and folded (F, dash lines) structures of TRPZIP4 at different temperatures.

forming α-helix increases for TRPZIP4 (see the comparison of the formation free energy of α-helix to that of β-hairpin in Figure 4a). Due to the unstable turn of this polypeptide, the α-helix indeed has lower energy than the β-hairpin structure (Figure 4a).] In contrast, the unfolded structures of TRPZIP4 at high temperatures and the unfolded structures of TRPZIP2 at both high and low temperatures (Figure 4b) are all of higher energies than their respective folded structures. Therefore, the unfolding of TRPZIP4 at high temperatures is primarily driven by the entropy increase. This entropic effect is also responsible for the heated denaturation of TRPZIP2. On the other hand, since the native β-structure has low energy (partly due to the very stable turn) and the unfolding enthalpy is positive, TRPZIP2 native structure becomes more stable as temperature decreases. Accordingly, the formation probability of α-helix decreases as temperature decreases for TRPZIP2.

Enthalpic versus Entropic Contributions in β-Structure Formation. The unfolding enthalpy and entropy as functions of temperature are shown in Figure 3b and c, respectively. It is seen from these figures that the four polypeptides show dramatically different features. The most noticeable difference is again between TRPZIP4 and TRPZIP2, while peptide 1 and GB1 show behaviors in between. At low temperatures, the unfolding entropy is large and negative for TRPZIP4 and is large but positive for TRPZIP2. These values are small for GB1 (slightly negative) and peptide 1 (slightly positive and showing a minimum at ∼300 K). At high temperatures, the unfolding entropy is positive for all four polypeptides, with TRPZIP4 and peptide 1 having much larger values than the other two. These results again show that the breaking of the native structure of TRPZIP4 is enthalpically driven at low temperatures and entropically driven at high temperatures. Since the behavior of TRPZIP4 resembles a class of proteins which exhibit both cold and heated denaturation of native structures, we performed further analysis on its non-native structures at both low and high temperatures. It is seen from the radius gyration distribution function that the non-native structures are indeed more compact at low temperatures than those at higher temperatures (Figure 5). The former are more and the latter are less compact than the room temperature native structures.

Temperature Dependence of the Folding Mechanism. As discussed earlier, one way of distinguishing different folding mechanisms is to examine the formation order of the different structural elements: the hydrophobic core, the turn, and the hydrogen-bond assembly. We thus show in Figure 6 the free energy landscapes of the folding of the four polypeptides as a function of the collective coordinates, e.g., the number of backbone hydrogen bonds formed (N_{HB}), the number of native hydrophobic contacts formed (N_{HC}), and the root-mean-square displacement of the turn segment ($rmsd_{turn}$) at different temperatures (273, 300, and 350 K). It is seen from these figures that the free energy profiles change with temperature in a similar way for all polypeptides. Although the folding pathways (as shown by the minimum free energy pathways, the dash lines in Figure 6a−d) of peptide 1 and TRPZIP2 are different from those of GB1 and TRPZIP4 at each temperature and the relative stability of the folded structures decrease with increasing temperature, the temperature dependence of the free energy profile shape is weak. In particular, the minimum free energy pathways only shift slightly for each polypeptide. These results show that the folding mechanisms of these polypeptides are largely determined by their sequences but are robust to the temperature change.

For GB1 and TRPZIP4 which possess a disfavored turn structure, the hydrophobic core formation is a barrierless process and thus is very easy to occur in the entire temperature range under study (see the free energy landscape as a function of N_{HC} and $rmsd_{turn}$ in Figure 6c and d). The turn formation, however, constantly associates with a free energy barrier and remains as the rate-limiting step (the energy barrier for the hydrogen-bond formation is smaller than that of the turn formation as shown in the free energy landscape as the function of N_{HB} and $rmsd_{turn}$ in Figure 6c and d).

On the other hand, in the folding of peptide 1 and TRPZIP2 which have the strongly favored turn structure, the hydrogen-bond formation is rate-limiting (see the free energy barrier in the free energy landscape as a function of N_{HB} and $rmsd_{turn}$ in Figure 6a and b). Moreover, the single local minimum in the free energy landscape as a function of N_{HC} and $rmsd_{turn}$ for peptide 1 shows that the turn formation is a barrierless process. The turn keeps stable once it is formed, whereas the hydrophobic interaction is weak. The turn structure of peptide 1 becomes unstable at the high temperature, as revealed by the broader local minimum in the same free energy landscape at the high temperature. The free energy landscape as a function of N_{HC} and $rmsd_{turn}$ for TRPZIP2 at low and middle temperatures (273 and 300 K) has two local minima, corresponding to one state at which only a portion of hydrophobic interactions are formed and the turn is not and the other state at which both the hydrophobic core and the turn are formed. This shows that the hydrophobic core formation is facilitated by the turn formation in TRPZIP4. Therefore at any temperature, the difference of the side-chain hydrophobicitiy and particularly the turn stability changes the shape of the folding free energy landscapes and leads to the different folding mechanism of β-hairpins. Nevertheless the temperature only changes the

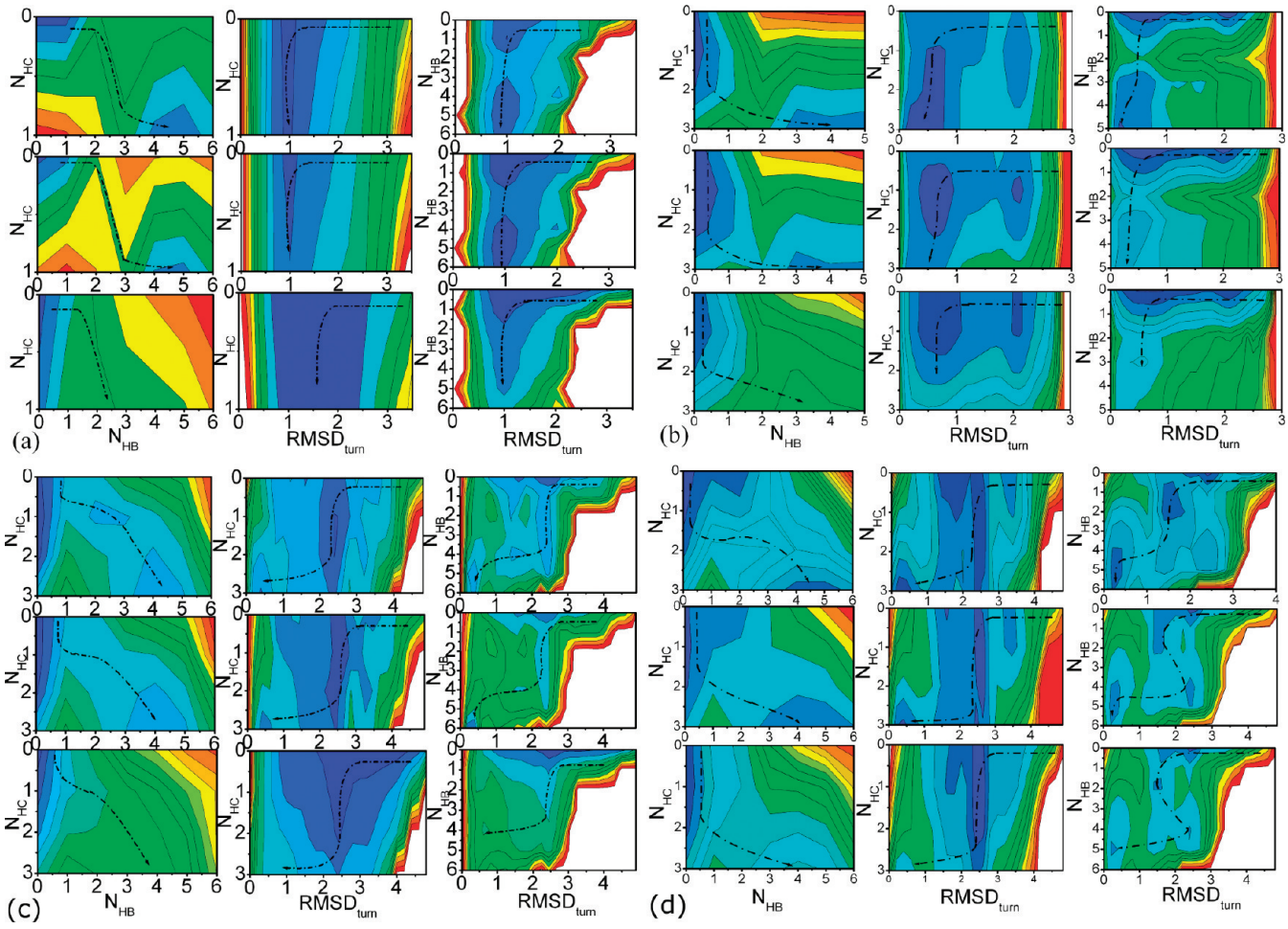


Figure 6. Free energy landscapes as a function of several collective coordinates including the number of backbone hydrogen bonds formed (N_{HB}), the number of native hydrophobic contacts formed (N_{HC}), and the root-mean-square displacement of the turn ($rmsd_{turn}$) for (a) peptide 1, (b) TRPZIP2, (c) GB1, and (d) TRPZIP4 at different temperatures (in all figures (a−d), top panel: 273, middle panel: 300, and bottom panel: 350 K).

stability of a β-hairpin structure but not its folding pathway. These results might be thought to support the usage of high-temperature unfolding in understanding the protein folding mechanism at room temperature.

Temperature Dependence of the Formation and Stability of Backbone Hydrogen Bonds. One interesting question in understanding the β-structure formation is on the sequence of backbone hydrogen-bond formation (e.g., "zip-in" versus "zip-out") and on its temperature dependence. To answer this question, we first calculated the average formation probabilities of individual hydrogen bonds for the four β-hairpins, and the results are shown in Figure 7(a−d). As demonstrated in these figures, the hydrogen-bond stability of the four polypeptides shows strong and different temperature dependence (each consistent with the corresponding free energy diagram shown in Figure 3a). In accordance with the high stability of their folded structures at low temperatures, the majority of the hydrogen bonds of GB1, peptide 1, and TRPZIP2 possess higher stabilities at lower temperatures, while the hydrogen bonds of TRPZIP4 are the most stable in the intermediate temperature region. There are also noticeable common features among the four polypeptides in the temperature dependence of their hydrogen-bond stability. For instance, it is seen from Figure 7 that the stability of

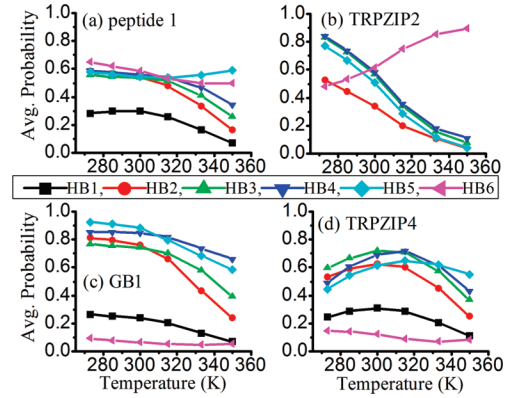


Figure 7. Average formation probability of individual backbone hydrogen bonds for (a) peptide 1, (b) TRPZIP2, (c) GB1, and (d) TRPZIP4 at different temperatures.

most inner hydrogen bonds (HB6 for GB1, TRPZIP2, and TRPZIP4 and HBs 5 and 6 for peptide 1) increases with temperature in the high-temperature range (>320 K). The stability of the other hydrogen bonds, in particular the terminal ones, decreases with temperature in this temperature range. The reason behind the difference between inner and outer hydrogen bonds will be discussed later.

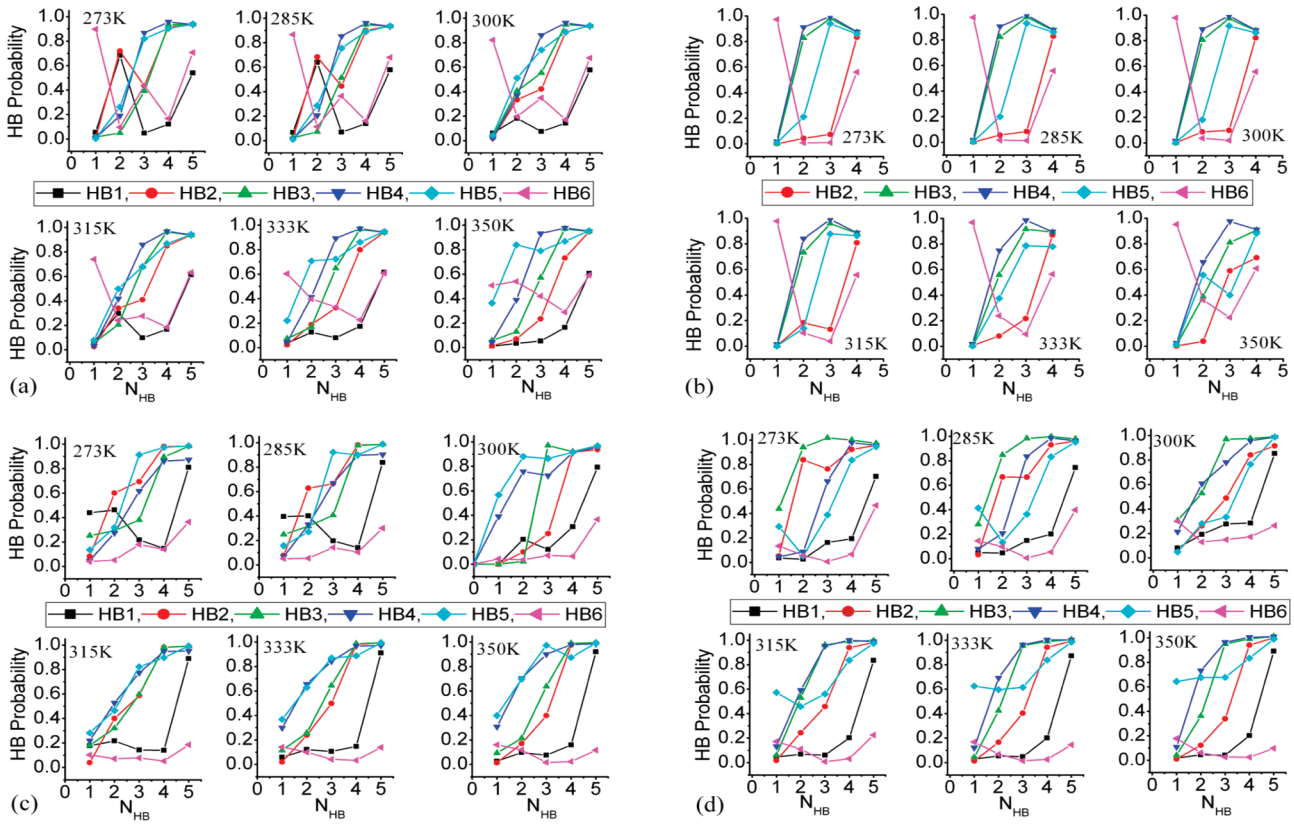


Figure 8. Formation probability of individual backbone hydrogen bonds as a function of the total number of formed hydrogen bonds for (a) peptide 1, (b) TRPZIP2, (c) GB1, and (d) TRPZIP4 at different temperatures.

In TRPZIP2, the stability of its hydrogen bonds except for the most inner one (HB6, whose stability increases with increasing temperature) markedly decreases with temperature. TRPZIP4, on the other hand, exhibits a maximum stability at the intermediate temperature range. Both its unfolding free energy (Figure 3a) and the formation probability of backbone hydrogen bonds, except for HB6, are bell-shaped. The temperatures at which individual hydrogen bonds show the highest stability increase systematically from the terminal (HB1 and HB2) to the inner positions. The stability of backbone hydrogen bonds of peptide1 and GB1, the two polypeptides with weak hydrophobic core clusters, depends on temperature roughly in the same fashion. The main difference between peptide 1 and GB1 is in HB6, which is much more stable in peptide 1 than in GB1 as a result of the more stable turn in the former. Moreover, it appears that HB6 has an effect on the stability of the neighboring HB5, which also shows slightly different temperature dependence in peptide 1 and GB1. Overall, from the above analysis, one concludes that the increase of temperature in the high temperature region destabilizes terminal hydrogen bonds that are further away from the turn and either stabilizes or has little effect on the hydrogen bonds close to the turn. The decrease of temperature in the low temperature region appears to increase the relative stability of the terminal hydrogen bonds and either destabilizes or has little effect on the ones near the turn.

In addition to the temperature dependence of the average stability of hydrogen bonds shown in Figure 7, we also investigated the formation probability of each individual hydrogen bond in the assembling of backbone hydrogen bonds during the formation of the native structure for each polypeptide. These data are shown in Figure 8 as a function of the total number of hydrogen bonds formed. The figure thus illuminates the stability and the formation order of backbone hydrogen bonds along the lowest free energy pathway in each polypeptide. The earlier the hydrogen bond appears in the high-formation probability region, the easier is it formed, namely the more preferential is its formation along the lowest free energy pathway in the assembling of backbone hydrogen bonds. As an example in Figure 8c, although the hydrogen-bond formation order of GB1 along the lowest free energy pathway is roughly 1 → 2 → 3, 4, 5 → 6 at low temperatures (the formation of HB6 is difficult due to the unstable turn of GB1; HB1 forms easily in the folding process, whereas becomes broken again during the assembling of the hydrogen bonds in the middle position), it roughly converts to the "zip-out" mechanism at high temperatures (5, 4 → 3 → 2 → 1→ 6). A similar temperature dependence of hydrogen-bond formation order along the lowest free energy pathway is also observed for the other three polypeptides. As shown in Figure 8a, at low temperatures, the appearance order of hydrogen bonds (from high to low probabilities) of peptide 1 is roughly 6 → 1, 2 → 4, 5 → 3 (HB6 forms easily as a result of the stable turn of peptide 1 but breaks down during the rest of the structure formation process). As temperature increases, the formation of terminal hydrogen bonds becomes more difficult, and at temperatures higher than 300 K, the formation order of hydrogen bonds along the lowest free energy pathway is 6 → 5 → 4 → 3 → 2 → 1, corresponding to a "zip-out" mechanism.

Table 1. Native Backbone Hydrogen Bonds along the Strands and Hydrogen Bonds within the Turn Region in GB1 and TRPZIP4 Polypeptides

backbone HBs		HBs in the turn region	
HB1	T15O−E2H	HBt1	D6Oδ−T9Hγ1
HB2	E2O−T15H	HBt2	T9Oγ1−T11Hγ1
HB3	T13O−T4H	HBt3	D6Oδ−T9H
HB4	T4O−T13H	HBt4	D6O−K10H
HB5	T11O−D6H	HBt5	D6Oδ−A8H
HB6	D6O−T11H	HBt6	D7O−K10H

TRPZIP2 and TRPZIP4 both possess a strong hydrophobic core that is close to the center of strands. As a result of the strong hydrophobic packing of these residues, the formation of hydrogen bonds in the middle of strands (HBs 2−4) is strongly favored compared to those in GB1 and peptide 1. Consistent with the facilitated hydrogen-bond formation in the middle, at low temperatures the hydrogen-bond formation along the lowest free energy pathway initiates from the middle of strands for both TRPZIP2 and TRPZIP4 (see the high formation probability of H3 and H4 in Figure 8b and d). For TRPZIP4, as temperature increases, the stability of the inner hydrogen bonds increases over the terminal ones as discussed earlier, and as a result, the formation of hydrogen bonds along the lowest free energy pathway follows a "zip-out" mechanism at high temperatures (5 → 4 → 3 → 2 → 1, HB6 is unstable due to the unstable turn structure of TRPZIP4). For TRPZIP2, at low temperatures, along the lowest free energy pathway, the formation of HB3 is easier than HB5, and at high temperatures a reverse order is seen, although in all temperatures HB4 appears to be the most easily formed. At all temperatures, the inner hydrogen bond HB6 forms easily as a result of the stable turn of TRPZIP2. However, only at high temperatures, it remains relatively stable during the assembling of the rest hydrogen bonds, again supporting a transition toward the "zip-out" mechanism.

Temperature Dependence of the Formation and Stability of Hydrogen Bonds in the Turn Region. Different from the type I′ turn in peptide 1 and TRPZIP2, the type I turn in GB1 and TRPZIP4 possesses a hydrogen-bond network covering the backbone and side chains of the turn residues (Asp6−Thr11),[24] as organized in Table 1 (HBt1−6) and shown in Figure 9. In addition, a salt bridge is formed between Asp7 and Lys10 side chains. As demonstrated by several novel experiments,[44,55] the hydrogen bonds in the turn region, particularly those formed between Asp6 and other residues, are crucial to the stability of the turn structure and the folding rate of GB1 and TRPZIP4. For instance, the T-jump IR experiment by Du et al. showed that the replacement of Asp6 by alanine decreases the folding rate of TRPZIP4 by ∼9 times, whereas the mutation of Asp7 by alanine only decreases the folding rate of the same polypeptide slightly.[44] Moreover, the NMR stability experiment on GB1 peptide demonstrated that the mutating of either Asp6, Lys10, or Thr9 with alanine destabilizes the turn structure, and the destabilization degree follows the order of Asp6> Lys10> Thr9.[55]

We calculated the average formation probabilities of individual hydrogen bonds in the turn region of GB1 and TRPZIP4, respectively, and the results are shown in Figure 9. As shown in this figure, the only two hydrogen bonds formed between the backbones in the turn region, HBt4 (D7O−K10H) and HBt6 (D7O−K10H), show much higher stabilities than other hydrogen bonds in both GB1 and TRPZIP4. Interestingly, these two hydrogen bonds have the totally opposite temperature dependence on their stability. On the contrary, Hbt2 formed between the side chains of Thr9 and Thr11 always has the extremely low stability in the whole temperature range under study. The left three hydrogen bonds (HBt1, HBt3, and HBt5), which are formed between the side chain of Asp6 and either backbone or side chain of other residues, have the middle stability, especially in TRPZIP4. Since HBt2 contributes least to the hydrogen-bond network in the turn region, the T9A mutation should lead to a minor change in the turn structure stability. On the other hand, the D7A or K10A mutation results in the cancellation of the salt bridge. Nevertheless, this side-chain configuration change will not largely affect the stability of the backbone−backbone hydrogen bonds. The D6A mutation, however, removes the three hydrogen bonds of HBt1, HBt3, and HBt5 which have the middle stability in the hydrogen-bond network in the turn region. As a result, only the D6A mutation will largely decrease the turn structure stability, which is consistent with the experimental observations mentioned above.[44,55]

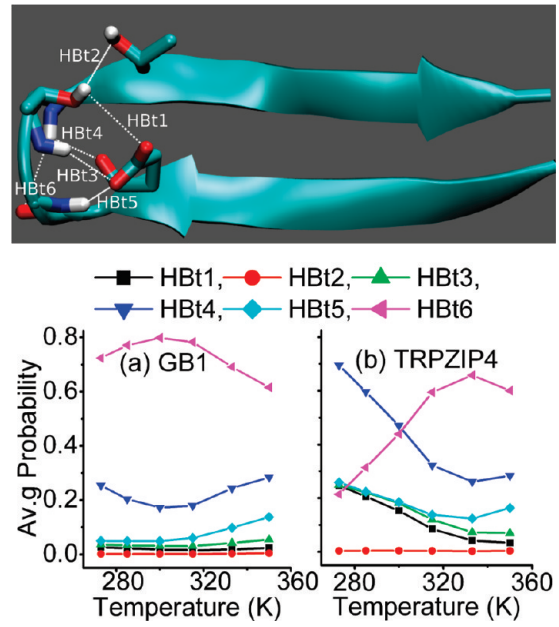


Figure 9. Average formation probability of individual hydrogen bonds in the turn region of (a) GB1 and (b) TRPZIP4. The top panel is the schematic representation of the hydrogen-bond network in GB1 peptide.

Conclusions

In this article, we performed detailed thermodynamics study on the temperature dependence in the folding of several polypeptides that form stable β-hairpin structures. The current study is expected to provide a qualitative understanding of the folding mechanism of β-hairpin structures and a fully atomic description of the structure formation process (par-

ticularly the backbone hydrogen-bond formation) for various amino acid sequences at different temperatures. Simulations were performed using an implicit solvent model, which was proven to yield free energy profiles in reasonable agreement with those obtained using explicit solvent models.[34] Cautions, however, have to be exercised in quantitative interpretation of the simulation data. For example, it will not be surprising that water contributes significantly (if not dominantly) to the folding energy of protein. Although the effects of water to the free energy were considered in a continuum model, there is no guarantee that the effects have been taken into account faithfully. Therefore, the current study should be understood in a more qualitative way. It provides a useful model system for the understanding of protein sequence and temperature dependence of protein folding mechanism. The results of the very recent isotope-edited two-dimensional infrared spectroscopy experiment on the temperature dependence of the backbone hydrogen-bond stability of TRPZIP2,[46] which are consistent with our predictions, demonstrate the validity of the present methodology in exploring the folding mechanism of β-hairpins.

Four polypeptides (peptide 1, TRPZIP2, GB1, and TRPZIP4), which differ at their side-chain hydrophobicity and turn stability, were used as the model systems in the present study. Based on the analysis of a variety of thermodynamics data, we showed that the folding of simple β-hairpin structures is highly sequence and temperature dependent. First, the formation order of the three important structural elements along the minimum free energy pathway in the free energy landscapes appears to be only sequence dependent and to be largely unaffected by the temperature change between 270 to 380 K (see Figure 6). The presence of the strong β-turn-promoting sequence in peptide 1 and TRPZIP2 leads to the folding of β-hairpins following the modified "hydrogen-bond centric" mechanism.[19] On the contrary, the presence of disfavored turn structure in GB1 and TRPZIP4 makes the hairpin folding more consistent with the "hydrophobic core centric" mechanism. Second, both the hydrophobic core cluster and the turn affect the nearby hydrogen bonds. The favored turn structure assists the formation of the inner hydrogen bond (HB6), whereas the strong hydrophobic core cluster strengthens the stability of hydrogen bonds (HBs 3 and 4) in the middle of strands. These effects lead to the different stability and formation order of individual hydrogen bonds at a given temperature. Finally, the pathway for the formation of backbone hydrogen bonds shows a strong dependence on temperature. At low temperatures, the formation of hydrogen bonds is likely initiated from the middle of the strands (at very low temperatures the formation of the terminal hydrogen bonds also become largely favored, but the terminal hydrogen bonds tend to be broken during the assembling of the rest hydrogen bonds, see the examples of GB1 and peptide 1 in Figure 8a and c, respectively). At high temperatures, however, there is a strong tendency for the hydrogen-bond assembling to be initiated from the turn position and to propagate through a "zip-out" mechanism. These results are easily understood in terms of entropy and enthalpy contributions in hydrogen-bond formation.

Without considering the side-chain interactions, it is easy to see that the breaking of terminal hydrogen bonds increases the configuration entropy to a larger extent than the breaking of inner hydrogen bonds does, simply because of their larger separation along the amino acid chain. As a result, one expects that under these hypothetical conditions, the formation of the terminal hydrogen bonds is less favored compared to the inner ones at high temperatures. At high temperatures, since the denatured states also likely have interrupted side-chain interactions, it is thus conceivable that the entropy effects mentioned above dominate the formation probability of the individual hydrogen bonds, and as a result, the stability of the hydrogen bonds increases from the turn to the terminal position, as observed in Figure 7.

At lower temperatures, the enthalpy effects also play important roles, and as a result, the hydrogen-bond stability is strongly affected by their local interactions, the simple order observed at high temperatures therefore disappears. In contrast, the hydrophobic cluster near the hydrogen bonds tends to shield them from solvent attack and effectively creates a low-dielectric environment. As a result, these hydrogen bonds can be expected to be more stable. This stabilization of the hydrogen bonds through this enthalpic effect is most easily seen from the hydrogen-bond stability of TRPZIP2 and TRPZIP4 at low temperatures. It is of great interest for the numerous predictions made through this study, for example, those on the temperature dependence of hydrogen-bond stability and those on the compactness of non-native states (see Figure 5), to be tested by experiments.

Acknowledgment. This study was supported financially by Peking University. We thank Professor Andrei Tokmakoff for discussions. Y.Q.G. is a 2008 Changjiang Scholar. Q.S. thanks China Postdoctoral Science Foundation funded project.

References

(1) Galzitskaya, O. V. *Mol. Biol.* **2002**, *36*, 607–612.

(2) Hughes, R. M.; Waters, M. L. *Curr. Opin. Struct. Biol.* **2006**, *16*, 514–524.

(3) Zhou, R. H.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12777–12782.

(4) Zhou, R. H.; Berne, B. J.; Germain, R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931–14936.

(5) Schaefer, M.; Bartels, C.; Karplus, M. *J. Mol. Biol.* **1998**, *284*, 835–848.

(6) Dinner, A. R.; Lazaridis, T.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9068–9073.

(7) Blanco, F. J.; Jimenez, M. A.; Pineda, A.; Rico, M.; Santoro, J.; Nieto, J. L. *Biochemistry* **1994**, *33*, 6004–6014.

(8) Blanco, F. J.; Rivas, G.; Serrano, L. *Nat. Struct. Biol.* **1994**, *1*, 584–590.

(9) Bonvin, A. M. J. J.; van Gunsteren, W. F. *J. Mol. Biol.* **2000**, *296*, 255–268.

(10) Munoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196–199.

(11) Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 5578–5583.

(12) Searle, M. S.; Ciani, B. *Curr. Opin. Struct. Biol.* **2004**, *14*, 458–464.

(13) Searle, M. S.; Zerella, R.; Dudley, D. H.; Packman, L. C. *Protein Eng.* **1996**, *9*, 559–565.

(14) Zerella, R.; Evans, P. A.; Ionides, J. M. C.; Packman, L. C.; Trotter, B. W.; Mackay, J. P.; Williams, D. H. *Protein Sci.* **1999**, *8*, 1320–1331.

(15) Espinosa, J. F.; Munoz, V.; Gellman, S. H. *J. Mol. Biol.* **2001**, *306*, 397–402.

(16) Munoz, V.; Henry, E. R.; Hofrichter, J.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5872–5879.

(17) Silva, R. A. G. D.; Sherman, S. A.; Keiderling, T. A. *Biopolymers* **1999**, *50*, 413–423.

(18) Kolinski, A.; Ilkowski, B.; Skolnick, J. *Biophys. J.* **1999**, *77*, 2942–2952.

(19) Tsai, J.; Levitt, M. *Biophys. Chem.* **2002**, *101*, 187–201.

(20) Garcia, A. E.; Sanbonmatsu, K. Y. *Proteins* **2001**, *42*, 345–354.

(21) Pande, V. S.; Rokhsar, D. S. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9062–9067.

(22) Nguyen, P. H.; Stock, G.; Mittag, E.; Hu, C. K.; Li, M. S. *Proteins* **2005**, *61*, 795–808.

(23) Wei, G. H.; Mousseau, N.; Derreumaux, P. *Proteins* **2004**, *56*, 464–474.

(24) Yoda, T.; Sugita, Y.; Okamoto, Y. *Proteins* **2007**, *66*, 846–859.

(25) Ono, S.; Nakajima, N.; Higo, J.; Nakamura, H. *J. Comput. Chem.* **2000**, *21*, 748–762.

(26) Zaman, M. H.; Shen, M. Y.; Berry, R. S.; Freed, K. F.; Sosnick, T. R. *J. Mol. Biol.* **2003**, *331*, 693–711.

(27) Higo, J.; Ito, N.; Kuroda, M.; Ono, S.; Nakajima, N.; Nakamura, H. *Protein Sci.* **2001**, *10*, 1160–1171.

(28) Kamiya, N.; Higo, J.; Nakamura, H. *Protein Sci.* **2002**, *11*, 2297–2307.

(29) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

(30) Zhou, R. H. *Proteins* **2003**, *53*, 148–161.

(31) Shell, M. S.; Ritterson, R.; Dill, K. A. *J. Phys. Chem. B* **2008**, *112*, 6878–6886.

(32) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383–394.

(33) Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156–169.

(34) Shao, Q.; Yang, L. J.; Gao, Y. Q. *J. Chem. Phys.* **2009**, *130*, 195104/1–195104/6.

(35) Ozkan, S. B.; Wu, G. A.; Chodera, J. D.; Dill, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 11987–11992.

(36) Gao, Y. Q. *J. Chem. Phys.* **2008**, *128*, 064105/1−064105/5.

(37) Gao, Y. Q.; Yang, L. J. *J. Chem. Phys.* **2006**, *125*, 114103/1–114103/5.

(38) Shao, Q.; Wei, H. Y.; Gao, Y. Q. *J. Mol. Biol.* **2010**, *402*, 595–609.

(39) 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.

(40) Santiveri, C. M.; Pantoja-Uceda, D.; Rico, M.; Jimenez, M. A. *Biopolymers* **2005**, *79*, 150–162.

(41) Gronenborn, A. M.; Clore, G. M. *Science* **1991**, *254*, 581–582.

(42) Santiveri, C. M.; Santoro, J.; Rico, M.; Jimenez, M. A. *J. Am. Chem. Soc.* **2002**, *124*, 14903–14909.

(43) Du, D. G.; Zhu, Y. J.; Huang, C. Y.; Gai, F. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 15915–15920.

(44) Du, D. G.; Tucker, M. J.; Gai, F. *Biochemistry* **2006**, *45*, 2668–2678.

(45) Yang, L. J.; Shao, Q.; Gao, Y. Q. *J. Phys. Chem. B* **2009**, *113*, 803–808.

(46) Smith, W. A.; Lessing, J.; Ganim, Z.; Peng, C. S.; Tokmakoff, A. *J. Phys. Chem. B* **2010**, *114*, 10913–10924.

(47) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.

(48) Yang, L. J.; Shao, Q.; Gao, Y. Q. *J. Chem. Phys.* **2009**, *130*, 124111/1–124111/8.

(49) Fesinmeyer, R. M.; Hudson, F. M.; Andersen, N. H. *J. Am. Chem. Soc.* **2004**, *126*, 7238–7243.

(50) Olsen, K. A.; Fesinmeyer, R. M.; Stewart, J. M.; Andersen, N. H. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 15483–15487.

(51) Huyghues-Despointes, B. M. P.; Qu, X. T.; Tsai, J.; Scholtz, J. M. *Proteins* **2006**, *63*, 1005–1017.

(52) Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C. *J. Phys. Chem. B* **2007**, *111*, 1846–1857.

(53) Hayer, N. R.; Singh, R. R. P.; Cox, D. L. arXiv:1009.0303v1 [q-bio.BM] published online: September 1, 2010.

(54) Nymeyer, H. *J. Phys. Chem. B* **2009**, *113*, 8288–8295.

(55) Kobayashi, N.; Honda, S.; Yoshii, H.; Munekata, E. *Biochemistry* **2000**, *39*, 6564–6571.

CT100436R

JCTC Journal of Chemical Theory and Computation

Topological Characterization of the Electron Density Laplacian in Crystals. The Case of the Group IV Elements

A. Otero-de-la-Roza* and Víctor Luaña*

Departamento de Química Física y Analítica, Facultad de Química, Universidad de Oviedo, 33006 Oviedo, Spain

Received May 21, 2010

Abstract: We discuss the rigorous characterization of the electron density Laplacian of crystals in terms of its topological properties: critical points (CPs), zero flux surfaces, and accumulation and depletion basins. Comparison with the atomic shell structure is exploited to characterize the numerous core critical points so that the important effort is applied to the more significant valence structure. Efficient algorithms are adapted or newly developed for the main tasks of topological study: finding the critical points, determining the 1D and 2D bundles of (3, − 1) and (3, + 1) CPs, and integrating well-defined properties within the accumulation and depletion basins. As an application of the tools and concepts developed we perform a quantitative analysis of chemical bonding on group IV semiconductors, mainly devoted to the properties of the diamond phase but also including the main effects of allotropy influences on these elements. The topological analysis of the Laplacian provides a complementary and very different image than the topology of the electron density. Whereas the Laplacian graphs show a qualitative agreement with Lewis classical model, the basin population analysis excludes direct quantitative relationships with Lewis pair and octet rules. In addition to the expected core and valence basins, all group IV elements show very important interstitial basins, that accumulate a large number of electrons and dominate the compressibility behavior of the crystals.

1. Introduction

The Quantum Theory of Atoms in Molecules (QTAIM) has been in use for some three decades now to determine chemical bonding properties as true observables of the electron wave function.[1−8] QTAIM studies are mainly based on the topological analysis of the electron density, $\rho(\boldsymbol{r})$, particularly the characterization of its critical points (ρ-CP), and the integration of every kind of quantum observable within the attraction basin of atomic nuclei. Both, theoretical and experimental electron densities have been the subject of such scrutiny and, in fact, QTAIM is the mainstream technique for the experimental analysis of chemical bonding.[3,9]

The name Quantum Chemical Topology (QCT) has been proposed by Popelier[10−12] to include the growing collection of methods inspired in the seminal work of Bader.[1] Studies based on the topological analysis of ρ, $\nabla^2\rho$, the electron localizing function (ELF),[13] the source function,[14] the momentum density,[15] the electron pair density,[16] and many other similar properties would be included under the umbrella of QCT. Beyond sharing many common techniques and language, the development of QCT is becoming a revolutionary perspective in the old quest for the chemistry holey grail: obtaining every bit of information regarding chemical bonding that exists in the experimentally or computationally available part of the wave function, with no recourse to unfounded simplifications. In other words, providing a strict physical foundation to the chemical bonding.

The value of the Laplacian of the electron density, $\nabla^2\rho(\boldsymbol{r})$, on the ρ-CPs has been used to distinguish between shared- and closed-shell bonding cases.[17] The Laplacian has also been instrumental in characterizing hydrogen bonding,[18−21] predict sites of nucleophilic and electrophilic attack, as well

* To whom correspondence should be addressed E-mail: alberto@carbono.quimica.uniovi.es; victor@carbono.quimica.uniovi.es.

10.1021/ct100269e
Published on Web 11/10/2010

as the reactivity propension,[22] used to follow chemical reactions[23,24] and to distinguish between Lewis nucleophilic and acidic zones of a molecule,[25,26] to name just some of its most prominent roles.

The electron density Laplacian has also received a lot of recent attention from the community of developers of exchange and correlation functionals. Quantum Monte Carlo investigations on the strongly inhomogeneous electron gas,[27,28] later extended to small molecules[29] and crystals,[30] have shown that "the nonlocal contributions to [the exchange-correlation energy density] contain an energetically significant component, the magnitude, shape, and sign of which are controlled by the Laplacian of the electron density".[27]

Such an important role sharply contrast with the fact that only a very small number of articles have been devoted to the full topological characterization of the electron density Laplacian,[6,11,31−34] and none of them has examined condensed matter systems. The full topological characterization of a three-dimensional (3D) scalar field like $\nabla^2\rho(\boldsymbol{r})$ requires, in our opinion, being able to complete, at least, three different tasks: (1) localizing efficiently the critical points (L-CPs in this case); (2) tracing the 1D (*field lines*) and 2D (*interbasin surfaces*) regions that start or end on the first- and second-order saddle points; and (3) integrating local properties within the 3D basins of the Laplacian maxima and minima. Whereas the first capability is included in the AIMPAC package[35] since the eighties and from this on many other molecular topological codes, Popelier's MORPHY[36] (since the 2001 version) is the only code that currently offers the three capabilities.

This article is devoted to the complete characterization of the electron density Laplacian in solids. In the next section, we examine the meanings and usages that are associated to the Laplacian. Section 3 considers briefly the consequences of the cusps that the nonrelativistic electron density shows at the nuclear positions under the Born−Oppenheimer approximation. Section 4 reviews the atomic shell structure that markedly influences the Laplacian topology in molecules and solids. This analysis of the shell structure will prove determinant for the difficult task of finding and classifying the abundance of critical points that can be found. Section 5 introduces the algorithms that we have adapted or created to complete the full topological analysis of the Laplacian. The analysis and discussion of our results in a representative set of crystals is the subject of section 6. We have selected the group IV elements for the first application of the new techniques: the examination of the five, C−Pb, elements on the same diamond phase will let us determine the influence of the number of electron shells, whereas the characteristic allotropy of these elements provides a window to the effect of crystal geometry on the topological properties. The article ends with a discussion of the main outcomes of our analysis and the prospect of the possible role of the presented techniques on the Quantum Chemical Topology studies.

2. Meaning of the Laplacian of the Electron Density

The most immediate meaning of the Laplacian comes from the geometrical interpretation: $\nabla^2\rho(\boldsymbol{r})$ provides the local curvature of the electron density at $\boldsymbol{r}$. Hence, if $\nabla^2\rho(\boldsymbol{r}) < 0$, the electron density at $\boldsymbol{r}$ is larger, on average, than in the differential region surrounding this point. In other words, the electron density is locally enhanced or accumulated at $\boldsymbol{r}$. Similarly, the electron density is locally depleted at those points where $\nabla^2\rho(\boldsymbol{r}) > 0$. This role is consequence of the Laplacian being the trace of the Hessian or curvature matrix: $\boldsymbol{H}(\boldsymbol{r}) = \nabla\otimes\nabla\rho(\boldsymbol{r})$. Accordingly, $\nabla^2\rho(\boldsymbol{r})$ represents the accumulated curvature of the three-dimensional neighborhood of $\boldsymbol{r}$.

This geometrical role of the Laplacian is stressed in the following equation, included by James C. Maxwell[37] in his *Treatise on Electricity and Magnetism* (1873):

$$\rho(\boldsymbol{r}_a) - \rho_{av}(\boldsymbol{r}_a) = -\frac{\tau^2}{10}\nabla^2\rho(\boldsymbol{r}_a) + \mathcal{O}(\tau^4) \qquad (1)$$

where $\rho_{av}(\boldsymbol{r}_a)$ represents the average value of the electron density for all the points within a sphere of radius τ centered on $\boldsymbol{r}_a$, and $\mathcal{O}(\tau^4)$ is a small term of the order of τ^4. This equation led Maxwell to propose calling $L(\boldsymbol{r}) = -\nabla^2\rho(\boldsymbol{r})$ the *"concentration of ρ at the point* $\boldsymbol{r}$, *because it indicates the excess of the value of ρ at that point over its mean value in the neighborhood of the point"*.[37]

The use of $L(\boldsymbol{r})$ instead of $\nabla^2\rho(\boldsymbol{r})$ has also been customary in the QTAIM literature at least since 1984.[38] Looking for a more intuitive comparison with the behavior of the density, maxima in $L(\boldsymbol{r})$ represent maximal concentration of density in a similar way to maxima in $\rho(\boldsymbol{r})$ that represent the maximal accumulation of electronic charge which are typical of nuclei. We will adhere to this tradition and our topological characterization will be referred from now on to $L(\boldsymbol{r})$ rather than to $\nabla^2\rho(\boldsymbol{r})$.

The importance of the Laplacian for the QTAIM theory is further evidenced by the fundamental relationships in which it appears. Of tantamount importance is the *local virial relationship*:[17,39,40]

$$\frac{\hbar^2}{4m}\nabla^2\rho(\boldsymbol{r}) = \mathcal{V}(\boldsymbol{r}) + 2\mathcal{G}(\boldsymbol{r}) \qquad (2)$$

where $\mathcal{G}(\boldsymbol{r})$ is the kinetic energy density and $\mathcal{V}(\boldsymbol{r})$ is the electronic potential energy density. $\mathcal{G}(\boldsymbol{r})$ is everywhere positive, and $\mathcal{V}(\boldsymbol{r})$ is everywhere negative, so the sign of $\nabla^2\rho(\boldsymbol{r})$ indicates which of the two contributions to the local virial theorem dominates at every point. Acidic regions, characterized by $\nabla^2\rho(\boldsymbol{r}) > 0$ show a kinetic energy dominance, whereas regions of basic character, $\nabla^2\rho(\boldsymbol{r}) < 0$, show the dominance of the electronic potential energy.

Requiring that the virial relationship holds in an arbitrary region Ω leads to the QTAIM characterization of atomic basins by bounding zero flux surfaces

$$\nabla\rho(\boldsymbol{r})\cdot\boldsymbol{n}(\boldsymbol{r}) = 0 \qquad (3)$$

where $\boldsymbol{n}(\boldsymbol{r})$ is the normal vector to the surface at $\boldsymbol{r}$. When and only when Ω is defined in this way then

$$\int_\Omega \nabla^2\rho(\boldsymbol{r})\mathrm{d}\boldsymbol{r} = \oint_{S(\Omega)} \nabla\rho(\boldsymbol{r})\cdot\boldsymbol{n}(\boldsymbol{r})\mathrm{d}\boldsymbol{r} = 0 \qquad (4)$$

In a similar way, $\nabla^2\rho$ can be shown to be the connection between different, but equally grounded, forms of the kinetic energy density. For instance[1,41]

$$\mathscr{K}(\boldsymbol{r}) = \mathscr{G}(\boldsymbol{r}) - \frac{\hbar^2}{4m}\nabla^2\rho(\boldsymbol{r}) \tag{5}$$

where

$$\mathscr{K}(\boldsymbol{r}) = -\frac{\hbar^2}{4m}\{\nabla^2 + \nabla'^2\}\Gamma^{(1)}(\boldsymbol{r},\boldsymbol{r}')|_{\boldsymbol{r}'\to\boldsymbol{r}'} \tag{6}$$

$$\mathscr{G}(\boldsymbol{r}) = \frac{\hbar^2}{2m}\{\nabla^2 \cdot \nabla'\}\Gamma^{(1)}(\boldsymbol{r},\boldsymbol{r}')|_{\boldsymbol{r}'\to\boldsymbol{r}'} \tag{7}$$

and $\Gamma^{(1)}(\boldsymbol{r},\boldsymbol{r}')$ is the nondiagonal one-electron density matrix. Only when Ω satisfies the zero flux condition $\mathscr{K}(\Omega) = \mathscr{G}(\Omega)$ and, in general, all the different forms of defining locally the kinetic energy density yield equivalent results.

The Ω regions defined in terms of the zero-flux surface condition constitute the electron density *basins* and represent the fundamental partition of the molecular and crystalline space according to the QTAIM theory. *Atomic attraction basins*, containing a single nuclear ρ-CP, and *minima repulsion basins*, containing a single cage ρ-CP, are two alternative partitions that exhaustively divide the crystal into nonoverlapping regions. Both are made, in fact, by joining appropriately *primary bundles*, mathematically defined as the space region made of the gradient lines joining together a particular nucleus with a particular cage ρ-CP.[42]

This partitioning of space is not exclusive of the electron density. Any differentiable C^2 scalar field provides a partition with similar topological properties, including $L(\boldsymbol{r})$. We define, for instance, the *accumulation basins* and the *depletion basins* as the regions bounded by zero flux surfaces of $\nabla L(\boldsymbol{r})$ and associated to maxima and minima of $L(\boldsymbol{r})$, respectively. The electron density, however, stands alone as the only partition that guarantees the correct behavior of all quantum mechanical operators on the local level. As a consequence, we will be able to integrate, within the $L(\boldsymbol{r})$ basins, local properties, like the volume and charge, but not such nonlocal properties as the kinetic energy.

3. Laplacian and Nuclear Cusps in the Electron Density

It is well-known that nonrelativistic Born−Oppenheimer electron densities exhibit a singularity or cusp at the fixed nuclear positions.[43] This is a consequence of the infinite asymptote of the Coulomb potential because of a fixed-point-like nucleus. This singularity, that would preclude the existence of derivatives, including $\nabla^2\rho(\boldsymbol{r})$, at the nuclear positions is usually discarded by assuming the mapping of a smooth function identical in value and properties to $\rho(\boldsymbol{r})$ except that the cusps are eliminated and substituted by some rounded shape.[1]

It should be noticed, however, that the cusps are removed if nuclei are modeled as small but finite-size particles, as it is routinely done in atomic[44] and solid-state relativistic calculations.[45,46] Cusps would also disappear if the electron density is the result of some statistical ensemble where the nuclear motion is taken into account.

Table 1. Rank and Signature ($\mathbf{r}$, σ) of the Hessian Matrix Can Be Used to Classify the Different Types of Critical Points[a]

($\mathbf{r}$, σ)	type	name	abbrev.	AD	RD
(3, −3)	maximum	nucleus	NCP	3D	0D
(3, −1)	saddle-1	bond	BCP	2D	1D
(3, +1)	saddle-2	ring	RCP	1D	2D
(3, +3)	minimum	cage	CCP	0D	3D
(2, −2)	2D-maximum			2D	0D
(2, +0)	2D-saddle			1D	1D
(2, +2)	2D-minimum			0D	2D
(1, −1)	1D-maximum			1D	0D
(1, +1)	1D-minimum			0D	1D

[a] The common name and the usual abbreviation is indicated as the third and fourth column, respectively. AD and RD are the dimensions of the attraction and repulsion basins, respectively, created by the critical point.

4. Topological Structure of $L(\boldsymbol{r}) = -\nabla^2\rho(\boldsymbol{r})$ and the Atomic-like Shell Structure

$L(\boldsymbol{r})$ induces a complete partition of the space into distinct and complementary regions by means of the gradient vector field, $\nabla L(\boldsymbol{r})$, described in terms of the critical points, $\boldsymbol{r}_c$:

$$\nabla L(\boldsymbol{r}_c) = \boldsymbol{0} \tag{8}$$

which can be classified by the rank and signature of the Hessian matrix:

$$\boldsymbol{H}(\boldsymbol{r}_c) = \nabla \otimes \nabla L(\boldsymbol{r}_c) \tag{9}$$

Table 1 gives a summary of the type and properties of the regular (i.e., 3D) and degenerated (2D or 1D) critical points of $L(\boldsymbol{r})$, and establishes the notation for the rest of the paper. We have decided to keep the same denominations already popular when describing the topology of $\rho(\boldsymbol{r})$. If confusion is possible, we will refer to ρ-CP or L-CP to distinguish between the critical points of both scalar fields.

Both, the electron density and the $L(\boldsymbol{r})$ scalar fields inherit their basic structure directly from the atoms. The electron density, peaked at the nuclear position, is formed by a collection of exponential arcs (one for each electronic shell, see Figure 1, left) connected by regions of larger curvature.

The shell structure is far more clearly revealed by the $L(\mathbf{r})$ function (see Figure 1, right). Starting from the nucleus, $L(\mathbf{r})$ shows a succession of maxima, zeros, minima, and zeros that we have labeled as K_+, KK, K_- (for the 1s electrons), KL, L_+, LL, L_- (the 2sp shell), LM, M_+, MM, M_- (the 3spd shell), and so on. These features in the radial $L(\mathbf{r})$ function correspond to spheres of degenerate critical points in the 3D $L(\boldsymbol{r})$ field for the isolated atom. The spherical symmetry is broken by the influence of the neighbor atoms in a molecule or solid, but there is a neat difference between the effects shown by the core and valence shells. Whereas the internal shells keep unaltered the distance to the nucleus of their topological spots, the outermost shell loses to a large part its atomic origin and it is determined by the competition between the neighbor atoms, like it happens to the electron density itself.

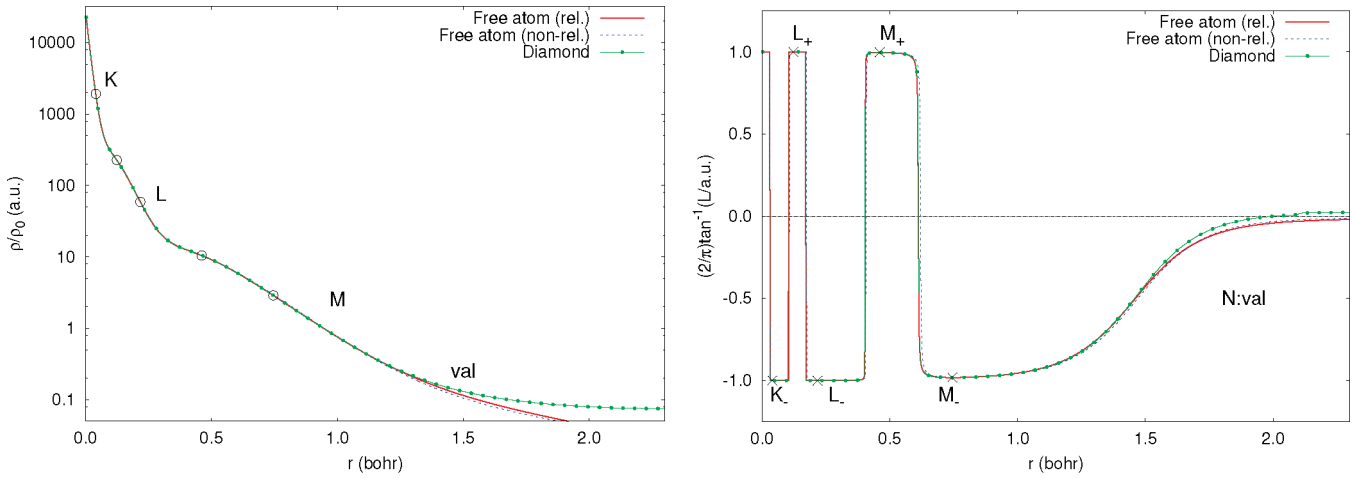


Figure 1. Radial structure of the electron density (left) and $L(\mathbf{r})$ (right) of Ge. The results from a nonrelativistic Hartree−Fock atomic calculation,[47] a relativistic Dirac−Fock atomic calculation,[44,48,49] and a relativistic FPLAPW calculation of the diamond phase of germanium are shown in both plots. The latter treats core states fully relativistically, while the scalar relativistic approximation is used for valence states.[49] The points K_+, L_+, are the maxima of $L(\mathbf{r})$ and K_-, $L-$, are the minima. Notice the log scale of the left plot. Similarly, an arctangent scale is used on the right plot, transforming the $(-\infty, +\infty)$ range of $L(\mathbf{r})$ into $[-1, 1]$, with a minimal distortion of the region close to zero, the most significative one from a chemical bonding perspective.

Figure 1 also shows that the core region experiences a negligible relative modification because of the environment, as well as the relativistic effects, both on $\rho(\mathbf{r})$ and $L(\mathbf{r})$. It is only the valence region that suffers the significant changes.

The association of the Laplacian of the electron density and the atomic shell structure was first formulated by Bader et al.[17,50] ref. 50 in particular, started the association of an electronic shell with a pair of spherical shells of alternating charge concentration and charge depletion, later accepted by most researchers. These initial studies, performed on light elements of the main groups, were later extended to heavier atoms by Sagar et al.,[51] Shi and Boyd,[52] and Kohout et al.[53] Those works have shown that, using Bader's definition, the Laplacian of ρ fails sometimes to resolve the valence from the inner shells, a problem that starts to occur in some elements of the fourth row of the periodic table and becomes more common as we progress to heavier elements. This was a common argument for the introduction of the Electron Localization Function (ELF)[13] that shows similar properties to $L(\boldsymbol{r})$ but maintains the core−valence difference up to the sixth row, at least. Following Eickerling and Reiher,[54] authors of the most extensive analysis to date of the Laplacian of relativistic multiconfigurational atomic calculations, we consider that all topological features of $L(\mathbf{r})$ should be considered: maxima, minima, and zeros. For elements $Z >$ 18, the assumed valence electronic shell is not observable as a local maximum in the positive region of $L(\mathbf{r})$, but the maximum remains even though in the negative region.

In any case, our main interest regarding the shell structure lies in the possibility of knowing in advance how a spherical shell will contribute to the topology of $L(\boldsymbol{r})$ in a molecule or crystal. This will help us in classifying and naming the large number of L-CPs that typically occur. The number and type of L-CPs in a unit cell is restricted by the Morse relationship,

$$n - b + r - c = 0 \tag{10}$$

where n, b, r, and c are the number of NCP, BCP, RCP, and CCP, respectively, per unit cell. The degeneracy of an atomic spherical shell is broken by the potential of the neighbor atoms. In the case of core shells, the resulting L-CPs move negligibly in the radial direction, and the radial curvature of $L(\boldsymbol{r})$ also changes negligibly. Moreover, the L-CPs produced by the symmetry breaking of a core shell are topologically equivalent to a polyhedron and the Euler relationship must be fulfilled:

$$\text{vertices} - \text{edges} + \text{faces} = 2 \tag{11}$$

When the core shell corresponds to a maximum of $L(\mathbf{r})$, the radial curvature is negative and the breaking of atomic symmetry can only produce NCP, BCP, and RCP, but not (3, +3) points. In this case, NCPs are the vertices, the 1D repulsion basins of BCPs form the edges, and the 2D repulsion basins of RCP's are the faces of the polyhedron. Accordingly, the radial maximum in $L(\mathbf{r})$ produces

$$n_+ - b_+ + r_+ = +2 \tag{12}$$

a contribution of +2 to the global Morse sum. Contrarily, a core minimum of $L(\mathbf{r})$, with a positive radial curvature, decomposes into CCPs (vertices), RCPs (edges), and BCPs (faces) but no NCPs. Their contribution to the Morse sum is, therefore

$$-b_- + r_- - c_- = -2 \tag{13}$$

Table 2 shows the mean radii of the mostly spherical shells for C−Pb in their cubic diamond phase. The K_+ position is essentially coincident with the nucleus and it is considered to remain a single point, although the behavior of the electron density at the nucleus depends, on relativistic calculations, of the type of model used to describe the nuclear charge. From our point of view it is enough to consider that the unsplit K_+ point contributes +1 to the Morse sum. Each pair of successive minimum plus maximum radial shells compensate to produce a null net Morse contribution. A last, uncompensated radial minimum shell would change the global core contribution from +1 to −1. This is what happens

Table 2. Mean Radius of the Spherical Shell Diamond Structure[a]

	C	Si	Ge	Sn	Pb	
K_+	0	0	0	0	0	K_+
KK	0.16844	0.07018	0.02915	0.01729	0.00804	KK
K_-	0.23042	0.09600	0.03992	0.02381	0.01134	K_-
KL	0.84696	0.27785	0.10452	0.06248	0.03426	KL
L_+	0.98123	0.32665	0.12327	0.07362	0.04015	L_+
LL		0.47154	0.17100	0.10058	0.05363	LL
L_-		0.59910	0.21708	0.12735	0.06825	L_-
LM		1.59407	0.40355	0.21606	0.11223	LM
M_+		1.83266	0.46067	0.24758	0.12846	M_+
MM			0.61172	0.31794	0.16148	MM
M_-			0.74538	0.38810	0.19615	M_-
MN			2.03817	0.65597	0.29462	MN
N_+				0.73175	0.33214	N_+
NN				0.92299	0.41801	NN
N_-				1.09627	0.49510	N_-
NO						NO
O_+					0.96772	O_+
OO						OO
O_-					1.11951	O_-
$R_{NN}/2$	1.45931	2.22190	2.31464	2.65579	2.88807	

[a] The horizontal lines mark the end of the core shell structure.

in C, Si, Ge, and Pb, but not in Sn, according to Table 2. Anyway, the consideration of what constitutes a core and what a valence shell depends upon the criterion used to accept that two Laplacian critical points have the same distance relative to the nucleus.

Our description in this section is not specific of the $L(\boldsymbol{r})$ function, but it can be applied to any scalar field showing an atomic shell structure. The ELF function comes immediately to mind, but many other similar fields can also be included.

5. Implementing the Topological Analysis of the Electron Density Laplacian in Crystals

The rich literature exploring the topology and properties of the Laplacian in molecules or for light elements has been discussed in previous sections. However, to the best of our knowledge, this is the first article devoted to the full topology of $L(\boldsymbol{r})$ in crystals and addressing solids with arbitrarily heavy elements. The reason behind this apparent neglect rests in the extreme behavior of the $L(\boldsymbol{r})$ scalar function. In this section, we describe the important modifications and new techniques that must be introduced in the usual QCT algorithms when analyzing $L(\boldsymbol{r})$. The principles and methods that will be described can be readily generalized to other scalar fields showing atomic shell structure, such as the ELF,[13,55] and the noninteracting electron pressure,[56] for instance.

As described previously, $L(\boldsymbol{r})$ displays a shell structure around the atoms: there exist regions surrounding the nuclei that present large value fluctuations, specially near the nucleus. If the atom were isolated, $L(\boldsymbol{r})$ would have a strict spherical symmetry. After the formation of the crystal, $L(\boldsymbol{r})$ is distorted, acquiring the symmetry of the local point group. However, the shell structure is maintained and the distortion is small, specially in the core region. The heavier elements show the sharpest oscillations. As an example, the value of $L(\boldsymbol{r})$ in the Pb atom varies more up to 14 orders of magnitude between the K_+ ($L \approx 10^8$) and L_- ($L \approx -10^6$) radial extrema. This example shows clearly the necessity for specialized algorithms to deal with shell-structured scalar fields.

The topological characterization of a scalar field $f(\boldsymbol{r})$, rests on three main tasks: the integration of the trajectories of ∇f, the localization of all the critical points of f, and the integration of properties on the attractor basins. The computational bottleneck is the latter by far.

Starting with the seminal work by Biegler-König,[57,58] a number of methods have been developed to deal with the problem of locating the interattractor surface (IAS) and integrating the property densities.[34,59−62] Most of them, however, are density-specific and not suitable for generic scalar fields, where the basins have different shapes[11] and geometrical properties. In the case of light molecules, Popelier has successfully applied a collection of strategies, including a specialized octree algorithm,[34] to the calculation of the basin properties of $L(\boldsymbol{r})$.

In this work, we have preferred to adopt the old and simple but robust bisection technique. Several reasons have lead us to use it: (a) the algorithm is general enough to deal with any basin shape, provided there are no multiple crossings of the ray and the IAS, and even that can be taken into account; (b) it is reasonably efficient if the gradient paths are traced sensibly (see below); and (c) it allows arbitrary precision of integration by increasing the number of rays, with an error given by the cubature employed and the precision of the IAS. Bisection depends on the efficiency of the gradient path tracing, and its performance is independent of the scalar field, so the three above-mentioned tasks of the QCT study reduce to two core routines: tracing gradient paths and finding the whole set of critical points.

5.1. Source of the $L(\boldsymbol{r})$ Function. First, let us examine the technical details of the analysis of $L(\boldsymbol{r})$ in solids. The $L(\boldsymbol{r})$ field is a quantum-mechanical observable, and as such, its features and the insights gained from its analysis are independent of the method used in its determination. In practice, however, both theory and experiments have shortcomings and the $L(\boldsymbol{r})$ field is obtained only as an approximation, and is subject to the limitations of the determination method.

Our approach to $L(\boldsymbol{r})$ in this article is based on the full-potential (linearized) augmented plane-waves method (FPLAPW)[63,64] as implemented in wien2k.[45,46] In the FPLAPW method, the real space is partitioned into regions, roughly corresponding to the core and valence zones of the solid. These regions are: the muffin tins, noncolliding spheres centered around each atom, and the interstitial space, that fills the rest of the crystal. The basis functions (APW or LAPW) and the density are split, behaving differently in each region. In particular, the density is expressed as

$$\rho(\boldsymbol{r}) = \begin{cases} \sum_{LM} \rho_{LM}(\boldsymbol{r}) Y_L^M(\hat{r}) & \boldsymbol{r} \in S_\alpha \\ \sum_{\boldsymbol{K}} \rho_{\boldsymbol{K}} e^{i\boldsymbol{K}\cdot\boldsymbol{r}} & \boldsymbol{r} \in 1 \end{cases} \tag{14}$$

In the muffin tin of the α nucleus (S_α), with radius R_{mt}, the density is expressed as a spherical harmonics (Y_L^M) expansion referred to its corresponding center (the position of the

nucleus), while in the interstitial region (I), ρ is written as a plane-wave expansion, where $\boldsymbol{K}$ is a reciprocal lattice vector. The Laplacian of the density has a similar form

$$\nabla^2\rho(\boldsymbol{r}) = \begin{cases} \sum_{LM} f_{LM}(r) Y_L^M(\hat{r}) & \boldsymbol{r} \in S_\alpha \\ -\sum_{\boldsymbol{K}} K^2 \rho_{\boldsymbol{K}} e^{i\boldsymbol{K}\boldsymbol{r}} & \boldsymbol{r} \in 1 \end{cases} \quad (15)$$

with

$$f_{LM}(r) = \rho''_{LM} + \frac{2}{r}\rho'_{LM} - \frac{L(L+1)}{r^2}\rho_{LM} \quad (16)$$

where primes represent differentiation in the radial coordinate. In both expressions, the expansion is carried not to the infinite set of local spherical harmonics and plane waves, but it is included only up to certain cutoff values $L_{\max}$ and $K_{\max}$. This truncation creates discontinuous gaps on the muffin tin surface that need to be dealt with by the topological algorithms. The absence of continuity in $\rho(\boldsymbol{r})$ and $L(\boldsymbol{r})$ is thus a basic, inescapable feature of the FPLAPW densities.

How does the lack of continuity at the muffin tin surface affect the results of the analysis? We have found that the discontinuity is invisible to both the CP localization method and to the gradient path tracer, provided there are no spurious critical points on the muffin tin surface. This applies for both the density and its Laplacian. The spurious CPs trap gradient path integrations in an anomalous way and, being the consequence of a discontinuous gap, their number do not fulfill the Morse sum criterion. Tuning of the calculation parameters seems to be the only way around the problem, being most sensitive to the variation of $R_{\mathrm{mt}}K_{\max}$ and R_{mt}. A very large value of $L_{\max}$ could, in principle, be effective but it is not possible to go beyond a hard-coded 10 value without a severe reprogramming of many parts of the wien2k[45,46] code.

There is a further consequence of the discontinuity that must be taken into account. The discontinuity introduces a surface term in the integral of $L(\boldsymbol{r})$ over the unit cell. Using Gauss theorem, it can be expressed as a flux of the density gradient across the muffin surfaces:

$$\int_{\mathrm{cell}} L(\boldsymbol{r})\mathrm{d}\boldsymbol{r} = -\sum_\alpha \oint_{S_\alpha} (\nabla\rho_\alpha - \nabla\rho_i)\cdot \mathrm{d}\boldsymbol{S} \quad (17)$$

where α runs over the atoms in the cell, ρ_i and ρ_α are the density function forms in the interstitial, and the α muffin (eq 14), respectively. This result has two important consequences: the $\mathscr{G}$ and $\mathscr{K}$ forms of the kinetic energy are not equivalent, the one entering the total energy expression being $\mathscr{G}$[65,66] and the integral of $L(\boldsymbol{r})$ within the topological ($\nabla\rho$) basins is not zero. Note that, although the sum in eq 17 can be easily computed to correct the integral of L over the cell, it is not possible to do the same to the atomic expectation values of L, except in the cases where no muffin crosses the interatomic surface.

To test our algorithms, we have selected a set of systems, containing an assortment of bonding characters and structures. These are listed in Table 3, along with their main calculation conditions. The number of k-points in the full first Brillouin zone (1BZ) was chosen so that the energy converged to the precision of the code in all the crystals but the metallic (Li, Mg, Na, Pb, and Sn). The FPLAPW calculations have been done using the Perdew−Burke−Ernzerhof[67] GGA functional. We have used the runwien text interface[68] to wien2k to carry out the calculations and a modified version of critic[69] to perform the QTAIM analysis.

Table 3. Test Cases for the Evaluation of the Algorithms Related to the Analysis of $L(\boldsymbol{r})$[a]

crystal	phase	$R_{\mathrm{mt}}K_{\max}$	k-points (1BZ)	$R_{\mathrm{mt},1}$	$R_{\mathrm{mt},2}$
AlN	blende	9.0	1000	1.77	1.30
AlN	wurzite	9.0	1000	1.77	1.30
AlP	blende	9.0	4000	1.50	2.10
BN	blende	10.0	1000	1.45	1.20
BP	blende	10.0	2000	1.59	2.00
C	graphite	9.0	60000	1.20	
C	diamond	9.0	60000	1.30	
GaN	blende	9.0	1000	1.70	1.40
GaN	wurzite	9.0	1000	1.70	1.40
GaP	blende	11.0	1000	2.00	2.00
Ge	diamond	11.0	8000	2.21	
Li	BCC	9.0	60000	2.20	
Mg	HCP	9.0	60000	2.90	
NaCl	rock salt	9.0	60000	2.30	2.30
Na	BCC	9.0	60000	2.20	
Pb	diamond	10.0	60000	2.30	
Pb	FCC	10.0	60000	2.30	
Si	diamond	11.0	8000	2.21	
Sn	diamond	10.0	8000	2.30	
Sn	tetragonal	10.0	14000	2.60	

[a] All the geometries correspond to the experimental structures,[70,71] with the exception of the diamond phase of Pb, that was optimized using FPLAPW and a Perdew−Burke−Erzenhof exchange-correlation functional, to a cell parameter of a = 13.339 bohr.

All the crystals in the Table 3 have been examined for the existence of spurious CPs by means of a direct and simple test: a number of $n_\theta \times n_\phi$ points are uniformly distributed in spheres of radii $R_{\mathrm{mt}} - \varepsilon$ and $R_{\mathrm{mt}} + \varepsilon$ around each atom, with $\varepsilon = 10^{-3}$ bohr. Consequently, one of the spheres is inside the muffin tin and the other is in the interstitial region. Every point in the inner sphere has a counterpart in the other one, at a distance 2ε. The radial component of the gradient of the scalar field, f_{r}, is then computed at each pair of points. Spurious CPs exist whenever a pair of points differ in the sign of their f_{r}. The calculation conditions were then modified for each crystal until some combination produced a density and Laplacian free from discontinuities on all tested directions. This extense exploration of calculation parameters has revealed that, in these systems, the occurrence of spurious CPs does not depend on the number of k-points, but it is affected heavily by the values of $R_{\mathrm{mt}}K_{\max}$ and R_{mt}. In most cases, the electron density was correct under a wide range of calculation conditions and it was the Laplacian the function posing real difficulties to the QTAIM analysis. The process of finding good parameters for III−V elements was specially painstaking, as no pattern for the occurrence of trouble was apparent.

5.2. Navigation in the $L(\boldsymbol{r})$ Surface. Now, we describe the computational details of the analysis of $L(\boldsymbol{r})$. The first step is the computation of the shell structure of the scalar

field around each atom. On bonding, the inner shells of $L(\boldsymbol{r})$ are largely unaffected, while the valence shells are distorted to accommodate the environment. The position of the radial maxima and minima are determined by bracketing and golden section search in a number of rays emerging from the nucleus and uniformly distributed. The resulting shells are classified into: (a) valence shells, distorted by the chemical environment and possibly not fulfilling the shell Euler sum, and (b) core shells, resembling the atomic shells and fully closed. The innermost core shells do not convey any chemical information and are the most difficult to treat from the point of view of the algorithms. Therefore, for the heavier elements, we define an effective nucleus, that is composed of the real nucleus and a number of shells up to, and including, a X_- shell, where $X = K, L, ...$ The structure within this effective nucleus is ignored by the algorithms, except that their actual contribution to the Morse sum is taken into account. The election of a shell that is a radial minimum as the frontier makes the effective nucleus a basin of $L(\boldsymbol{r})$, easy to integrate as a sphere.

The localization of the critical points of $L(\boldsymbol{r})$ is based on the Newton−Raphson (NR) method with two modifications to take into account the shell structure of $L(\boldsymbol{r})$. First, the same seeding scheme as in the electron density is used.[42,69] Namely, the irreducible wedge of the WS cell (IWS) is built by applying the local symmetry of the origin to the full WS cell. The IWS is split into disjoint tetrahedra and each of them undergoes a barycentric subdivision process to determine the starting points for the NR exploration. This method allows the rapid localization of the symmetry-forced CPs of $L(\boldsymbol{r})$. This scheme is inherently suited for solids, and far superior in efficiency to its molecular counterparts based on the search between pairs, triplets and quartets of atoms. In this particular scalar field, this strategy allows the localization of all the valence CPs, but requires high subdivision levels to locate the shell CPs. Therefore, we have added a new set of seed points, placed at the spherical shells of each atom, so as to locate the in-shell CPs.

The second necessary modification to NR consists of switching to spherical coordinates near the nuclei, at distances lower than the largest core shell. When the NR sequence of points falls into one of the inner shells, the transformation to spherical coordinates effectively decouples the radial from the in-shell (angular) coordinates. The Hessian matrix is thus approximately blocked. Special care must be taken regarding the numerical errors in the computation of the elements of the gradient (f_i, $i = r, \theta, \phi$) and the Hessian (f_{ij}, $i,j = r, \theta, \phi$). The radial components (f_r and f_{rr}) are much larger than the ones involving the angular coordinates, so that the transformation of derivatives from Cartesian (x_k values) to spherical (s_i) coordinates:

$$\frac{\partial}{\partial s_i} = \sum_k \frac{\partial}{\partial x_k}\frac{\partial x_k}{\partial s_i} \tag{18}$$

$$\frac{\partial^2}{\partial s_i \partial s_j} = \sum_k \frac{\partial}{\partial x_k}\frac{\partial^2 x_k}{\partial s_i \partial s_j} + \sum_{kl} \frac{\partial^2}{\partial x_k \partial x_l}\left(\frac{\partial x_l}{\partial s_i}\right)\left(\frac{\partial x_k}{\partial s_j}\right) \tag{19}$$

is subject to cancellation errors. In FPLAPW densities, a workaround to this problem is calculating the nonradial terms of the gradient and Hessian in the muffin tin by using an expression of $L(\boldsymbol{r})$ where the spherical term $L = 0,\ M = 0$ has not been summed. This eliminates the dominant spherical contribution to the value of $L(\boldsymbol{r})$ and prevents the cancellation errors in the transformation to spherical coordinates.

Additionally, a modified stop criterion for NR is necessary in the shells. Usually, a CP is located whenever $|\nabla f(\boldsymbol{r})| < \varepsilon$ where ε is customarily set to 1×10. However, it is too difficult to find a radial component of the gradient below that threshold, because of the rapidly oscillatory character of $L(\boldsymbol{r})$ in the core region. Therefore, when in core shells, the norm of the gradient is calculated using only the angular coordinates.

The topologies of the test cases have been determined with the modified NR method. All the CPs of $L(\boldsymbol{r})$ have been located in less than two minutes on a typical desktop PC (see Table 4). The topologies fulfill the global and shell Morse conditions. If smaller effective nuclei are considered, the success of the modified NR algorithm in the innermost shells varies with the atom involved. For example, all the CPs of Ge are located, even if the effective nucleus is shrunk to only one shell, but this is not possible for chlorine (both in NaCl and in Cl_2), for which the K_-, L_+, and L_- shells do not fulfill the local Morse sum using the default parameters of our modified NR. Popelier[32] and Gatti[72] have published finding the full topology of $L(\boldsymbol{r})$ by using the eigenvector-following method,[73] which we have also implemented using a transformation to spherical coordinates in core shells, as described above. By comparing to our modified NR method, we have found that eigenvector-following does not improve on the results of our method in terms of efficiency or success in locating CPs of $L(\boldsymbol{r})$, so we have opted for the simpler NR approach.

Table 4. Summary of the Analysis of $L(\boldsymbol{r})$ in the Test Cases[a]

crystal	phase	$L(\boldsymbol{r})$ topology	t_{top} (s)
AlN	blende	4(52)|6(152)|6(144)|5(44)	32.0
AlN	wurzite	7(26)|13(76)|14(74)|8(24)	97.1
AlP	blende	5(56)|8(192)|8(192)|5(56)	31.4
BN	blende	3(48)|6(128)|7(128)|6(48)	15.6
BP	blende	6(92)|8(216)|8(184)|6(60)	19.7
C	graphite	4(20)|10(70)|10(70)|5(20)	29.5
C	diamond	2(64)|4(224)|5(208)|3(48)	9.7
GaN	blende	4(52)|6(152)|6(144)|5(44)	25.3
GaN	wurtzite	7(26)|14(78)|13(72)|6(20)	90.7
GaP	blende	5(56)|7(168)|7(168)|5(56)	16.0
Ge	diamond	2(40)|3(176)|4(192)|3(56)	10.6
Li	BCC	2(28)|4(120)|4(114)|3(22)	3.6
Mg	HCP	4(14)|8(56)|7(58)|3(16)	2.5
NaCl	rock salt	4(64)|6(200)|5(200)|3(64)	5.6
Na	BCC	2(14)|6(78)|4(112)|1(48)	6.4
Pb	diamond	2(40)|3(96)|3(96)|2(40)	15.4
Pb	FCC	2(28)|3(96)|3(112)|3(44)	4.0
Si	diamond	4(168)|6(352)|6(256)|3(72)	11.6
Sn	diamond	2(40)|3(176)|4(192)|3(56)	15.8
Sn	tetragonal	2(20)|6(80)|7(104)|4(44)	19.8

[a] The full topology of $L(\boldsymbol{r})$ is shown in *n*|*b*|*r*|*c* format. Each of these fields is of the form *x*(*y*) where *x* and *y* are the number of CPs in the asymmetric and conventional unit cells respectively. All the topologies fulfill the global and shell Morse sum conditions. The cpu times correspond to a typical desktop PC. The computational cost increases in lower symmetry systems (e.g., wurtzite) because more NR search seeds are used.[69]

5.3. Integration on the $L(\boldsymbol{r})$ Basins. The other fundamental task in the analysis of $L(\boldsymbol{r})$ is the integration of gradient paths (GP). The basin integration method that we have chosen is bisection so the purpose of tracing of gradient trajectories boils down to two tasks: (a) locating the terminal points of the paths originating at a given gradient source and (b) depicting the behavior of the gradient vector field. In both of them, the primary concern is not the extreme accuracy of the paths but the computational efficiency, for GP tracing is the bottleneck of the integration of atomic properties. Consequently, we have chosen a simple explicit Euler method and included some modifications, similar to those introduced in the NR method, to provide for the special shell structure of $L(\boldsymbol{r})$.

As in NR, the integration of the GPs near the nuclei is done in spherical coordinates. For sufficiently inner shells, the adaptive step shrinks to the point of making the navigation impracticable, even with high order one-step methods. To avoid this problem, we have eliminated the radial coordinate from the GP tracing algorithm, provided several conditions are met: (a) the trajectory is traversing one of the known shells, (b) the shell has the correct curvature ($f_{rr} > 0$ if the trajectory goes downward, $f_{rr} < 0$ if upward), (c) the step size is smaller than 10^{-3} bohr, and (d) the absolute value of the radial curvature (f_{rr}) is greater than a certain value (in the $L(\boldsymbol{r})$ scalar field, this value is 1 au). For safety, the radial coordinate is optimized whenever the Newton-like step $|f_r/f_{rr}|$ is greater than 10^{-3} bohr. For core shells, this is seldom the case, as they remain approximately spherical on bonding. The radial coordinates are considered again whenever one of the above conditions, except (c), is not met, thereby taking into account the possibility of partial shells of $L(\boldsymbol{r})$.

If only the ending critical points of the GPs are needed, we have found that the use of β-spheres (the *atomic trust spheres*[74]) accelerates the assignment of the terminal atom by 4−6 times. The β-spheres are centered around each atom, with their radii being initially set to 75% of the distance of the atom to its closest bond critical point. The election of this radii follows from the ready availability of the complete list of CPs. When a GP enters a β-sphere, the terminal atom is automatically assigned to the owner of the sphere. This method prevents the expensive tracing of the gradient path near the nucleus, where the step size shrinks to prevent the gradient path from bouncing around the critical point. The case where one of the IAS retraces into one of the β-spheres is rare but possible, and a clear indication of this situation is the basin limit being assigned incorrectly to the surface of a β-sphere. Should this happen, the β-sphere radius is decreased by a factor and the basin limit is recalculated. We have checked that, in all the systems we examined, the value of the integrated atomic properties is not affected by the use of β-spheres.

6. Topological Analysis of Group IV Allotropes

As a first application of the newly developed topological tools we are going to analyze the diamond phase of the group IV elements: from C to Pb. We will also examine, as a term of comparison, some allotropes that compete in stability with the diamond phase or are even the most stable phase under normal pressure and temperature: the graphite phase of C, the white or β-Sn, and the fcc or α-Pb.

Table 5. wien2k Calculation Parameters Used for the Topological Analysis

crystal	KPTS	RKMAX	RMT
C	60000	9.0	1.30
C (graphite)	60000	9.0	1.20
Si	8000	11.0	2.21
Ge	5000	11.0	2.10
Sn	8000	10.0	2.30
Sn (white)	14000	10.0	2.60
Pb	60000	10.0	2.70
Pb (fcc)	60000	10.0	2.30

Table 6. Calculated (First Row) and Experimental (Second Row) Equilibrium Properties

crystal	*a* (Å)	*c* (Å)	*B* (GPa)
C	3.5827		444.3
	3.568		444.0
Si	5.4797		91.2
	5.4307		99.2
Ge	5.5934		75.1
	5.6574		77
Sn	6.6563		49.1
	6.49		53
Pb	7.0589		28.7
Sn (white)	5.9213	3.2280	63.7
	5.8197	3.1749	57.9
Pb (fcc)	5.0550		36.0
	4.9502		43.2

6.1. Electronic Structure Calculations. The electronic structure of all the crystals has been obtained from FPLAPW calculations using the wien2k[45,46] code with the new runwien text interface.[68] All calculations have been done using the Perdew-Burke-Ernzerhof[67] exchange and correlation functional. Care has been taken to converge the calculations with respect to all relevant internal parameters, in particular the muffin tin radius (RMT), the number of planewaves used as basis set (controlled by RKMAX), and the grid used to integrate the first Brillouin zone (controlled by KPTS). Table 5 shows the value of these essential parameters used in our calculations. It is important to notice that our requirements are somewhat different from those of a typical wien2k calculation. In most cases the FPLAPW codes are run using a muffin tin zone as large as possible to diminish the computational effort. In our case, we have to play carefully with the parameters to avoid, as much as possible, the discontinuity of the electron density and its derivatives at the muffin boundaries.

We have determined the equilibrium properties of the crystals, as a check of the calculations, and the most relevant results are collected in Table 6. In general, the equilibrium cell lengths are obtained within 1−2% of the experimental values, and the bulk moduli within 3% for the light elements and 10% for Sn and Pb. We have not included the graphite equilibrium geometry as it is well-known that typical GGA functionals are defective in their representation of intermolecular interactions and the graphene sheets are too loosely bound.

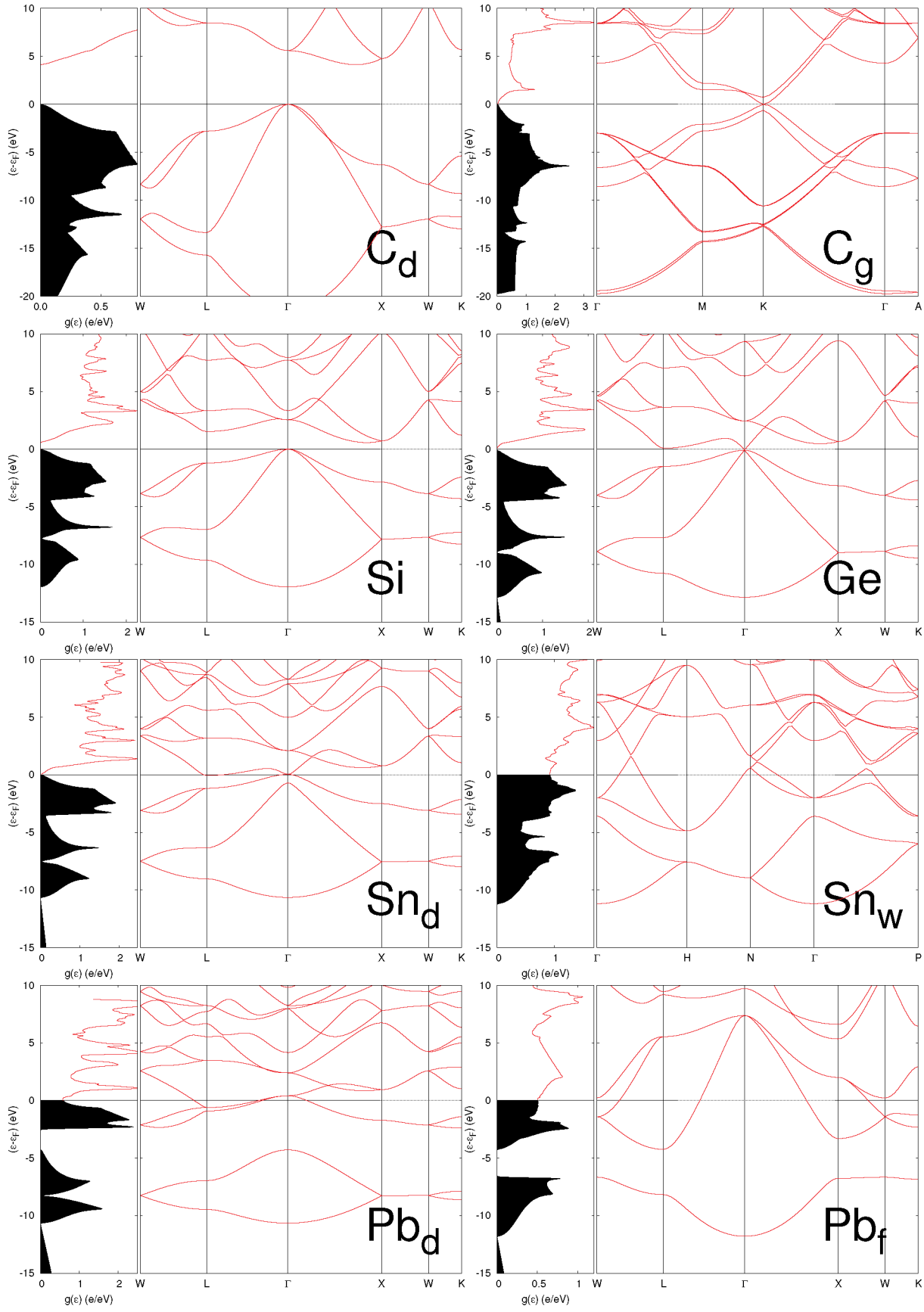


Figure 2. Band structure and density of states (DOS) for the diamond phases of C−Pb, graphite C, white Sn, and fcc Pb. The wien2k GGA calculations correspond to the experimental geometry, when available, and to the predicted equilibrium geometry otherwise.

C and Si are predicted to be semiconductors in the diamond phase, with an indirect band gap of 4.13 and 0.58 eV, respectively. These band gaps are significantly smaller than the experimental values of 5.5 and 1.1 eV, following

Table 7. Topology of the Electron Density for the Diamond Structures[a]

CP	Wyckoff	position	$\rho(\boldsymbol{r}_c)$	$\nabla^2\rho(\boldsymbol{r}_c)$
C	*8a*	(1/8, 1/8, 1/8)	1.28217×10^2	
b	*16c*	(0, 0, 0)	0.24005	−0.55144
r	*16d*	(1/2, 1/2, 1/2)	0.21507	+0.10522
c	*8b*	(3/8, 3/8, 3/8)	0.01323	+0.07595
Si	*8a*	(1/8, 1/8, 1/8)	1.89360×10^3	
b	*16c*	(0, 0, 0)	0.08369	−0.12344
r	*16d*	(1/2, 1/2, 1/2)	0.00559	+0.01353
c	*8b*	(3/8, 3/8, 3/8)	0.00324	+0.00897
Ge	*8a*	(1/8, 1/8, 1/8)	3.12582×10^4	
b	*16c*	(0, 0, 0)	0.07502	−0.03556
r	*16d*	(1/2, 1/2, 1/2)	0.00436	+0.01021
c	*8b*	(3/8, 3/8, 3/8)	0.00257	+0.00652
Sn	*8a*	(1/8, 1/8, 1/8)	1.89122×10^5	
b	*16c*	(0, 0, 0)	0.05427	+0.00270
r	*16d*	(1/2, 1/2, 1/2)	0.00260	+0.00510
c	*8b*	(3/8, 3/8, 3/8)	0.00148	+0.00315
Pb	*8a*	(1/8, 1/8, 1/8)	2.96523×10^6	
b	*16c*	(0, 0, 0)	0.03885	+0.03082
r	*16d*	(1/2, 1/2, 1/2)	0.00166	+0.00258
c	*8b*	(3/8, 3/8, 3/8)	0.00099	+0.00162

[a] Wyckoff positions correspond to the $Fd\bar{3}m$ group. All values in atomic units.

Table 8. Topology of the Electron Density for the Non-Diamond Allotropes[a]

CP	Wyckoff	position	$\rho(\boldsymbol{r}_c)$	$\nabla^2\rho(\boldsymbol{r}_c)$	*x*
C1	*2b*	(0, 0, 1/4)	1.28333×10^2		
C2	*2c*	(1/3, 2/3, 1/4)	1.28325×10^2		
b_1	*2a*	(0, 0, 0)	0.00588	+0.01767	
b_2	*6h*	(*x*, 2*x*, 1/4)	0.30199	−0.88798	0.83316
r_1	*6g*	(1/2, 1/2, 0)	0.00410	+0.01400	
r_2	*2c*	(2/3, 1/3, 3/4)	0.02271	+0.13476	
c	*4f*	(1/3, 2/3 x)	0.00340	+0.01317	0.47142
Sn	*4a*	(0, 0, 0)	1.89128×10^5		
b_1	*4b*	(0, 0, 1/2)	0.02838	+0.01723	
b_2	*8c*	(0, 1/4, 1/8)	0.03751	+0.01222	
r_1	*8d*	(0, 1/4, 5/8)	0.01303	+0.01733	
r_2	*16f*	(*x*, 1/4, 1/8)	0.00805	+0.01397	0.29807
c	*16g*	(*x*, *x*, 0)	0.00803	+0.01391	0.28456
Pb	*4a*	(0, 0, 0)	2.96493×10^6		
b	*24d*	(1/4, 1/4, 0)	0.01788	+0.02222	
r	*32f*	(*x*, *x*, *x*)	0.01287	+0.01518	0.31837
c_1	*8c*	(1/4, 1/4, 1/4)	0.01208	+0.01361	
c_2	*4b*	(1/2, 1/2, 1/2)	0.00674	+0.00904	

[a] The Wyckoff positions correspond to the space groups $P6_3/mmc$ (graphite), $I4_1/amd$ (white Sn), and $Fm\bar{3}m$ (α-Pb).

the well-known trend of DFT GGA calculations. Ge and Sn in the diamond phase, and the graphite phase of C have a null band gap in our calculations, as it can be observed in Figure 2, but the Fermi level occurs in the limit of two bands, so the electronic density of states (DOS) is essentially zero at the Fermi level. White Sn and the fcc and diamond phases of Pb are completely different, with the Fermi level occurring in a energy region with a large DOS that confers a clear metallic behavior to these crystals.

The wave functions that we will analyze in the next sections correspond to the calculations performed at the experimental geometry. We have also examined the topological properties of the wave functions obtained from DFT LCAO calculations with the crystal code,[75,76] but given that the results are essentially equivalent, they will not be discussed again.

6.2. Topology of the Electron Density. Tables 7 and 8 describe the position and main properties of the critical points of the electron density for the diamond and nondiamond crystal structures of C−Pb. One of the most relevant observations is that all the diamond structures show the same topology, with a single type for each of *n*, *b*, *r*, and *c* critical points (NCP, BCP, RCP, and CCP, respectively), all of them occupying symmetry-defined positions within the unit cell. Once revealed this uniformity, the properties of the CPs clearly show important differences for each element. The electron density at the nuclear position shows a markedly correlation with the atomic number: $\rho_n = 0.11378Z^{3.7415}$, with a linear correlation coefficient of corr(ρ, Z) = 99.4%, similar to the law cited by Bader[1] for nonrelativistic calculations. The properties at the BCP are particularly significant. The electron density at the BCP decreases as the cubic cell length increases: $\rho_b = 6.412656a^{-2.57024}$ (corr = −99.5%). The BCP Laplacian shows a well-defined trend in the C−Pb sequence, increasing from the negative −0.55144 e/bohr^5 of C, typical of a highly covalent bond, to the small but positive +0.03082 e/bohr^5 of Pb, closed-shell like and similar to the values found in many metals. This negative/positive $\nabla^2\rho_b$ difference separates C, Si, and Ge on the covalent side, and Sn and Pb on the closed-shell group.

The non-diamond allotropes offer some fine aspects for contrast and comparison to the above crystals. Graphite, for instance, shows two different types of C−C BCPs: a strong bond that keeps together the graphene sheets, and a much weaker BCP gluing together the sheets. The first BCP has a larger electron density and a more negative Laplacian than the diamond structure, close, in fact, to the values shown by the C−C BCP in benzene. The β-Sn and α-Pb structures show a marked difference in topology with respect to their diamond crystals: the BCP is weaker (smaller ρ_b and more positive $\nabla^2\rho_b$) and the electron density is globally flatter[77] ($f = \rho_c/\rho_b$ is 21.4% in β-Sn and 37.7% in α-Pb versus 2.7% and 2.5% in their respective diamond structures).

All together, we can see that the crystalline structure has a strong influence on the electron density topology, and that a clear group trend can be observed only after we examine the different elements on a common crystal phase.

6.3. Topology of the $L(\boldsymbol{r})$ Field. One of the most remarkable aspects of the topology of the $L(\boldsymbol{r})$ field is that the total number of critical points increases heavily with the atomic number of the atoms involved but the overall complexity, once the core CP's are discounted, does not follow this trend but rather it appears to depend on the nature of the bonding and, in general, it tends to diminish in going from the light to the heavy elements. This effect is evident in the topologies presented in Tables 9 and 10. The $L(\boldsymbol{r})$ are grouped into core subshells when their distance to the closest atomic nucleus is quite close to the minima and maxima of the radial $L(\mathbf{r})$ function. The core character of those CPs can be confirmed by the fact that the attraction basins of all core NCP's form a small sphere, as it will be discussed later.

The $L(\boldsymbol{r})$ critical points (L-CPs) can be classified into core, valence and interstitial. A core subshell is formed by the breaking of one minimum or maximum of the atomic radial $L(r)$ function, all CP's in the subshell keep an almost identical

Table 9. Topology of the $L(\mathbf{r})$ Field for the Diamond Structures[a]

		core: *t, n*		total: *t, n*		type/Wyckoff			$\mathbf{r}_{cp}$	$\rho(\mathbf{r}_{cp})$	$\nabla^2\rho(\mathbf{r}_{cp})$	*x* coord.	*z* coord.
C	n	1	8	3	48	n	I	*8b*	(3/8, 3/8, 3/8)	0.013 24	+0.075 95		
	b	1	32	5	208	n	val	*32e*	(*x*, *x*, *x*)	0.277 42	−0.895 63	0.042 41	
	r	1	48	4	224	b	bond	*16c*	(0, 0, 0)	0.240 05	−0.551 44		
	c	1	32	2	64	b	I	*96g*	(*x*, *x*, *z*)	0.067 70	+0.172 08	0.078 94	0.840 82
	last		K_			b	I	*16d*	(1/2, 1/2, 1/2)	0.021 51	+0.105 22		
						b	val	*48f*	(1/8, 1/8, *x*)	0.213 00	−0.197 58	0.269 23	
						r	val	*32e*	(*x*, *x*, *x*)	0.178 75	+0.030 93	0.210 77	
						r	I	*96h*	(*x*, \bbar*x*\ebar\, 0)	0.067 27	+0.173 36	0.137 86	
						r	I	*48f*	(1/8, 1/8, *x*)	0.080 81	+0.184 42	0.374 41	
						c	I	*32e*	(*x*, *x*, *x*)	0.078 47	+0.206 77	0.254 24	
Si	n	2	40	4	80	n	I	*8b*	(3/8, 3/8, 3/8)	0.003 24	+0.008 97		
	b	3	112	7	288	n	val	*32e*	(*x*, *x*, *x*)	0.086 56	−0.138 22	0.024 69	
	r	3	128	7	400	b	bond	*16c*	(0, 0, 0)	0.083 69	−0.123 44		
	c	2	64	4	192	b	I	*96g*	(*x*, *x*, *z*)	0.020 75	+0.027 13	0.189 91	0.822 96
	last		L_			b	I	*16d*	(1/2, 1/2 1/2)	0.055 89	+0.013 54		
						b	val	*48f*	(1/8, 1/8, *x*)	0.051 04	−0.009 31	0.301 72	
						r	val	*32e*	(*x*, *x*, *x*)	0.036 73	+0.017 55	0.231 47	
						r	I	*96h*	(*x*, \bbar*x*\ebar\, 0)	0.021 51	+0.027 49	0.364 84	
						r	I	*48f*	(1/8, 1/8, *x*)	0.022 02	+0.027 93	0.386 10	
						r	I	*96g*	(*x*, *x*, *z*)	0.021 23	+0.027 99	0.058 03	0.360 72
						c	I	*96g*	(*x*, *x*, *z*)	0.021 46	+0.027 99	0.322 40	0.628 91
						c	I	*32e*	(*x*, *x*, *x*)	0.021 52	+0.028 79	0.257 43	
Ge	n	3	72	5	96	n	I	*8b*	(3/8, 3/8, 3/8)	0.002 57	+0.006 52		
	b	5	192	7	304	n	bond	*16c*	(0, 0, 0)	0.075 02	−0.035 56		
	r	5	208	6	304	b	I	*96g*	(*x*, *x*, *z*)	0.025 51	+0.029 23	0.065 84	0.868 20
	c	3	96	3	96	b	I	*16d*	(1/2, 1/2, 1/2)	0.004 36	+0.010 21		
	last		M_			r	I	*96h*	(*x*, \bbar*x*\ebar\, 0)	0.025 55	+0.029 27	0.286 01	
Sn	n	4	104	6	128	n	I	*8b*	(3/8, 3/8, 3/8)	0.001 48	+0.003 15		
	b	8	304	10	416	n	bond	*16c*	(0, 0, 0)	0.054 27	+0.002 70		
	r	7	336	8	432	b	I	*96g*	(*x*, *x*, *z*)	0.022 12	+0.018 93	0.057 79	0.884 87
	c	4	144	4	144	b	I	*16d*	(1/2, 1/2, 1/2)	0.025 98	+0.005 01		
	last		N_			r	I	*96h*	(*x*, \bbar*x*\ebar\, 0)	0.022 10	+0.018 99	0.400 13	
Pb	n	5	136	6	144	n	I	*8b*	(3/8, 3/8, 3/8)	0.000 99	+0.001 62		
	b	10	384	11	400	b	I	*16d*	(1/2, 1/2, 1/2)	0.001 66	+0.002 58		
	r	8	416	9	432	r	bond	*16c*	(0, 0, 0)	0.038 85	+0.030 82		
	c	5	176	5	176								
	last		O_										

[a] The left part of the table resumes the type (*t*) and number (*n*) of CPs included in the core and those in the whole crystal unit cell. *Last* is the identity of the last core subshell. The right part of the table is a detailed description of the valence CPs, classified into valence, bond (a kind of valence CP placed in the line between two atoms bonded by the $\rho(\mathbf{r})$ field), and interstitial (I, not recognizable as belonging to any single atom or pair of atoms). In the case of Pb, the algorithm fails to detect a RCP of multiplicity 96 in the L_ core shell, but the complete topology can be recovered from the invariance laws that rule shell and cell CPs.

distance to the originating nucleus, and the number of CP's fulfill the Euler relationship (eq 11). In addition, the attraction basins of the set of core NCP's form a small sphere around the nucleus, as it will be discussed later.

Valence L-CPs are originated from the outermost or perhaps the two outermost extrema of the atomic $L(\mathbf{r})$ radial function. The distance to the nucleus, however, is not so closely maintained as in the core case, and some CPs can be displaced toward the interatomic space, or even merged with the CP's from other nuclei, so the Euler relationship is not necessarily fulfilled by a valence subshell.

Interstitial L-CPs, finally, cannot be assigned to a single atom, but they lie well into the interatomic space. Interstitial NCP's, typically have a negative *L* value, at difference from core and valence NCP's. In other words, interstitial NCP's do not show an increased concentration of electron density relative to their differential neighborhood. This counterintuitive property turns to be one of the most prominent features of interstitial regions.

6.3.1. Graphs for the L(r) Topology. The complexity of the $L(\mathbf{r})$ topology is difficult to examine without an appropriate map. Aray et al.[78−82] and Popelier et al.[6,11,31,33] have taken great advantage of special chemical graphs for that purpose. Drawing a significant *L* graph is not trivial nor automatic, but it involves some creative decisions about what information is relevant and what should be left out to avoid cluttering.

Figures 3 and 4 shows our interpretation of the relevant $L(\mathbf{r})$ topology for the diamond and non-diamond structures. Some of the most relevant features correspond to the organization of the valence L-BCPs and the bond paths that connect them to the NCPs. Many of those *L*-bond paths are quite curved lines. In particular, bond paths that connect NCPs in the same subshell and thus are equidistant to the generating nucleus are almost circular arcs. This sharply contrast with the ρ-bond paths that are typically straight lines and do only curve away from the internuclear axis in such cases as the occurrence of steric stress or electron deficient bonding.[1]

Table 10. Topology of the $L(\boldsymbol{r})$ Field for the Non-diamond Allotropes[a]

		core: *t, n*		total: *t, n*		type/Wyckoff			$\boldsymbol{r}_{cp}$	$\rho(\boldsymbol{r}_{cp})$	$\nabla^2\rho(\boldsymbol{r}_{cp})$	*x* coord.	*z* coord.
graphite	n	2	4	5	20	n	I	*4f*	(1/3, 2/3, *z*)	0.003 48	+0.012 72	0.482 23	
	b	2	12	10	70	n	val	*6h*	(*x*, 2*x*, 1/4)	0.323 37	−1.132 02	0.881 25	
	r	2	12	10	70	n	val	*6h*	(*x*, 2*x*, 1/4)	0.322 96	−1.126 87	0.785 52	
	c	2	8	4	20	b	I	*6h*	(*x*, 2*x*, 1/4)	0.301 99	−0.887 98	0.833 29	
	last		K_			b	I	*4f*	(1/3, 2/3, *z*)	0.068 83	+0.113 31	0.628 53	
						b	I	*12k*	(*x*, 2*x*, *z*)	0.071 78	+0.117 92	0.175 76	0.860 34
						b	val	*12k*	(*x*, 2*x*, *z*)	0.179 63	−0.096 52	0.393 24	0.680 88
						b	I	*2c*	(1/3, 2/3, 1/4)	0.022 71	+0.134 76		
						b	val	*12k*	(*x*, 2*x*, *z*)	0.180 45	−0.091 11	0.064 66	0.182 65
						b	I	*4e*	(0, 0, *z*)	0.068 60	+0.114 73	0.128 69	
						b	I	*6g*	(1/2, 0, 0)	0.004 10	+0.014 00		
						r	val	*6h*	(*x*, 2*x*, 1/4)	0.210 27	−0.015 96	0.544 96	
						r	I	*12j*	(*x*, *y*, 1/4)	0.079 57	+0.235 37	0.560 98	0.667 06
						r	val	*6h*	(*x*, 2*x*, 1/4)	0.210 88	−0.020 57	0.121 58	
						r	val	*4f*	(1/3, 2/3, *z*)	0.166 11	−0.084 69	0.670 25	
						r	val	*4e*	(0, 0, *z*)	0.164 37	−0.076 57	0.170 07	
						r	I	*12k*	(*x*, 2*x*, *z*)	0.074 52	+0.118 05	0.206 44	0.640 14
						r	I	*12k*	(*x*, 2*x*, *z*)	0.074 67	+0.118 47	0.118 31	0.860 22
						r	I	*2a*	(0, 0, 0)	0.005 88	+0.017 67		
						c	I	*6h*	(*x*, 2*x*, 1/4)	0.092 22	+0.252 29	0.475 98	
						c	I	*6h*	(*x*, 2*x*, 1/4)	0.092 10	+0.251 88	0.190 86	
β-Sn	n	4	52	7	80	n	I	*16g*	(*x*, *x*, 0)	0.008 03	+0.013 91	0.285 99	
	b	11	144	15	208	n	bond	*8c*	(0, 1/4, 1/8)	0.037 51	+0.012 23		
	r	10	144	13	184	n	bond	*4b*	(0, 0, 1/2)	0.028 38	+0.017 23		
	c	4	56	4	56	b	I	*16h*	(0, *y*, *z*)	0.014 33	+0.017 03	0.199 36	0.518 18
	last		N_			b	I	*16h*	(0, *y*, *z*)	0.021 95	+0.018 52	0.098 24	0.497 51
						b	I	*16f*	(*x*, 1/4, 1/8)	0.008 05	+0.013 97	0.296 43	
						b	I	*16h*	(0, *y*, *z*)	0.022 07	+0.018 92	0.212 11	0.362 60
						r	I	*8d*	(0, 1/4, 5/8)	0.013 03	+0.017 34		
						r	I	*16f*	(*x*, 1/4, 1/8)	0.019 99	+0.020 12	0.138 57	
						r	I	*16g*	(*x*, *x*, 0)	0.020 38	+0.018 88	0.422 20	
α-Pb	n	5	132	7	144	n	I	*4b*	(1/2, 1/2, 1/2)	0.006 74	+0.009 04		
	b	9	336	10	368	n	I	*8c*	(1/4, 1/4, 1/4)	0.012 08	+0.013 61		
	r	9	336	10	360	b	I	*32f*	(*x*, *x*, *x*)	0.012 72	+0.015 29	0.336 65	
	c	5	136	5	136	r	I	*24d*	(0, 1/4, 1/4)	0.017 88	+0.022 22	0.833 29	
	last		O_										

[a] See description in Table 9.

Another striking difference between the *L* and ρ graphs is that while the five elements, C to Pb, shows identical ρ graph in the diamond phase, their *L* graph can be grouped into three quite different models. C and Si form the first model, that closely resembles a prototypical Lewis image for covalent bonding: each atom is surrounded by a curved tetrahedron frame with NCP's at the vertices. Each NCP is then connected through a *L*-bond path to the NCP of a nearest neighbor (NN) atom. As a consequence, the middle point between two NN atoms (the 16c Wyckoff position in Table 7 and Table 9) is simultaneously a ρ-BCP and a L-BCP. Furthermore, this double BCP occurs in a region of significant local charge accumulation, thus completing the characterization of the C and Si *L* graph as a prototype of covalently bonded system.

This Wyckoff *16c* position has the key to characterize the three types of *L* graphs in the diamond structure. The second type is shown by Ge and Sn, where the *16c* point continues to be a ρ-BCP but it is now a L-NCP, connected through *L* bond paths to six other equivalent *16c* positions. At the same time, the outermost valence shell of each atom is now made of a tetrahedron of CCPs rather than NCPs.

Pb diamond phase shows the third kind of *L* graph. The Pb core is quite large, and the only noncore NCPs have moved into the interstitial region. The *16c* position is now a ring CP and corresponds to a corner in which four interstitial NCP basins intersect. The valence electron density has been maximally delocalized and transferred to the interstitial region.

The nondiamond phases behave like their diamond equivalents. C shows again in the graphite phase a covalent pattern, with two bonded NCPs along the internuclear axis, the ρ-BCP and the L-BCP occurring at the same position, within a region of increased electron density concentration. Of course, rather than a tetrahedral pattern, NCPs form now a flat triangle surrounding each C nucleus. In Mulliken terms, the sp^3 arrangement has been converted into a sp^2 one. White Sn and fcc Pb also show the same pattern as their corresponding diamond phases.

As described in section 4, the shape of the valence electron density is affected by the relativistic treatment of the crystal. To measure the sensitivity of *L* graphs to this effect, we have recalculated Sn and Pb in both diamond and experimental phases without the scalar relativistic correction. The *L* topology of Sn is unaffected in the diamond phase, while the changes in the β-Sn phase are minor: a L-BCP bonding two interstitial maxima and a L-RCP are displaced to a lower symmetry position, without any significant consequence on the preceding discussion. The effect is more pronounced on both phases of Pb. In the fcc phase, a number of new CP

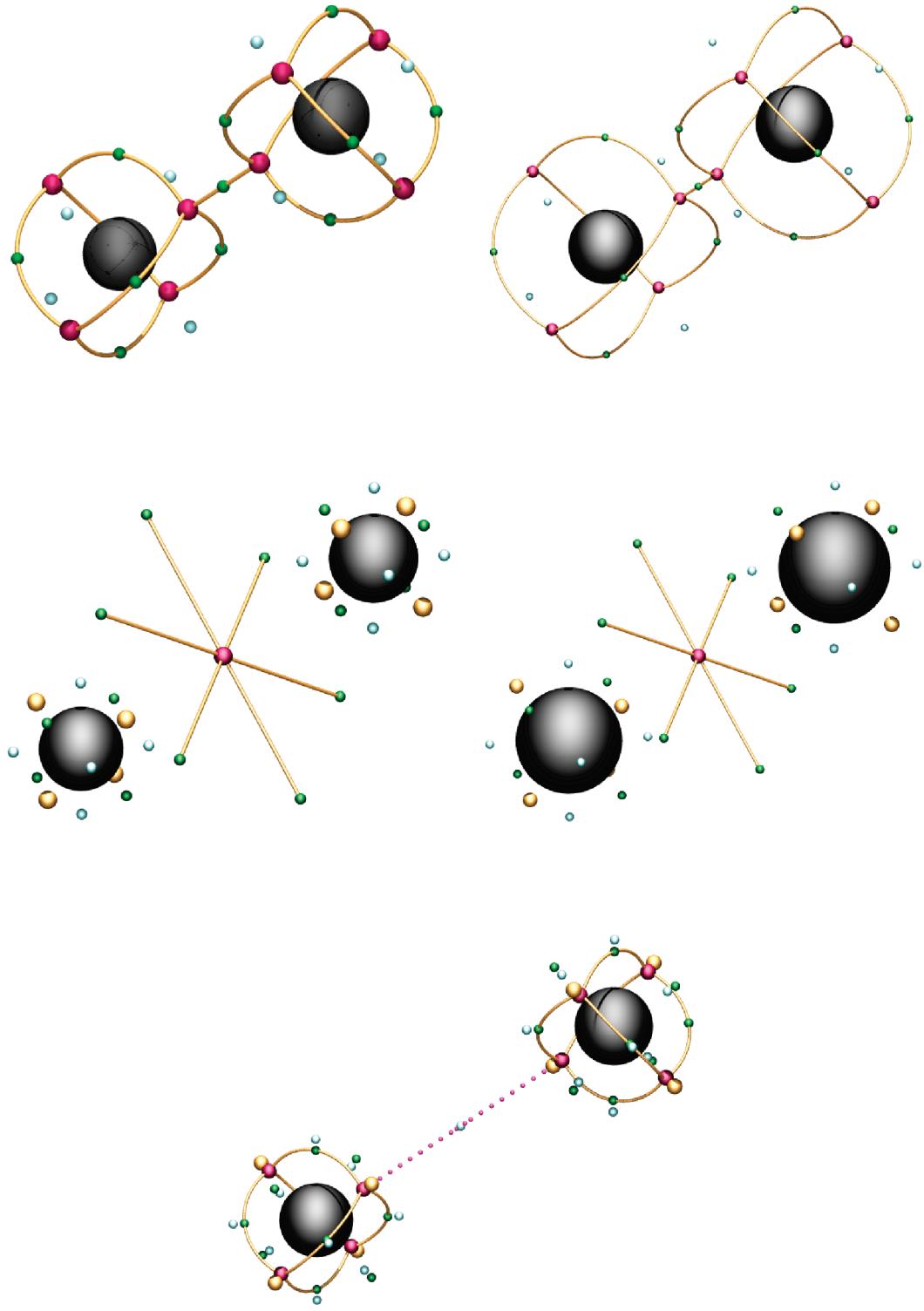


Figure 3. *L* graphs for the diamond structures (top, C and Si; middle, Ge and Sn; bottom, Pb). The very large black spheres contain the nuclei and most of the core structure. Far smaller, L-CPs can be discriminated by the size and color of the spheres that represent them: NCPs are large and dark (red), BCPs smaller and dark (green), RCPs small and light (light blue), and CCPs large and light (yellow). *L*-bond paths are sometimes represented by thin golden lines. The Pb-diamond graph shows the *L* ring path, created by the RCP lying in the middle of two nearest neighbors Pb atoms, as a discontinuous line of very thin spheres (pink).

appear, the most important of them being a new interstitial maximum at the (0, 1/4, 1/4) (24d) position. In the diamond phase, the *16c* maximum reappears, resulting in Pb having a graph equivalent to Ge and Sn. The greater effects of the relativistic corrrection in heavier elements couples in this case with the well-known lability of the topology of ρ (and hence L) in metals:[83] the flatness of the interstitial part induces that even small density changes rearrange the valence topology completely.

Far less important is the effect of changing the exchange-correlation functional. The topologies of all the crystals examined are unaffected when calculated at the LDA level.

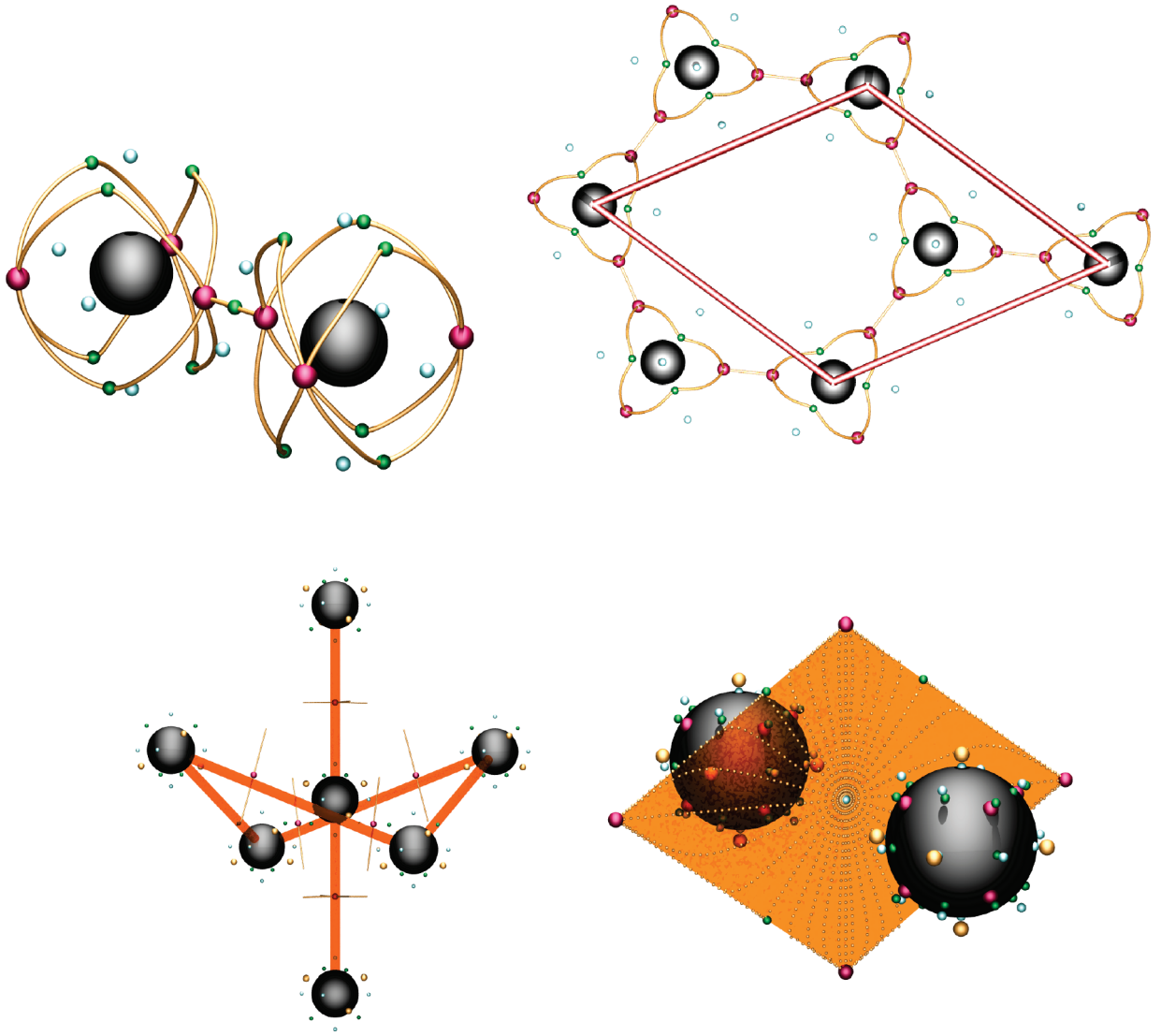


Figure 4. *L* graphs for the non-diamond structures (upper row, graphite; lower row, white Sn and fcc Pb). The zenithal view of graphite shows the crystal unit cell using thick cylinders. The Pb-fcc graph shows the equivalent RCP surface and the graph lines that form this surface.

6.3.2. Local Properties of the L-NCP Basins. The number, arrangement, and properties of NCPs is the central issue of the $L(\boldsymbol{r})$ topology. The *L* graphs analyzed previously fail to communicate the shape, size, and relative importance of the NCP basins. All L-NCPs in the diamond phase are placed along the cube diagonal, that is, along the (x, x, x) crystal direction. We have taken advantage of this coincidence to produce the illustration in Figure 5. The plots show clearly the difference between the three types of *L* graphs described in the previous subsection (6.3.1). The twin L-NCPs between two NN atoms observed on C and Si, get converted into a single NCP at the NN midpoint on Ge and Sn, and finally the internuclear axis is simply the common edge of four interstitial NCPs on Pb.

Figure 5 is also a qualitative demonstration of the growing importance of the interstitial NCP as the atomic number increases. The core region increases too and, more interesting, the core subshells form concentric spheres that surround the atomic nucleus. That a given shell keeps the spherical shape typical of a free atom can be seen as the ultimate evidence that it belongs in the core. Contrarily, a significant deformation from sphericity is a direct proof of the participation of the corresponding electrons in the valence chemical bonding.

Going beyond the qualitative requires integrating properties within the *L* basins. In this regard it must be clear that there is a fundamental difference between the space partition induced by $\rho(\boldsymbol{r})$ and the partition due to $L(\boldsymbol{r})$. All kind of quantum mechanical observables can be integrated within the basins determined by the zero flux of $\rho(\boldsymbol{r})$ condition (eq 3). This is not true for the basins determined by the topology of $L(\boldsymbol{r})$ and, for instance, it is not correct to determine the contribution of a L-NCP basin to the kinetic energy. There is no problem, however, with the integration of strictly local properties like the volume, charge, electrostatic field, or multipolar moments.

Table 11 presents the volume and electronic populations of the L-NCP basins. We have classified the basins into five different groups according to its properties: nucleus, core, valence, bond, and interstitial. The nucleus can include some L-NCP so internal that we have preferred not to distinguish them from the atomic position. The "valence" L-NCPs correspond to the twin maxima situated along a NN inter-

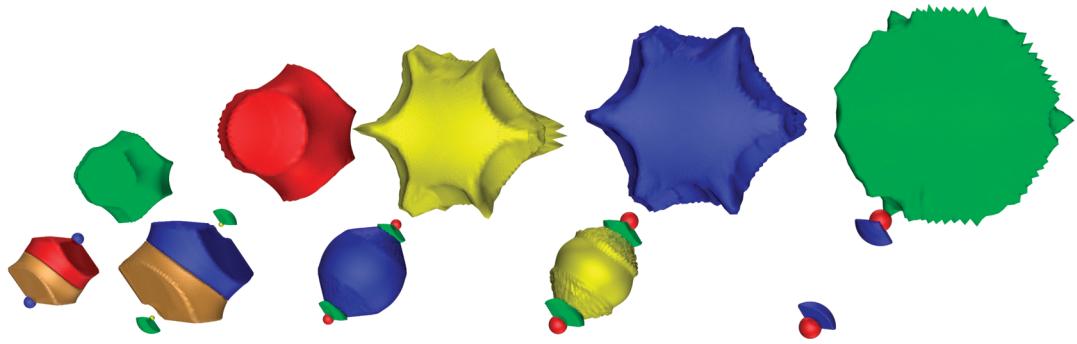


Figure 5. From left to right: L-NCP basins of C, Si, Ge, Sn, and Pb in the diamond phase. All plots are made in the same scale, and use the same viewpoint, so the apparent relative size corresponds to the actual size of the basins. The two small spheres appearing in all the plots correspond to the atomic core regions; between them we can find the single or double "bond" L-NCP, absent in the case of Pb; the uppermost basin is, in all cases, the interstitial L-NCP.

Table 11. Volume (V_Ω, bohr^3) and Electronic Population (Q_Ω, e) of the L-NCP Basins[a]

diamond phase											
type	Wyckoff	C: $Q\Omega$	$V\Omega$	Si: $Q\Omega$	$V\Omega$	Ge: $Q\Omega$	$V\Omega$	Sn: $Q\Omega$	$V\Omega$	Pb: $Q\Omega$	VΩ
nuc.	*8a*	0.988	0.051	1.049	0.004	8.060	0.043	24.078	0.245	51.264	0.508
core	*32e*			1.561	0.224	3.428	0.423	4.017	1.319	4.658	1.343
val.	*32e*	1.078	5.144	1.511	18.628						
bond	*16c*					3.163	22.210	2.331	22.789		
I	8*b*	0.707	17.630	0.658	59.722	3.931	106.627	5.260	179.806	11.959	290.608
total		48.053	306.050	111.946	1081.074	256.226	1222.259	400.534	1847.233	654.836	2371.890
cell		48.000	306.217	112.000	1080.847	256.000	1221.900	400.000	1845.744	656.000	2373.614
error		0.053	−0.167	−0.054	0.227	0.226	0.359	0.534	1.488	−1.164	−1.724
%		0.111	−0.055	−0.048	0.021	0.088	0.029	0.133	0.081	−0.177	−0.073

non-diamond phase											
graphite		$Q\Omega$	$V\Omega$	white Sn		$Q\Omega$	$V\Omega$	fcc Pb		$Q\Omega$	$V\Omega$
C	2*b*	0.988	0.051	Sn	*4a*	40.113	5.513	Pb	*4a*	51.281	0.509
C	2*c*	0.988	0.051	bond	*4b*	0.966	11.184	core	*32f*	2.342	0.688
val.	6*h*	1.348	5.891	bond	*8c*	1.816	20.004		*8c*	2.401	36.549
val.	6*h*	1.348	5.891	I	*16g*	1.344	31.188	I	*4b*	7.174	125.556
I	4*f*	0.960	41.259								
total		23.969	235.931	total		200.350	725.832	total		327.969	818.660
cell		24.000	236.048	cell		200.000	725.647	cell		328.000	818.736
error		−0.031	0.116	error		0.350	0.186	error		−0.031	−0.076
%		−0.130	0.049	%		0.175	0.026	%		−0.009	−0.009

[a] Notice that the calculation of the total properties has been made using partial data with more digits than those shown in the table. The properties reported correspond to single basins of each type and must be multiplied by the Wyckoff multiplicity to determine the contribution to the cell property.

nuclear axis. They are characterized by a positive value of $L(\boldsymbol{r})$ and belong, accordingly, to the region called VSCC (Valence Shell Charge Concentration) by Bader.[1] The single "bond" L-NCP, situated midway between NN atoms, can have a positive (Ge) or negative (Sn) $L(\boldsymbol{r})$ value. The last type of L-NCPs, finally, are disconnected from the L graphs of nuclei, belong to the interstitial region and have, in all cases, a negative $L(\boldsymbol{r})$ value.

This classification of L-NCP basins is quite relevant for the analysis of the local properties. The nuclear and core L -NCP's occupy a very small volume but contain a significant number of the electrons. Valence (C and Si) and bond (Ge and Sn) NCP's occupy a part of the cell and contain some 2−3 electrons for each NN pair of atoms. The interstitial NCP's, finally, represent most of the cell volume and contain an electron population that grows from 0.7 e in C and Si to a shocking 12 e per NCP in Pb.

The important electron population of interstitial regions is not particular to crystals, but it was already observed by Malcolm and Popelier[34] in molecules like NH_3 and H_2O. This fact, which Malcolm and Popelier elude to interpret, is one of the problematic features if we try to explain the L populations in terms of a simple Lewis model.

In a classical Lewis description, each C in the diamond structure uses four electrons to form the same number of covalent bonds to its NN, remaining two nonbonding electrons on each C core. Our topological analysis of $L(\boldsymbol{r})$ shows a small excess of 0.08 e involved per C on each covalent bond, but a donation of 0.71 e per C toward the interstitial space.

The comparison with the Lewis model is even worse in Si. Now each Si atom uses 1.5 e to form each of its four covalent Si−Si bonds. The donation to the interstitial space is quite similar to the diamond case, however, 0.66 e per Si

atom. Ge and Sn also accumulate an extra number of electrons on the NN internuclear space: 3.16 *e* (Ge) and 2.33 *e* (Sn) rather than the two electrons expected for a classical Lewis single covalent bond. The donation to the interstitial space is significantly increased: 3.93 *e* in Ge and 5.26 *e* in Sn. Finally, Pb lacks the features that could be described as covalent bonds. Contrarily, all the electrons removed from the nuclear and core regions now belong to the interstitial zone: a record 11.96 *e* per Pb atom.

It is tempting to explain this behavior of group IV elements as successive steps in the conversion from covalent to metallic bonding. The small population of the interstitial zones, similar to the values previously reported for some covalent molecules,[34,34] could be regarded as the minimal background. Some more cases, and more diverse crystals and molecules, should be analyzed before this conjecture can be accepted, however.

The changes induced by the scalar relativistic treatment of valence electrons are apparent by comparing the Sn and Pb results to its nonrelativistic counterparts. It was mentioned in section 4 that the *L* -graphs are not compatible, so the comparison is not direct. However, several observations can be made about these differences: (1) The shell radii are contracted; the nucleus plus the inner shells loses approximately 0.1 electrons in Sn, and almost 1 electron in Pb. (2) The bond basins shrink and lose electrons, the charge smearing out to the interstitial basins.

The effect of changing the exchange-correlation potential is, again, not as significant. The cores are almost unaffected, with a difference in core population that peaks at 0.03 electrons in Pb, with an analogous behavior of the outer core basins. Regarding the valence and interstitial basins, LDA assigns less charge to the valence and bond basins, with slightly larger and more populated interstitial basins than GGA. The differences between both functionals increase on advancing to the heavier elements of the group. Exchange-correlation and relativistic effects are certainly interesting and, hopefully, will be addressed in a future work, once the utility of the present methodology is established.

The comparison between the two different phases of C, Sn and Pb opens an important window for analyzing the transferability of L-NCP basin properties among distinct structures and compounds. Diamond and graphite are very dissimilar in their bonding pattern and cell volume per atom (38.3 vs 59.0 bohr^3, respectively), but the nuclear L-NCP is almost identical in both crystals ($Q = 0.9875$ vs 0.9882 *e*, $V = 0.0512$ vs 0.0512 bohr^3), the single bond L-NCP has a similar number of electrons per C atom (4.31 vs 4.04 *e*) even though this region occupies a larger volume in the more dense diamond phase than in the less dense graphite (20.6 vs 17.7 bohr^3) and, finally, the larger differences occur between the corresponding interstitial L -NCP ($Q = 0.71$ vs 0.96 *e*, $V = 17.6$ vs 41.3 bohr^3).

This scheme is repeated on Sn and Pb. The transferability of properties between different structures is almost exact for the nuclear and core L-NCP basins. Bond and valence L-NCP show significant regularities, although we need a larger set of compounds to extract the organizing principles. The interstitial L-NCP basins, the most unexpected topological feature of the Laplacian, is also the most variable element, and much study is required before its role can be clarified.

6.3.3. Local Compressibilities of the L-NCP Basins. The tetrahedral arrangement of covalently bonded C atoms has been usually called to explain the extreme hardness of diamond. The same arrangement, however, does not explain the large hardness differences between the isostructural group IV elements. We can gain some insight into the phenomenon by determining the contribution of the several L-NCP basins to the elastic properties. We will follow the method proposed by Martn Pendás et al. on the analysis of $\rho(\boldsymbol{r})$[84,85] and recently applied by Recio et al. to the ELF function.[86]

The static compressibility (κ) and bulk modulus (B) of a crystal are defined as

$$\kappa = \frac{1}{B} = -\frac{1}{V}\left(\frac{\partial V}{\partial p}\right) \tag{20}$$

Using in these definitions the partition of the cell volume into L-NCP basin contributions, $V = \Sigma_\Omega V_\Omega$, we can write[84]

$$\kappa = \sum_\Omega f_\Omega \kappa_\Omega \text{ and } \frac{1}{B} = \sum_\Omega f_\Omega \frac{1}{B_\Omega} \tag{21}$$

where $f_\Omega = V_\Omega/V$ is the fraction of the cell volume occupied by the Ω basin, and

$$\kappa_\Omega = \frac{1}{B_\Omega} = -\frac{1}{V_\Omega}\left(\frac{\partial V_\Omega}{\partial p}\right) \tag{22}$$

The local compressibility of a basin is thus defined in the same way that the compressibility of the whole cell, and the global value of the crystal is the result of averaging the local compressibilities in such a way that the contribution of a basin is proportional to the volume fraction of the basin in the crystal cell.

To determine the local compressibilities of the group IV crystals we have followed the static model, in which the vibrational entropy is neglected by assuming a temperature of zero Kelvin, and the zero point vibrational energy is also neglected. Under these conditions the pressure is given by

$$p = -\left(\frac{\partial A}{\partial V}\right)_T \approx -\left(\frac{\partial E}{\partial V}\right) \tag{23}$$

where V and E are the cell volume and energy, result from the quantum mechanical calculation, and $A = E + A_{\text{vib}}(T, V) \approx E$ would have been the Helmholz free energy.

The actual sequence of calculations goes as follows. First, some 11−15 points of the $E(V)$ curve are determined, with the volume bracketing a range of ±10% around the experimental geometry. Once verified that this range effectively contains the equilibrium volume, the pressure is obtained from eq 23, using a polynomial or a Birch−Murnaghan function to fit the calculated $E(V)$ points. Simultaneously, a topological analysis is performed on each wave function and the volumes of the L-NCP basins are determined. The $V(p)$ and $V_\Omega(p)$ data are used to evaluate the crystal and the local compressibilities, with a polynomial fitting to the data being again instrumental in obtaining the derivatives. The results from this analysis are presented in Table 12.

Table 12. Volume Fraction (f_Ω, %) and Local Compressibility (κ_Ω, TPa^{-1}) of the L-NCP Basins[a]

diamond phase:											
type	Wyckoff	C: $f\Omega$	$\kappa\Omega$	Si: $f\Omega$	$\kappa\Omega$	Ge: $f\Omega$	$\kappa\Omega$	Sn: $f\Omega$	$\kappa\Omega$	Pb: $f\Omega$	$\kappa\Omega$
nuc.	*8a*	0.133	0.008	0.003	0.004	0.026	0.021	0.098	0.004	0.172	0.019
core	*32e*			0.648	0.018	1.043	0.005	2.116	0.083	1.811	0.210
val.	*32e*	54.126	1.255	55.447	5.123						
bond	*16c*					28.356	8.074	18.739	15.157		
I	*8b*	45.834	3.876	44.180	20.563	70.812	22.665	79.211	31.504	98.356	42.614
κ_t			2.456		11.926		18.339		27.797		41.917
κ_0			2.330		11.977		19.275		29.406		39.531
B_t			407.210		83.853		54.528		35.976		23.857
B_0			429.240		83.492		51.882		34.006		25.297

non-diamond phase									
type	Wyckoff	C: $f\Omega$	$\kappa\Omega$	Wyckoff	Sn: $f\Omega$	$\kappa\Omega$	Wyckoff	Pb: $f\Omega$	$\kappa\Omega$
nuc.	*2b*	0.043	0.013	*4a*	0.127	0.002	*4a*	0.234	0.012
nuc.	*2c*	0.043	0.013						
core				*4b*	5.761	0.853	*32f*	2.535	0.149
core				*16h*	2.741	0.183			
val.	*6h*	14.861	1.926	8*c*	20.806	9.149			
val.	*6h*	14.874	1.936						
I	*4f*	70.146	4.461	*16g*	70.384	28.597	*8c*	36.127	28.784
I							*4b*	61.246	24.185
κ_t			3.704			22.085			25.215
κ_0			3.662			20.660			28.226
B_t			270.000			45.280			39.659
B_0			273.070			48.403			35.429

[a] The bulk moduli are given in GPa. The topological κ_t and B_t values are obtained using eq 21. The values of κ_0 and B_0 are obtained from the total cell energy and volume according to eq 20. The difference between B_t and B_0 is consequence of small errors in the determination of the basin volumes.

The results in Table 12 provide an excellent confirmation of the classification of the L-NCPs. Nuclear and core NCPs fill a small fraction of the cell volume and, more significantly, have a very small compressibility, less than 1% of the value of κ for the whole crystal. In sharp contrast, the values of κ_Ω for the valence, bond, and interstitial L-NCPs are of the same order of magnitude than κ, smaller in the case of valence and bond regions and larger in the case of interstitial ones.

Restricting our analysis to the diamond phase, the extreme hardness of C is the consequence of the small compresibility of all the L-NCP regions. C and Si show a similar contribution balance, $f_{val} = 54-55\%$ and $f_I = 46-44\%$, but the κ_Ω values are several times larger in Si than in C. Ge and Sn show both the progressive increase in the coefficient of the interstitial zone and the increase of the κ_Ω values for all the zones. Finally, the soft metal character of Pb is almost exclusively because of the large and quite compressible interstitial regions.

Graphite shows clearly the importance of the crystalline structure. Compared with the diamond phase, graphite presents an interstitial zone larger and more compressible. The atomic number, however has a similar or larger influence, so the search for truly hard compounds can be restricted to the lightest elements. A topological partition of the shear modulus could be of interest as it would show the graphene sheets hard compared to the larger compressibility perpendicular to the sheets. We are working toward achieving a practical way of partitioning the elastic constants of arbitrary crystals.

β-Sn and α-Pb, being significatively more dense than their diamond allotropes, show also a larger value for the bulk modulus, discarding the exclusive influence of the tetrahedral coordination on the hardness of group IV elements. To be fair in this conclusion, we should remember that β-Sn and α-Pb show a metallic rather than covalent behavior and the hardness of single crystals is mostly controlled by the shear and not by the bulk modulus.

Diamond and nondiamond phases agree, anyway, on the fundamental importance of the interstitial regions in determining the bulk modulus and compressibility of crystals.

7. Conclusions and Perspectives for Future Work

The topology of $L(\boldsymbol{r})$ is far more complex than the topology of $\rho(\boldsymbol{r})$. First, $L(\boldsymbol{r})$ retains the shell structure inherited from the isolated atoms. Second, the range of $L(\boldsymbol{r})$ goes from $-\infty$ to $+\infty$, giving rise to maxima, for instance, with $L < 0$ and others with $L > 0$, both having a different chemical interpretation. Third, $L(\boldsymbol{r})$ has more critical points than $\rho(\boldsymbol{r})$, and their number increases heavily with the atomic number of the element. Fourth, L basins tend to be more irregular, and the source or sink point of the basin can be separated from the geometrical center, thus leading to a more difficult integration of the basin properties.

Taking advantage of the shell structure is important in designing and adapting efficient algorithms, like the radial navigation method presented in section 5. It is also particularly important for simplifying the analysis and presentation of the $L(\boldsymbol{r})$ topology by removing unimportant core features in a controlled way. Basin plots (Figure 3) and L graphs (Figure 5) have been found to be fundamental instruments

to understand the organization and qualitative importance of the topological features of $L(\boldsymbol{r})$.

The electron density Laplacian provides a perspective that complements and is not directly available from the electron density. This is clearly observed in the group IV diamond phases. Whereas all, C to Pb, elements show identical topology for $\rho(\boldsymbol{r})$, the analysis of $L(\boldsymbol{r})$ evidence deep differences between the three groups formed by C and Si (the covalent group), Ge and Sn (the semilocal group), and Pb (the most delocalized one). This difference between the elements is transferred to other phases, showing the dominant influence of the nature of the element on $L(\boldsymbol{r})$, rather than the effect of the crystal geometry, more akin to influence the $\rho(\boldsymbol{r})$ topology.

Our topological methodology is mature enough for application to general crystals. Further work should now be directed to examine the $\rho(\boldsymbol{r})$ and $L(\boldsymbol{r})$ topologies in a diverse set of crystal types with an objective pointed toward solving some of the puzzles observed on group IV crystals. In particular, the role of interstitial regions and the contrast between the Lewis model and the real electron population on each basin.

Acknowledgment. We thank the Spanish Ministerio de Ciencia e Innovación (MICINN), grants CTQ2006-02976 and CTQ2009-08376, and the Malta/Consolider initiative CSD2007-00045. A.O.R. is indebted to the Spanish Ministerio de Educación y Ciencia (MEC) for a FPU grant.

References

(1) Bader, R. F. W. *Atoms in Molecules. A Quantum Theory*; Oxford University Press: Oxford, U.K., 1990.

(2) Popelier, P. L. A. *Atoms in Molecules: An Introduction*; Prentice Hall: London, 2000; p 188.

(3) Boyd, R.; Matta, C. *The Quantum Theory of Atoms in Molecules: From Solid State to DNA and Drug Design*; Wiley-VCH: Weinheim, Germany, 2007.

(4) Bader, R. F. W. *Chem. Rev.* **1991**, *91*, 893–928.

(5) Bader, R. F. W. *Theor. Chem. Acc.* **2001**, *105*, 276–283.

(6) Popelier, P. L. A. *Coord. Chem. Rev.* **2000**, *197*, 169–189.

(7) Cortés-Guzmán, F.; Bader, R. F. W. *Coord. Chem. Rev.* **2005**, *249*, 633–662.

(8) Merino, G.; Vela, A.; Heine, T. *Chem. Rev.* **2005**, *105*, 3812–3841.

(9) Coppens, P. *X-Ray Charge Densities and Chemical Bonding*; IUCr and Oxford U. P.: New York, 1997.

(10) Popelier, P. L. A. *Struct. Bonding (Berlin)* **2005**, *115*, 1–56.

(11) Malcolm, N. O. J.; Popelier, P. L. A. *Faraday Discuss.* **2003**, *124*, 353–363.

(12) Popelier, P. L. A.; Aicken, F. M. *ChemPhysChem* **2003**, *4*, 824–829.

(13) Silvi, B.; Savin, A. *Nature* **1994**, *371*, 683–686.

(14) Gatti, C. Solid State Applications of QTAIM and the Source Function—Molecular Crystals, Surfaces, Host-Guest Systems and Molecular Complexes. In *The Quantum Theory of Atoms in Molecules: From Solid State to DNA and Drug Design*; Boyd, R., Matta, C., Eds.; Wiley-VCH: Weinheim, Germany, 2007; pp 165−206.

(15) Simas, A. M.; Smith, V. H.; Thakkar, A. J. *Int. J. Quantum Chem.* **1984**, S18.

(16) Bader, R. F. W.; Heard, G. L. *J. Chem. Phys.* **1999**, *111*, 8789–8798.

(17) Bader, R. F. W.; Essen, H. *J. Chem. Phys.* **1984**, *80*, 1943–1960.

(18) Carroll, M. T.; Chang, C.; Bader, R. F. W. *Mol. Phys.* **1988**, *63*, 387–405.

(19) Silvi, B. *Phys. Rev. Lett.* **1994**, *73*, 842–845.

(20) Popelier, P. L. A. *J. Phys. Chem. A* **1998**, *102*, 1873–1878.

(21) Espinosa, E.; Souhassou, M.; Lachekar, H.; Lecomte, C. *Acta Cryst. B* **1999**, *55*, 563–572.

(22) Bader, R. F. W.; MacDougall, P. J. *J. Am. Chem. Soc.* **1985**, *107*, 6788–6795.

(23) Shi, Z.; Boyd, R. J. *J. Phys. Chem.* **1991**, *95*, 4698–4701.

(24) Calvo-Losada, S.; Sánchez, J. J. Q. *J. Phys. Chem. A* **2008**, *112*, 8164–8178.

(25) Bader, R. F. W.; Popelier, P. L. A.; Chang, C. *J. Mol. Struct. (Theochem)* **1992**, *255*, 145–171.

(26) Carroll, M. T.; Cheeseman, J. R.; Osman, R.; Weinstein, H. *J. Phys. Chem.* **1989**, *93*, 5120–5123.

(27) Nekovee, M.; Foulkes, W. M.; Needs, R. J. *Phys. Rev. Lett.* **2001**, *87*, 036401.

(28) Nekovee, M.; Foulkes, W. M.; Needs, R. J. *Phys. Rev. B* **2003**, *68*, 235108.

(29) Hsing, C. R.; Chou, M. Y.; Lee, T. K. *Phys. Rev. A* **2006**, *74*, 032507.

(30) Cancio, A. C.; Chou, M. Y. *Phys. Rev. B* **2006**, *74*, 081202.

(31) Malcolm, N. O. J.; Popelier, P. L. A. *J. Phys. Chem. A* **2001**, *105*, 7638–7645.

(32) Malcolm, N. O. J.; Popelier, P. L. A. *J. Comput. Chem.* **2003**, *24*, 437–442.

(33) Popelier, P. L. A.; Burke, J.; Malcolm, N. O. J. *Int. J. Quantum Chem.* **2003**, *92*, 326–336.

(34) Malcolm, N. O. J.; Popelier, P. L. A. *J. Comput. Chem.* **2003**, *24*, 1276–1282.

(35) Keith, T. A.; Laidig, K. E.; Krug.; P.; Cheeseman.; J. R.; Bone.; R. G. A.; Biegler-König.; F. W.; Duke.; J. A.; Tang.; T.; Bader.; R. F. W. The AIMPAC95 programs, 1995, Available from: able from: http://www.chemistry.mcmaster.ca/aimpac/ (accessed Oct 14, 2010).

(36) Popelier, P. L. A. *Comput. Phys. Commun.* **1996**, *93*, 212–240.

(37) Maxwell, J. C. *A Treatise on Electricity and Magnetism*; Oxford at Clarendon Press: Oxford, U.K., 1873.

(38) Bader, R. F. W.; Essén, H. *J. Chem. Phys.* **1984**, *80*, 1943–1960.

(39) Bader, R. F. W. *J. Chem. Phys.* **1980**, *73*, 2871–2883.

(40) Bader, R. F. W.; Nguyen-Dang, T. T. *Adv. Quantum Chem.* **1981**, *14*, 63–124.

(41) Bader, R. F. W.; Preston, H. J. T. *Int. J. Quantum Chem.* **1969**, *3*, 327–347.

(42) Pendás, A. M.; Costales, A.; Luaña, V. *Phys. Rev. B* **1997**, *55*, 4275–4284.

(43) Kato, T. *Commun. Pure Applied. Math.* **1957**, *10*, 151–177.

(44) Desclaux, J. P. *Comput. Phys. Commun.* **1975**, *9*, 31–45.

(45) Schwarz, K.; Blaha, P.; Madsen, G. K. H. *Comput. Phys. Commun.* **2002**, *147*, 71–76.

(46) Schwarz, K.; Blaha, P. *Comput. Mater. Sci.* **2003**, *28*, 259–273.

(47) Koga, T.; Watanabe, S.; Kanayama, K.; Yasuda, R.; Thakkar, A. J. *J. Chem. Phys.* **1995**, *103*, 3000–3005.

(48) Desclaux, J. P. *Comput. Phys. Commun.* **1970**, *1*, 216–222.

(49) Blaha, P.; Schwarz, K.; Madsen, G.; Kvasnicka, D.; Luitz, J. *WIEN2k user's guide*; Techn. Universität Wien: Vienna, 2001; http://www.wien2k.at/ reg_user/textbooks (accessed Oct 14, 2010).

(50) Bader, R. F. W.; MacDougall, P. J.; Lau, C. D. H. *J. Am. Chem. Soc.* **1984**, *106*, 1594–1605.

(51) Sagar, R. P.; Simas, A. M.; Smith, V. H.; Ku, A. C. T. *J. Chem. Phys.* **1988**, *88*, 4367–4374.

(52) Shi, Z.; Boyd, R. J. *J. Chem. Phys.* **1988**, *88*, 4375–4377.

(53) Kohout, M.; Savin, A.; Preuss, H. *J. Chem. Phys.* **1991**, *95*, 1928–1942.

(54) Eickerling, G.; Reiher, M. *J. Chem. Theory Comput.* **2008**, *4*, 286–296.

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

(56) Tao, J.; Vignale, G.; Tokatly, I. V. *Phys. Rev. Lett.* **2008**, *100*, 206405.

(57) Biegler-König, F. W.; Bader, R. F. W.; Duke, A. J.; Nguyen-Dang, T. T.; Tal, Y. *J. Phys. B* **1981**, *14*, 2739–2751.

(58) Biegler-König, F. W.; Bader, R. F. W.; Tang, T. H. *J. Comput. Chem.* **1982**, *3*, 317–328.

(59) Cioslowski, J.; Nanayakkara, A.; Challacombe, M. *Chem. Phys. Lett.* **1993**, *203*, 137–142.

(60) Cioslowski, J.; Stefanov, B. B. *Mol. Phys.* **1995**, *84*, 707–716.

(61) Popelier, P. L. A. *Comput. Phys. Commun.* **1998**, *108*, 180–190.

(62) Sanville, E.; Kenny, S. D.; Smith, R.; Henkelman, G. *J. Comput. Chem.* **2007**, *28*, 899–908.

(63) Andersen, O. K. *Phys. Rev. B* **1975**, *12*, 3060–3083.

(64) Madsen, G. K. H.; Blaha, P.; Schwarz, K.; Sjostedt, E.; Nordstrom, L. *Phys. Rev. B* **2001**, *64*, 195134.

(65) Sjöstedt, E.; Nordström, L.; Singh, D. J. *Solid State Commun.* **2000**, *114*, 15.

(66) Brouder, C. *Phys. Rev. B* **2005**, *72*, 085118.

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

(68) Otero-de-la Roza, A.; Luana, V. *Comput. Phys. Commun.* **2009**, *180*, 800–812; Source code distributed by the CPC program library. http://cpc.cs.qub.ac uk/summaries/AECM_v1_0.html (accessed Oct 14, 2010).

(69) Otero-de-la Roza, A.; Blanco, M. A.; Martn Pendás, A.; Luaña, V. *Comput. Phys. Commun.* **2009**, *180*, 157–166; Source code distributed by the CPC program library. http://cpc.cs.qub.ac.uk/summaries/AECB_v1_0.html (accessed Oct 14, 2010).

(70) Gražulis, S.; Chategnier, D.; Downs, R. T.; Yokochi, A. F. T.; Quirós, M.; Lutterolli, L.; Manakova, E.; Butkus, J.; Moeck, P.; Le Bail, A. *J. Appl. Crystallogr.* **2009**, *42*, 726–729; http://www.crystallography.net/ (accessed Oct 14, 2010).

(71) Wyckoff, R. W. G. *Crystal Structures*; Robert E. Krieger: Malabar, 1986.

(72) Gatti, C. Private communication.

(73) Popelier, P. L. A. *Chem. Phys. Lett.* **1994**, *228*, 160–164.

(74) Rodríguez, J. I.; Koester, A. M.; Ayers, P. W.; Santos-Valle, A.; Vela, A.; Merino, G. *J. Comput. Chem.* **2009**, *30*, 1082–1092.

(75) Dovesi, R.; Saunders, V. R.; Roetti, C.; Causà, M.; Harrison, N. M.; Orlando, R.; Aprà, E. *CRYSTAL User's Manual*; Università di Torino: Turin, 1996.

(76) Orlando, R.; Dovesi, R.; Roetti, C.; Saunders, V. R. *J. Phys.: Condens. Matter* **1990**, *2*, 7769–7789.

(77) Mori-Sánchez, P.; Martn Pendás, A.; Luana, V. *J. Am. Chem. Soc.* **2002**, *124*, 14721–14723.

(78) Aray, Y.; Rodríguez, J.; Rivero, J. *J. Phys. Chem. A* **1997**, *101*, 6976–6982.

(79) Aray, Y.; Rodríguez, J.; López-Boada, R. *J. Phys. Chem. A* **1997**, *101*, 2178–2184.

(80) Aray, Y.; Rodríguez, J.; Vega, D. *J. Phys. Chem. B* **2000**, *104*, 5225–5231.

(81) Aray, Y.; Rodríguez, J.; Vega, D.; Coll, S.; Rodrguez Arias, E. N.; Rosillo, F. *J. Phys. Chem. B* **2002**, *106*, 13242–13249.

(82) Aray, Y.; Vega, D.; Rodríguez, J.; Vidal, A. B.; Grillo, M. E.; Coll, S. *J. Phys. Chem. B* **2009**, *113*, 3058–3070.

(83) Luana, V.; Mori-Sánchez, P.; Costales, A.; Blanco, M. A.; Martn Pendás, A. *J. Chem. Phys.* **2003**, *119*, 6341–6350.

(84) Pendás, A. M.; Costales, A.; Blanco, M. A.; Recio, J. M.; Luaña, V. *Phys. Rev. B.* **2000**, *62*, 13970–13978.

(85) Recio, J. M.; Franco, R.; Pendás, A. M.; Blanco, M. A.; Pueyo, L.; Pandey, R. *Phys. Rev. B* **2001**, *63*, 184101−1–7.

(86) Contreras-García, J.; Mori-Sánchez, P.; Silvi, B.; Recio, J. M. *J. Chem. Theory Comput.* **2009**, *5*, 2108–2114.

CT100269E

JCTC Journal of Chemical Theory and Computation

Computational Studies on Polarization Effects and Selectivity in K^+ Channels

Christopher J. R. Illingworth, Simone Furini, and Carmen Domene*

Physical and Theoretical Chemistry Laboratory, Department of Chemistry, University of Oxford, Oxford OX1 3QZ, United Kingdom

Received May 26, 2010

Abstract: Umbrella sampling in combination with a polarizable QM/MM model have been used to study the role of electrostatics and polarization in the translocation and selectivity properties of two K^+ channels, KcsA and KirBac, with ions traversing the channel according to an ion−water−ion mechanism. Analysis of electrostatic interaction energies shows an increased electrostatic gradient within the KirBac channel relative to KcsA. Quantitative measurements of polarization effects induced by ions and water molecules in the channel suggest a decreased interaction with K^+ and Rb^+ close the S2 binding site. This effect cannot be explained solely by the geometry of the polarizable region, or by conformational changes in the filter, but appears to be due to the polarization of the valine residue of the TVGYG selectivity filter motif. We observe that the presence of an ion in the S2 site, and the absence of an ion at the S3 site, where there is a water molecule instead, depolarizes valine and, hence, decreases the interaction energy between that residue and the ion in S2. Our results suggest that the incorporation of polarization effects can make an observable difference to the potential experienced by an ion in the channel.

Introduction

Potassium channels have an essential role in controlling the electric potential across cell membranes, making a vital contribution to the functioning of neurons and cardiac muscle cells, among others. An important landmark in the understanding of these proteins was the first derivation by crystallography of the 3D structure of a potassium channel, namely KcsA,[1] followed by other potassium channel structures, including that of the inwardly rectifying channel KirBac1.1.[2]

These two channels share many structural features. Each channel is a tetramer, composed of identical subunits, and contains the TVGYG sequence motif characteristic of potassium channels[3] (Figure 1a). This sequence motif is called the selectivity filter, and it allows for the fast diffusion of K^+ ions, occurring at a rate approaching 10^8 ions per second,[4] while providing a very strong selectivity for K^+ over Na^+. Experimental studies measuring the conductance of the KcsA channel under symmetrical solution conditions show a more moderate preference for K^+ over Rb^+, such that the rates of diffusion can be represented as $K^+ > Rb^+ \gg Na^+$ [5]. Crystallographic studies have shown that within the selectivity filter, K^+ ions are coordinated by the backbone carboxyl groups of the protein, binding in five locations.[6] These binding sites are often denoted S0 to S4, the number increasing with distance into the intracellular region.

Two challenges in modeling potassium channels have been noted in the literature. First, the average time required for the permeation of an ion is long relative to the time scales typically employed for classical molecular dynamics (MD) simulations. This has led to the use of a range of techniques in order to obtain information about the selectivity and the ion permeation, including umbrella sampling,[7,45] steered molecular dynamics,[8] free energy perturbation,[9] or metadynamics.[10] Second, studies of K^+ and other ion channels have highlighted the potential importance of polarization effects in systems of this nature,[11] and the limitations in using common classical force fields where these effects are often neglected. One solution to this question, followed to an extent in this paper, is the use of quantum mechanical methods to model the ions and the selectivity filter.[12] For example, *ab*

* Corresponding author telephone: +44-1865285401; fax: +44 1865 275410; e-mail: carmen.domene@chem.ox.ac.uk.

10.1021/ct100276c
Published on Web 11/12/2010

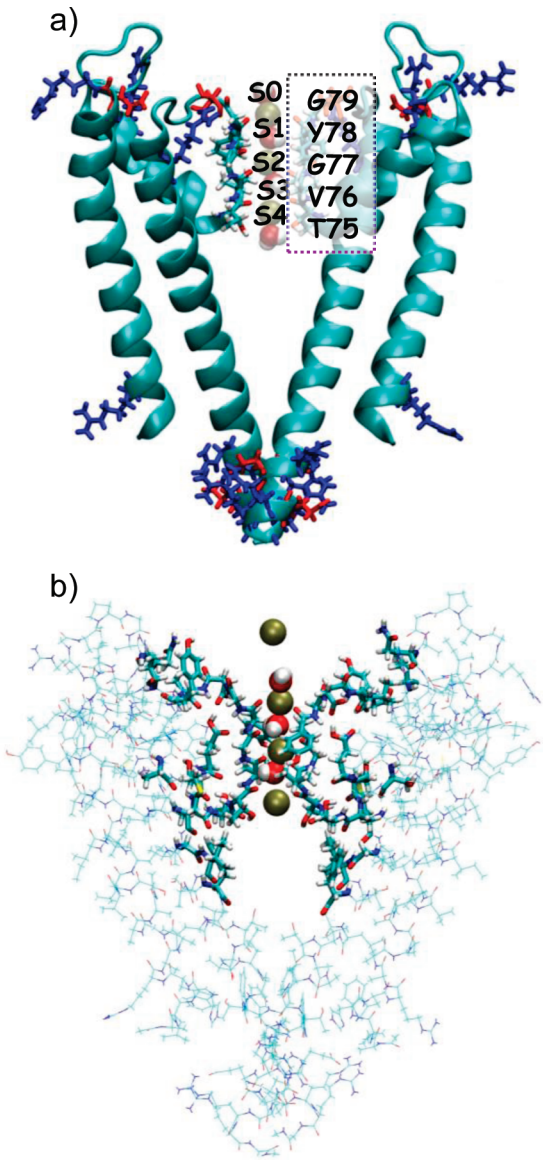


Figure 1. (a) Representation of the selectivity filter of KcsA. Just two of the four chains of the protein are shown for simplicity. The selectivity motif TVGYG is shown in licorice representation, colored by atomic type. Charged residues in the protein are shown in licorice representation, colored by charge, with positively charged residues in blue and negatively charged residues in red. The remainder of the protein is drawn in cartoon representation. The ion−water−ion−water pattern can be seen in the channel in VM representation, with water molecules and ions labeled according to binding site. (b) Representation of the model system in KcsA. Just two of the four chains of the protein are shown. Atoms in the QM region are shown in VDW representation. Atoms in the polarizable MM region are shown in licorice representation. Protein atoms in the nonpolarizable MM region are shown in line representation.

initio geometry optimizations have been used to model protein−ion interactions at the ends of the selectivity filter,[13] and the intracellular entrance of the KcsA channel.[14] Current technology allows for systems of over 400 atoms to be modeled in this way, for example, in demonstrating the role of water in the protein cavity in selectivity between K^+, Na^+, and other ions.[15] In a popular approach to studies of selectivity, *ab initio* calculations are carried out in this manner on a representative structure of either a single binding site[13,16] or another region in the channel,[14,15] containing a single K^+ or Na^+ ion. A second approach that has been taken to quantum mechanical modeling employs snapshots of the entire channel from classical molecular dynamics trajectories, using them as a basis for single-point quantum mechanics calculations. Applied to models of the selectivity filter, this has been used in conjunction with the Merz−Kollman method of assigning partial charges to highlight the shortcomings of standard classical force field methods, in that the atomic partial charges derived for residues in the selectivity filter part of KcsA differ markedly from those used in classical methods.[17] Classical MD simulations have also been used to generate starting points for integrated Quantum Mechanics/Molecular Mechanics (QM/MM) dynamics, for example, in the application of Car−Parrinello molecular dynamics to KcsA.[12,18] In this model, close to 100 atoms are included in a QM region at the center of a classically modeled protein, a few picoseconds of Car−Parrinello dynamics are obtained via a QM/MM method, and Wannier function centers are calculated to evaluate the polarization in the channel.

Quantum mechanical models offer advantages over classical models, avoiding issues of parametrization.[12] However, they require significant amounts of computational time and, so far as dynamics is concerned, do not allow for simulations of a length sufficient to model the transport of ions in K^+ channels.

A promising approach to the modeling of polarization in protein systems is found in the number of methods that have been proposed in which polarization is incorporated into a classical system, or into the classical part of a QM/MM framework.[19] An example of this approach is the Moving Domain-QM/MM method.[20] In this method, a series of ONIOM (Our own N-layered Integrated molecular Orbital + Molecular mechanics)[21] calculations are carried out, with different residues placed in the high level system in each calculation. This is applied in an iterative fashion to modify the charges in the classical low-level system, representing the effect of polarization. Applied to residues from a portion of the KcsA channel, this method gives good agreement with *ab initio* calculations in the electrostatic potential of an ion moving along the filter.

Electrostatic and geometrical properties have each been suggested as important factors in ion selectivity.[22] The hypothesis that K^+ prefers to occupy higher coordination states than Na^+, making it more suited to the 8-fold coordination state observed in K^+ channels, has been a source of debate. MD simulation of K^+ and Na^+ complexes in aqueous solution suggested K^+ to prefer an 8-fold coordination,[23] while *ab initio* studies suggested a common preferred 4-fold coordination for both K^+ and Na^+.[24] Recently, application of Car−Parrinello molecular dynamics suggested an increased preferred coordination number for K^+ compared to that of Na^+,[25] indicating that coordination number is indeed an important factor in selectivity.

A second geometrical hypothesis is the snug-fit theory, which suggests that passage through the channel of the smaller Na^+ ion requires a distortion in the filter residues,

the energetic cost of which mitigates against ion conduction.[26] Although MD simulations have shown considerable flexibility in the selectivity filter, of a magnitude greater than the difference in size between the Na^+ and K^+ ions,[10,27] isothermal titration calorimetry methods suggest that the size of the ion is indeed of importance.[28]

Electrostatic effects have also been proposed as a mechanism for selectivity.[29] In a recent work on selectivity,[30] using small models of the selectivity filter of KcsA and that of the nonselective NaK channel, the authors propose that the pore's selectivity for K^+ over Na^+ increases with an increasing hydration number of K^+ relative to that of Na^+, increasing number of K^+ or Na^+-coordinating dipoles, and decreasing magnitude of the coordinating dipoles provided by the pore.

Therefore, the conclusions emerging from all of these studies are that ion selectivity is achieved by a combination of several factors, not exclusively structural or energetic, involving both the ion and the ion-coordinating ligands, either water or protein, the dehydration penalty of the permeating cations, the electrostatic interactions, and redistribution of charge between the cation and the channel dipoles, the architecture of the ion binding site, and the pore size and flexibility.

Polarization effects have often been suggested as being of importance in K^+ channels.[11,31] However, they have not often been included in computational models of these systems. Therefore, the aim of this work is to establish the potential importance of polarization effects in selectivity and translocation.

In our model, umbrella sampling is used to generate representative sets of structures simulating the passage of Na^+, K^+, and Rb^+ through each of the channels. These sets are then used as a basis for QM/MM calculations. In contrast to methods described in the literature in which ions are displaced along a fixed selectivity filter,[17a] here fluctuations of the protein in response to different positions of the ions are allowed. Representing the ions as QM entities removes concerns which have been raised about the parametrization of ions within a classical framework.[11c] Although the QM/MM method applied here does not allow for charge transfer between the QM ions and waters, and the MM channel, it allows for calculations on the whole protein to be carried out in a feasible amount of time and has the advantage of allowing the calculation of a discrete polarization energy for an ion. In this manner, we aim to derive a comparison between the KirBac and KcsA channels and to evaluate the effect of polarization in these systems.

Materials and Methods

Model Definitions. The atomic structures of KcsA and KirBac were based on the protein data bank entries 1K4C[32] and 1P7B,[33] respectively. Only the transmembrane pore regions were included in the channel models, i.e., amino acids A23−G123 for KcsA and amino acids A40−R151 for KirBac. N termini were acetylated, and an N-methylamide group was added to the C termini. The amino acid E71 of KcsA was modeled in the protonated state,[34] to form a diacid hydrogen bond with D80. The analogous residue in KirBac (E106) was also modeled in the protonated state. Default ionization states were used for the remaining amino acids. Four water molecules were placed at the back of the selectivity filter, in agreement with crystallographic data and previous MD simulations. The channels were embedded in a pre-equilibrated lipid bilayer of 256 1-palmitoyl,2-oleoyl-sn-glycero-3-phosphocholine (POPC) molecules. The channel axis was aligned to the bilayer normal, and the extracellular aromatic belt (amino acids Y45 in KcsA and amino acids Y82 in KirBac) was aligned to the bilayer surface. Lipid molecules closer than 1.0 Å to protein atoms were removed. The atomic systems were solvated using the *Solvate* plug-in of VMD,[35] and then water molecules within 1.2 Å of protein and lipid atoms were removed. Ions corresponding to a concentration of 150 mM of KCl, NaCl, or RbCl were added to neutralize the systems. For convenience, in comparison, the residues of the KirBac structure were renumbered to align the sequence with that of KcsA, such that the residues in the selectivity filter motif TVGYG had residue numbers 75−79. This convention is maintained throughout the remaining sections of this paper.

Molecular Dynamics Simulations. The KcsA and KirBac system models were validated by MD simulations. The same protocol was used for all of the models. Harmonic restraints with a force constant of 20 kcal mol^{-1} $Å^{-2}$ were applied to the protein backbone atoms in the first 500 ps. Then, 20 ns of unrestrained dynamics were performed. The CHARMM27 force field was used for lipids and with CMAP correction[36] for proteins, together with TIP3P model for water molecules.[37] Parameters for ions inside the selectivity filter were selected according to ref 11c, while default CHARMM parameters were used for ions in bulk solution. The particle mesh Ewald algorithm was used to treat the electrostatic interactions.[38] van der Waals forces were smoothly switched off at 10−12 Å. Bonds with hydrogen atoms were restrained by the SETTLE algorithm,[39] in order to use a 2 fs time step. The multi-time-step algorithm r-RESPA[40] was used to integrate the equation of motion. Nonbonded short-range forces were computed every time step, while electrostatic forces were updated every two time steps. MD simulations were performed in the NPT ensemble. Pressure was kept at 1 atm by the Nose−Hoover Langevin piston method,[41] with a damping time constant of 100 fs and a period of 200 fs. The temperature was kept at 300 K by coupling to a Langevin thermostat, with a damping coefficient of 5 ps^{-1}. Calculations were performed using version 2.6 of NAMD.[42] A common assumption dating back to early structural work,[43] adopted in many computational studies,[7,22a,44] is that ions traverse the channel in an ion−water−ion−water fashion. While the potential existence of other permeation pathways has been suggested elsewhere,[45] the ion−water−ion−water pattern was here chosen as the basis for simulation.

Umbrella Sampling Simulations. To obtain representative structures of the selectivity filter with ions at all positions along it, umbrella sampling simulations were carried out. The four ions involved in the conduction process were named I1 (outermost ion, extracellular side) to I4 (innermost ion, intracellular side). Three independent biasing potentials were

applied in order to control the positions along the channel axis. The center of the biasing potential acting on I4 moved from the intracellular cavity to the binding site S4, while the biasing potential acting on I1 moved from the binding site S0 to the extracellular milieu. The position along the axis of I2 and I3 was controlled by a harmonic potential acting on the center of mass of the pair, with the center of the biasing potential acting moving from a situation with ions at the binding site S4 and S2 to a situation with ions at binding sites S2 and S0. Harmonic potentials were updated in 0.5 Å steps. The force constant of the harmonic potentials was set to 20 kcal mol^{-1} $Å^{-2}$. Over 400 windows were run of 120 ps each, representing a total of 48 ns of simulation time to achieve convergence. The first 20 ps of each window were discarded as an equilibration period.

For the structures including Na^+, four sets of umbrella sampling simulations were run, with the Na^+ ion taking the place of one of the K^+ ions in the filter, respectively I1−I4 in the four sets. Thus, a total of 12 sets of umbrella sampling simulations were run, representing KirBac and KcsA with four K^+ ions, with four Rb^+, and with the various orderings of three K^+ and one Na^+ ion.

The last frame of each umbrella sampling calculation was used as an input to a QM/MM calculation in which induced charges were used to model polarization in the filter region. In this way, a comprehensive picture of the passage of ions through each of the two channels has been generated.

QM/MM Calculations. Modeling of electrostatics and, more specifically, polarization in the channel was carried out using a variant of an induced charge method,[46] adapted for modeling larger (more than 1000 atom) systems. Here, the QM region was defined as the four ions in the channel, in addition to the three water molecules closest to the filter, the waters being identified according to the minimum atom−atom distance between a water molecule and either of the middle two ions in the filter. In a previous application of the induced charge model, an enzyme−substrate system was defined in two regions, with a QM ligand surrounded by a few chosen residues of the protein represented as polarizable MM entities.[47] Here, however, the entirety of the protein structure was included, with a polarizable MM region set within a fixed-charge MM representation of the protein. Polarization effects are short-range in nature, and calculations on a charged ligand in explicit aqueous solvent[47] showed more than 75% of the polarization energy being captured by a 5 Å cutoff. Here, the cutoff was defined such that all residues with at least one atom within 9 Å of one of the middle two filter ions in at least one of the structures generated by the umbrella sampling was included in the polarizable MM region. An example of the model system is shown in Figure 1b. More details of the residues included in each case are contained within the Supporting Information. Initial charges for the MM region were taken from the CHARMM force field, as used in the umbrella sampling calculations.

Two simplifications of the induced charge model, appropriate to the study of a large protein system, were made. In small molecule systems, incorporation of polarization due to the interaction between classical parts of the system led to QM−MM interaction energies closer to those in which the MM system was represented by QM-derived charges. Here, in order to isolate polarization effects caused by the movement of ions, and in common with earlier applications of the method to protein−ligand systems,[48] this classical−classical term was omitted. Second, whereas in earlier work four or five iterations of the induced charge method were applied, the vast majority of the change in charges is captured by the first iteration, and as such, only a single iteration was applied here.

Following the induced charge model, polarization was modeled using a series of QM/MM calculations. First, a single-point calculation was carried out on the QM atoms, described above, in the presence of the protein, modeled as a set of point charges without polarization. Second, polarization was incorporated into the charges of atoms in the polarizable MM region. The electrostatic field of the QM region of the system, calculated in the presence of the MM charges using Gaussian 03,[49] was represented as a set of atom-centered multipole series, generated using the GDMA software package.[50] These were used to calculate induced multipole series on the polarizable MM atomic centers, which were then converted into modified point charges using the mulfit software package.[51] Under this process, the overall charge on each molecule within the MM system is conserved. Finally, another single-point calculation was carried out on the QM atoms in the presence of the protein, this time with polarization incorporated into the point charges. Comparing the result of this calculation with the result of the original single-point calculation gave a measure of the polarization energy. In this manner, two key values were calculated, the electrostatic interaction energy between a QM water or ion and the unpolarized channel and the polarization energy, equal to the change in the interaction energy with the addition of polarization to the polarizable region of the channel.

All QM calculations were carried out at the HF/LANL2DZ level, and some representative sets were repeated at the B3LYP/LANL2DZ and MP2/LANL2DZ levels of theory for KcsA for comparison. The LANL2DZ basis set was chosen as being roughly equivalent in quality to the 6-31G* basis set, while having been parametrized for Na^+, K^+, and Rb^+.

Ion Position Definitions. In order to compare energies from different snapshots, it was necessary to obtain a definitive measure of the position of the ions in the channel. Use of a straightforward z coordinate is potentially misleading, due to changes in the internal structure of the channel with changes of ion position, even following alignment of the residues in this region. Therefore, an alternative coordinate, referred to as the relative channel coordinate (RCC), was defined, expressing the ion position in relation to the protein binding sites S0 to S4. Details are presented in the Supporting Information.

Theoretical Basis. For each of the structures on which calculations were performed, if the set of QM atoms is labeled A, then the electrostatic interaction energy of the atoms in A with the channel, $E_{es}(A)$ was calculated as the change in the energy of the QM atoms resulting from the interaction with the unpolarized channel:

$$E_{es}(A) = E_{QMMM1}(A) - E_{QM}(A) - E_{pc1} \quad (1)$$

where $E_{QMMM1}(A)$ is the total energy of the initial QM/MM system, $E_{QM}(A)$ is the energy of the QM atoms without the surrounding MM atoms, and E_{pc1} is the self-energy of the initial point charges in the MM region. The total polarization energy $E_{pol}(A)$ was calculated as the difference between the energies of interaction, calculated as above, for the polarized and unpolarized channel:

$$E_{pol}(A) = [E_{QMMM2}(A) - E_{QM}(A) - E_{pc2}] - E_{es}(A) \quad (2)$$

where $E_{QMMM2}(A)$ is the total energy of the polarized QM/MM system and E_{pc2} is the self-energy of the polarized point charges in the MM region. That is, $E_{pol}(A)$ is the difference in the energy of interaction made by the inclusion of polarization into the MM region. Note that point charges of MM atoms outside of the polarizable region were kept constant.

In order to understand the behavior of ions and water molecules passing through the channel, it was necessary to calculate energies of interaction for individual molecules in the QM region. The interaction energy of a molecule was calculated by performing QM/MM calculations on a system in which that molecule was omitted. For any molecule m, the electrostatic interaction energy $E_{es}(m)$ of that molecule with the rest of the system was defined as

$$E_{es}(m) = E_{es}(A) - E_{es}(A_m) \quad (3)$$

where A_m is the set of QM atoms excluding molecule m and $E_{es}(A_m)$ is calculated according to eq 1.

Applied to an ion, this energy is equal to the sum of the electrostatic energies of interaction between the ion and the other ions and QM waters, and that between the ion and the channel itself. In order to derive the interaction energy of an ion or water molecule with the channel excluding ion−ion effects, the QM−QM interaction energy $E_{QM}(m)$ was defined as the *in vacuo* interaction of a molecule in the QM region with the rest of the QM region and was subtracted from the total interaction energy $E_{es}(m)$ (defined in eq 3). The interaction energy of a molecule m with the channel is thus

$$E_{channel}(m) = E_{es}(m) - E_{QM}(m) \quad (4)$$

Similarly, the polarization energy between the molecule m with the channel, $E_{pol}(m)$, was defined as

$$E_{pol}(m) = E_{pol}(A) - E_{pol}(A_m) \quad (5)$$

where $E_{pol}(A_m)$ was calculated according to eq 2.

By the separation of individual energies of interaction outlined above, and the consideration of multiple structures with multiple ion and water positions relative to the channel, it is possible to obtain an interaction energy profile describing the passage of ions through the channel. Energy values for ions and waters were averaged over bins of RCC length 0.1 in order to calculate this. It is important to note that this interaction energy profile describes the energy of interaction between the ions and the channel. Effects such as desolvation are omitted in order to isolate terms arising from the interaction of the ions and the channel. These interaction energies should not be confused with free energy differences.

Calculations for the KcsA channel with K^+ ions were repeated using the B3LYP/LANL2DZ and MP2/LANL2DZ levels of theory, generating electrostatic interaction and polarization energy profiles for ions moving through the channel. The electrostatic interaction in each case was virtually identical to that found at the HF level of theory. The overall shape of the polarization energy profiles was preserved.

By means of similar calculations, it was possible to carry out residue-by-residue decompositions of the electrostatic and polarization energies. The electrostatic interaction between a QM molecule and a residue within the protein was calculated as described in eq 3, with the MM region containing only the residue in question. This method was used to measure the contributions of different residues to selectivity in the S2 site, the S2 binding site exhibiting the maximal selectivity between Na^+ and K^+ ions.[16a] The polarization energy of interaction between a QM molecule and a residue within the protein was similarly calculated as described in eq 5, with the MM region only containing the residue in question. For polarization calculations, the polarized charges for the residue were taken from the polarized charges generated for the entire QM/MM system. Note that due to the symmetry of the system, when a single residue is considered, this in fact corresponds to four identical amino acids, one from each of the identical protein chains.

Modeling polarization in this way is complicated by the fact that the polarizable region is finite and undergoes small geometrical changes between snapshots, and by the fact that ions move relative to this region. As the polarization energy is dependent on the region over which it is measured, this leads to an inherent variation in the calculated polarization energy as an ion moves through the polarizable region. In order to model the effect of these relative changes in geometry on the polarization energy, a geometrical factor was calculated, giving an indication of the expected polarization energy at each ion position relative to the channel. Modeling the ion, *i*, as a unit charge and placing unit dipoles at each of the atoms *j* in the polarizable region gave the sum of charge−dipole interactions:

$$P_G(i) = \sum_{j=1}^{N} \frac{1}{r_{ij}^2} \quad (6)$$

where r_{ij} is the distance between *i* and *j*. This factor gave a reference against which calculated polarization energies could be compared.

Results

In order to analyze the QM/MM data, results for single ions were grouped by RCC in bins of width 0.1, equal to a tenth of the distance between binding sites. Average values of the electrostatic interactions were calculated for each of the RCC bins. In addition to a combined energy, representing the interaction of an ion with the remainder of the system, the interaction energies with the QM system, and with the protein, were calculated separately. Figure 2 shows mean

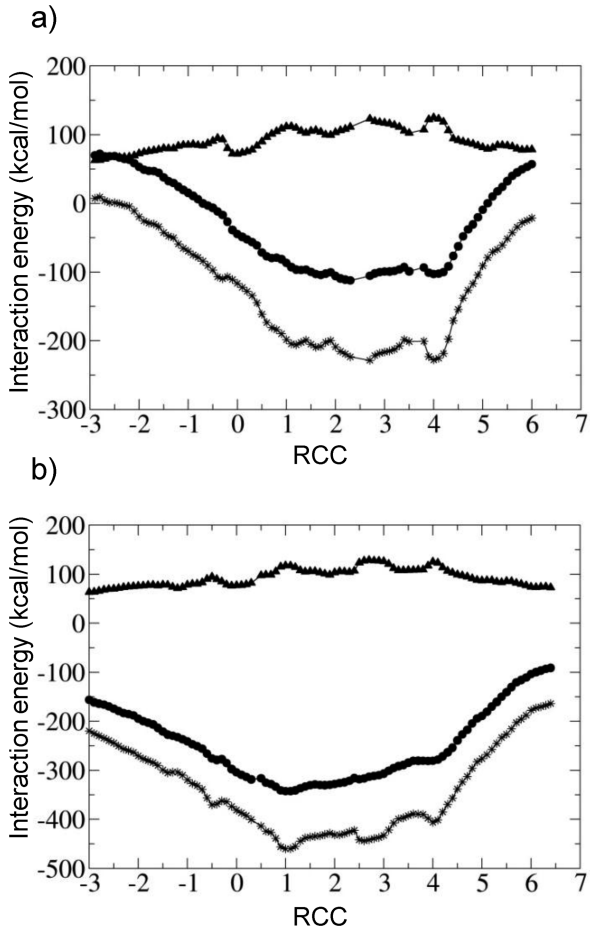


Figure 2. Averaged electrostatic interaction energies of K^+ ions moving through the selectivity filters of (a) KcsA and (b) KirBac. For each system, the total interaction energy (circles), the QM−QM interaction energy (triangles), and the ion−protein interaction energy (stars) are plotted.

electrostatic interaction energies for K^+ ions passing through the two different channels. Equivalent data for Na^+ and Rb^+ are given in Supporting Information Figure S1. Each point in the figure represents a collection of structures. Note that the energies, representing the interaction of the ion with the channel, exclude factors such as desolvation and should not be confused with free energy differences. While the position of the ion in question is denoted by the RCC, plotted on the horizontal axis, the remaining three ions can occupy a range of positions, leading to variations in the energy that are in general removed by averaging over the snapshots in each bin.

Absolute electrostatic interaction energies, by contrast to free energy differences, are large, on the order of hundreds of kilocalories per mole.[52] Examination of the these energies revealed that Na^+ ions have a stronger electrostatic interaction with the channel than do K^+ or Rb^+. In order to compare the behaviors of the different ions, mean electrostatic interaction energies were calculated for ions for which the RCC was between 1 and 4 (i.e., in the filter region). The differences between these energies were then corrected, subtracting the differences between the experimentally observed desolvation energies of the ions.[53] In KcsA, the corrected mean interaction with K^+ was 2.1 kcal/mol greater than that of Rb^+, and 6.5 kcal/mol greater than that for Na^+, while for KirBac, the K^+ interaction was 2.0 kcal/mol greater than that for Rb^+, and 12.8 kcal/mol greater than that for Na^+. Accurate characterization of ion selectivity requires the balancing of many subtle factors and is not easy with either classical or quantum mechanical approaches. In this case, a greater level of sampling at the QM/MM level would inevitably bring a greater degree of accuracy, though the ordering of energies found here ($K^+ > Rb^+ > Na^+$) was consistent with experimental measurements.[5]

Table 1. Mean Electrostatic Interaction Energies for Ions in Binding Sites S1 to S4 with the Amino Acids of the TVGYG Motif and Another Residue in KcsA and KirBac[a]

residue			mean interaction energy (kcal/mol)	
KcsA	KirBac	binding site	KcsA	KirBac
Thr75	Thr75	S1	−2.6 ± 1.4	−4.6 ± 1.2
Thr75	Thr75	S2	−15.6 ± 4.9	−15.8 ± 3.5
Thr75	Thr75	**S3**	**−58.9 ± 6.5**	**−59.3 ± 4.5**
Thr75	Thr75	**S4**	**−109.5 ± 8.0**	**−101.9 ± 10.9**
Val76	Val76	S1	−16.3 ± 5.1	−22.3 ± 1.4
Val76	Val76	**S2**	**−43.0 ± 7.3**	**−55.8 ± 8.0**
Val76	Val76	**S3**	**−36.9 ± 7.9**	**−47.0 ± 5.3**
Val76	Val76	S4	−10.8 ± 2.6	−14.2 ± 2.0
Gly77	Gly77	**S1**	**−62.7 ± 3.6**	**−58.1 ± 4.8**
Gly77	Gly77	**S2**	**−64.9 ± 4.6**	**−59.6 ± 3.7**
Gly77	Gly77	S3	−23.3 ± 3.9	−19.1 ± 2.3
Gly77	Gly77	S4	−6.5 ± 1.1	−7.2 ± 1.3
Tyr78	Tyr78	**S1**	**−27.0 ± 7.2**	**−45.4 ± 3.9**
Tyr78	Tyr78	S2	−4.9 ± 1.9	−8.5 ± 2.9
Tyr78	Tyr78	S3	−2.6 ± 1.2	−5.1 ± 1.2
Tyr78	Tyr78	S4	0.5 ± 0.7	−0.9 ± 0.9
Gly79	Gly79	S1	−0.4 ± 1.4	−6.9 ± 2.6
Gly79	Gly79	S2	4.4 ± 1.1	2.5 ± 2.2
Gly79	Gly79	S3	3.9 ± 0.6	2.3 ± 0.8
Gly79	Gly79	S4	3.3 ± 0.2	2.7 ± 0.7
Arg52	Ala52	S1	49.6 ± 0.7	0.2 ± 0.1
Arg52	Ala52	S2	48.1 ± 0.6	0.1 ± 0.2
Arg52	Ala52	S3	46.2 ± 0.6	−0.2 ± 0.1
Arg52	Ala52	S4	44.0 ± 0.5	−0.3 ± 0.1

[a] Each energy value is calculated from a representative set of structures; the standard deviation gives a measure of the variability of the energy in each case. Interaction energies making the largest contributions to the electrostatic interaction with ions in different sites are highlighted in bold.

Sets of structures of KcsA and KirBac containing ions with an RCC equal to 1.0, 2.0, 3.0, or 4.0 were selected, and average residue-by-residue electrostatic interaction energies were calculated for ions in each system at each of these points. Some representative results for these interactions are shown in Table 1. Unsurprisingly, for both protein systems, the largest residue differences between ion positions occurred for residues of the selectivity filter, which comprise the ionic binding sites. For each system, differences between the residue electrostatic interaction energies with ions in S1 and in S4 were summed across all residues, giving an approximation to the overall difference in binding energies between ions in S1 and ions in S4. As can be seen in Figure 1, for structures analyzed here, the interaction energy between the KcsA channel and ions at S4 is slightly stronger than that at S1, while the interaction energy between KirBac and ions at S4 is significantly weaker than that at S1. Residue-by-residue comparison of the sequences, carried out on the basis of the alignment in ref 33, suggests that a number of factors contribute to this effect. First, the electrostatic interaction between the ion in the S1 site and the residues

Table 2. Breakdown of Electrostatic Interaction Energies between Na^+, K^+, and Rb^+ Ions in the S2 Site of KcsA and KirBac[a]

	KcsA			KirBac		
	Na^+	K^+	Rb^+	Na^+	K^+	Rb^+
sum of residue interactions for all residues (kcal/mol)	−113.7	−104.2	−100.7	−335.1	−333.2	−319.2
relative binding energy of ion with all residues, corrected for desolvation (kcal/mol)	+7.2	0.0	−1.3	+14.8	0.0	+9.3
sum of residue interactions for residues TVGYG (kcal/mol)	−135.8	−124.2	−120.1	−139.5	−137.2	−125.8
relative binding energy of ion with residues TVGYG, corrected for desolvation (kcal/mol)	+5.1	0.0	−0.7	+14.5	0.0	+6.7

[a] Relative energies are calculated by subtracting experimental values for desolvation from the interaction energies and are given relative to that for K^+.

Y78 and G79 is significantly stronger in KirBac than in KcsA (as mentioned earlier, for convenience, in comparison, the residues of the KirBac structure were renumbered to align the sequence with that of KcsA, such that the residues in the selectivity filter motif TVGYG had residue numbers 75−79). Second, KcsA and KirBac have a different distribution of charged and uncharged residues. In both proteins, charged residues are clustered away from the center of the cell membrane. However, in the extracellular region, KcsA contains the positively charged residues R52, R64, and R89, in addition to the negatively charged residues E51 and D80. In the intracellular region, KcsA contains the positively charged R117, R121, R122, and R27 and the negatively charged residues E118 and E120 (see Figure 1a). No charged residues occur toward the center of the cell membrane. Therefore, in KcsA, ions being transported out of the cell move from a region of the protein with an overall charge of +8 to a region with an overall charge of +4. In KirBac, however, the distribution is different. The extracellular region contains the negatively charged residues D51 and D80, with E95 slightly buried into the membrane region and E71 in a neutral protonation state, while the intracellular region contains the positively charged residues R14, K22, R113, and R116 and the negatively charged D15. Thus, ions being transported out of the cell move from a region of the protein with an overall charge of +12 to a region with an overall charge of −12. This creates a differing energy gradient which can be observed in the residue differences. As illustrated in Table 1, the movement of the positively charged ion away from, for instance, R52 in KcsA contributes to the favorability of S4 over S1, an effect that is not present in the equivalent neutral residue in KirBac. Electrostatic interaction energies with ions in the S2 and S3 positions are intermediate to those for S1 and S4, corresponding to the positions of the ions relative to these charged residues. While the observed difference in the interaction energies with charged residues would be reduced by the internal dielectric constant of the protein, the sum of the long-range interactions with charged residues is non-negligible.

Residue-by-residue interactions were also calculated for structures of KcsA and KirBac containing either a Na^+ or Rb^+ ion with RCC equal to 2.0. Results are shown in Table 2. Taking the sum of all of the residue interactions in each case, Na^+ bound more strongly than K^+. A larger difference in binding energy was also observed for Rb^+ in KirBac, although in KcsA, Rb^+ bound more strongly in the S2 site than did K^+. In each case, the residues in the TVGYG filter motif made a substantial contribution to the relative binding energies of the ions, accounting for between 50 and 98% of the calculated energy difference.

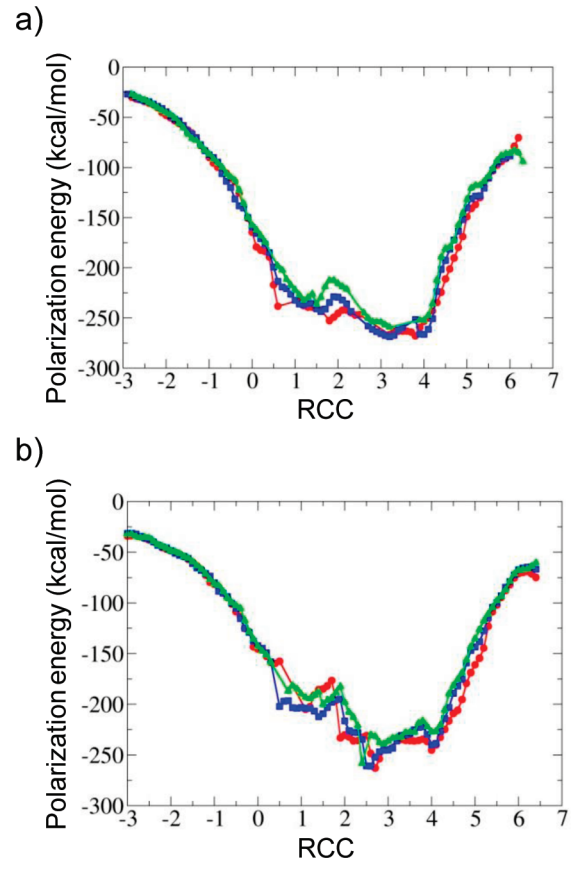


Figure 3. Averaged polarization energies of K^+ ions (blue squares), Na^+ ions (red circles), and Rb^+ ions (green triangles) moving through the selectivity filter of (a) KcsA and (b) KirBac.

Average values of the polarization energy were calculated for each bin in a similar manner to the electrostatic energies (Figure 3). As already noted, the polarization energy is dependent on the position of the ion relative to the polarizable region of the protein, and therefore, it is more difficult to compare energies at different ion positions. However, this does not affect the comparison of different ions at identical positions in the filter, and in this measure, a pattern similar to that in the electrostatic energies was observed.

Averaged polarization energies were calculated for ions in the filters of each KcsA and KirBac. No consistent pattern

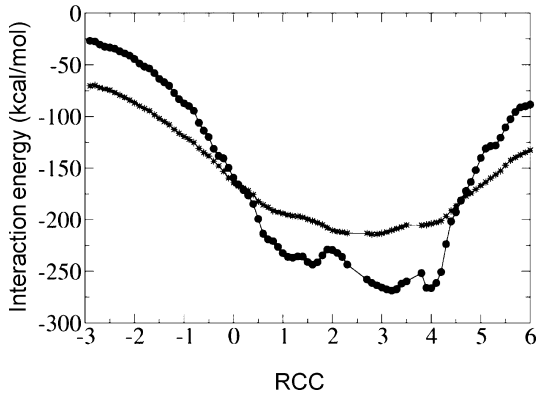


Figure 4. Measured (circles) and expected (stars) polarization energies of K^+ ions moving through the KcsA selectivity filter. Linear scaling has been applied to the latter measure in order to give identical mean values, thereby aiding comparison of the shapes of the graphs.

was observed for the relative polarizations of K^+ and Na^+. In KcsA, Na^+ had slightly greater polarization energy than K^+, while in KirBac the polarization of Na^+ was lower than that of K^+. In each case, the polarization energy of K^+ tended to be stronger than that of Rb^+.

An interesting feature of the polarization energy is the presence of a peak representing low polarization energy at a relative channel coordinate of around 2, seen for both K^+ and Rb^+ ions in both KcsA and KirBac. A similar feature is also observed for Na^+ ions, albeit less pronounced and at a relative channel coordinate slightly greater than 2 in KcsA, and slightly less than 2 in KirBac. Calculation of expected polarization energies suggested that this feature is not simply an artifact of the geometry of the polarizable region. A comparison of the expected polarization energy with the measured averaged polarization energies for K^+ in KcsA is shown in Figure 4. A constant term has been added to the former measure in order to give identical mean values, thereby aiding comparison of the shapes of the graphs.

The expected values here demonstrate that the movement of the ions through the polarizable region has a significant effect on the measured polarization energy. Toward the edges of the polarizable region, the expected polarization decreases due to the decrease in polarizable volume in the immediate vicinity of the ion. This contributes in part to the increase in the measured polarization energy in the filter region relative to that in the cavity and external regions. However, the decreased polarization measured at S2 cannot be explained by the movement of the ion relative to the polarizable region.

An additional factor considered as a possible source of lowered polarization energy in the S2 site was the presence of changes in the conformation of the backbone of the channel residues. In some of the structures analyzed, flips in one of the Val76 residues were observed, such that the carbonyl group of the residue pointed away from the center of the channel. Flipping of carbonyl ligands is an ordinary observation in K^+ channels, as described in various computational studies for different K^+ channels.[54] Average interaction and polarization energies were calculated as above for a set of structures edited so that these flips did not occur, and for a randomly edited set of structures, where an identical number of randomly chosen structures were removed (results shown in Supporting Information Figure S2). A difference between the sets was observed close to S2, with an increase in both the electrostatic interaction energy and the polarization energy when the flips were removed, likely due to the interaction with the additional carbonyl group present when no flip occurs. Despite this increase, however, the overall character of the polarization energy profile was unaffected.

A residue-by-residue breakdown of polarization energy was carried out for structures from the umbrella sampling calculations for KcsA and KirBac with four K^+ ions. Residue-by-residue polarization energies were calculated for structures containing ions with an RCC of 1.0, 1.6, 1.7, 1.8, 1.9, 2.0, 3.0, or 4.0, thereby providing detail of the polarization interaction at each binding site. In each case, the vast majority of the polarization energy was captured by the interaction of ions with the residues in the TVGYG selectivity motif, with other residues in the polarizable region contributing 10% or less of the total polarization energy. Table 3 shows mean polarization energies for residues in

Table 3. Residue-by-Residue Breakdown of Single Ion Polarization Energies for K^+ Ions at Different Locations in the Selectivity Filter (TVGYG Motif) of KcsA and KirBac[a]

	mean polarization energy (kcal/mol), ion and RCC									
	K^+	K^+	K^+	K^+	K^+	K^+	K^+	K^+	Na^+	Rb^+
residue	1.0	1.6	1.7	1.8	1.9	2.0	3.0	4.0	2.0	2.0
					KcsA					
T75	−23.6	−34.4	−33.8	−48.2	−62.3	−62.5	−96.2	−197.5	−45.9	−61.2
V76	−53.1	−91.6	−89.8	−81.2	−68.6	−75.9	−122.6	−45.0	−84.3	−69.4
G77	−63.7	−61.2	−65.3	−58.9	−59.6	−60.3	−19.3	−6.8	−82.7	−51.8
Y78	−38.5	−37.9	−33.6	−29.3	−15.3	−17.8	−13.5	0.1	−22.7	−18.0
G79	−56.2	−24.1	−18.4	−17.1	−16.8	−17.7	−15.6	−8.1	2.2	−12.7
total	−240.5	−252.6	−248.3	−239.6	−231.8	−242.2	−277.3	−280.5	−266.3	−222.2
					KirBac					
T75	−26.4	−34.9	−37.9	−42.6	−52.9	−51.3	−94.4	−171.8	−33.2	−56.3
V76	−73.9	−98.8	−98.9	−105.9	−97.4	−101.2	−146.9	−72.1	−87.7	−86.5
G77	−58.9	−68.8	−66.2	−54.8	−52.3	−65.5	−15.9	−9.2	−79.0	−50.6
Y78	−93.5	−60.0	−45.9	−40.1	−32.2	−17.9	−21.1	0.4	−51.9	−25.9
G79	1.2	1.9	5.3	3.7	7.7	−4.5	−0.4	−1.0	0.5	−1.3
Total	−244.8	−252.9	−239.7	−237.9	−226.2	−238.2	−276.2	−254.9	−245.9	−218.9

[a] Values for Na^+ and Rb^+ are given for the center of the S2 binding site, this site exhibiting the maximal selectivity between ions.

the selectivity filter, in structures containing ions with specific relative channel coordinates.

Certain patterns can be observed in the residue-by-residue polarization energy data. For example, in KcsA, the polarization energy due to the residue T75 steadily increases in magnitude as the RCC of a K^+ ion increases from 1.0 to 4.0, this effect being easily explained by the increasing proximity of the ion to this residue as the RCC increases. Simultaneously, the components of the polarization energy due to the interactions with Y78 and G79 exhibit a constant decrease in magnitude with increasing RCC, as the distance between the ion and these residues increases. The polarization energy due to the interactions with V76 and G77 are more complex. As might be expected, the interaction with V76 achieves its greatest value when an ion has a RCC of 3.0, in which case the carboxyl group of the residue is directly involved in binding the ion. However, when the RCC of the ion is increased from 1.6 to 2.0, moving toward this residue, there is a decrease in the magnitude of the polarization interaction. Figure 5 shows the mean changes in charge with polarization in the T75 and V76 residues for structures of KcsA containing a K^+ ion with an RCC of either 1.6 or 2.0 (details for other residues given in Supporting Information Figures S4 and S5). Changes in charge have been averaged over the four copies, one per chain, of each residue. In general, the magnitude of the changes in the atom charges reflects
the magnitude of the polarization energies recorded in Table 3. For example, the polarization energy captured by the T75 residue increases in magnitude as the RCC of the ion increases from 1.6 to 2.0. Corresponding with this movement, the change in the charge of the α carbon atom of this residue also increases in magnitude, from −0.15 to −0.29. Examination of the charges in the V76 residue shows that at each ion position, the polarization increases the magnitude of the negative charge on the carbonyl oxygen, as would be expected for an interaction with a set of positively charged ions. However, as the ion moves closer to this carbonyl group, its RCC increasing from 1.6 to 2.0, the amount of polarization decreases, a surprising result. Some explanation for this can be seen in Figure 6, which shows the distribution of the RCC of the ion adjacent to that with an RCC between 1.6 and 2.0, moving into the cell. For the purposes of the figure, the RCC of the second ion is rounded to the nearest integer value. When the first ion is located further into the channel, its RCC increasing from 1.6 to 2.0, the location of this next ion also changes so that the RCC of this second ion is 4 or above in all of the observed structures. Hence, the increase in the RCC of an ion from 1.6 to 2.0 is associated with a second ion leaving the S3 binding site, away from the carbonyl oxygen of the V76 residue. This decreases the polarization of this residue, lowering the energy of polarization between V76 and the ion further up the channel. Hence, binding of an ion in the S2 binding site is affected by polarization effects that result from the interaction with other ions further into the channel. Repeat calculations performed for a single Na^+, K^+, or Rb^+ ion moving through KcsA or KirBac, for which the other binding sites

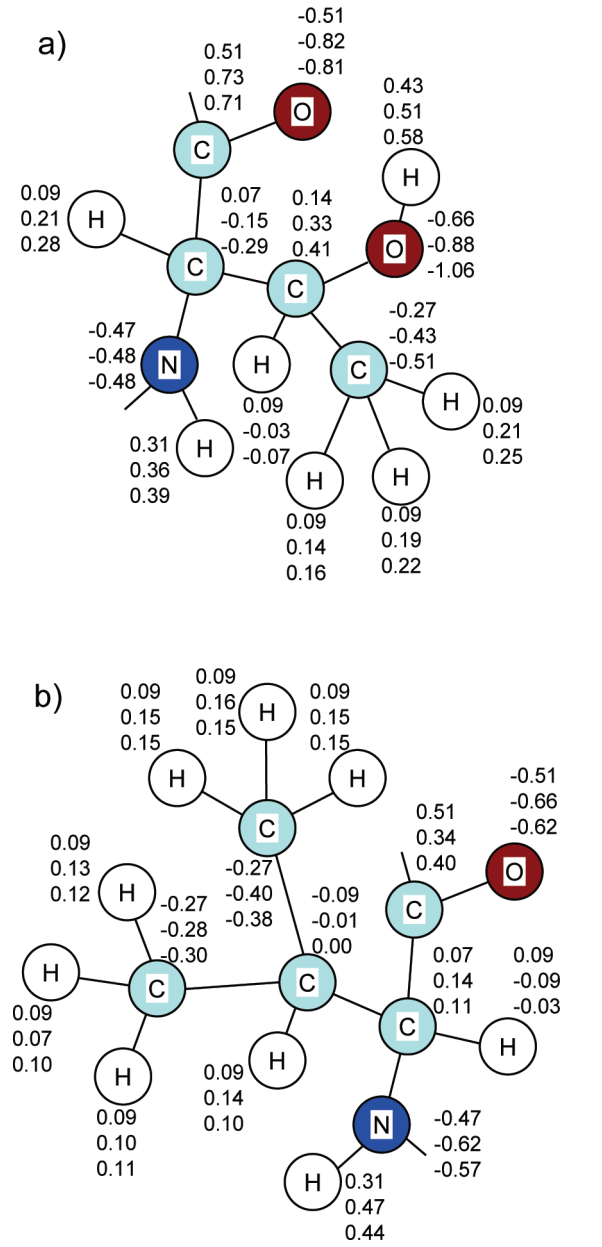


Figure 5. Mean atomic charges in the (a) T75 and (b) V76 residues of KcsA before and after polarization of the channel in structures. Charges are given for the unpolarized system (top), and for structures with K^+ ions at RCC values of 1.6 (middle number) and 2.0 (bottom number). Changes are averaged over the atoms in the four V76 residues in each case.

were occupied by water molecules, did not show the same effect at S2.

In KirBac, patterns broadly similar to those seen in KcsA can be observed, with differences in the polarization interactions of V76 and G77 leading to a drop in the magnitude of polarization at S2 as the RCC of a K^+ ion increases from 1.6 to 2.0. Analysis of changes in charges again showed a decrease in the magnitude of polarization of the carbonyl oxygen of V76 occurring as a result of this movement. As with KcsA, this can be associated with interactions taking place further down the channel. A difference between KcsA and KirBac occurs in residues Y78 and G79. In KcsA, the

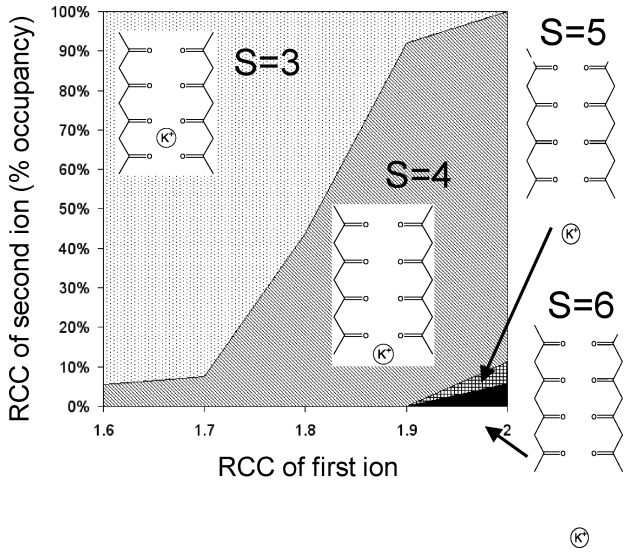


Figure 6. Distribution of the RCC of the next ion into the KcsA channel, for structures containing an ion with an RCC between 1.6 and 2.0. The RCC of the first ion is given on the horizontal axis, while the vertical axis shows the distribution of the RCC of the second ion. The RCC of the second ion is rounded to the nearest integer and takes values of 3 (dotted shading), 4 (diagonal shading), 5 (hashed shading), or 6 (solid shading). Inset figures indicate the position of the second ion. As the RCC of the first ion increases from 1.6 to 2.0, the second ion leaves the S3 site and moves into S4 or a location further into the channel.

polarization of these residues is split between the two residues, while in KirBac, it is concentrated in changes in the charges of the tyrosine residue.

Electrostatic interaction and polarization energies calculated for water molecules in the channels were much smaller in magnitude than those calculated for the ions.

Discussion and Conclusions

By means of a series of umbrella sampling calculations, we have obtained a complete set of structures of Na^+, K^+, and Rb^+ ions passing through the KcsA and KirBac selectivity filters. The umbrella sampling method provides a representative set of sample structures for the QM/MM calculations with a number of advantageous properties. First, ions can be located in a range of different places within the filter region, not being constrained to the positions of lowest energy. Second, the freedom of the ions to move relative to each other allows for the sampling of a range of different ion configurations given that one ion is in a specific point in the channel. Third, throughout the simulations, the protein is unfixed, allowing for the channel to respond to the movements of the ions.

Whereas, in some previous work, QM calculations have been carried out on some averaged channel structures[17a] or by running dynamics calculations from a small number of starting structures,[12] we have extended the induced charge method of polarization to carry out polarizable QM/MM calculations on a complete representative set of classically generated structures. The use of quantum mechanical methods has the advantage of avoiding questions of parametrization of ions, while the incorporation of polarization into a classical force field has the advantage of allowing for the calculation of electrostatics and polarization in a large number of structures in a reasonable amount of time. Although our QM/MM methodology does not incorporate charge transfer, potentially leading to a slight inflation in the interaction energies with the selectivity filter, we do not anticipate that this omission would alter the fundamental nature of the results obtained.

Energetic calculations carried out in a QM/MM framework showed mean binding energies consistent with those observed experimentally, with a slight preference in binding for K^+ over Rb^+, and a more significant difference between K^+ and Na^+. Comparison of the electrostatic interaction energies for KirBac and KcsA highlight the existence of an electrostatic gradient in the KirBac filter not present in KcsA, with a significantly higher energy for an ion in the S4 binding site than in the S1 binding site. Two factors are presented in explanation of this: first, an energy contribution resulting from the residues of the TVGYG selectivity motif and, second, a distribution of charged residues in the intra- and extra-cellular regions of KirBac disfavoring to an extent the passage of ions from S1 to S4. The second of these carries the implication that residues relatively far from the selectivity filter can affect the electrostatic energies of K^+ and Rb^+ ions moving through it. Changes in the electrostatic energy may modify the conductance properties of the channels. Thus, the conductance variability observed in the K^+ channel family may partially be explained by an electrostatic effect of nonconserved residues surrounding the selectivity filter.

Calculations of the polarization energy of each ion were made using a QM/MM induced charge method applied to a classical representation of the protein. Modeled in this way, polarization was observed to be a short-range effect, with 90% or more of the polarization energy caused by the ions being captured by the residues in the TVGYG filter motif. As noted in the electrostatic interaction energies calculated above, calculations that neglect polarization can recreate certain experimentally observed characteristics of the channel, for example, the K^+/Na^+ selectivity. However, calculations of polarization energy indicate that polarization plays a nontrivial role in the energetics of ion diffusion, with structural features of the channel influencing the binding energy. Notably, we observed a reduction in the magnitude of the polarization energy in the S2 binding site, this pattern being repeated for KirBac and KcsA with K^+ and Rb^+ ions, and in a less consistent way for Na^+. This effect, at least in part, appears to be due to the influence of a second ion polarizing the V76 residue, affecting the interaction energy of the ion near S2. This has potential implications for the modeling of these systems. While substantial progress toward understanding the mechanism of K^+ channels has been made through single-ion, single-binding site models of the selectivity filter, by modeling multiple ion positions and including polarization effects, it has been shown that the interplay between an ion and a binding site in the channel can be affected by interactions taking place in other sites in the channel. In order to derive a fully accurate picture of the behavior of an ion in one binding site, the behavior of ions

in the remainder of the channel should be considered. In addition, while many excellent studies of free energy differences in selection and permeation have been carried out using nonpolarizable classical force fields, the observation here of multi-ion polarization effects cannot be ignored. Whether the differences caused by polarization would translate into energetic or geometrical differences in a fully polarizable model of dynamics is a question for future research. However, our results support the consensus of opinion that polarizable models are of great importance in modeling ion channel systems.

Acknowledgment. C.D. thanks The Royal Society for a University Research Fellowship. This work was supported by grants from The Leverhulme Trust and the EPSRC. The Oxford Supercomputing Center and HECToR are acknowledged for providing computational resources.

Supporting Information Available: (1) Details of the RCC method for defining ion channel positions. (2) Lists of residues included in the polarized MM region of the calculations on KcsA and KirBac. (3) Figure S1. (4) Figure S2. (5) Figure S3. (6) Figure S4. (7) Figure S5. (8) List of residues included in the polarized MM region of the calculations. (9) Details of smoothing algorithm used for plotting energy data. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Doyle, D. A.; Cabral, J. M.; Pfuetzner, R. A.; Kuo, A. L.; Gulbis, J. M.; Cohen, S. L.; Chait, B. T.; MacKinnon, R. The structure of the potassium channel: Molecular basis of K^+ conduction and selectivity. *Science* **1998**, *280* (5360), 69–77.

(2) Kuo, A. L.; Gulbis, J. M.; Antcliff, J. F.; Rahman, T.; Lowe, E. D.; Zimmer, J.; Cuthbertson, J.; Ashcroft, F. M.; Ezaki, T.; Doyle, D. A. Crystal structure of the potassium channel KirBac1.1 in the closed state. *Science* **2003**, *300* (5627), 1922–1926.

(3) MacKinnon, R.; Cohen, S. L.; Kuo, A. L.; Lee, A.; Chait, B. T. Structural conservation in prokaryotic and eukaryotic potassium channels. *Science* **1998**, *280* (5360), 106–109.

(4) Hille, B. *Ion channels of excitable membranes*, 3rd ed.; Sinauer Associates: Sunderland, MA, 2001; pp 1−814.

(5) LeMasurier, M.; Heginbotham, L.; Miller, C. KcsA: It's a potassium channel. *J. Gen. Physiol.* **2001**, *118* (3), 303–313.

(6) Zhou, Y. F.; Morais-Cabral, J. H.; Kaufman, A.; MacKinnon, R. Chemistry of ion coordination and hydration revealed by a K^+ channel-Fab complex at 2.0 angstrom resolution. *Nature* **2001**, *414* (6859), 43–48.

(7) Berneche, S.; Roux, B. Energetics of ion conduction through the K^+ channel. *Nature* **2001**, *414* (6859), 73–77.

(8) Leech, J.; Prins, J. F.; Hermans, J. SMD: Visual steering of molecular dynamics for protein design. *IEEE Comput. Sci. Eng.* **1996**, *3* (4), 38–45.

(9) Aqvist, J.; Luzhkov, V. Ion permeation mechanism of the potassium channel. *Nature* **2000**, *404* (6780), 881–884.

(10) Domene, C.; Klein, M. L.; Branduardi, D.; Gervasio, F. L.; Parrinello, M. Conformational changes and gating at the selectivity filter of potassium channels. *J. Am. Chem. Soc.* **2008**, *130* (29), 9474–9480 .

(11) (a) Allen, T. W.; Kuyucak, S.; Chung, S. H. Molecular dynamics study of the KcsA potassium channel. *Biophys. J.* **1999**, *77* (5), 2502–2516. (b) Kuyucak, S.; Andersen, O. S.; Chung, S. H. Models of permeation in ion channels. *Rep. Prog. Phys.* **2001**, *64* (11), 1427–1472. (c) Roux, B.; Berneche, S. On the potential functions used in molecular dynamics simulations of ion channels. *Biophys. J.* **2002**, *82* (3), 1681–1684.

(12) Bucher, D.; Raugei, S.; Guidoni, L.; Dal Peraro, M.; Rothlisberger, U.; Carloni, P.; Klein, M. L. Polarization effects and charge transfer in the KcsA potassium channel. *Biophys. Chem.* **2006**, *124* (3), 292–301.

(13) Ban, F. Q.; Kusalik, P.; Weaver, D. F. Density functional theory investigations on the chemical basis of the selectivity filter in the K^+ channel protein. *J. Am. Chem. Soc.* **2004**, *126* (14), 4711–4716.

(14) Kariev, A. M.; Znamenskiy, V. S.; Green, M. E. Quantum mechanical calculations of charge effects on gating the KcsA channel. *BBA Biomembr.* **2007**, *1768* (5), 1218–1229.

(15) (a) Kariev, A. M.; Green, M. E. Quantum mechanical calculations on selectivity in the KcsA channel: The role of the aqueous cavity. *J. Phys. Chem. B* **2008**, *112* (4), 1293–1298. (b) Kariev, A. M.; Green, M. E. Quantum calculations on water in the KcsA channel cavity with permeant and nonpermeant ions. *BBA Biomembr.* **2009**, *1788* (5), 1188–1192.

(16) (a) Varma, S.; Rempe, S. B. Tuning ion coordination architectures to enable selective partitioning. *Biophys. J.* **2007**, *93* (4), 1093–1099. (b) Thomas, M.; Jayatilaka, D.; Corry, B. The predominant role of coordination number in potassium channel selectivity. *Biophys. J.* **2007**, *93* (8), 2635–2643.

(17) (a) Huetz, P.; Boiteux, C.; Compoint, M.; Ramseyer, C.; Girardet, C. Incidence of partial charges on ion selectivity in potassium channels. *J. Chem. Phys.* **2006**, *124*, 4. (b) Compoint, M.; Ramseyer, C.; Huetz, P. Ab initio investigation of the atomic charges in the KcsA channel selectivity filter. *Chem. Phys. Lett.* **2004**, *397* (4−6), 510–515.

(18) Guidoni, L.; Carloni, P. Potassium permeation through the KcsA channel: a density functional study. *BBA* **2002**, *1563*, 1–6.

(19) Jorgensen, W. L. Special issue on polarization. *J. Chem. Theory Comput.* **2007**, *3* (6), 1877–1877.

(20) Gascon, J. A.; Leung, S. S. F.; Batista, E. R.; Batista, V. S. A self-consistent space-domain decomposition method for QM/MM computations of protein electrostatic potentials. *J. Chem. Theory Comput.* **2006**, *2* (1), 175–186.

(21) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. ONIOM: A multilayered integrated MO+MM method for geometry optimizations and single point energy predictions. A test for Diels-Alder reactions and Pt(P(t-Bu)(3))(2)+H-2 oxidative addition. *J. Phys. Chem.* **1996**, *100* (50), 19357–19363.

(22) (a) Luzhkov, V. B.; Aqvist, J. K+/Na+ selectivity of the KcsA potassium channel from microscopic free energy perturbation calculations. *BBA Protein Struct. Mol. Enzymol.* **2001**, *1548* (2), 194–202. (b) Noskov, S. Y.; Berneche, S.; Roux, B. Control of ion selectivity in potassium channels by electrostatic and dynamic properties of carbonyl ligands. *Nature* **2004**, *431* (7010), 830–834. (c) Bostick, D. L.; Brooks, C. L. Selectivity in K+ channels is due to topological control of the permeant

ion's coordinated state. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (22), 9260–9265 .

(23) Harding, M. M. Metal-ligand geometry relevant to proteins and in proteins: sodium and potassium. *Acta Crystallogr., Sect. D* **2002**, *58*, 872–874.

(24) Varma, S.; Rempe, S. B. Coordination numbers of alkali metal ions in aqueous solutions. *Biophys. Chem.* **2006**, *124* (3), 192–9.

(25) Bucher, D.; Guidoni, L.; Carloni, P.; Rothlisberger, U. Coordination Numbers of K^+ and Na^+ Ions Inside the Selectivity Filter of the KcsA Potassium Channel: Insights from First Principles Molecular Dynamics. *Biophys. J.* **2010**, *98* (10), L47–L49.

(26) Bezanilla, F.; Armstrong, C. M. Negative conductance caused by entry of sodium and cesium ions into potassium channels of squid axons. *J. Gen. Physiol.* **1972**, *60* (5), 588–&.

(27) (a) Furini, S.; Beckstein, O.; Domene, C. Permeation of water through the KcsA K^+ channel. *Proteins* **2009**, *74* (2), 437–448. (b) Domene, C.; Sansom, M. S. P. Potassium channel, ions, and water: Simulation studies based on the high resolution X-ray structure of KcsA. *Biophys. J.* **2003**, *85* (5), 2787–2800.

(28) Lockless, S. W.; Zhou, M.; MacKinnon, R. Structural and thermodynamic properties of selective ion binding in a K+ channel. *Plos Biol.* **2007**, *5* (5), 1079–1088.

(29) Kraszewski, S.; Boiteux, C.; Langner, M.; Ramseyer, C. Insight into the origins of the barrier-less knock-on conduction in the KcsA channel: molecular dynamics simulations and ab initio calculations. *Phys. Chem. Chem. Phys.* **2007**, *9* (10), 1219–1225.

(30) Dudev, T.; Lim, C. Determinants of K+ vs Na+ Selectivity in Potassium Channels. *J. Am. Chem. Soc.* **2009**, *131* (23), 8092–8101.

(31) Illingworth, C. J.; Domene, C. Many-body effects and simulations of potassium channels. *Proc. R. Soc. London, Ser. A* **2009**, *465* (2106), 1701–1716.

(32) Zhou, Y.; Morais-Cabral, J. H.; Kaufman, A.; MacKinnon, R. Chemistry of ion coordination and hydration revealed by a K^+ channel-Fab complex at 2.0 Å resolution. *Nature* **2001**, *414*, 43–48.

(33) Kuo, A.; Gulbis, J. M.; Antcliff, J. F.; Rahman, T.; Lowe, E. D.; Jochen, Z.; Cuthbertson, J.; Ashcroft, F. M.; Ezaki, T.; Doyle, D. A. Crystal structure of the potassium channel KirBac1.1 in the closed state. *Science* **2003**, *300*, 1922–1926.

(34) Berneche, S.; Roux, B. The ionization state and the conformation of Glu-71 in the KcsA K^+ channel. *Biophys. J.* **2002**, *82* (2), 772–780.

(35) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14* (1), 33.

(36) (a) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616. (b) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimisation, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

(37) Jorgensen, W. L.; Chandresekhar, 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.

(38) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593.

(39) Miyamoto, S.; Kollman, P. A. SETTLE - an analytical version of the Shake and Rattle algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13* (8), 952–962.

(40) Tuckerman, M.; Berne, B. J.; Martyna, G. J. Reversible multiple time scale molecular-dynamics. *J. Chem. Phys.* **1992**, *97* (3), 1990–2001.

(41) (a) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant pressure molecular-dynamics algorithms. *J. Chem. Phys.* **1994**, *101* (5), 4177–4189. (b) Feller, S. E.; Zhang, Y. H.; Pastor, R. W.; Brooks, B. R. Constant-pressure molecular dynamics simulation- the langevin piston method. *J. Chem. Phys.* **1995**, *103* (11), 4613–4621.

(42) (a) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802. (b) Kale, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Philips, J.; Shinozaki, A.; Varadarajan, K.; Schuten, K. Molecular dynamics programs design - NAMD2: Greater Scalability for Parallel Molecular Dynamics. *J. Comp. Phys.* **1999**, *151*, 283–312.

(43) Morais-Cabral, J. H.; Zhou, Y. F.; MacKinnon, R. Energetic optimization of ion conduction rate by the K+ selectivity filter. *Nature* **2001**, *414* (6859), 37–42.

(44) (a) Allen, T. W.; Bliznyuk, A.; Rendell, A. P.; Kuyucak, S.; Chung, S. H. The potassium channel: Structure, selectivity and diffusion. *J. Chem. Phys.* **2000**, *112* (18), 8191–8204. (b) Guidoni, L.; Carloni, P. Potassium permeation through the KcsA channel: a density functional study. *BBA Biomembr.* **2002**, *1563* (1−2), 1–6. (c) Domene, C.; Vemparala, S.; Furini, S.; Sharp, K.; Klein, M. L. The role of conformation in ion permeation in a K+ channel. *J. Am. Chem. Soc.* **2008**, *130* (11), 3389–3398.

(45) Furini, S.; Domene, C. Atypical mechanism of conduction in potassium channels. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106* (38), 16074–16077.

(46) Illingworth, C. J. R.; Gooding, S. R.; Winn, P. J.; Jones, G. A.; Ferenczy, G. G.; Reynolds, C. A. Classical polarization in hybrid QM/MM methods. *J. Phys. Chem. A* **2006**, *110* (20), 6487–6497.

(47) Illingworth, C. J. R.; Parkes, K. E.; Snell, C. R.; Marti, S.; Moliner, V.; Reynolds, C. A. The effect of MM polarization on the QM/MM transition state stabilization: application to chorismate mutase. *Mol. Phys.* **2008**, *106* (12−13), 1511–1515.

(48) Illingworth, C. J. R.; Morris, G. M.; Parkes, K. E. B.; Snell, C. R.; Reynolds, C. A. Assessing the Role of Polarization in Docking. *J. Phys. Chem. A* **2008**, *112* (47), 12157–12163.

(49) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, J. 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. J.; 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, 2009.

(50) Stone, A. J. *J. Chem. Theory Comput.* **2005**, *1*, 1128–1132.

(51) Ferenczy, G. G.; Reynolds, C. A. Modeling polarization through induced atomic charges. *J. Phys. Chem. A* **2001**, *105* (51), 11470–11479.

(52) Bliznyuk, A. A.; Rendell, A. P. Electronic effects in biomolecular simulations: Investigation of the KcsA potassium ion channel. *J. Phys. Chem. B* **2004**, *108* (36), 13866–13873.

(53) Marcus, Y. The thermodynamics of solvation of ions. 2 The enthalpy of hydration at 298.15 K. *J. Chem. Soc., Faraday Trans. 1* **1987**, *83*, 339–349.

(54) (a) Berneche, S.; Roux, B. Molecular dynamics of the KcsA K^+ channel in a bilayer membrane. *Biophys. J.* **2000**, *78* (6), 2900–2917. (b) Khalili-Araghi, F.; Tajkhorshid, E.; Schulten, K. Dynamics of K+ ion conduction through Kv1.2. *Biophys. J.* **2006**, *91* (6), L72−4. (c) Miloshevsky, G. V.; Jordan, P. C. Conformational changes in the selectivity filter of the open-state KcsA channel: An energy minimization study. *Biophys. J.* **2008**, *95* (7), 3239–3251. (d) Grottesi, A.; Domene, C.; Sansom, M. Permeation and gating in KirBac: Molecular dynamics simulations. *Biophys. J.* **2004**, *86* (1), 178A–178A.

CT100276C

JCTC Journal of Chemical Theory and Computation

Another Coarse Grain Model for Aqueous Solvation: WAT FOUR?

Leonardo Darré,[†] Matías R. Machado,[†] Pablo D. Dans,[†] Fernando E. Herrera,[†,‡] and Sergio Pantano*,[†]

Institut Pasteur de Montevideo, Calle Mataojo 2020, CP 11400, Montevideo, Uruguay, and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Avda. Rivadavia 1917 - CP C1033AAJ - Cdad. de Buenos Aires, Argentina

Received July 7, 2010

Abstract: Biological processes occur on space and time scales that are often unreachable for fully atomistic simulations. Therefore, simplified or coarse grain (CG) models for the theoretical study of these systems are frequently used. In this context, the accurate description of solvation properties remains an important and challenging field. In the present work, we report a new CG model based on the transient tetrahedral structures observed in pure water. Our representation lumps approximately 11 WATer molecules into FOUR tetrahedrally interconnected beads, hence the name WAT FOUR (WT4). Each bead carries a partial charge allowing the model to explicitly consider long-range electrostatics, generating its own dielectric permittivity and obviating the shortcomings of a uniform dielectric constant. We obtained a good representation of the aqueous environment for most biologically relevant temperature conditions in the range from 278 to 328 K. The model is applied to solvate simple CG electrolytes developed in this work (Na^+, K^+, and Cl^-) and a recently published simplified representation of nucleic acids. In both cases, we obtained a good resemblance of experimental data and atomistic simulations. In particular, the solvation structure around DNA, partial charge neutralization by counterions, preference for sodium over potassium, and ion mediated minor groove narrowing as reported from X-ray crystallography are well reproduced by the present scheme. The set of parameters presented here opens the possibility of reaching the multimicroseconds time scale, including explicit solvation, ionic specificity, and long-range electrostatics, keeping nearly atomistic resolution with significantly reduced computational cost.

Introduction

Computer simulation of biological systems is continuously experiencing a tremendous expansion urged by the ever-growing computer power that allows for the treatment of always more complex systems and for time scales that continuously approach biological relevancy.[1] Parallel to this, the greediness to achieve structural and dynamical descriptions of yet longer and bigger sized systems has prompted the scientific community to develop simplified models of molecular assemblies that mimic arbitrarily intricate molecular systems with a lower degree of complexity. These simplified or coarse grain (CG) representations reduce significantly the computational demands but still capture the physical essence of the phenomena under examination.[2,3] Starting from the pioneering simplified models used to describe protein folding,[4,5] a huge number of successful applications covering a wide range of biomolecular and nanotechnologically relevant applications have been presented.[6−18] For an exhaustive review of this area, the book *Coarse-Graining of Condensed Phase and Biomolecular Systems*[19] is recommended. In this context, the accurate treatment of solvent effects is still a challenging issue. In fact, many CG approaches use a uniform dielectric constant, which may produce an incorrect partition of

* Corresponding author. Tel.: +598-2522 0910. Fax: +598-2522 0910. E-mail: spantano@pasteur.edu.uy.

[†] Institut Pasteur de Montevideo.

[‡] CONICET.

10.1021/ct100379f
Published on Web 11/04/2010

hydrophilic molecules in a hydrophobic medium. Recently, elaborated and/or systematic developments of CG models for simulating water, Hbond (hydrogen bond) bound, and/or ionic liquids with high accuracy have been presented.[20−23] Here, we present a new and simple CG model for water derived from elementary physicochemical concepts and fitting the interaction parameters to reproduce some characteristic features of liquid water. The main advantage of our model is that all of the interactions are described by a typical Hamiltonian for classical simulations, explicitly including long-range electrostatics. This model is composed of four interconnected beads arranged in a tetrahedral conformation (Figure 1). Each bead carries an explicit partial charge. In this way, the liquid generates its own dielectric permittivity, avoiding the use of a constant dielectric medium. The model achieves a reasonable reproduction of some common properties of liquid water in the range of temperatures relevant for most biological applications.

As examples of the potentiality of the model, we study first the solvation of CG monovalent electrolytes developed in this work (Na^+, K^+, and Cl^-). Then, we present molecular dynamics (MD) simulations of a recently published CG model for DNA.[24] This model was shown to provide nearly atomistic resolution information of the structure and dynamics of double-stranded DNA under the generalized Born model approach for implicit solvation. In this contribution, we present an extension of that model for explicit solvation.

We show that this CG scheme is able to reproduce solvation spines, electrolyte specificity, and cation-driven narrowing of the minor groove. These examples illustrate the usefulness of the model in incorporating electrostatic effects in a physiological medium, keeping the chemical details of the different ionic species within CG simulations and overcoming the drawbacks of implicit solvation.

Methods

Description of the Model. The underlying idea of the model is that, due to its molecular characteristics, pure water behaves as a structured liquid forming (among other structural arrangements) transient tetrahedral clusters.[25] These clusters are composed of a central water coordinated by four identical molecules that form an elementary tetrahedral arrangement (Figure 1A). In this arrangement, the central molecule is buried and unable to interact with any other particle outside of the cluster. Our working hypothesis is that, owing to the replication of this structure in the bulk, the central molecule of any tetrahedron can be taken into account implicitly passing from a highly packed (atomistic) to a more granular (CG) liquid (Figure 1B). Aimed at reproducing the structural organization of the liquid, we generated a molecular topology in which four "covalently bound" beads are placed on the geometric positions of four oxygen atoms at the corners of an ideal tetrahedron (hence, the name WAT-FOUR or WT4 for short). The proposed topology implies that within an elementary cluster, the Hbond interactions that hold together the atomistic liquid water are represented by spring constants linking four beads (Figure 1B). The interactions between elementary clusters are taken into account by normal vdW and electrostatic terms in the classical Hamiltonian (Table S1, Supporting Information). These forces reproduce the overall tetrahedral ordering of water, allowing the elementary clusters to diffuse freely.

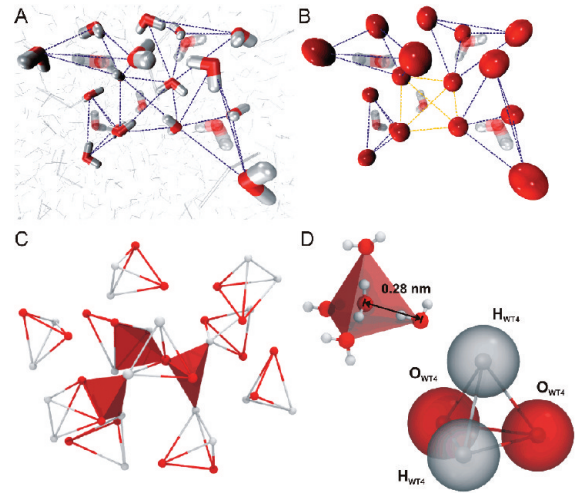


Figure 1. From atomistic to CG water. (A) Snapshot taken from a MD simulation showing the typical ordering of bulk water molecules. Gray molecules represent the liquid bulk. The structural organization is illustrated with a few opaque, thick water molecules which occupy the corners of irregular tetrahedrons. They saturate the Hbond capacity of a (semitransparent) molecule located in the center of each tetrahedron. Hbonds are indicated with dashed lines. (B) The positions of each of the oxygen atoms at the corners of the tetrahedra in A are now indicated with red beads. The concept behind the WAT FOUR (WT4) model is that those elementary tetrahedral clusters can be represented by four harmonically linked beads. The covalent bonds included in the WT4 model are represented by dark dashed lines, while intercluster interactions (vdW and electrostatics) are indicated with light dashed lines. The model implies that a number of water molecules are taken into account implicitly (represented as semitransparent molecules). Notice that water molecules can be implicitly represented even between noncovalently bound beads (take, for example, the central water molecule in the picture). The positions of all of the elements in A and B are identical. (C) Structural organization of WT4 in the bulk solution taken from a MD snapshot. The model reproduces higher-granularity tetrahedral organization in the space through noncovalent interactions. Red planes evidence the presence of rough tetrahedrons formed between different WT4 molecules comprising an implicit water molecule. (D) The ideal organization of a tetrahedral water cluster leads to the geometry of WT4. The separation of 0.28 nm between the water oxygen located at the center of the tetrahedron and its corners corresponds to the oxygen−oxygen (first neighbor) distance. This elementary cluster can be mapped to a WT4 molecule (bottom) composed by four harmonically bonded beads. White and red beads (hydrogen-like, H_{WT4}, and oxygen-like, O_{WT4}) carry positive and negative partial charges of 0.41*e*, respectively.

In analogy with the nearly tetrahedral water molecule that promotes a tetrahedral ordering in the surrounding space, a WT4 molecule recreates a roughly similar arrangement with a higher granularity (Figure 1C). Indeed, the structure of a WT4 molecule is replicated in its neighborhood, leaving holes that can be regarded as atomistic waters implicitly taken

Table 1. Interaction Parameters of the CG Models for Water and Ions[a]

					bond parameters	
	mass (au)	charge (e)	σ[b] (nm)	ε (kJ mol^{-1})	d_{eq} (nm)	K_{bond}[c] (kJ mol^{-1} nm^{-2})
SPC[27]	Ow:16 Hw:1	Ow:−0.82 Hw:+0.41	0.3166	0.650	0.1[d]	172500
TIP3P[28]	Ow:16 Hw:1	Ow:−0.8340 Hw:+0.4170	0.315061	0.6364	0.09572[d]	251208
WT4	O_{WT4}:50 H_{WT4}:50	O_{WT4}:−0.41 H_{WT4}:+0.41	0.42	0.55	0.45[e]	2092
NaW^+	130.99	1	0.58	0.55		
KW^+	147.1	1	0.645	0.55		
ClW^-	143.45	−1	0.68	0.55		

[a] The parameters of two common atomistic water models (SPC and TIP3P) are included for comparison. [b] Distance from the atomic center to the minimum of the vdW function. [c] Corresponds to a harmonic approximation of the form $E_{bond} = K_{bond}(d - d_{eq})^2$. [d] Hydrogen−oxygen distance. [e] Interbead distance.

into account by the CG scheme. These implicit waters can be present not only within the four bonded beads but also between tetrahedrons formed by beads belonging to a different molecule (Figure 1B and C). This suggests that the WT4 molecules in the bulk solution have the capacity to form interactions alike to Hbond networks.

The distance between the central oxygen of a tetrahedral water cluster (Figure 1D) and any other oxygen is ∼0.28 nm, as determined from diffraction experiments.[26] Taking this into account and the geometry of a perfect tetrahedron, the equilibrium distance between beads was set to 0.45 nm. The bond stretching force constant was set to mimic the interaction energy involved in typical hydrogen bonds. We tried harmonic constants within a range from 837 kJ/mol nm^2 to 4184 kJ/mol nm^2 (2 kcal/mol Å^2 to 10 kcal/mol Å^2). A value of 2092 kJ/mol nm^2 (5 kcal/mol Å^2) was chosen, as it results in a better fit of different water properties. This weak link confers the molecule a certain degree of structural plasticity, resulting in small deviations from a perfect tetrahedron upon temperature effects. These deformations could be identified with the nonperfect tetragonal ordering present in liquid water at room temperature. Given the tetrahedral symmetry, only these two bonded parameters for intramolecular interactions are needed (Table 1 and Figure 1D).

Intermolecular nonbonded interactions are ruled by normal van der Waals and electrostatic parameters, listed in Table 1. Partial charges were assigned considering that the central water molecule in a given atomistic tetrahedral cluster neutralizes the atomic charges of the waters in the corners by Hbond formation. If the water atomic charges are q for the hydrogen and $-2q$ for the oxygen, this yields two positive corners with charge q (alike to Hbond acceptors) and two negative corners with charge $-q$ (alike to Hbond donors, Figure 1D). The assignment of partial charges is a largely unsolved issue in classical force fields. In the particular case of water, this task has been addressed in many different ways, from adjusting parameters to reproduce experimental quantities in the liquid or gas phase to ab initio potentials derived from calculations using small clusters of molecules. However, no available model is capable of reproducing all of the water properties with good accuracy. Given the roughness of our model, we just sought to keep the electrostatic interactions engaged by CG beads comparable to atomistic Hbonds. Therefore, we simply tried the same atomic charge values used in common three-point water models (Table 1). Among several atomistic three-point water models tried, the charge distribution that better fit the experimental values was that of the SPC model.[27] The van der Waals radii and well deepness were used as free parameters. Intramolecular nonbonded interactions were excluded.

The mass of each bead was assigned to fit the density of liquid water. To this task, we used a computational box containing 497 WT4 molecules simulated at 300 K and 1 bar. The mass per bead necessary to match a density close to 1 g/mL resulted in 50 au. Taking into account that the mass of each atomistic water molecule is 18 au, it is implied that each WT4 bead represents ∼2.8 water molecules (50 au/18 au). This corresponds on average to about 11 real waters per WT4 molecule. Namely, we assume that each WT4 molecule represents 11 real water molecules in the CG scheme. Therefore, whenever we compare with physicochemical properties, a renormalization factor of 11 is taken into account (see below).

The packing factor of the WT4 spheres calculated as the volume of the cubic box that contains the WT4 molecules divided into the excluded volume of beads is ∼0.47, close to the 0.5 calculated for the SPC model. These values are significantly lower than the ideal 0.74 expected for the hexagonal closest packing (the maximum compaction for rigid spheres). This suggests that the bulk structure of WT4 leaves a number of interstitial cavities in a slightly higher proportion than in the SPC model.

CG Model for the Ions. Three ionic species were developed to represent, at the CG level, the hydrated states of Na^+, K^+, and Cl^- (hereafter called NaW^+, KW^+ and ClW^-, respectively).

Ions were developed considering that six water molecules are always attached to them[29] (i.e., roughly considering an implicit first solvation shell). Therefore, their masses were set as the sum of the ionic mass plus that of six water molecules (Table 1). Partial charges were set to unitary values. The van der Waals radii were adopted to match the first minima of the radial distribution function (RDF, also known as $g(r)$) of hydrated ions as obtained from neutron diffraction experiments.[30] The deepness of the well was set to the same values as the WT4 beads. This was done to ensure compatibility since when a WT4 molecule contacts a CG ion it interacts with its first solvation shell,

Table 2. Description of the Simulated Systems

system		water model	number of molecules	ionic species (number of ions)[a]	solute	temperature (K)	simulation time (ns)	ionic pair concentration (M)
AA[b]	S_1^{AA}	SPC	2483[c]			278−323	45	
AA	S_2^{AA}	SPC	5368[c]	Na^+(1) Cl^-(1)		300	20	0.01
AA	S_3^{AA}	SPC	5368[c]	K^+(1) Cl^-(1)		300	20	0.01
AA	S_4^{AA}	TIP3P	2483[c]			278−323	45	
AA	S_5^{AA}	TIP3P	7612[c]	Na^+(34) Cl^-(34)		300	15	0.5
CG[d]	S_1^{CG}	WT4	497[e]			300	100	
CG	S_2^{CG}	WT4	497[e]			278−328	200	
CG	S_3^{CG}	WT4	268[e]			300	3	
CG	S_4^{CG}	WT4	268[e]			300	3	
CG	S_5^{CG}	WT4	473[e]	NaW^+(1) ClW^-(1)		300	100	0.01
CG	S_6^{CG}	WT4	473[e]	KW^+(1) ClW^-(1)		300	100	0.01
CG	S_7^{CG}	WT4	456[e]	NaW^+(44) ClW^-(44)		300	30	0.5
CG	S_8^{CG}	WT4	456[e]	KW^+(44) ClW^-(44)		300	30	0.5
CG	S_9^{CG}	WT4	174[e]	NaW^+(7) ClW^-(7)		300	200	0.2
CG	S_{10}^{CG}	WT4	174[e]	KW^+(7) ClW^-(7)		300	200	0.2
CG	S_{11}^{CG}	WT4	170[e]	NaW^+(11) ClW^-(11)		300	200	0.3
CG	S_{12}^{CG}	WT4	170[e]	KW^+(11) ClW^-(11)		300	200	0.3
CG	S_{13}^{CG}	WT4	655[e]	NaW^+(34) ClW^-(34)		300	100	0.5
CG	S_{14}^{CG}	WT4	655[e]	KW^+(34) ClW^-(34)		300	100	0.5
CG	S_{15}^{CG}	WT4	506[e]	NaW^+(19) KW^+(19) ClW^-(16)	CG-DNA	300	4000	0.15[f]

[a] Parameters from Berendsen et al.[44] and van Gunsteren et al.[45] In system S_5^{AA}, the CHARMM PARAM27 parameters[46] were used. [b] AA: all atoms. [c] Atomistic water molecules. [d] CG: coarse grain. [e] WT4 molecules. [f] Not considering 22 neutralizing counterions.

which is implicitly considered. A list of nonbonded interaction parameters for the CG monovalent ions is detailed in Table 1.

CG Model for DNA. The CG system used for DNA was essentially the same as that previously presented by us.[24] This CG model reduces the complexity of the atomistic picture to six beads per nucleobase (see Supporting Information Figure S1 for the coarse graining scheme). This mapping keeps the "chemical sense" of specific Watson−Crick recognition allowing the 5′−3′ polarity. Similarly to the approach taken here for water and ions, molecular interactions are evaluated using a classical Hamiltonian. The beads used in this representation carry partial charges, which permits the use of explicit electrostatics

Minor changes have been introduced to the interaction parameters to improve the stability of the double strand using a time step of 20 fs. Back mapping of the atomic coordinates during the trajectory permitted an evaluation of the overall structural quality of the DNA dodecamer in terms of helical parameters (Supporting Information Figure S2). This new parameter set reproduces equally well the structural features of the double-stranded helix.

The complete set of new parameters for DNA is listed in Supporting Information Table S1.

A similarity index between the present implementation and that using the GB model for implicit solvation was calculated from the covariance matrices obtained from the trajectories performed in the present work and that performed in Dans et al.[24] for the Drew−Dickerson dodecamer.

Molecular Dynamics. MD simulations were performed using Gromacs 4.0.5[31−34] in the NPT ensemble unless otherwise stated. The temperature was coupled using the Nose−Hoover thermostat,[35,36] while pressure was kept at 1 bar by means of a Parrinello−Rahman[37,38] barostat, with coupling times of 1 and 5 ps, respectively. A cutoff for nonbonded interactions of 1.2 nm was used, while long-range electrostatics were evaluated using the Particle Mesh Ewald approach.[39,40] A time step of 2 fs was used in all-atom (AA) simulations, while in the CG simulations the time step was set to 20 fs. In order to ensure that the use of such a relatively long integration step does not introduce energy conservation problems, we performed a series of simulations at constant energy (NVE ensemble) using such a time step and varying the cutoff. For an acceptable accuracy in the integration of the equations of motion, one should expect the fluctuations of the total energy to be lower than one-fifth (20%) of the kinetic or potential energy components of the system.[41] According to our results, this criterion is well fulfilled with total energy fluctuations representing 5% of potential or kinetic energy fluctuations, using cutoff values of 1.0, 1.2, and 1.5 nm (Supporting Information Table S2). It was decided to use a cutoff of 1.2 nm, which besides ensuring energy conservation also includes direct nonbonded interactions up to the second neighboring WT4 molecule in solution. Additionally, NVT simulations were performed for some systems in order to compute the WT4 surface tension and the ionic osmotic pressure as detailed below.

All of the interactions (i.e., WT4−WT4, WT4−ion, ion−ion, ion−DNA, WT4−DNA, and DNA−DNA) were straightforwardly calculated within the pairwise Hamiltonian of Gromacs 4.0.5, which is common to many popular MD packages. The van der Waals cross interactions were calculated using the Lorentz−Berthelot combination rules.

Five atomistic (S_{1-5}^{AA}) and 15 CG systems (S_{1-15}^{CG}) were constructed to evaluate different properties of interest (see Table 2). Atomistic simulations were used to obtain reference properties to be compared with the CG models for water and ions. Systems S_1^{AA} and S_4^{AA} were used to compute density and diffusion coefficient profiles in a relevant range of temperatures (see Table 2). The temperature scan was carried out raising the reference temperature by 5° in steps of 5 ns.

Both radial distribution functions (ion−Ow) and electrostatic potential (on the line connecting both ions) were

calculated from systems S_2^{AA} and S_3^{AA}, where the cation−anion distance was kept fixed at 3.6 nm during the whole simulation. This last property was also calculated for system S_1^{AA} at room temperature in order to use it as a reference state for pure water. System S_5^{AA} was used to validate the methodology for measuring the osmotic pressure (described in the Supporting Information).

Regarding the CG simulations, bulk water properties under room conditions were obtained from system S_1^{CG}. The behavior of the model in the range of temperatures from 278 to 328 K was assessed using system S_2^{CG}. The temperature scan was carried out as in the corresponding atomistic simulations (S_1^{AA} and S_4^{AA}) but using time windows of 20 ns instead of 5 ns.

Surface tension and isothermal compressibility at the CG level were computed from systems S_3^{CG} and S_4^{CG}, respectively, according to the following steps, as proposed elsewhere.[42] First, an initial configuration at 300 K and 1 bar (generated by a short NPT equilibration of a simulation box containing 268 WT4 molecules) underwent a 0.1 ns equilibration in the NVT ensemble. The resulting configuration was used, on one hand, to construct system S_3^{CG} by adding vacuum slabs above and below the water bulk, so the box length in the z direction was tripled. A 3 ns production NVT simulation was conducted in such a system at 300 K, from which the surface tension was computed from the pressure tensors:

$$\gamma = \frac{L_z}{2}\left\langle P_{zz} - \left(\frac{P_{xx} + P_{yy}}{2}\right)\right\rangle \quad (1)$$

On the other hand, the NVT equilibrated configuration was also used as the starting structure (system S_4^{CG}) for a 3 ns NPT simulation at 300 K and 1 bar, from which the isothermal compressibility was computed according to[43]

$$\kappa = \frac{\langle V^2\rangle - \langle V\rangle^2}{\langle V\rangle k_B T} \quad (2)$$

Radial distribution functions (CG ion−WT4) and an electrostatic potential profile (obtained in the same way as in the atomistic system) were calculated for systems S_5^{CG} and S_6^{CG} and compared with systems S_2^{AA} and S_3^{AA}, respectively, in order to assess the ability of the CG model to reproduce atomistic results.

Systems S_7^{CG} and S_8^{CG} were used to compute radial distribution functions (CG ion-WT4) using an ionic concentration of roughly 0.5 M, in order to compare them with experimental data.[30]

Bjerrum (λ_B) and Debye (κ^{-1}) lengths were calculated as

$$\lambda_B(\rho) = \frac{(1.0 \times 10^9)\beta\, e^2}{4\pi\varepsilon_0\, \varepsilon_r(\rho)} \quad (3)$$

$$\kappa^{-1}(\rho) = \left(\frac{2(1.0 \times 10^{-15})F^2\rho}{RT\varepsilon_0\, \varepsilon_r(\rho)}\right)^{-1/2} \quad (4)$$

where β is the thermal energy, $\varepsilon_0 = 8.85 \times 10^{-12}$ C^2 J^{-1}m^{-1}, $F = 96485.3399$ C mol^{-1}, and $R = 8.314472$ J mol^{-1} K^{-1}.

The dielectric constants of the electrolyte aqueous solutions (Na^+Cl^- and K^+Cl^-) at different concentrations (0.2 and 0.3 M) were obtained from simulations of systems S_{9-12}^{CG} (see Table 2).

The osmotic pressure measurement was based on the methodology presented by Roux and Luo[47] (systems $S_{13,14}^{CG}$). The idea behind it is to simulate an aqueous solution where the ions are restrained to stay only in one-half of the simulation box and from the force exerted by the restraints, calculate the osmotic pressure. To accomplish this, we used a restraining strategy previously developed in our group called BRIM[48] (see the Supporting Information for a more exhaustive explanation).

Finally, we performed a 4 μs unconstrained simulation (S_{15}^{CG}) of a CG version[24] of a double-stranded DNA using the Drew−Dickerson sequence 5′-d(CGCGAATTCGCG)-3′ in an octahedron box filled with WT4 and CG ions (see Table 2). Global DNA structural behavior, DNA hydration, and specific DNA−water and DNA−ion interactions were evaluated. Helical parameters for DNA were computed using the Curves+ software.[49] The cation-induced narrowing of the minor groove was studied. Such structural changes were estimated from the average interphosphate distance between opposite strands measured for the following pairs: {(*5*, 24), (*6*, 23), (*7*, 22), (*8*, 21), (9, *20*), (10, *19*), (11, *18*), (12, *17*)} (italics indicate the residue numbers at the AT track). Cations were considered to be bound to the minor groove if their distance to the phosphate groups of both opposite strands was below 0.5 nm.

Results

In the following paragraphs, we describe the performance of the WT4 model to reproduce some common parameters of pure water. Comparisons are made, whenever possible, against experimental data. However, some of the calculated properties are also confronted with the results obtained from popular atomistic water models just to provide a reference frame for our results against well established AA models used by the broad scientific community. Subsequently, we analyze the solvation structure of simple electrolyte representations. Finally, to provide an example of application to a biologically relevant system, we briefly present a simulation of a CG DNA double helix in the presence of explicit solvent and mixed salts. A more detailed study on different properties of DNA (flexibility, breathing, DNA−solvent interactions on the multi-microsecond time scale, etc.) will be published elsewhere.

WT4 in the Bulk. A characteristic feature of water is its intrinsic ordering. A good reproduction of the oxygen−oxygen radial distribution function is a common goal for most water models in atomistic detail. The shape of the radial distribution function (RDF) at points far from the first spheres of hydration may furnish an idea of the liquid character of the substance under study. While for a liquid the RDF is expected to converge to a unitary value after a certain point, repetitive behavior is indicative of a crystalline state.

Although the RDF obtained with our model retains some characteristic features of liquid water, comparison of the RDF

obtained for WT4 with other atomistic models reveals some dissimilarities. The most evident difference with respect to the RDF calculated for SPC or TIP3P simulations (systems S_1^{AA} and S_4^{AA}) is the complete lack of the first solvation peak. Owing to the size and topology of the beads, WT4 presents a void space from the center of each bead up to the distance corresponding to the second solvation shell of real water. In this sense, the WT4 representation can be considered a second shell solvation model. In fact, the position of the first maximum in WT4 corresponds to the second peak of atomic water[50] (Figure 2A). It is important to notice that the normalization to the bulk value and the more granular character of the CG model generates a difference in the relative heights of the probability distribution of WT4 with respect to real water. Furthermore, the harmonic bonds existing within the tetrahedron translate into an overestimation of the probability of finding the first neighbor in the WT4 solution. After this global maximum, the relatively large size of WT4 generates some residual ordering that extends up to ~1.2 nm. The radial distribution function converges to one (bulk density) beyond 1.3 nm.

An important property for models of liquid water is their capability to reproduce the correct water diffusion. Clearly, the diffusivity of the WT4 molecules is much lower than that of atomistic water. At 300 K, we obtained a value of 2.03×10^{-6} cm^2 s^{-1}. However, the displacement of a WT4 molecule implicitly represents the movement of the center of mass of ~11 water molecules. Taking into account that the average mean squared displacement of the center of mass of N molecules is N times slower than the average mean squared displacement of N molecules diffusing separately,[53,54] we can conclude that the self-diffusion coefficient for the water molecules represented by the CG model at room temperature is 2.23×10^{-5} cm^2 s^{-1}, which is in good agreement with the experimental value (Table 3).

The WT4 model includes the explicit treatment of the electrostatic interactions as each bead carries a point charge (Table 1). This gives rise to a dielectric permittivity without imposing a continuum dielectric medium. The dielectric permittivity simulated by WT4 is 110.[55] Although this value is nearly 30% higher than that of real water, it must be noticed that this has been a problematic point even for more sophisticated atomistic models of water, and values ranging from 53[56] to 116[57] have been reported.

An important issue regards the long-range ordering of the WT4 molecules. In fact, some CG models for water present a freezing point very close to room temperature.[41] Therefore, we sought to perform a temperature scan over a range from 278 to 328 K. This range of temperatures covers most of the potential and biologically relevant applications of the model. Calculation of the RDF along the studied temperature range suggests that WT4 retains its liquid character, as no significant changes are found between 278 and 328 K (Figure 2A, inset).

The density of the WT4 model was set to match the value of pure water at 300 K. However, a reasonably good reproduction of the variations of the density versus temperature is also desirable. From the qualitative point of view, we obtained the expected reduction of the density with the

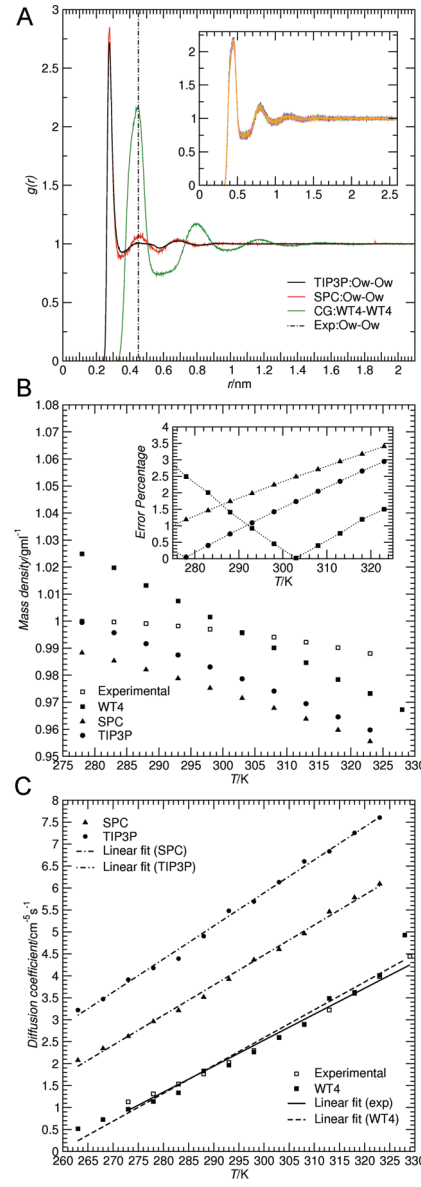


Figure 2. Bulk properties of WT4. (A) RDF calculated over all of the WT4 beads at room temperature from system S_1^{CG} (green line). Comparison is made with the oxygen−oxygen RDF calculated from TIP3P and SPC atomistic simulations as obtained from systems S_1^{AA} and S_4^{AA} at 298 K (black line and red line, respectively). The position of the second solvation peak obtained from experiments[50] is also shown (vertical, dot-dashed line). The inset shows the behavior of the RDF upon temperature variations (system S_2^{CG}) in the range from 278 to 328 K. No significant changes are observed. (B) The variation of the CG water mass density with the temperature (filled squares) calculated from system S_2^{CG} as compared with experimental data (empty squares)[51] and simulations of SPC (triangles) and TIP3P (circles) systems (S_1^{AA} and S_4^{AA}, respectively). The inset shows the relative error of the WT4, SPC, and TIP3P models compared to the experimental data. (C) The dependence of the diffusion coefficient on temperature is compared between the WT4, SPC, and TIP3P models (S_2^{CG} (filled squares), S_1^{AA} (triangles), and S_4^{AA} (circles), respectively) and experimental data[52] (empty squares). All four profiles present an almost linear trend, as revealed by the corresponding linear fits.

Table 3. Bulk Water Properties at Room Conditions for Atomistic Water Three-Point Models (SPC and TIP3P), WT4, and Experimental Data

	dielectric constant	diffusion coefficient (10^{-5} cm^2 s^{-1})	expansion coefficient (10^{-4} K^{-1})	mass density (g mL^{-1})	number density[a] ($\times$ 10^{22} mL^{-1})	surface tension (mN m^{-1})	isothermal compressibility (GPa^{-1})
WT4	110	2.23	11.6	1.0001	0.3	17	2.43
SPC	65[58]	3.85[59]	7.3[60]	0.9705[61]	3.2	53.4[62]	0.53[63]
TIP3P	82[56]	5.19[59]	9.2[64]	1.002[64]	3.4	49.5[62]	0.58[63]
Exp.	78.4[65]	2.27[66]	2.53[51]	0.9970[51]	3.3	71.2[67]	0.46[68]

[a] Calculated from the corresponding mass density, considering the molar mass of water (18 g mol^{-1}) and WT4 (200 g mol^{-1}). Accordingly, the number density for the atomistic models and real water corresponds to the number of water molecules per milliliter, while for WT4 it corresponds to the number of WT4 molecules (~11 water molecules) per milliliter.

temperature and an almost perfectly linear behavior of the system's density against temperature in the explored range (Figure 2B). Although the functional dependence of real water against temperature is certainly not linear, it is a good approximation within the temperature range chosen. In fact, the relative error of the WT4 density with respect to that of the real water in this temperature window remains always below 3%, with the higher deviations near the critical point of real water (Figure 2B). This behavior is comparable with those of the SPC and TIP3P atomistic models (Figure 2B).

Following the volume changes upon thermal variations at constant pressure allows also for the calculation of the isobaric expansion coefficient of our model. We obtain an overestimation of this quantity at 298 K (Table 3). The expansion coefficient of WT4 gives a value of 11.6×10^{-4} K^{-1} as compared with the experimental value of 2.53×10^{-4} K^{-1}.[51] Although overestimated, it is comparable with the values reported for widely used three-point water models (Table 3).

Another frequently calculated property for CG models is the surface tension. In our case, we obtained a value of 17 mN m^{-1}, which is nearly 4 times smaller than the experimental value. Similarly, we found a 5 fold higher isothermal compressibility as compared with the experimental value (Table 3). These discrepancies are very frequently found in CG models that lump a number of water molecules into one single entity.[42] The origin of this effect may be the loss of fully atomic interactions that decrease the cohesive forces and increase the granularity of the system.

A more stringent test for our representation comes from the calculation of the diffusion coefficient. Clearly, a rise in the diffusion must occur upon heating. Experimental data indicates that pure water experiences a nearly linear increase in the diffusion coefficient between 278 and 328 K. The model shows the correct dependence of the diffusion coefficient on temperature. Indeed, it shows good agreement with the experimental behavior within the explored range (Figure 2C).

Taking into account the above results, the range of validity of the model may be delimited by the following considerations: the lower limit should not go below 278 K. Applications at lower temperatures are strongly discouraged since ice formation implies quantum effects that can, obviously, not be achieved by simplified models. On the upper limit, the relative error in the renormalized diffusion coefficient arrives at ~11% at 328 K, suggesting that simulations at higher temperatures could require some reparameterization to keep the accuracy at acceptable levels.

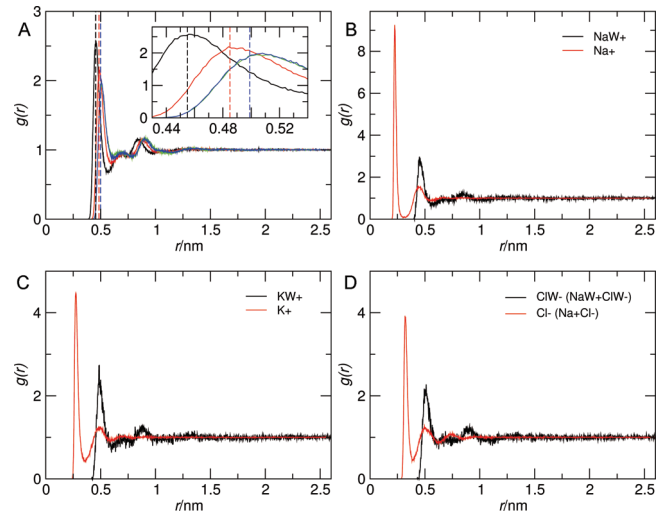


Figure 3. Ionic solvation. (A) The RDF of WT4 around CG electrolytes computed for systems S_7^{CG} and S_8^{CG} (NaW, black; KW, red; ClW, blue for NaW^+ClW^- and green for KW^+ClW^-). Vertical dashed lines indicate the position of the second solvation peak as determined from neutron scattering experiments.[30] The inset shows a closeup on the region between 0.43 and 0.54 nm allowing for a more precise comparison. (B, C, and D) Comparison between RDFs obtained from atomistic and CG simulations (systems S_2^{AA}, S_3^{AA}, S_5^{CG}, and S_6^{CG}). The plot corresponding to the solvation structure around chlorine ions in the presence of potassium is similar to that shown for the case of sodium. It is omitted for brevity.

Ionic Solvation. The characteristics of the WT4 model open the possibility to study the solvation properties of systems in which electrostatics are dominant. In this context, we developed the CG parameters of three simple electrolytes: Na^+, K^+, and Cl^-. Since we can imagine WT4 as a second solvation shell model, we represent ions together with their first sphere of hydration. Aimed at exploring the solvation structure generated by the WT4 model on the CG ions, simulations were conducted at roughly similar ionic concentrations to those reported in neutron diffraction experiments.[30] As depicted in Figure 3A, there is good correspondence, especially for the cations, between the first solvation maximum found for WT4 and the second hydration shell estimated from the experimental data.[30] A second solvation peak is found at nearly 0.9 nm, which has to be considered mainly as an artifact of the geometry of the model since the beads located in the last peak are harmonically linked to those

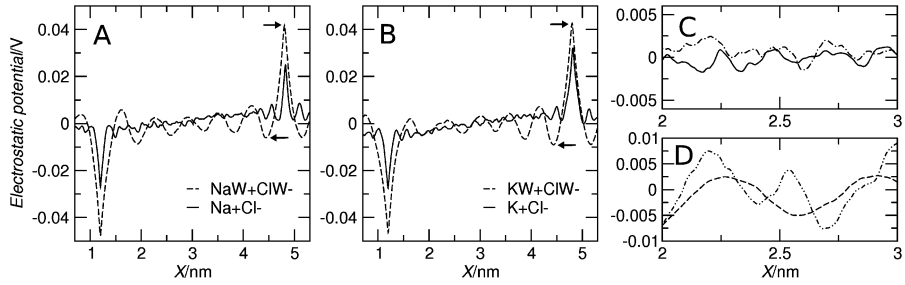


Figure 4. Profiles of electrostatic potential. (A) Electrostatic potential calculated along the line connecting the ionic pairs Na^+Cl^- (filled line, S_2^{AA}) and NaW^+ClW^- (dashed line, S_5^{CG}). Arrows indicate the points where the differences in the electrostatic potential were calculated. (B) Same as A for systems K^+Cl^- (S_3^{AA}) and KW^+ClW^- (S_6^{CG}). (C) Comparison of the electrostatic potential between the central portion of panel A (filled line, S_2^{AA}) against the analogous quantity calculated along a box containing pure SPC water (dot-dashed line, S_1^{AA}). (D) Same as C but for the CG systems (dashed line, S_5^{CG}, and double-dot-dashed line, S_1^{CG}).

of the first. After that point, the RDF converges to the bulk value in all cases.

Unfortunately, experimental data for ionic solvation is only available at high electrolyte concentrations. To explore lower (and more physiological) concentrations for which no experimental data are available, we tried a comparison with atomistic simulations confronting systems S_2^{AA} and S_3^{AA} with systems S_5^{CG} and S_6^{CG}, respectively, both having an ionic concentration of 0.01 M (Figures 3B, C, and D).

In close analogy with the case of pure WT4, the RDF of WT4 around CG ions shows a complete lack of the first solvation shell. A good reproduction of the position of the second solvation peak is observed, confirming the behavior of WT4 as a second solvation shell solvent. As expected, WT4 is not able to reproduce the third solvation shell. The relevance of this inaccuracy is, however, uncertain and could only be relevant in the case of chlorine ions, where such a shell is slightly more pronounced.[30]

Electrostatic Potential. Having analyzed the hydration structure of simple electrolytes, we turned our attention to the profiles of electrostatic potential and the screening properties. This was done by comparing the results of systems S_2^{AA} and S_3^{AA} against those of S_5^{CG} and S_6^{CG}, respectively. These systems consist of an ionic pair of Na^+Cl^- (or K^+Cl^-) kept at a fixed position during the simulation. The separation between both ions was 3.6 nm. Atomistic ionic pairs were immersed in a computational box containing an equivalent number of water molecules. This setup allowed us to compare under similar conditions atomistic and CG simulations as well as the behavior of the different ionic species. In order to assign the proper weight to the perturbations introduced by the electrolytes, we also made comparisons with the fluctuations produced by pure solvent (atomistic and CG) in the profiles of the electrostatic potential. In this way, it is possible to separate the observed features into two components: intrinsic bulk fluctuations and ionic perturbations. Furthermore, this approach gives an idea about the relaxation of the ionic potential at increasing distances from the ion and compares it with pure water and electrolyte solution.

A comparative view of the atomistic versus CG simulation can be acquired from Figure 4A. The first noticeable difference regards the height of the peaks centered on the positions of the ions. Owing to its smaller size, the SPC waters can get closer to the atomistic ion generating a more pronounced electrostatic screening. In the CG counterpart, the corresponding first solvation shell, which is implicit in the NaW^+ and ClW^- ions, only serves to create a void space without screening properties. This translates to a higher electrostatic potential induced by the CG ion. The implicit consideration of the first solvation shell in the CG ions implies that the first minimum observed in the atomistic system is absent in the CG system (Figure 4A). Furthermore, the position of the first minimum observed in the CG simulation (second solvation shell) roughly corresponds to the position of the second minimum around the ion in the atomistic system. Clearly, this effect derives from the solvation structure around the electrolytes; i.e., the first and second minima around the position of the ions (both, Na^+ and Cl^-) shown in Figure 4A correspond to the position of the oxygen atom in the first and second solvation shells shown in Figure 3B and D. Similar features are observed for the cases of K^+Cl^- and KW^+ClW^- ionic pairs (Figure 4B).

The distinctive characteristics of both cations evidenced by the solvent organization around NaW^+ and KW^+ (Figure 3A) can also be obtained from the calculation of the difference in electrostatic potential measured at the position of the cation with respect to that of its first minima (Figure 4A,B). This difference was about 10% higher for the case of KW^+ with respect to NaW^+, in qualitatively good agreement with the ~25% obtained from the atomistic case. This behavior may reflect the fact that water around potassium is bound in a more disorderly fashion than around sodium,[29] probably generating a less marked electrostatic screening in the case of potassium.

As seen from Figure 4A and B, the CG scheme presents higher fluctuations in the potential than the atomistic system. Aimed at excluding the possibility of a spurious ordering of WT4 molecules around the electrolytes, we compared the perturbations in the electrostatic potential introduced by the ions against those observed for pure solvent (both, atomistic and CG). This was assessed by computing the electrostatic potential along an arbitrary axis in two simulation boxes containing pure SPC and WT4 (systems S_1^{AA} and S_1^{CG},

Table 4. Thermodynamic Properties of Electrolyte Solutions

		Bjerrum length (nm)		Debye length (nm)		osmotic pressure[a] (bar)
ρ (molarity)		0.2 M	0.3 M	0.2 M	0.3 M	0.5 M
NaW^+ClW^-/WT4		0.57	0.61	0.76	0.6	35 (s.d. 15)
KW^+ClW^-/WT4		0.55	0.61	0.78	0.6	33 (s.d. 16)
Exp.[b]	NaCl	0.75	0.77	0.66	0.53	~25 (taken from Roux and Luo[47])
	KCl	0.74	0.76	0.67	0.54	

[a] The value obtained in the atomistic simulations using CHARMM PARAM 27 was 37 (s.d. 9) bar. [b] The function $\varepsilon_r(\rho) = \varepsilon(0)/(1 + A\rho)$ (NaCl, $A = 0.27$; KCl, $A = 0.24$), which results from fitting to experimentally obtained dielectric constants,[69] was used to estimate $\varepsilon_r(\rho)$ at the desired concentration, which is necessary for the computation of both Bjerrum and Debye lengths.

respectively). Superposition of both profiles (Figure 4C and D) suggests that both pure water systems show important fluctuations in the electrostatic potential of nearly the same magnitude as those observed in the region between the ions in the ionic solution. This indicates that the perturbations observed in those regions are not an effect induced by the ions but correspond to variations in the electrostatic potential, which are intrinsic to the pure solution. According to this, the difference in the amplitude of the fluctuations observed between the atomistic and CG models (Figure 4A and B) are explained by the augmented granularity of the CG model. An estimation of such a difference is obtained from the approximate amplitudes observed in both atomistic (~0.002 V) and CG (~0.018 V) simulations. This indicates that the oscillations in the CG system have amplitudes nearly 1 order of magnitude higher than the ones in the atomistic system.

Bulk Electrolytic Properties. The vast majority of empirical parametrizations for single ions are typically developed to fit single ion properties, such as those examined in the previous sections. In order to complement the structural description of the CG aqueous solutions we studied some thermodynamic properties regarding ion−ion interactions: in particular, the Bjerrum and Debye lengths. The first represents the separation between two elementary charges at which the electrostatic interaction is comparable in magnitude to the thermal energy, and the second provides information regarding the distance at which the electrostatic potential of one ion is screened by the ionic strength of the surrounding medium. From the qualitative point of view, we retrieved the correct tendency in Bjerrum and Debye lengths upon changes in the ionic concentration (Table 4). Calculation of the Bjerrum and Debye lengths at 0.2 and 0.3 M gave values within a maximum error of 13% with respect to experimental values (Table 4). We obtained an underestimation of the Bjerrum length and, correspondingly, an overestimation of the Debye length, which is indicative of a slightly higher global electrostatic screening in the bulk solution, independent of the salt considered in the simulation (i.e., NaW^+ClW^- or KW^+ClW^-).

A direct measurement of the strength of the effective solvent-mediated interaction between ions is also very relevant, and it can be obtained from the osmotic pressure. For the case of NaW^+ClW^- at an ionic concentration of 0.5 M, we obtained a value of 35 bar (33 bar for KW^+ClW^-), which is essentially identical to that obtained by atomistic simulations using the CHARMM force field. Despite the large standard deviations, these values are in agreement with experimental reports (Table 4), suggesting a satisfactory balance in ion−ion and ion−WT4 interactions.

CG Solvation of Double-Stranded DNA. As a final example of application, we analyzed the explicit solvation of a dodecameric segment of double-stranded DNA. For this task, we used the already published CG scheme for simulating nucleic acids within the framework of the generalized Born model for implicit solvation.[24] In this contribution, the same system was simulated in the presence of explicit solvent and added salts. Both approaches furnish a similar picture of the structural and dynamical behavior of the double-helical segment of DNA with a maximum pairwise RMSD between both average structures of 0.25 nm. This is in agreement with the good reproduction of helical parameters obtained upon backmapping from the DNA simulation in explicit CG solvent (Figure S2, Supporting Information). Furthermore, the superposition of the covariance matrices calculated along the MD trajectories of the Drew−Dickerson dodecamer performed using implicit and explicit solvation gives an identity of 84%. This strongly suggests that both approaches sample nearly equivalent conformational spaces.

During the dynamics in the presence of explicit CG solvent, the global structure of the DNA dodecamer was fairly well conserved with an average RMSD of 0.25 nm from the starting (canonical) conformer. This can be inferred from the good superposition of snapshots taken at different times of the simulation (Figure 5A). Moreover, a good agreement is also obtained at the atomistic level upon backmapping. The all atoms RMSD of those shapshots compared with the X-ray structure 1BNA resulted in values of 0.35 nm (blue), 0.39 nm (green), and 0.34 nm (orange) (Figure S3, Supporting Information).

The WT4 molecules and cations closely interact with the CG nucleobases. It can be observed that the ordering of the WT4 molecules around the DNA qualitatively resembles the hydration features encountered in atomistic systems at both experimental and theoretical levels.[70−75] Conical arrangements of WT4 beads form around the phosphate groups (Figure 5B). The molecules of WT4 acquire an orientation guided by the electrostatic attraction between the positive (hydrogen-like) beads and the negatively charged phosphate superatoms. This results in the formation of structures alike to hydration cones (Figure 5B). In this kind of solvent arrangement around the backbone, WT4 molecules can be replaced by cations from the solution (Figure 5B) as observed experimentally.[76] Furthermore, ions can also remain transiently bound to the DNA visiting different positions within the minor groove. Extended hydration of the major groove and the formation of hydration spines in the minor groove are also observed, as illustrated in Figure 5C. A comprehensive picture of the hydration structure can be obtained from the WT4 beads' occupancy density map projected in the plane perpendicular to the DNA axis placed at the AT step (Figure 5D).

A more quantitative view of the solute/solvent interaction can be obtained from the cumulative RDF of the different species around the phosphate superatoms (Figure 5E). The

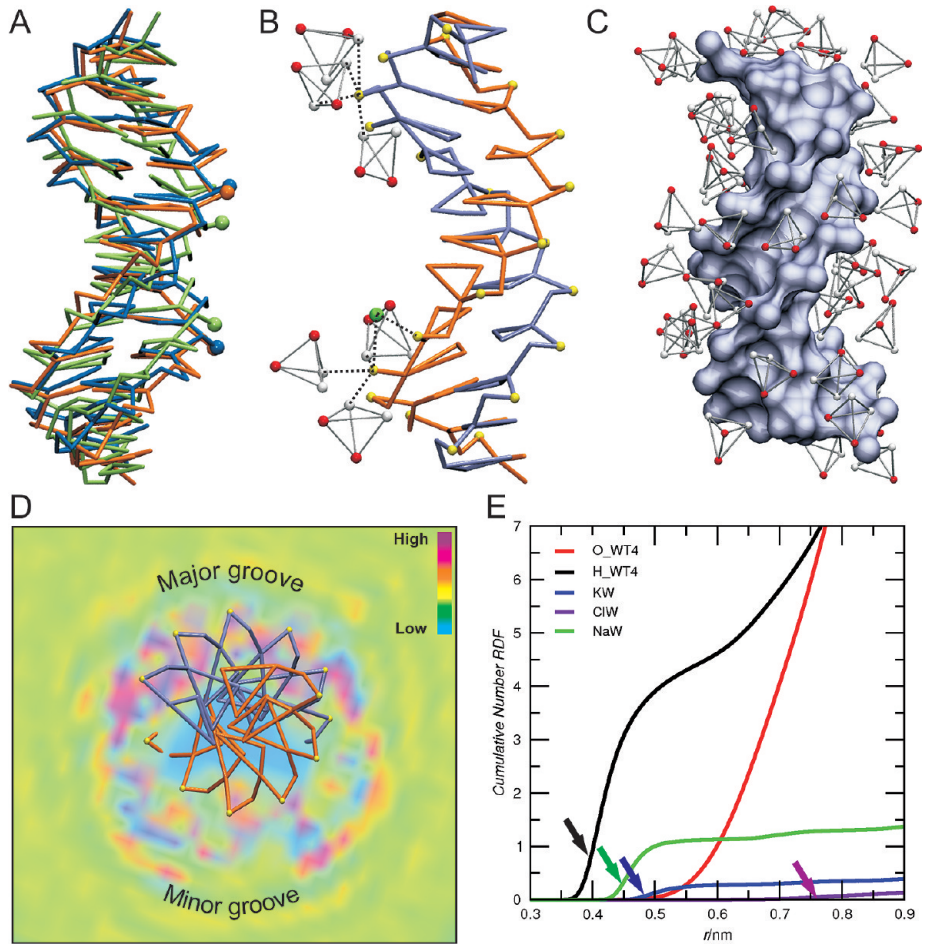


Figure 5. DNA and solvation structure. (A) Superposition of the DNA conformers taken from the first (blue), middle (green), and last (orange) frame of the simulation of system S_{15}^{CG}. Spheres indicate a pair of phosphate superatoms from opposite strands, which are highlighted in order to show the minor groove narrowing. An atomistic view of this superposition obtained from backmapped CG coordinates can be seen in Figure S6, Supporting Information. (B) WT4 and NaW specific interaction with the phosphate groups taken from a representative MD snapshot. The dashed lines highlight the conical arrangement of WT4 beads around phosphates (top) and the competition for the phosphate groups by WT4 and NaW (bottom). (C) WT4 solvation in the major and minor grooves from a random frame. The extensive hydration of the major groove and spines of hydration within the minor groove are evident. (D) WT4 occupancy density map projected onto a plane orthogonal to the DNA axis, located in the central AT step. The color scale represents the occupancy level, with a color range from cyan (low occupancy) to purple (high occupancy). Differences in major and minor groove are plain. Notice also the more punctuated location of WT4 within the minor groove indicative of solvation spines. (E) Cumulative number (integral of the RDFs) of negative and positive WT4 beads (red and black, respectively), NaW^+ (green), KW^+ (blue), and ClW^- (violet), with respect to the phosphate groups. Arrows indicate inflection points, which correspond to the first maxima of each RDF.

directionality in the WT4− phosphate interaction is evident from the right shift observed in the integral of the RDF corresponding to the oxygen-like beads' position with respect to that of the hydrogen-like beads (compare red and black lines in Figure 5E). The position of the first WT4 solvation shell forming conical structures lies at 0.4 nm from the phosphate superatom. This distance is in good agreement with the 0.38 nm found in atomistic simulations[24] and is comparable to the minimum distance of 0.32 observed in X-ray structures.[77]

Our model can also take into account the specificity in the DNA−cation interactions. As expected, sodium ions are more prone than potassium to interact with the solute. As mentioned above, sodium is frequently found in the close neighborhood of the phosphate moieties and even within the minor groove.[76,78,79] The closest sodium shell is localized at 0.45 nm from the phosphate, as compared with the 0.5 nm found for the bulkier potassium. In contrast, the radial distribution of the chlorine ions is much more shifted to the right with a first peak at 0.76 nm (Figure 5E).

The fraction of DNA charge neutralized within a cylinder of 0.9 nm from the exterior surface of the double-stranded helix is 0.75. This is in good agreement with the fraction of condensed counterions calculated within the condensation volume using Manning's counterion condensation theory for polyelectrolytes.[80] Moreover, this number is comparable with a fraction of 0.76 obtained by previous atomistic simulations using the same DNA sequence.[81] Among the fraction of condensed counterions, 76% corresponds to sodium and 24% to potassium; this is in qualitative agreement with a series of experimental and theoretical studies (see Savelyev and Papoian[82] and references therein).

While the global distribution of cations around the DNA contributes to the stability of the double helix, the specific

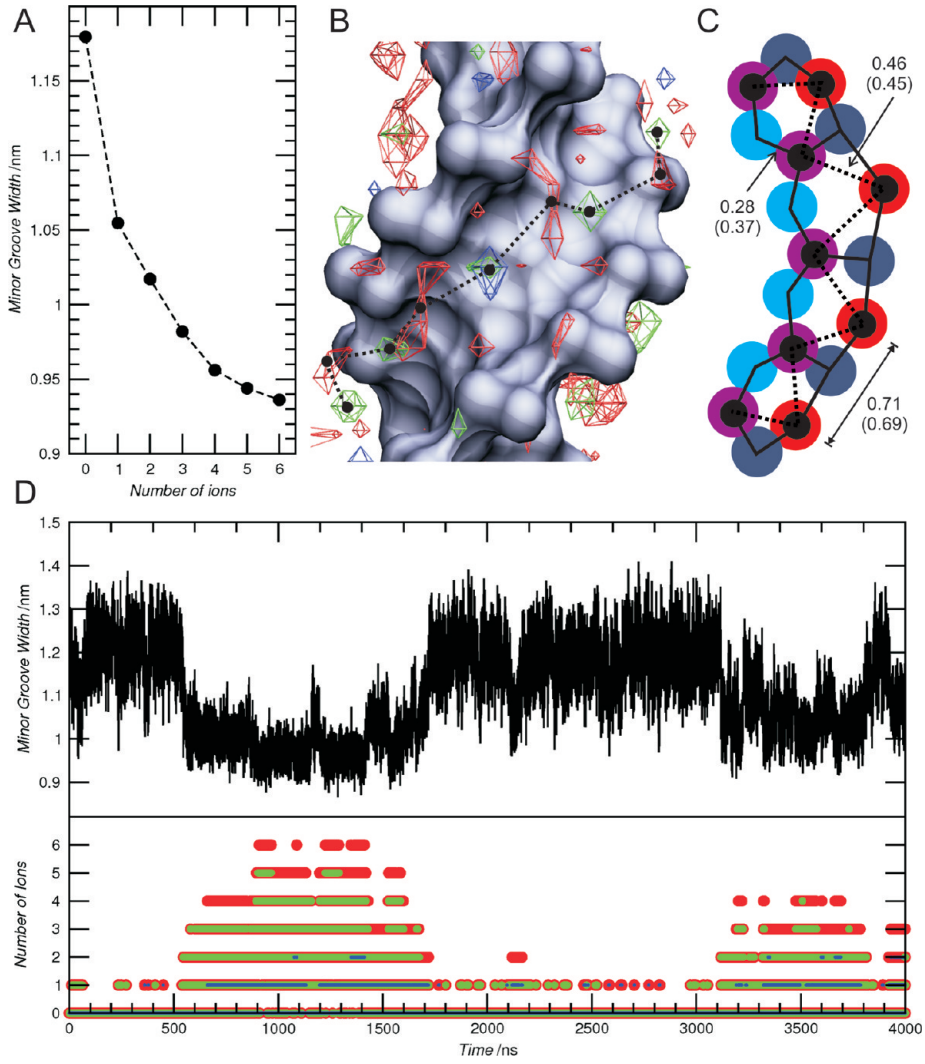


Figure 6. Binding of cations within the minor groove. (A) The minor groove width averaged over all frames with an equal number of bound cations is plotted against the number of bound ions (according to the criteria explained in the Methods section). (B) WT4 (red), NaW^+ (green), and KW^+ (blue) occupancy isosurfaces located in the minor groove of the AT track. Dashed path connecting black points indicate the zig-zag motif formed by the cations and WT4 beads in the minor groove. (C) Scheme showing the superimposition of the zig-zag motif (black circles connected by dashed line) observed in the CG simulation over the fused hexagon motif (continuous line) formed by the solvent sites (cyan, violet, gray, and orange circles) experimentally observed. These sites can by occupied by both water or cations.[77] The distances between corresponding solvation sites in the fused hexagon motif are shown and compared to the corresponding ones in the zig-zag motif (parentheses). (D) Minor groove width (top) and number of bound cations plotted against time (bottom). The number of cations is shown as the number of NaW^+ (green), number of KW^+ (blue), and total number of cations (sum of the number of NaW^+ and number of KW^+, in red).

interaction of cations with DNA has been related to local structural distortions. In particular, the binding of sodium ions within the minor groove has been proposed to mediate a narrowing in the minor groove.[83] In agreement with this proposal, we observed a clear correlation between the width of the minor groove and the binding of cations. Moreover, there seems to be a cumulative effect between these two events; i.e, a higher number of bound ions induces a more pronounced narrowing. This is clear from a measure of the average width of the minor groove with respect to the total number of bound ions (Figure 6A). The binding of one single ion is enough to induce a sensible change in the minor groove. Upon the successive incorporation of ions, the narrowing becomes more marked, reaching a minimum when six ions are concomitantly bound. Experimental studies on the same dodecamer also reveal a high occupancy of cations in the minor groove, leading to its narrowing.[78] Furthermore, a highly ordered structure is formed when cations and water interact with the AT track of the DNA. Such a structure is organized in four layers of solvent sites and resembles a series of fused hexagonal motifs.[78] Figure 6B shows the 3D occupancy map of WT4 and cations around the minor groove of the AT track. This map reveals sites highly occupied by WT4 (red wire mesh), NaW^+ (green wire mesh), and KW^+ (blue wire mesh) that resembles a zig-zag structure (dashed path connecting black points in Figure 6B). When such a zig-zag structure is superimposed onto the experimentally observed fused hexagon motif, good agreement is obtained for the second and fourth solvent-site layers, as confirmed by the inter solvent-site distances (Figure 6C).

The minor groove narrowing process appears to take place on two different time scales. The first is related to the binding

of one or two ions for up to a few dozen nanoseconds, while the second corresponds to the simultaneous binding of three to six ions for a period of nearly 1 μs (Figure 6D). This last induces a more marked and persistent but always reversible structural distortion with an average minor groove width of 0.98 nm (three bound cations) to 0.94 nm (six bound cations). The magnitude of this DNA distortion is in very good agreement with the average value of 0.96 nm obtained experimentally.[79]

It is worth noticing that temporal scales for sodium binding are coincident with the faster events found in this study have been reported for MD simulations at the atomistic level.[83−85] Unfortunately, the longest atomistic simulation reported in this system was carried out for 1.2 μs.[83] Although only nanosecond binding events were reported in that work, the agreement of the position of the binding sites and DNA distortion with X-ray data[78,79] may allow for speculation that a lack of longer binding events in the atomistic simulation could be related to insufficient sampling. Clearly, longer simulation times that go beyond the introductory scope of this paper would be needed to properly sample these long lasting events. This issue will be addressed in a forthcoming publication.

There is a marked selectivity for sodium against potassium. Indeed, while the simultaneous binding of more than two sodium ions is very frequent, only two potassium ions were present within the minor groove simultaneously, and this rather rare event was detected only five times in the 4 μs trajectory (Figure 6D and Figure S4, Supporting Information).

Finally, to complete the picture regarding the ionic structure around DNA, we analyzed the ionic distribution at longer distances from the double helix. This was done by calculating the number density of the three types of ions present in the system at increasing distances from DNA. In good agreement with the prediction from Poisson−Boltzmann theory,[86] the amount of electrolytes along a direction perpendicular to the DNA principal axis follows an exponential decay (Figure S5, Supporting Information).

Discussion and Conclusions

In this work, we have presented a model for simulating water at a coarse grain level. The WT4 model presented here is based on the transient tetrahedral structure adopted by water molecules in solution, preserving the molecular characteristics of the atomistic liquid. Due to the large number and heterogeneity of the CG models proposed in the literature, it is difficult to establish a fair comparison in terms of a computational speedup obtained with WT4. However, a comparison is more straightforward if we restrict it to the simplest models that condense three or four water molecules into a single bead.[53,54,87] This implies a coarse graining factor from 9 to 12, as compared to the value of ∼8 obtained for WT4. In addition to a similar coarse graining factor, our model offers some advantages, like the capacity to interact via explicit short- and long-range electrostatic interactions, and a dielectric permittivity. This grants the model the ability to reproduce some of the characteristic properties of water and electrolytic solutions.

The bead's masses were assigned to fit the water density at 300 K. Although this may raise some concerns about the suitability of the model at different temperatures, the relative error for the WT4 density with respect to the experimental determination of pure water remains below 3% in the range of 278 to 328 K (Figure 2B).

A strong assumption of the model is the fact that the existences of these five-member water clusters are supposed to be permanent, while their average lifetime in real water is on the order of picoseconds. This defect is partially compensated by setting a loose harmonic constraint between the beads of our representation. This allows for bond stretching variations of about 10% in the bond lengths, conferring a large plasticity to the WT4 molecules and the possibility of adapting their conformation according to its molecular environment.

The use of the WT4 model to solvate simple ions reproduces their hydration structure and some thermodynamic properties such as osmotic pressure, which is often considered a quality gauge of the parametrization.

We notice that important properties such as the isothermal compressibility and surface tension are poorly described by WT4. This may be of particular relevance in the study of self-assembly phenomena, and for such treatment special caution is advised. Despite this deviation from ideal behavior, the description of the double-stranded DNA segment does not seem to be compromised. This suggests that the long−medium range screening properties of the solvation model are suitable for overcoming the strong electrostatic repulsion generated by the negatively charged phosphate groups of DNA. In fact, the addition of explicit solvation and different ionic species highly enhanced the description of the DNA dynamics, allowing, for instance, the reproduction of the cation-mediated narrowing of the minor groove that could not be studied within the implicit solvation approach. While the implicit solvation approach can provide a good and faster description of sequence dependent effects on the structural and dynamical stability, inclusion of the explicit solvent can allow for the study, for instance, of the influence of intrinsic versus extrinsic sources of DNA flexibility, solvent mediated effects, ionic specificity, etc. Furthermore, the use of periodic boundary conditions and explicit electrostatics permits a more realistic consideration of long-range effects.

WT4 together with the CG electrolyte model represent correctly the gross solvation structure around DNA, as noted by the percentage of DNA charge neutralized at 0.9 nm that closely resembles that of atomistic simulations and that predicted by counterion condensation theory. Moreover, DNA hydration features like the extensive major groove hydration, minor groove hydration spines, and conical arrangement around phosphate groups that resembles the hydration cones observed in atomistic simulations and experimental data are well reproduced. It is important to note in this context that the development of interaction parameters has always been carried out within the philosophy of fitting structural properties of water, ionic solvation, and DNA. In this respect, we first developed the representation for WT4 in the bulk and then added the description of simple CG

electrolytes. Finally, the existing DNA parameters for implicit solvation were slightly modified to further refine its structural description when embedded in explicit solvent. In this sense, good agreement with experimental determinations can be considered emerging properties of the model because no specific fittings of cross interaction potentials have been performed.

The simulation scheme presented here allows for runing at a rate of ~1 μs per day (S_{15}^{CG}) on a dual quad core PC (Intel Xeon 2.66 GHz). This performance along with the nearly atomic resolution achievable upon backmapping of the coordinates in our DNA model[24] make the millisecond time scale reachable. This would effectively bridge the gap between the time scales feasible to MD and those that are biologically relevant.

Finally, we would like to stress the point that the model presented here computes all of the interactions using a typical Hamiltonian function, avoiding *ad hoc* code modifications/recompilations. Topologies, interaction parameters, and coordinate files for GROMACS implementation are available from the authors upon request.

Acknowledgment. This work was supported by ANII—Agencia Nacional de Investigación e Innovación, Programa de Apoyo Sectorial a la Estrategia Nacional de Innovación—INNOVA URUGUAY (Agreement n8 DCI - ALA/2007/19.040 between Uruguay and the European Commission) and Grant FCE_60-2007. L.D. and M.R.M. are beneficiaries of the National Fellowship System of ANII.

Supporting Information Available: Mapping scheme between atomistic and CG model for DNA. Helical parameters of the backmapped trajectory. Table containing interaction parameters sets for DNA. Energy fluctuations analysis. Cations in the minor groove frequency. Supperposition of backmapped DNA structures. Details on osmotic pressure calculation. Long range ionic structure. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

(2) Klein, M. L.; Shinoda, W. Large-scale molecular dynamics simulations of self-assembling systems. *Science* **2008**, *321*, 798–800.

(3) Murtola, T.; Bunker, A.; Vattulainen, I.; Deserno, M.; Karttunen, M. Multiscale modeling of emergent materials: biological and soft matter. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1869–1892.

(4) Levitt, M.; Warshel, A. Computer simulation of protein folding. *Nature* **1975**, *253*, 694–698.

(5) Tanaka, S.; Scheraga, H. A. Statistical Mechanical Treatment of Protein Conformation. I. Conformational Properties of Amino Acids in Proteins. *Macromolecules* **1976**, *9*, 142–159.

(6) Yin, Y.; Arkhipov, A.; Schulten, K. Simulations of membrane tubulation by lattices of amphiphysin N-BAR domains. *Structure* **2009**, *17*, 882–892.

(7) Arkhipov, A.; Yin, Y.; Schulten, K. Four-scale description of membrane sculpting by BAR domains. *Biophys. J.* **2008**, *95*, 2806–2821.

(8) Wee, C. L.; Gavaghan, D.; Sansom, M. S. P. Interactions between a voltage sensor and a toxin via multiscale simulations. *Biophys. J.* **2010**, *98*, 1558–1565.

(9) Ayton, G. S.; Voth, G. A. Hybrid coarse-graining approach for lipid bilayers at large length and time scales. *J. Phys. Chem. B* **2009**, *113*, 4413–4424.

(10) Treptow, W.; Marrink, S.; Tarek, M. Gating motions in voltage-gated potassium channels revealed by coarse-grained molecular dynamics simulations. *J. Phys. Chem. B* **2008**, *112*, 3277–3282.

(11) Yefimov, S.; van der Giessen, E.; Onck, P. R.; Marrink, S. J. Mechanosensitive membrane channels in action. *Biophys. J.* **2008**, *94*, 2994–3002.

(12) Periole, X.; Huber, T.; Marrink, S.; Sakmar, T. P. G protein-coupled receptors self-assemble in dynamics simulations of model bilayers. *J. Am. Chem. Soc.* **2007**, *129*, 10126–10132.

(13) Durrieu, M.; Bond, P. J.; Sansom, M. S. P.; Lavery, R.; Baaden, M. Coarse-grain simulations of the r-snare fusion protein in its membrane environment detect long-lived conformational sub-states. *Chem. Phys. Chem.* **2009**, *10*, 1548–1552.

(14) Srinivas, G.; Discher, D.; Klein, M. Self-assembly and properties of diblock copolymers by coarse-grain molecular dynamics. *Nature* **2004**, *3*, 638–644.

(15) Nielsen, S.; Lopez, C.; Srinivas, G.; Klein, M. Coarse grain models and the computer simulation of soft materials. *J. Phys.: Condens. Matter* **2004**, *16*, R481–R512.

(16) Arkhipov, A.; Freddolino, P. L.; Schulten, K. Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure* **2006**, *14*, 1767–1777.

(17) Srinivas, G.; Klein, M. Molecular dynamics simulations of self-assembly and nanotube formation by amphiphilic molecules in aqueous solution: a coarse-grain approach. *Nanotechnology* **2007**, 18.

(18) Zhou, J.; Thorpe, I. F.; Izvekov, S.; Voth, G. A. Coarse-grained peptide modeling using a systematic multiscale approach. *Biophys. J.* **2007**, *92*, 4289–4303.

(19) Voth, G. A. *Coarse-graining of condensed phase and biomolecular systems*, 1st ed.; Taylor & Francis Group: New York, 2009; pp 1−455.

(20) DeMille, R. C.; Molinero, V. Coarse-grained ions without charges: reproducing the solvation structure of NaCl in water using short-ranged potentials. *J. Chem. Phys.* **2009**, *131*, 034107.

(21) Molinero, V.; Moore, E. B. Water modeled as an intermediate element between carbon and silicon. *J. Phys. Chem. B* **2009**, *113*, 4008–4016.

(22) Savelyev, A.; Papoian, G. A. Molecular renormalization group coarse-graining of electrolyte solutions: application to aqueous NaCl and KCl. *J. Phys. Chem. B* **2009**, *113*, 7785–7793.

(23) Yesylevskyy, S. O.; Schäfer, L. V.; Sengupta, D.; Marrink, S. J. Polarizable water model for the coarse-grained MARTINI force field. *PLoS Comput. Biol.* **2010**, *6*, e1000810.

(24) Dans, P.; Zeida, A.; Machado, M.; Pantano, S. A coarse grained model for atomic-detailed DNA simulations with explicit electrostatics. *J. Chem. Theory Comput.* **2010**, *6*, 1711–1725.

(25) Head-Gordon, T.; Hura, G. Water structure from scattering experiments and simulation. *Chem. Rev.* **2002**, *102*, 2651–2670.

(26) Narten, A. H.; Danford, M. D.; Levy, H. A. X-ray diffraction study of liquid water in the temperature range 4−200°C. *Discuss. Faraday Soc.* **1967**, *43*, 97–107.

(27) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. *Intermolecular forces*; Pullman, B., Ed.; D. Reidel Publishing Company: Dordrecht, The Netherlands, 1981; pp 331−342.

(28) 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.

(29) Mancinelli, R.; Botti, A.; Bruni, F.; Ricci, M. A.; Soper, A. K. Hydration of sodium, potassium, and chloride ions in solution and the concept of structure maker/breaker. *J. Phys. Chem. B* **2007**, *111*, 13570–13577.

(30) Mancinelli, R.; Botti, A.; Bruni, F.; Ricci, M. A.; Soper, A. K. Perturbation of water structure due to monovalent ions in solution. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2959–2967.

(31) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. Gromacs: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

(32) Lindahl, E.; Hess, B.; van der Spoel, D. Gromacs 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001**, *7*, 306–317.

(33) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. Gromacs: fast, flexible and free. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(34) Bekker, H.; Berendsen, H. J. C.; Dijkstra, E. J.; Achterop, S.; van Drunen, R.; van der Spoel, D.; Sijbers, A.; Keegstra, H.; Reitsma, B.; Renardus, M. K. R. *Gromacs: A parallel computer for molecular dynamics simulations*; de Groot, R. A., Nadrchal, J., Eds.; World Scientific: Singapore, 1993.

(35) Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **1984**, *52*, 255–268.

(36) Hoover, W. Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(37) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(38) Nosé, S.; Klein, M. L. Constant pressure molecular dynamics for molecular systems. *Mol. Phys.* **1983**, *50*, 1055–1076.

(39) Darden, T.; York, D.; Pedersen, L. Particle mesh ewald: an n-log(n) method for ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(40) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. A smooth particle mesh ewald potential. *J. Chem. Phys.* **1995**, *103*, 8577–8592.

(41) Winger, M.; Trzesniak, D.; Baron, R.; van Gunsteren, W. F. On using a too large integration time step in molecular dynamics simulations of coarse-grained molecular models. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1934–1941.

(42) He, X.; Shinoda, W.; DeVane, R.; Klein, M. L. Exploring the utility of coarse-grained water models for computational studies of interfacial systems. *Mol. Phys.* **2010**, *108*, 2007–2020.

(43) Herrero, C. P. Compressibility of solid helium. *J. Phys.: Condens. Matter* **2008**, *20*, 295230.

(44) van Buuren, A. R.; Marrink, S. J.; Berendsen, H. J. C. A molecular dynamics study of the decane/water interface. *J. Phys. Chem.* **1993**, *97*, 9206–9212.

(45) Mark, A. E.; van Helden, S. P.; Smith, P. E.; Janssen, L. H. M.; van Gunsteren, W. F. Convergence properties of free energy calculations: alpha-cyclodextrin complexes as a case study. *J. Am. Chem. Soc.* **1994**, *116*, 6293–6302.

(46) Beglov, D.; Roux, B. Finite representation of an infinite bulk system: solvent boundary potential for computer simulations. *J. Chem. Phys.* **1994**, *100*, 9050–9063.

(47) Luo, Y.; Roux, B. Simulation of osmotic pressure in concentrated aqueous salt solutions. *J. Phys. Chem. Lett.* **2010**, *1*, 183–189.

(48) Herrera, E. F.; Pantano, S. Salt induced asymmetry in membrane simulations by partial restriction of ionic motion. *J. Chem. Phys.* **2009**, *130*, 195105–195114.

(49) Lavery, R.; Sklenar, H. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.* **1988**, *6*, 63–91.

(50) Soper, A. K. The radial distribution functions of water and ice from 673 K and at pressures up to 400 MPa. *Chem. Phys.* **2000**, *258*, 121–137.

(51) Kell, G. S. Density, thermal expansivity, and compressibility of liquid water from 0° to 150°C: correlations and tables for atmospheric pressure and saturation reviewed and expressed on 1968 temperature scale. *J. Chem. Eng. Data* **1975**, *20*, 97–105.

(52) Holz, M.; Heil, S. R.; Sacco, A. Temperature-dependent self-diffusion coefficients of water and six selected molecular liquids for calibration in accurate H-1 NMR PFG measurements. *Phys. Chem. Chem. Phys.* **2000**, *2*, 4740–4742.

(53) 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.

(54) Groot, R. D.; Rabone, K. L. Mesoscopic simulation of cell membrane damage, morphology change and rupture by nonionic surfactants. *Biophys. J.* **2001**, *81*, 725–736.

(55) The value of the permittivity can vary with the condition of the simulation, size of the computational box, etc.

(56) Kusalik, P. G.; Svishchev, I. M. The spatial structure in liquid water. *Science* **1994**, *265*, 1219–1221.

(57) van Maaren, P. J.; van der Spoel, D. Molecular dynamics of water with novel shell-model potentials. *J. Phys. Chem. B* **2001**, *105*, 2618–2626.

(58) van der Spoel, D.; van Maaren, P. J.; Berendsen, H. J. C. A systematic study of water models for molecular simulation: derivation of water models optimized for use with a reaction field. *J. Chem. Phys.* **1998**, *108*, 10220–10230.

(59) Mahoney, M. W.; Jorgensen, W. L. Diffusion constant of the TIP5P model of liquid water. *J. Chem. Phys.* **2001**, *114*, 363–366.

(60) Yu, H.; van Gunsteren, W. F. Charge-on-spring polarizable water models revisited: from water clusters to liquid water to ice. *J. Chem. Phys.* **2004**, *121*, 9549–9564.

(61) Yu, H.; Hansson, T.; van Gunsteren, W. F. Development of a simple self-consistent polarizable model for liquid water. *J. Chem. Phys.* **2003**, *118*, 221–234.

(62) Chen, F.; Smith, P. E. Simulated surface tensions of common water models. *J. Chem. Phys.* **2007**, *126*, 221101–221104.

(63) Wang, H.; Junghans, C.; Kremer, K. Comparative atomistic and coarse-grain study of water: what do we lose by coarse-graining. *Eur. Phys. J. E* **2009**, *28*, 221–229.

(64) 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*, 8910–8922.

(65) Murrell, J. N.; Jenkins, A. D. *Properties of liquids and solutions*, 2nd ed.; John Wiley & Sons: Chichester, U. K., 1994; pp 1−299.

(66) Eisenberg, D.; Kauzmann, W. *The Structure and Properties of Water*; Oxford University Press: Oxford, U.K., 1969; pp 1−308.

(67) Dilmohamud, B. A.; Seeneevaseen, J.; Rughooputh, S. D. D. V.; Ramasami, P. Surface tension and related thermodynamic parameters of alchohols using the Traube stalagmometer. *Eur. J. Phys.* **2005**, *26*, 1079.

(68) Rodnikova, M. N. A new approach to the mechanism of solvophobic interactions. *J. Mol. Liq.* **2007**, *136*, 211–213.

(69) Kalcher, I.; Horinek, D.; Netz, R. R.; Dzubiella, J. Ion specific correlations in bulk and at biointerfaces. *J. Phys.: Condens. Matter* **2009**, *21*, 424108.

(70) Shotton, M. W.; Pope, L. H.; Forsyth, V. T.; Langan, P.; Grimm, H.; Rupprecht, A.; Denny, R. C.; Fuller, W. A high-angle neutron fiber diffraction study of the hydration of B-DNA. *Physica B* **1998**, *243*, 1166–1168.

(71) Young, M. A.; Ravishanker, G.; Beveridge, D. L. A 5-ns molecular dynamics trajectory for B-DNA: analysis of structure, motions, and solvation. *Biophys. J.* **1997**, *73*, 2313–2336.

(72) Cheatham, T. E., 3rd.; Srinivasan, J.; Case, D. A.; Kollman, P. A. Molecular dynamics and continuum solvent studies of the stability of polyG-polyC and polyA-polyT DNA duplexes in solution. *J. Biomol. Struct. Dyn.* **1998**, *16*, 265–280.

(73) Duan, Y.; Wilkosz, P.; Crowley, M.; Rosenberg, J. M. Molecular dynamics simulation study of DNA dodecamer d(CGCGAATTCGCG) in solution: conformation and hydration. *J. Mol. Biol.* **1997**, *272*, 553–572.

(74) Feig, M.; Pettitt, B. M. Modeling high-resolution hydration patterns in correlation with DNA sequence and conformation. *J. Mol. Biol.* **1999**, *286*, 1075–1095.

(75) Young, M. A.; Beveridge, D. L. Molecular dynamics simulations of an oligonucleotide duplex with adenine tracts phased by a full helix turn. *J. Mol. Biol.* **1998**, *281*, 675–687.

(76) Kochoyan, M.; Leroy, J. L. Hydration and solution structure of nucleic acids. *Curr. Opin. Struct. Biol.* **1995**, *5*, 329–333.

(77) Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. E. Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 2179–2183.

(78) Shui, X.; Sines, C. C.; McFail-Isom, L.; VanDerveer, D.; Williams, L. D. Structure of the potassium form of CGC-GAATTCGCG: DNA deformation by electrostatic collapse around inorganic cations. *Biochemistry* **1998**, *37*, 16877–16887.

(79) Shui, X.; McFail-Isom, L.; Hu, G. G.; Williams, L. D. The B-DNA dodecamer at high resolution reveals a spine of water on sodium. *Biochemistry* **1998**, *37*, 8341–8355.

(80) Manning, G. S. The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. *Q. Rev. Biophys.* **1978**, *11*, 179–246.

(81) Ponomarev, S. Y.; Thayer, K. M.; Beveridge, D. L. Ion motions in molecular dynamics simulations on DNA. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14771–14775.

(82) Savelyev, A.; Papoian, G. A. Electrostatic, steric, and hydration interactions favor Na+ condensation around DNA compared with K+. *J. Am. Chem. Soc.* **2006**, *128*, 14506–14518.

(83) Pérez, A.; Luque, F. J.; Orozco, M. Dynamics of B-DNA on the microsecond time scale. *J. Am. Chem. Soc.* **2007**, *129*, 14739–14745.

(84) McConnell, K. J.; Beveridge, D. L. Molecular dynamics simulations of B-DNA: sequence effects on A-tract-induced bending and flexibility. *J. Mol. Biol.* **2001**, *314*, 23–40.

(85) Feig, M.; Pettitt, B. M. Sodium and chlorine ions as part of the DNA solvation shell. *Biophys. J.* **1999**, *77*, 1769–1781.

(86) Fuoss, R. M.; Katchalsky, A.; Lifson, S. The potential of an infinite rod-like molecule and the distribution of the counter ions. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 579–589.

(87) Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Bandyopadhyay, S.; Klein, M. L. A coarse grain model for phospholipid simulations. *J. Phys. Chem. B* **2001**, *105*, 4464–4470.

CT100379F

JCTC Journal of Chemical Theory and Computation

Third-Generation Hydrogen-Bonding Corrections for Semiempirical QM Methods and Force Fields

Martin Korth*

Theory of Condensed Matter Group, Cavendish Laboratory, 19 J J Thomson Avenue, Cambridge CB3 0HE, United Kingdom

Received July 21, 2010

Abstract: Computational modeling of biological systems is a rapidly evolving field that calls for methods that are able to allow for extensive sampling with systems consisting of thousands of atoms. Semiempirical quantum chemical (SE) methods are a promising tool to aid with this, but the rather bad performance of standard SE methods for noncovalent interactions is clearly a limiting factor. Enhancing SE methods with empirical corrections for dispersion and hydrogen-bonding interactions was found to be a big improvement, but for the hydrogen-bonding corrections the drawback of breaking down in the case of substantial changes to the hydrogen bond, e.g., proton transfer, posed a serious limitation for its general applicability. This work presents a further improved hydrogen-bonding correction that can be generally included in parameter fitting procedures, as it does not suffer from the conceptual flaws of previous approaches: hydrogen bonds are now treated as an interaction term between electronegative acceptor and donor atoms, "weighted" by a function of the position of H atoms between them, and multiplied with a damping function to correct the short- and long-range behavior. The performance of the new approach is evaluated for PM6, AM1, OM3, and SCC-DFTB as well as several force-field (FF) methods for a number of standard benchmark sets with hydrogen-bonded systems. The new approach is found to reach the same accuracy as the second-generation hydrogen-bonding correction with less parameters, while it avoids among other issues the conceptual problem with electronic structure changes. SE methods augmented this way reach the accuracy of DFT-D approaches for a large number of cases investigated, while still being about 3 orders of magnitude faster. Moreover, the new correction scheme is transferable also to FF methods that were shown to have serious problems with hydrogen-bonding interactions.

1. Introduction

Many promising applications of computational methods like computer-aided drug design are related to large-scale simulations of biologically relevant molecular systems. While significant successes have already been achieved, e.g., in computer-aided drug lead generation and optimization,[1,2] the field is still confronted with serious challenges, especially considering the effects of protein flexibility and solvation.[3] As a possibly very valuable tool for tackling these problems, semiempirical quantum chemical (SE) methods have come into the focus of several groups in recent years.[4−12] SE methods offer a compromise between the accuracy of "full" ab initio treatments and the speed of force field (FF) approaches. This way, SE methods allow for extensive sampling of large systems, while keeping the ability to describe the effects of electronic structure changes. The latter point is of high importance, because customary used FF point charge models ignore effects such as charge transfer and polarization[13] that are likely to be quite important in biomolecular modeling applications.[14]

But biomacromolecules are dominantly influenced by noncovalent interactions like dispersion and hydrogen bonding that generally need very high-level quantum chemical methods to be modeled with sufficient accuracy.[15] In this sense, it comes to no surprise that standard SE methods perform rather poorly for these types of interaction. A big

* E-mail: mk642@cam.ac.uk and dgd@uni-muenster.de.

10.1021/ct100408b
Published on Web 11/01/2010

$cos(\theta)^n$ like **DH1**, FS1	H-bond	f_{geom} like **DH2**, DH+
-4.18	N...HN bond	-2.78
-0.17	cross-molecule	none
-0.54	intra-molecule	-0.02
-0.04	cross-molecule	none
-0.18	cross-molecule	none
-0.48	cross-bond	none
-0.24	cross-molecule	0.00
-0.84	cross-bond	0.00
-0.25	O...HN bond	-2.01
0.01	cross-molecule	-0.01
-0.62	cross-molecule sum	-0.01

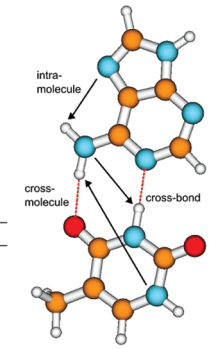
H-bond Correction Contributions for a 3225 Atom Protein[a,b]

type	No. of H-bonds[c]	overall correction
$cos(\theta)^n$ like DH1, FS1	≈ 1500	≈ 630 kcal/mol
$cos(\theta)^n$ w/o 'cross-molecule'	≈ 1200	≈ 530 kcal/mol
f_{geom} like DH2, **DH+**	≈ 230	≈ 120 kcal/mol

[a] all numbers already with long-range damping, i.e. no contributions beyond 10.5 Å

[b] in bold letters the method with which the actual numbers are produced

[c] i.e. number of contributions larger than 0.1 kcal/mol



Red dots indicate actual H-bonds, black arrows indicate unphysical contributions: Non-existing or very weak intra-molecule H-bonds, interactions with parallel H-bonds ('cross-bond'), or across the molecule core ('cross-molecule').

Figure 1. Illustration of the importance to go beyond a simple $\cos(\theta)^n$ term for hydrogen-bonding correction terms. See the text for further explanation.

step forward for the description of biomolecular systems with SE methods was made with the inclusion of empirical dispersion corrections (e.g., PM3-D, AM1-D[16]), similar to the ones used for DFT (e.g., refs 17 and 18) methods. But in contrast to DFT methods that perform acceptably well for hydrogen-bonding interactions,[19] SE methods remain as deficient as commonly used FF approaches[20] for systems beyond pure dispersion interactions. While this disadvantage has been known for many years[21−24] and several attempts to cure this remedy were successful up to a certain point,[23−29] only the inclusion of force-field-type terms for empirical hydrogen-bonding corrections was able to improve the accuracy of SE methods to a level near that of DFT-D approaches. Because of conceptual problems with the initial PM6[30,31]-based approach,[32] the method was redesigned in a physically more sound way and is now publicly available in Mopac2009[33] as the "-DH2" add-on method. One remaining major drawback of DH2 is the breakdown of the correction in the case of an acceptor-atom change, a problem that is, among other issues, addressed in the following.

2. Empirical H-Bonding Corrections for Semiempirical Quantum Chemical Methods

First-Generation and Second-Generation Corrections. The first-generation correction, termed "DH" and later on "DH1", (eq 1) made use of the charges (q) on the acceptor (A) and hydrogen (H) atoms, the distance (r) between these atoms, and a cosine term that promotes a 180° bonding situation for the A···H−D (with the donor atom D) angle:

$$E_{\mathrm{H-bond}} = a\left[\frac{q_{\mathrm{A}} \times q_{\mathrm{H}}}{r^2} \times \cos(\theta) + b \times c^r\right] \qquad (1)$$

This design led to a number of problems, with the possibility of a large number of unphysical contributions to the correction from nonexisting hydrogen bonds, e.g., through the back of acceptor atoms, because the orientation of the acceptor atom is not taken into account (see below for a detailed discussion). Other problems include large discontinuities (because only the attractive but not the repulsive term is multiplied with the angular dependency), a high number of unsystematic parameters, and unphysical cutoffs.

The second-generation correction (eq 2) is a complete redesign of this approach, with the most important change being the inclusion of the missing information about the sterical arrangement of the acceptor side of the hydrogen bonds (see ref 34 for a detailed explanation). This "H2" correction uses the same distance r, the two angles A···H−D (termed Θ) and R_2−A···H (termed Φ, with R_2 being a donor "base atom"), and the corresponding three torsional angles, of which only one directly influences the H-bond interaction energy, R_1R_2A···H (termed Ψ):

$$E_{\mathrm{H-bond}} = \left[a \times \frac{q_{\mathrm{A}} \times q_{\mathrm{H}}}{r^b} + c \times d^r\right] \times \cos(\theta) \times \cos(\phi) \times \cos(\psi) \qquad (2)$$

with ϕ and ψ as the deviations of the R_2−A···H angle and R_1R_2A···H torsion angle from the idealized optimal H-bond values. This redesign also allowed for keeping terms and parameters more physically sound (e.g., avoiding the above-mentioned large discontinuities), led to a much smaller number of now systematic parameters, and made the correction transferable to other SE methods. Large systematic gains in accuracy for hydrogen-bonded complexes were possible with only one overall parameter; the final accuracy using eight fitted parameters reached the DFT-D level for a large number of investigated cases.[34]

Figure 1 illustrates how important it is to go beyond a simple $\cos(\theta)^n$ ansatz for the geometrical definition of hydrogen-bond correction terms: even in a rather small system like the Watson−Crick bound adenine/thymine base pair, the number and impact of unphysical contributions (indicated in the picture with black arrows) is quite large when the simple $\cos(\theta)^n$ term is used. These contributions sum up to enormous interaction energies for larger systems, as shown for a medium-sized protein. Please note that our numbers should be considered rather conservative estimates, because we already used a long-

range damping function, so that no interactions beyond 10.5 Å contribute to the shown values.

The major remaining drawback of the second-generation correction is the kept direct dependence on the distance between the hydrogen and the acceptor atom (and the corresponding parametrization to acceptor atom types) that requires a constant bonding situation between acceptor, hydrogen, and donor atoms, making it (unlike the common empirical dispersion corrections) a bond-type term, with all the disadvantages attached: if the acceptor atom changes (e.g., in the case of proton transfer from the donor to the acceptor), then the correction is likely to break down (see below for example data). Other problems include the need for charge derivatives for analytical gradient calculations (ignored in the published DH2 version, but on the order of tenths of a kilocalorie per mole even for small systems in the case of strong H bonds), a problematic repulsive term including a distance cutoff to prevent problems with the optimization of strong H bonds, and a partially adjusted dispersion correction (with a modified C_6 for sp^3 carbon and a changed van der Waals radius for hydrogen) that could now profit from recent developments of dispersion corrections with system-dependent C6 coefficients (something not addressed in this paper, but under development). A final issue is the long-range behavior of previously published hydrogen-bonding corrections: As it is not generally clear what the exact long-range behavior should be (among other reasons because semiempirical methods already account to some extent for hydrogen-bonding interactions), we think the most reasonable thing to do is to design the correction to be of a rather short-ranged nature. This seems to be the safest way to go in the sense that no correction is safer than a correction from a huge sum of very tiny and most likely wrong contributions.

Shortly before we finished our manuscript, Foster and Sohlberg published another hydrogen-bonding correction scheme for the SE method AM1, termed FS1,[35] which they compare to PM6-DH but unfortunately not to the likewise earlier published DH2 scheme. FS1 has the same basic outline as PM6-DH1, with the repulsive term replaced by a damping function (as previously suggested as an alternative approach for DH2[34]) and the bond-type parametrization replaced with four general, fitted parameters, which somewhat spoils the accuracy but partly solves the problem with electronic structure changes. The authors nevertheless have to admit that a safe treatment of such changes would need a further modified (effectively doubled) version of their ansatz and, therefore, do not recommend using their published version in such cases. As we have tried a doubling of terms for DH2 before publishing it, we believe to have good reasons to question that such a modification will not further diminish the accuracy of the FS1 scheme. Independently of this issue, we consider the FS1 scheme to be a first-generation hydrogen-bonding correction, because it does not take the complete geometric information into account (and accordingly suffers from all of the related problems), as discussed above when comparing DH2 to the earlier PM6-DH1.

Third-Generation Correction. The third-generation correction (eq 3) does not make the assumption of a specific acceptor/hydrogen/donor binding situation but instead takes the hydrogen bond as a charge-independent atom−atom term between two atoms capable of serving as an acceptor or donor part (e.g., O, N), weighted by a function that accounts for the steric arrangement of the two fragments to each other and the preferably favorable positioning of a H atom somewhere between them (a definition of coordinates analogous to the above, with A and B being the two possible acceptor/donor atoms and C_A and C_B the corresponding hydrogen-bonding correction parameters from Table 1 for semiempirical and Table 2 for force field methods), multiplied with a damping function to correct the short- and long-range behaviors:

Table 1. Hydrogen-Bonding Correction Parameters $C_{element}$ for Semiempirical QM Methods

element	OM3	PM6	AM1	DFTB
N	−0.05	−0.16	−0.29	−0.21
O	−0.07	−0.12	−0.29	−0.08

Table 2. Hydrogen-Bonding Correction Parameters $C_{element}$ for Force Field Methods

element	MM2*	MM3*	AMBER*	OPLS*	OPLSAA	MMFF94
N	−0.64	−0.63	−0.21	−0.24	−0.25	−0.21
O	−0.08	−0.17	−0.03	−0.00	−0.00	−0.05

$$E_{\mathrm{H-bond}} = \frac{C_{\mathrm{AB}}}{r_{\mathrm{AB}}^2} \cdot f_{\mathrm{geom}} \times f_{\mathrm{damp}} \tag{3}$$

$$f_{\mathrm{geom}} = \cos(\theta_{\mathrm{A}})^2 \times \cos(\phi_{\mathrm{A}})^2 \times \cos(\psi_{\mathrm{A}})^2 \times \cos(\phi_{\mathrm{B}})^2 \times \cos(\psi_{\mathrm{B}})^2 \times f_{\mathrm{bond}} \tag{4}$$

$$f_{\mathrm{bond}} = 1 - \frac{1}{1 + \exp[-60(r_{\mathrm{XH}}/1.2 - 1)]} \tag{5}$$

$$f_{\mathrm{damp}} = \left(\frac{1}{1 + \exp[-100(r_{\mathrm{AB}}/2.4 - 1)]}\right) \times \left(1 - \frac{1}{1 + \exp[-10(r_{\mathrm{AB}}/7.0 - 1)]}\right) \tag{6}$$

$$C_{\mathrm{AB}} = \frac{C_{\mathrm{A}} + C_{\mathrm{B}}}{2} \tag{7}$$

The damping functions can be chosen as a "safe bet", so that no fitting is necessary for them (albeit the long-range cutoff could in principle be taken as a fit parameter, e.g., if it turns out that the structures of very large molecules are found to be too dense): the f_{damp} function is switched on between a donor−acceptor distance of 2.3 and 2.5 Å (safe choice for the assumption of no H bonds below 2.5 Å) and slowly switched off between 3.5 and 10.5 Å (safe choice for the assumption of full H-bond strength up to 3.5 Å and no strength anymore at three times this distance). Figure 2 shows this damping function and a resulting example energy profile for the overall correction. The f_{bond} function brings the correction to zero if the hydrogen wanders away too far from both electronegative atoms (with r_{XH} being the smaller one of the two distances r_{AH} and r_{BH}): It is switched off between 1.15 and 1.25 Å (safe choice for the assumption of a maximum distance of 1.15 for a covalent hydrogen bond).

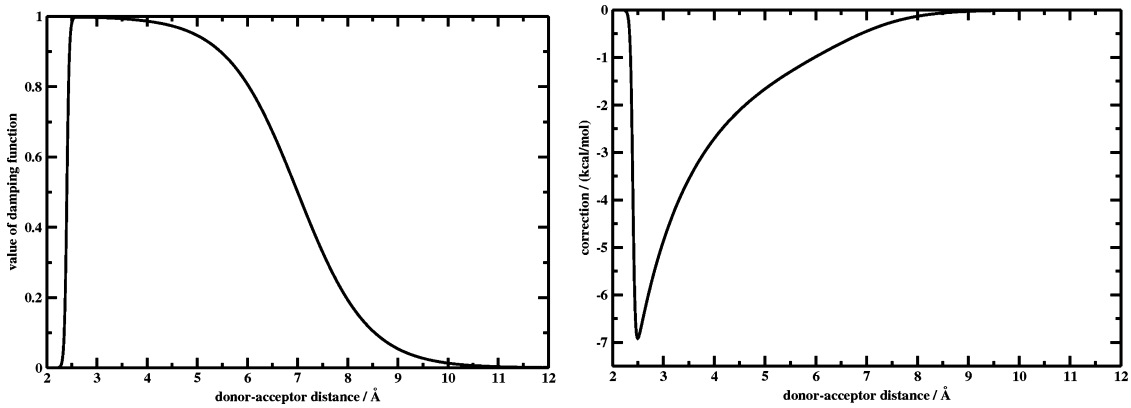


Figure 2. The f_{damp} function (left) and an example for the overall correction energy (right).

Table 3. Comparison of First-, Second-, and Third-Generation Hydrogen-Bonding Correction Schemes

generation	scheme	transferability	generality[a]	full geometric information[b]	safe long-range behavior[c]	no charges used[d]	number of fitted parameters
1	DH1	PM6	no	no	no	no	24
	FS1	AM1	yes/no[e]	no	no[f]	no	4
2	DH2	SE methods	no	yes	no	no	8
3	DH+	SE and FF methods	yes	yes	yes	yes	2

[a] Robust scheme, does not break down for electronic structure changes, can be generally included in parameter fits for new semiempirical QM (SE) and force field (FF) methods. [b] Uses full geometric information, not just a cosine term, which is likely to lead to problems with larger systems (see Figure 1). [c] Shows safe long-range behavior by avoiding huge sums of very small (and likely wrong) contributions. [d] Allows for affordable analytical gradients, as no gradients with respect to charges are required. [e] Usage for proton transfer "effectively possible" in a suggested, modified (doubled) version, but not recommended by the authors of AM1-FS1. [f] Damping function with bad long-range behavior (significantly above zero at long distances).

The (torsion) angles of the f_{geom} function are defined similarly to those of the DH2 correction,[34] with ϕ and ψ now symmetrically used for both the donor and acceptor atoms. This ansatz is not a doubling or "double-potential" version of the DH2 correction (which was tested by the author before the publication of DH2 but found to be quite problematic): the important difference is the change from the use of the hydrogen−acceptor distance (with its requirement of a hydrogen−donor bond definition) to the core−core interaction picture that results from using the donor−acceptor distance instead of the hydrogen−acceptor distance. Through this change, the implicit hydrogen−donor bond definition (also still present in the recently published AM1-FS1 method, see above) can be avoided. The target angles can nevertheless be kept as for the second-generation correction,[34] which of course has to be the case for "text-book" ideal values.

This way, the new scheme accounts for the major drawback of the (first- and) second-generation correction, i.e., the problem of a substantial change to the hydrogen bond, but several additional benefits are gained as side effects: while keeping the high accuracy of the "DH2" scheme, the number of fitted parameters can be reduced from eight to two. As charges are no longer used, no charge-derivative terms are needed for the analytical gradient. The repulsive term can be replaced by a damping function, and cutoff distances are no longer needed for an accurate description of nonequilibrium structures. (In the development of DH2, an unphysical short-distance cutoff was introduced to avoid problems with very strong hydrogen bonds where the correction was much too high, because strong partial charges and short H-bond distances both increase the value of the DH2 correction.) At the same time, the damping function greatly improves the long-range behavior in the sense that we think it to be preferable to have no (rather than very likely wrong) long-range contributions from hydrogen-bonding corrections. Finally, we found that the new scheme is also well suited for the application to force field methods. This is of great importance as it was recently shown how strongly common force fields underestimate hydrogen-bonding interactions (while they actually perform very well for dispersion interactions)[20] and a possible improvement, e.g., of water models, could have major impact for biomolecular modeling in general. We should mention though that while our straightforward implementation of the third-generation correction is about 2 orders of magnitude faster than the underlying semiempirical methods for midsized proteins (as is "DH2"), the application to force fields will require a more sophisticated approach to avoid slowing down the force field calculations by about 1 order of magnitude.

The new correction was parametrized on the hydrogen-bonded complexes of the S26 + S22x4 set for the AM1, PM6, OM3, and SCC-DFTB methods (all enhanced with standard dispersion corrections, see Table 1) and of the S22 set only for several FF methods (see Table 2). Optimization of the parameters with respect to the mean unsigned error (MUE) and the root-mean-square error (RMSE) over all reactions led to nearly identical parameter values; the final parameters are taken from the MUE optimizations. We call our approach "H+", to indicate the conceptual difference of "DH+" from the first- and second-generation "DHn" approach. Table 3 summarizes the developments from the first to the second and third generation H-bonding corrections explained in this section.

Table 4. Results for the H-Bonded Complexes of the S26 Set[a]

	OM3-D	-DH2	-DH+	PM6-D	-DH2	-DH+	AM1-D[b]	-DH2	-DH+	DFTB-D	-DH2	-DH+
MSE	−1.75	−0.66	−0.36	−2.82	0.03	0.13	−5.58	0.25	0.81	−2.86	0.14	−0.11
MUE	1.75	0.66	0.62	2.82	0.19	0.66	5.58	0.73	2.28	2.86	0.88	1.01
RMSE	2.22	0.96	0.84	3.56	0.27	0.88	7.57	0.95	2.60	3.15	1.01	1.20
Δ	5.16	2.35	3.10	6.20	0.92	3.18	17.30	3.33	8.58	4.88	2.93	3.69

[a] Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values are in kilocalories per mole. [b] AM1-D refers to standard AM1 with a standard empirical dispersion correction, unlike AM1-D.

Table 5. Results for the H-Bonded Complexes of the S26 Set, Optimized with Each Method[a]

	OM3- D[b]	-DH 2[b]	-DH+[b]	PM6-D	-DH2	-DH+	AM1-D[c]	-DH2	DH+
MSE	4.74	5.28	4.48	−2.63	0.68	0.45	−3.68	1.16	2.31
MUE	5.18	6.18	5.03	2.63	1.10	0.75	3.68	1.21	2.42
RMSE	11.52	12.17	7.70	3.20	1.56	0.91	5.02	1.56	2.84
Δ	38.83	40.56	20.20	4.94	5.36	2.85	10.45	3.43	5.80

[a] Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values are in kilocalories per mole. [b] Already, the OM3 method itself (without D or H corrections) has a serious problem with the very strongly bound formic acid dimer, which is the main reason why the errors are much larger than for the other examples. [c] Refers to standard AM1 with a standard empirical dispersion correction, unlike AM1-D.

Table 6. Results for the H-Bonded Complexes of the S26 and S22x4 Sets[a]

	OM3-D	-DH2	-DH+	PM6-D	-DH2	-DH+	AM1-D[b]	-DH2	-DH+	DFTB-D	-DH2	-DH+
MSE	1.25	0.21	−0.02	2.35	0.01	−0.39	4.91	0.27	−0.97	2.83	0.33	0.22
MUE	1.45	0.91	1.03	2.35	0.24	0.81	4.91	1.42	2.88	2.83	0.74	0.80
RMSE	2.05	1.36	1.46	3.20	0.34	1.00	7.76	2.55	3.54	3.33	0.89	1.01
Δ	7.92	6.30	7.45	7.88	1.65	4.14	26.18	14.38	15.46	8.12	3.19	4.51

[a] Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values are in kilocalories per mole. [b] Refers to standard AM1 with a standard empirical dispersion correction, unlike AM1-D.

Table 7. Results for the 105 Small, H-Bonded Complexes of the PM6-DH1 Fit Set[a]

	OM3-D	-DH2	-DH+	PM6-D	-DH2	-DH+	AM1-D[b]	-DH2	-DH+	DFTB-D	-DH2	-DH+
MSE	−0.88	−0.51	0.03	−1.66	−0.43	0.46	−2.55	−0.12	1.85	−2.33	−0.40	−0.14
MUE	0.91	0.66	0.46	1.77	1.15	1.21	2.71	1.59	2.40	2.36	0.85	1.07
RMSE	1.14	0.86	0.59	2.35	1.54	1.44	4.04	2.12	2.87	2.79	1.06	1.44
Δ	6.52	4.48	3.71	9.61	7.37	6.18	22.64	12.14	13.08	10.47	5.15	8.64

[a] Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values are in kilocalories per mole. [b] Refers to standard AM1 with a standard empirical dispersion correction, unlike AM1-D.

3. Computational Details

Semiempirical PM6 and AM1 calculations applying the MOZYME algorithm were done with MOPAC2009,[33] OM3 calculations with MNDO2005, and SCC-DFTB calculations with DFTB+.[36] AM1-D* refers to standard AM1[16] with a standard Jurecka-type[18] empirical dispersion correction (see ref 34 for details), not AM1-D, which is additionally based on a refit of 18 AM1 parameters. B3-LYP[37,38] DFT calculations with empirical dispersion corrections of the Jurecka type[18] were done with Turbomole 5.9[39] using TZVP[40] and QZVP[41] Gaussian AO basis sets and the RI approximation[42,43] for two-electron integrals.

Energies and analytical gradients for our new "H+" hydrogen-bonding correction are implemented as a standalone program that is freely available from the author upon request. Preparations to make the correction available within the open source FF code GROMACS are underway.

4. Results and Discussion

Tables 4−9 show results of OM3, PM6, AM1, and SCC-DFTB (shortened to "DFTB" in the tables) calculations with dispersion and second- and third-generation hydrogen-bonding corrections for the hydrogen-bonded complexes of the S26[44] (Table 4, again in Table 5 with structures optimized at each level of theory) and S26 + S22x4[34] (Table 6) benchmark sets; the PM6-DH1 training set of 105 small hydrogen-bonded complexes[32] (Table 7); the 37 noncharged, H-bonded DNA base pair complexes from the JSCH2005 set[15] (Table 8); and the 13 noncharged, H-bonded peptide structures from the JSCH2005 test set (Table 9). (We do not supply DFTB data in Table 5 because our interface does not allow for DFTB geometry optimizations with DH+ yet.) The geometries of these benchmarks are optimized at the MP2/cc-pVTZ level or higher (S22, S26, S22x4, PM6-DH1 training set, JSCH2005 partly) or represent experimental data (JSCH2005 partly); see the references given above for details.

Table 8. Results for the JSCH2005 H-Bonded DNA Base Pairs[a]

	OM3-D	-DH2	-DH+	PM6-D	-DH2	-DH+	AM1-D[b]	-DH2	-DH+	DFTB-D	-DH2	-DH+
MSE	−2.41	−0.81	−0.49	−6.10	−0.54	−0.87	−10.16	0.11	−0.05	−5.49	2.64	0.90
MUE	2.50	1.22	1.02	6.10	1.76	1.35	10.16	2.29	1.68	5.49	3.06	1.64
RMSE	2.79	1.44	1.29	6.30	2.23	1.59	10.91	2.87	2.40	5.80	3.48	1.97
Δ	6.81	4.52	5.74	7.67	7.94	5.94	16.31	12.46	11.78	6.85	8.53	7.20

[a] Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values are in kilocalories per mole. [b] Refers to standard AM1 with a standard empirical dispersion correction, unlike AM1-D.

Table 9. Results for the JSCH2005 H-Bonded Peptides[a]

	OM3-D	-DH2	-DH+	PM6-D	-DH2	-DH+	AM1-D[b]	-DH2	-DH+	DFTB-D	-DH2	-DH+
MSE	0.33	0.36	0.36	−0.07	−0.00	−0.00	1.37	1.45	1.50	−0.84	−0.75	−0.76
MUE	0.60	0.62	0.62	0.65	0.69	0.68	1.49	1.56	1.60	0.92	0.83	0.85
RMSE	0.80	0.81	0.82	0.85	0.88	0.86	1.94	2.03	2.11	1.07	0.98	0.99
Δ	2.81	2.79	2.80	3.45	3.42	3.40	4.30	4.27	4.60	2.91	2.84	2.85

[a] Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values are in kilocalories per mole. [b] Refers to standard AM1 with a standard empirical dispersion correction, unlike AM1-D.

Table 10. Results for the H-Bonded Complexes of the S22 Set[a]

	MM2*	-H+	MM3*	-H+	AMBER*	-H+	OPLS*	-H+	OPLSAA	-H+	MMFF94	-H+	B3LYP-D/TZVP
MSE	−8.77	−0.63	−10.90	−1.37	−3.47	−0.74	−2.76	−0.20	−3.27	−0.61	−3.01	0.05	0.74
MUE	8.77	2.12	10.90	3.61	3.93	2.63	3.30	2.03	3.55	1.73	3.15	0.84	0.74
RMSE	10.68	2.70	13.13	5.00	5.29	3.60	4.42	2.57	4.49	2.53	3.88	1.19	0.84
Δ	−17.18	−8.75	−19.07	−16.17	−11.65	−9.83	−9.10	−8.09	−8.46	−7.70	−7.57	−4.30	1.17

[a] Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values are in kilocalories per mole.

Structures and reference energies for these test sets can be obtained online from the Benchmark Energy and Geometry DataBase BEGDB, see http://www.begdb.com. Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are given in kilocalories per mole. Table 10 gives the same statistical error measures for the hydrogen-bonded complexes of the S22 set, this time for a number of force field methods without and with augmentation by our third-generation hydrogen-bonding correction. Force field interaction energies for the "frozen" geometries of the S22 and S22x4 sets were kindly provided by the authors of ref 20, from their extensive study on the performance of force field methods for noncovalent interactions. The force fields are the MacroModel implementations of MM2*,[45] MM3*,[46] AMBER*,[47−49] and OPLS*[50] and native versions of OPLSAA[51] and MMFF94.[52] Further details can be found in the original publication.[20]

Perusing Tables 4−9, the following conclusions can be drawn: All six tables illustrate that even dispersion-corrected semiempirical QM methods perform quite badly for hydrogen-bonding interactions (a known issue, see Introduction). While OM3 is doing rather well, AM1 especially gives large errors for H-bond interaction energies. Tables 4 and 6−9 also show that the inclusion of the second-generation "H2" correction (in combination with standard dispersion corrections such as "DH2") consistently improves the accuracy of all methods, but unfortunately DH2 suffers from several conceptual problems (as explained in the theory section above). Our new "H+" correction (in combination with standard dispersion corrections such as "DH+") is able to reach the same overall accuracy as the DH2 correction, while it avoids all of the conceptual problems connected with the DH2 ansatz: Tables 4 and 7−9 show that, e.g., MUEs are improved by a factor of 1.5 to 3 for all sets (with significantly strong H-bonding interactions, unlike the peptide set in Table 9) and methods.

That this is also the case for nonequilibrium structures can be seen in Table 6, and that this conclusion still holds for structures optimized at the corresponding level of theory is shown in Table 5. [The used benchmark sets are designed to be used with the given benchmark geometries, because the goal is a correct energetic description within the correct geometrical arrangement. For a comparison of interaction energies of structures optimized at different levels, it is necessary to carefully check and compare all final geometries and resulting energetic effects. To allow for comparison with earlier work, we present this data here for the S22 set—where for both DH2 and DH+ no substantial change of binding motifs occurs—but stick with the intended use of the benchmark sets in the other cases.]

While the overall accuracies of DH2 and DH+ are very similar, DH2 seems to do better for AM1, while DH+ seems to be better suited for OM3 (see Tables 7 and 8). The performance for DFTB is better with DH2 for the DH1 training set (Table 7) but better with DH+ for the JSCH2005 set, presumably because the latter one includes multiple hydrogen bonds. In addition, while DH2 works exceptionally well for the S26 and S26+S22x4 fit sets (as three parameters for each method were added to achieve exactly this goal), this additional gain in accuracy is not transferred to systems beyond the fit sets, e.g., the diverse hydrogen-bonded structures of the PM6-DH1 set (Table 7), where DH+ is at the same level, and especially not the DNA base pairs (Table

Table 11. Results for Several Methods for the Hydrogen-Bonded Complexes of the S26 Set (S22 Set for Force Field Methods)[a]

methods	MUE	average error per H bond
MM2*	8.77	5.0
MM2*-H+	2.12	1.2
MMFF94	3.15	1.8
MMFF94-H+	0.84	0.5
SCC-DFTB-D	1.75	1.2
SCC-DFTB-DH+	1.01	0.7
PM6-D	2.82	1.9
PM6-DH+	0.66	0.4
OM3-D	1.75	1.2
OM3-DH+	0.62	0.4
B3LYP-D	0.74	0.5

[a] Mean unsigned errors (MUEs) and average errors per H bond with respect to the benchmark CCSD(T)/CBS interaction energies are presented. DFT methods with TZVP basis sets. All values in kiloclaories per mole.

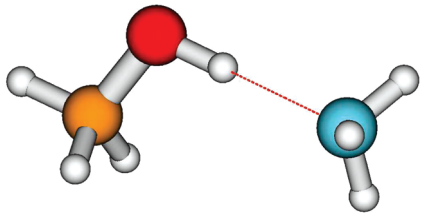


Figure 3. Simple Model System for Proton Transfer.

8), where DH+ (two fit parameters) even outperforms DH2 (eight fit parameters). The peptide systems in Table 9 were added to show that the DH+ correction (as the DH2 correction) does not worsen the energies of complexes with very weak hydrogen bonds by introducing unphysical contributions.

Table 10 shows that the overall good performance of the "H+" correction is also transferable to FF methods, with MUEs for the hydrogen-bonded systems of the S22 set nearly comparable to much more sophisticated computational approaches. Tests additionally including the nonequilibrium structures from the S22x4 set showed a reduction of the mean unsigned error (MUE) from 6.5 to 2.5 kcal/mol for MM2* and from 2.1 to 0.9 kcal/mol for MMFF94, quite comparable to the gain shown in Table 10 for equilibrium structures only. More thorough studies are in preparation but are beyond the scope of this work.

For a large number of investigated cases, the new "DH+" correction reaches the accuracy of DFT-D approaches, while being several orders of magnitude faster and now free of the conceptual problems of the older DH2 correction scheme. This is again summarized in Table 11 where different force field and semiempirical QM methods as well as a "standard" DFT-D approach are compared for the MUE and average error per H bond over the hydrogen-bonded systems of the S26 set (s22 set for force field methods).

To illustrate the practical applicability of DH+ to model proton transfer reactions, we have looked at the simple model reaction of methanol and ammonia visualized in Figure 3. Starting from a MP2/TZVP optimized structure, all coordinates are kept frozen except the O−H distance, which is varied between 1.0 and 1.9 Å in steps of 0.025 Å, corresponding to a proton transfer from methanol to ammonia.

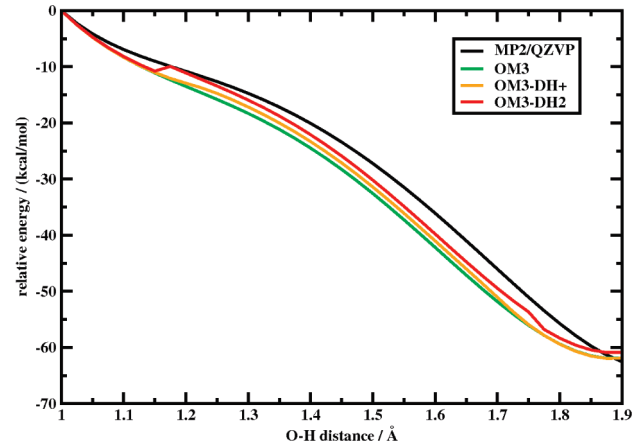


Figure 4. OM3(-DH2/-DH+) proton transfer energetics for the model system from Figure 3, with MP2/QZVP reference data (energies at 1 Å taken as a reference point).

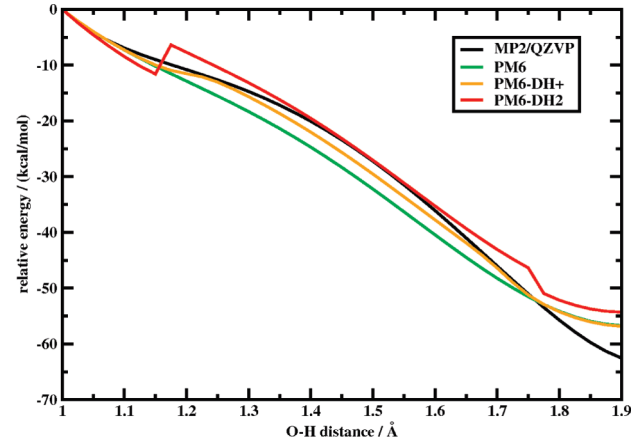


Figure 5. PM6(-DH2/-DH+) proton transfer energetics for the model system from Figure 3, with MP2/QZVP reference data (energies at 1 Å taken as a reference point).

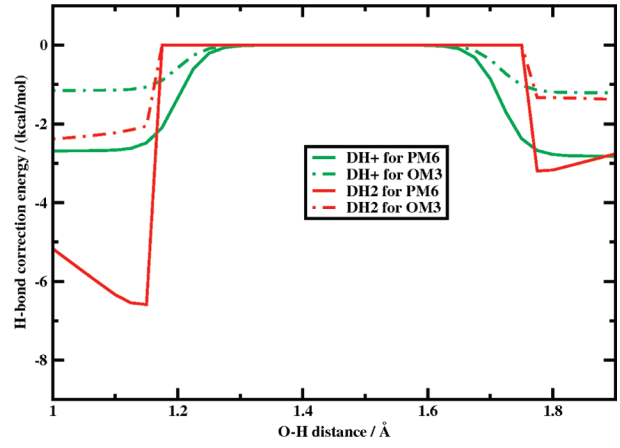


Figure 6. DH2 and DH+ hydrogen-bonding correction energies for the model system from Figure 3 (with parametrization corresponding to the OM3 and PM6 methods).

Figures 4 and 5 show the resulting proton transfer energetics for OM3 and PM6 without and with the DH2 and DH+ corrections, illustrating that in opposition to DH2, DH+ does not break down in the case of proton transfer. Apart from that, the impact of both corrections is small in comparison with the overall reaction energy, with a cor-

Table 12. Selected Hydrogen-Bond Interaction Energies from Linear, Hydrogen-Bonded Formamide Chains, As Well As B3LYP/D95**, MP2/TZVP, and MP2/QZVP Reference Data[a]

	OM3	OM3-DH2	OM3-DH+	PM6	PM6-DH2	PM6-DH+	B3LYP/D95**[b]	MP2/TZVP	MP2/QZVP
dimer	−5.13	−6.31	−6.57	−5.36	−6.71	−7.81	−7.31	−6.65	−6.47
hexamer terminal	−6.84	−8.27	−8.27	−7.17	−8.82	−9.56	−10.04	−8.66	
hexamer central	−9.05	−10.82	−10.39	−9.27	−11.33	−11.45	−13.20	−11.26	

[a] All values are in kilocalories per mole. [b] From ref 53.

respondingly small, indirect effect on the "barrier" height through the energetic lowering of reactants and products. Figure 6 shows a direct comparison of the DH2 and DH+ correction energies for our simple model reaction (using the corresponding the OM3 and PM6 parameters), illustrating the problems of DH2 and the conceptual improvement of DH+ in more detail.

Also of some importance is the performance of DH+ for hydrogen-bonding cooperativity, because such a type of cooperativity has been shown to exist in β sheets, which makes it important for the accurate modeling of large proteins, where the stability of secondary structures might be influenced.[53] Table 12 shows interaction energies for selected hydrogen bonds in linear formamide chains of lengths two and six, a model system taken from the work of Dannenberg and co-workers.[53,54] Besides OM3(-DH2/DH+) and PM6(-DH2/DH+) data, DFT and MP2 reference values are given. (Following ref 53, interaction energies are calculated by simple subtraction; e.g., the energy of the terminal H-bond in the hexamer is taken to be the energy of the hexamer less the combined energies of the pentamer and the monomer.)

First of all, Table 12 illustrates the strong cooperative nature of the H-bond interactions (emphasized already in the above-mentioned work by Dannenberg); i.e., the central interaction in the hexamer is predicted to be nearly two times as much as the dimer interaction. Comparing OM3 and PM6 with MP2, it looks as if semiempirical methods seem to be rather well capable of modeling such hydrogen-bond cooperativity effects, especially also concerning the ratio of interaction strengths (with a factor of 1.6 and 1.8 between the dimer and the central hexamer H-bond strength for all approaches). DH2 and DH+ show again a very similar performance, systematically improving the underlying SE methods, with DH2 being slightly more advantageous at least for PM6, presumably because DH2 has one parameter specially dedicated to amide interactions and fitted to the formamide dimer. Overall, empirical correction schemes seem to also work surprisingly well with semiempirical QM methods for hydrogen-bonding cooperativity effects.

5. Conclusions

This work presents a further improved, "third-generation" hydrogen-bonding correction scheme that can now be generally included in parameter fits of semiempirical QM and force field methods, as it does not suffer any longer from several conceptual limitations of previous approaches in this direction: hydrogen bonds are now treated as an interaction term between electronegative acceptor and donor atoms, "weighted" by a function of the positioning of H atoms between them. This way, the new correction scheme improves over existing ones with regard to the following issues: Electronic structure change (e.g., proton transfer) becomes generally possible; a safe long-range behavior is enforced; exact analytical gradients are affordable;, transferability to force field methods is achieved, and straightforward extendability for other hydrogen- (and halogen-)bonding types is given; and the same (high) overall accuracy can be achieved with significantly less parametrization. Our new correction scheme consistently improves the accuracy of the semiempirical QM methods PM6, AM1, OM3, and SCC-DFTB as well as the MM2*, MM3*, AMBER*, OPLS*, OPLSAA, and MMFF94 force field methods for several benchmark sets of hydrogen-bonding interactions by up to 1 order of magnitude at the cost of a force-field-type calculation.

Acknowledgment. The author would like to thank Pavel Hobza for introducing the author to hydrogen-bonding correction schemes and Jonathan Goodman for kindly supplying force field data and advice. This work was supported by Grant LPDS-2009-19 from the German National Academy of Science Leopoldina.

Supporting Information Available: Geometries for the proton transfer and hydrogen-bond cooperativity model systems. This material is available free of charge via the Internet at http://pubs.acs.org

References

(1) Jorgensen, W. L. *Acc. Chem. Res.* **2009**, *42*, 724.

(2) Jorgensen, W. L. *Science* **2004**, *303*, 1813.

(3) Klebe, G. *Drug Discovery Today* **2006**, *11*, 580.

(4) Möhle, K.; Hofmann, H.-J.; Thiel, W. *J. Comput. Chem.* **2001**, *22*, 509.

(5) Elstner, M.; Jalkanen, K. J.; Knapp-Mohammady, M.; Frauenheim, T.; Suhai, S. *Chem. Phys.* **2001**, *263*, 203.

(6) Nikitina, E.; Sulimov, V.; Zayets, V.; Zaitseva, N. *Int. J. Quantum Chem.* **2004**, *97*, 747.

(7) Vasilyev, V.; Bliznyuk, A. *Theor. Chem. Acc.* **2004**, *112*, 313.

(8) Villar, R.; Gil, M. J.; Garcia, J. I.; Martinez-Merino, V. *J. Comput. Chem.* **2005**, *26*, 1347.

(9) Raha, K.; Merz, K. M., Jr. *J. Med. Chem.* **2005**, *48*, 4558.

(10) Nikitina, E.; Sulimov, V.; Grigoriev, F.; Kondakova, O.; Luschekina, S. *Int. J. Quantum Chem.* **2006**, *106*, 1943.

(11) Raha, K.; Peters, M. B.; Wang, B.; Yu, N.; Wollacott, A. M.; Westerhoff, L. M.; Merz, K. M., Jr. *Drug Discovery Today* **2007**, *12*, 725.

(12) Thiriot, E.; Monard, G. *THEOCHEM* **2009**, *898*, 31.

(13) Wollacott, A. M.; Merz, K. M., Jr. *J. Chem. Theory Comput.* **2007**, *3*, 1609.

(14) van der Vaar, A.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **1999**, *121*, 9182.

(15) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.

(16) McNamara, J. P.; Hillier, I. H. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2362.

(17) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1436.

(18) Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555.

(19) Rao, L.; Ke, H.; Fu, G.; Xu, X.; Yan, Y. *J. Chem. Theory Comput.* **2009**, *5*, 86.

(20) Paton, R. S.; Goodman, J. M. *J. Chem. Inf. Model.* **2009**, *49*, 944.

(21) Dannenberg, J. J. *THEOCHEM* **1997**, *401*, 279.

(22) Csonka, G. I.; Angyan, J. G. *THEOCHEM* **1997**, *393*, 31.

(23) Clark, T. J. *THEOCHEM.* **2000**, *530*, 1.

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

(25) Bernal-Uruchurtu, M. I.; Ruiz-Lopez, M. F. *Chem. Phys. Lett.* **2000**, *330*, 118.

(26) Monard, G.; Bernal-Uruchurtu, M. I.; Van Der Vaart, A.; Merz, K. M., Jr.; Ruiz-Lopez, M. F. *J. Phys. Chem. A* **2005**, *109*, 3425.

(27) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601.

(28) Yang, Y.; Yu, H.; York, D.; Cui, Q.; Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 10861.

(29) Wang, Q.; Bryce, R. A. *J. Chem. Theory Comput.* **2009**, DOI: 10.1021/ct9002674.

(30) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173.

(31) Stewart, J. P. *J. Mol. Model.* **2009**, *15*, 765.

(32) Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1749.

(33) OPENMOPAC. www.openmopac.net (accessed Aug 31, 2009).

(34) Korth, M.; Pitonak, M.; Rezac, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 344.

(35) Foster, M. E.; Sohlberg, K. *J. Chem. Theory Comput.* **2010**, *6*, 2153.

(36) DFTBplus. http://www.dftb-plus.info (accessed Aug 31, 2009).

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

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

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

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

(41) Weigend, F.; Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *119*, 12753.

(42) Eichhorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *242*, 652.

(43) Eichhorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119.

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

(45) Allinger, N. L. *J. Am. Chem. Soc.* **1977**, *99*, 8127.

(46) Allinger, N. L.; Yuh, Y. H.; Lii, J.-H. *J. Am. Chem. Soc.* **1989**, *111*, 8551.

(47) McDonald, D. Q.; Still, W. C. *Tetrahedron Lett.* **1992**, *33*, 7743.

(48) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S., Jr.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765.

(49) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *J. Comput. Chem.* **1986**, *7*, 230.

(50) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657.

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

(52) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490.

(53) Kobko, N.; Paraskevas, L.; del Rio, E.; Dannenberg, J. J. *J. Am. Chem. Soc.* **2001**, *123*, 4348.

(54) Kobko, N.; Dannenberg, J. J. *J. Phys. Chem. A* **2003**, *107*, 10389.

CT100408B

JCTC Journal of Chemical Theory and Computation

Conformational Energies of DNA Sugar−Phosphate Backbone: Reference QM Calculations and a Comparison with Density Functional Theory and Molecular Mechanics

Arnošt Mládek,[†] Judit E. Šponer,[†] Petr Jurečka,[‡] Pavel Banáš,[‡] Michal Otyepka,[‡] Daniel Svozil,*[,†,§] and Jiří Šponer*[,†,∥,⊥]

Institute of Biophysics, Academy of Sciences of the Czech Republic, Královopolská 135, 612 65 Brno, Czech Republic, Department of Physical Chemistry, Faculty of Science, Palacký University, 771 46 Olomouc, Czech Republic, Laboratory of Informatics and Chemistry, Faculty of Chemical Technology, Institute of Chemical Technology, Technická 3, 166 28 Prague 6, Czech Republic, Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Flemingovo náměstí 2, 166 10 Prague 6, Czech Republic, and National Centre for Biomolecular Research, Faculty of Science, Masaryk University, 611 37 Brno, Czech Republic

Received August 17, 2010

Abstract: The study investigates electronic structure and gas-phase energetics of the DNA sugar−phosphate backbone via advanced quantum chemical (QM) methods. The analysis has been carried out on biologically relevant backbone conformations composed of 11 canonical BI-DNA structures, 8 pathological structures with α/γ torsion angles in the g+/t region, and 3 real noncanonical γ-trans structures occurring in the loop region of guanine quadruplex DNA. The influence of backbone conformation on the intrinsic energetics was primarily studied using a model system consisting of two sugar moieties linked together via a phosphodiester bond (SPSOM model). To get the conformation of the studied system fully under control, for each calculation we have frozen majority of the dihedral angles to their target values. CCSD(T) energies extrapolated to the complete basis set were utilized as reference values. However, the calculations show that inclusion of higher-order electron correlation effects for this system is not crucial and complete basis set second-order perturbation calculations are sufficiently accurate. The reference QM data are used to assess performance of 10 contemporary density functionals with the best performance delivered by the PBE-D/TZVPP combination along with the Grimme's dispersion correction, and by the TPSS-D/6-311++G(3df,3pd) augmented by Jurečka's dispersion term. In addition, the QM calculations are compared to molecular mechanics (MM) model based on the Cornell et al. force field. The destabilization of the pathological g+/t conformers with respect to the reference canonical structure and the network of intramolecular CH···O interactions were investigated by means of natural bond orbital analysis (NBO) and atoms-in-molecules (AIM) Bader analysis. Finally, four additional model systems of different sizes were assessed by comparing their energetics to that of the SPSOM system. Energetics of smaller MOSPM model consisting of a sugar moiety linked to a phosphate group and capped with methyl and methoxy group on the 5′- and 3′-ends, respectively, is fairly similar to that of SPSOM, while the role of undesired intramolecular interactions is diminished.

10.1021/ct1004593
Published on Web 11/16/2010

Introduction

Nucleic acids (NA; DNA and RNA) consist of linear chains of covalently bound sugar−phosphate units to which aromatic nucleic acid bases are attached. Nucleic acids form an astonishing variability of tertiary structures (ranging from simple double helices to complex ribonucleoprotein particles) that determine their function. Structural dynamics of nucleic acids result from delicate balance of numerous contributions. Among them, conformational preferences of the sugar−phosphate backbone belong to the most important ones. The sugar−phosphate backbone is chemically monotonous (sequence-independent). It contains a number of consecutive single bonds, which allow a substantial freedom for dihedral rotations. Thus, it has often been assumed that the backbone plays a rather passive role in structuring nucleic acids, while interactions involving the nucleobases are decisive (the base-centered view of NA structure).[1,2] On the other hand, there have also been suggestions that the internal backbone conformational preferences are decisively important.[3−5]

The conformation of the backbone is defined by a number of torsion angles called α, β, γ, δ, ε, and ζ (Figure 1). In addition to these, the precise conformation of inherently nonplanar (puckered) five-membered deoxyribose sugar ring is fully specified by five internal torsions $\tau_0-\tau_4$, the set of which can be simplified to only two internal degrees of freedom (pseudorotation P and amplitude τ_{max}). As the individual bases in nucleic acids are considered to be flat, the last degree of freedom in nucleotide is represented by the glycosidic bond linking a deoxyribose sugar and a base, the rotation around which is characterized by the torsion angle χ. Steric restrictions confine the values of these structural descriptors to discrete ranges.[6] A common convention for describing these backbone angles is to term values of ∼60° as gauche+ (g+), ∼ −60° as gauche− (g−), and ∼180° as trans (t). The typical (i.e., average) values of torsion angles of individual conformers can be obtained by careful analysis of NMR or crystal structures.[7] Such studies identify not only ranges individual torsions can adopt, but also describe numerous correlations involving pairs of the backbone torsion angles, as well as sugar pucker and glycosidic angle. The X-ray database contains hundreds of high-resolution DNA X-ray structures revealing thousands of individual dinucleotide backbone topologies. Using advanced bioinformatics tools, it is possible to cluster the backbone topologies into typical conformation families and to determine their representative (i.e., averaged) geometries with a high degree of confidence.[7] The existence of correlations is important as it means that the atomic motions in nucleotides follow concerted pattern of interdependence. Between the most important correlations belong the correlation between sugar pucker and glycosidic angle χ, the correlation between γ and α torsions, and the correlation between sugar pucker and δ angle (this correlation is rather strong, as it is given by the fact that one of the sugar internal torsions represents the rotation around the same bond as does δ).

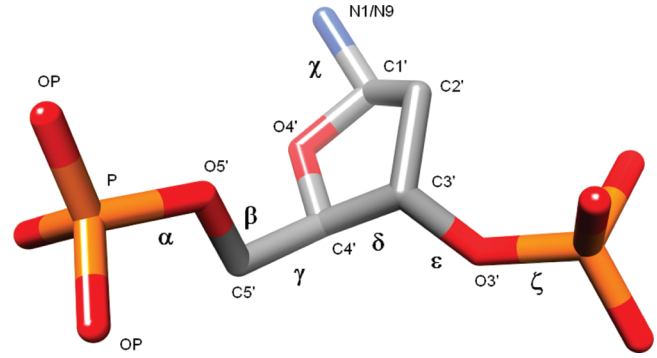


Figure 1. Atomic numbering and definition of the deoxyribonucleotide backbone torsion angles. The nucleotide backbone is described by the P−O5′−C5′−C4′−C3′−O3′ linkage. The torsion angles represent the rotation around the given bond. It is conventional to describe the backbone torsion angles of ∼60° as gauche+ (g+), of ∼300° as gauche− (g−), and of ∼180° as trans (t). The standard progression of NA chain is the 5′→3′ direction, which is from the left to the right in this particular figure.

Because of the inherent conformational flexibility of the polynucleotide backbone, there exists a wide range of different double helical conformations. The most common form is the antiparallel right-handed B-DNA double helix.[8] This conformation, often referred to as canonical one, is characterized by the following set of typical torsion angles: α = 299° (g−), β = 179° (t), γ = 48° (g+), δ = 133°, ε = 182° (t), and ζ = 263°. Another possible right-handed form, A-DNA, is similar to B-DNA, but with different sugar conformation leading to different base position with respect to the helical axis. Z-DNA, a left-handed form, was also prepared, although its biological relevance is still the subject of investigation.[9] The structural variability of DNA is critical for recognition between DNA and proteins, which plays a crucial role in such essential processes as replication or transcription. Characterization of the backbone conformational space is therefore highly important for understanding of DNA recognition.

However, while intrinsic energetics of interbase interactions has been widely studied,[10,11] very little is known about the backbone electronic structure and energetics. Their study is considerably more difficult[12−22] due to the high flexibility and various correlations between the individual torsion angles. The ability to uniquely assign and compare energies of individual biomolecular conformers is indispensable to clarify their conformational preferences. The intrinsic conformational preferences are established by analysis of the relation between molecular structures and molecular energies. Using computational methods, we can derive potential energy surfaces (PES) by assigning corresponding electronic energy to each single geometry, and thus evaluating energy as an unambiguous function of the molecular structure. The resultant potential function that drives a biomolecular system arrangement in natural environment can be generally ex-

* Corresponding author e-mail: daniel.svozil@gmail.com (D.S.), sponer@ncbr.chemi.muni.cz (J.S.).
† Institute of Biophysics, Academy of Sciences of the Czech Republic.
‡ Palacký University.
§ Institute of Chemical Technology.
|| Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic.
⊥ Masaryk University.

pressed as a sum of three distinct terms. The first term, intrinsic electronic energy of the system in vacuum, can be acquired exclusively via a highly accurate ab initio QM treatment. The intrinsic energy component can also be approximated by means of balanced MM force fields, where the electronic effects are mimicked by effective classical potential functions. The second contribution to the potential energy function arises from the interaction of the studied molecular entity with its natural environment, that is, mainly the solvation effects and the overall context of the NA molecule. The last term represents coupling between the intrinsic and environmental contributions.

The present work is focused on the first energy component, for it influences the resulting conformational space occupation to a considerable extent. We provide an extensive QM characterization of the intrinsic conformational preferences of the sugar−phosphate unit of DNA. It is, however, important to bear in mind that the environment and coupling terms significantly affect the resulting energetics too, and thus should not be omitted when predicting structural preferences in real environments. This is the main limitation of our study.

The present study has several key features. First, we derive reference QM data; that is, we push the theoretical calculations to the highest limits achievable by contemporary computational tools. These calculations are then used for comparison with a wide range of less expensive QM methods and also variants of the Cornell et al. MM force field,[23] to assess their performance and to obtain basic physical chemistry insights into the systems under study.

We compare five model systems of different complexity, to establish sensitivity of the results to the choice of the model system. The very first critical step in a theoretical investigation of biologically relevant macromolecules is to select a convenient model system. The size and the overall structural complexity of biological units to be described inherently delimit the set of appropriate model systems. The model should be large enough to capture all important conformational and electronic characteristics of the studied biomolecule. Too large model can make rigorous ab initio high-level QM investigation intractable, while too small model system may be chemically irrelevant. As the backbone conformation is unambiguously determined by six strongly coupled torsion angles α, β, γ, δ, ε, and ζ (Figure 1),[7] the energetic assessment of the complete conformational space constitutes a complex six-dimensional problem. Furthermore, the sugar pucker defined by two independent variables and coupled to a certain degree with the δ torsion angle also influences PES of the model compound. In addition, the correct description of the electronic distribution along the negatively charged backbone is a rather demanding task. To properly analyze the polarizable anionic nature of the backbone, at least moderate size basis set with diffuse functions must be utilized. This makes high-level QM computations difficult even for quite small (e.g., one nucleotide) DNA fragments. Besides, the diffuse electron density due to anionic character causes slower density matrix convergence. However, there is yet another reason that complicates investigations of larger model systems, even when they are computationally tractable per se. The larger is the model system, the more prone it is to adopt geometries that are biased, for example, by intramolecular H-bonds and other interactions that are not relevant to complete solvated biomolecules. The complexity of the conformational spaces increases dramatically with the size of the system. Additional issues that preclude studies of larger systems are uncompensated charges of multiple phosphate groups that would dominate the gas-phase electrostatics and also an artifact known as intramolecular basis set superposition error (BSSE; for more details, see the QM calculations paragraph below).

In contrast to our preceding study,[12] we modified the computations in such a way that basically we always keep all backbone dihedral angles frozen at predefined values. This has been necessitated by the fact that when freezing only very few dihedral angles the remaining free dihedral angles can adopt numerous combinations as local minima. This substantially biases conformational scans and often leads to unrealistic geometries.

We concentrate our efforts to three DNA backbone geometrical substates: canonical B-DNA conformation and two conformations with γ angle in trans and α angle in gauche+ region. The first one corresponds to α/γ topology that normally should not occur in free B-DNA;[24] however, it accumulated in longer simulations with an earlier version of the Cornell et al. MM force field. With the original parametrization (ff94[23] and ff99[25] force fields), irreversible sequence-independent α/γ flips occur in long molecular dynamics (MD) simulations of B-DNA duplexes.[26−28] The accumulation of α/γ transitions then causes entire B-DNA structure degradation.[29] The backbone γ torsional profile was recently reparametrized in the parmbsc0 force field,[29] which allows for long time scale simulations of B-DNA duplexes without any significant loss of helical structure. Although the parmbsc0 force field represents a decisive progress in MD simulations of DNA, further refinement would be still useful as not all imbalances are yet treated sufficiently.[30−32] It can be illustrated with the second γ-trans geometry investigated in this study, which has been experimentally observed in the first nucleotide of single-stranded loops of human telomere guanine quadruplex (G-DNA) and thus corresponds to real substate of DNA backbone.[33,34] However, it has been destabilized by the parmbsc0 force field.[31]

Methods and Model Systems

Model Systems. To fully capture the conformational behavior of the sugar−phosphate backbone, at least the dinucleotide unit must be utilized. However, the complexity of the potential model systems is limited by the fact that, to avoid strong electrostatic repulsion between two anionic phosphate groups, which is otherwise screened by the solvent and counterions in the physiological environment, only one phosphate group is conceivable. In addition, stacking interactions between nucleobases in dinucleotide model systems would result in a significant conformational bias, and the bases were thus excluded from the model systems (see the discussion below). The model systems used in the present study can be divided into two groups.

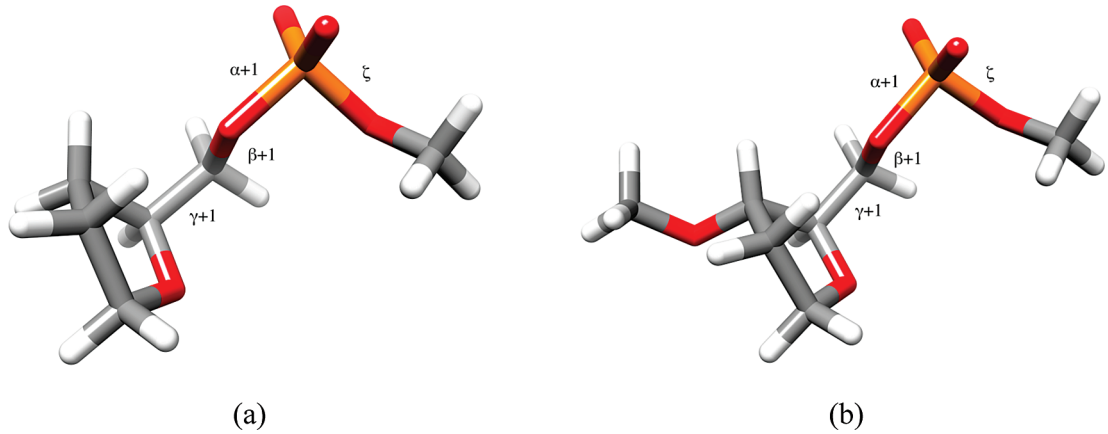


Figure 2. (a) SPM (sugar−phosphate−methyl) and (b) MOSPM (methyl−oxygen−sugar−phosphate−methyl) Group I model compounds. Labeling of the bonds is according to the standard nucleic acid nomenclature. The ζ torsion angle is defined as C3′−O3′−P−O5′(n+1), the α+1 torsion angle as O3′−P−O5′(n+1)−C5′(n+1), the β+1 torsion angle as P−O5′(n+1)−C5′(n+1)−C4′(n+1), and the γ+1 torsion is defined as O5′(n+1)−C5′(n+1)−C4′(n+1)−C3′(n+1).

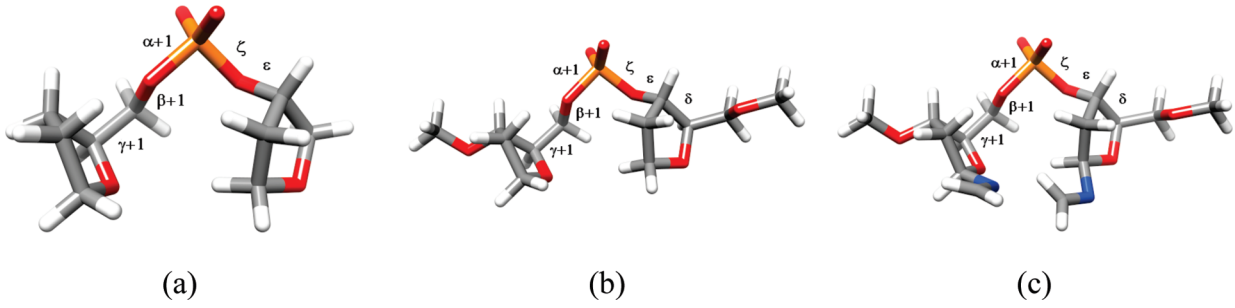


Figure 3. (a) T3PS (tetrahydrofuran with 3′ phosphate with a capping sugar), (b) SPSOM (sugar−phosphate−sugar with capping methoxy groups), and (c) SPSOM-NCH2 (sugar−phosphate−sugar with capping methoxy groups and with $-N{=}CH_2$ groups) Group II models. Labeling of the bonds is according to the standard nucleic acid nomenclature. The δ torsion angle is defined as C5′−C4′−C3′−O3′, the ε torsion angle as C4′−C3′−O3′−P, the ζ torsion as C3′−O3′−P−O5′(n+1), the α+1 torsion angle as O3′−P−O5′(n+1)−C5′(n+1), the β+1 torsion angle as P−O5′(n+1)−C5′(n+1)−C4′(n+1), and the γ+1 torsion is defined as O5′(n+1)−C5′(n+1)−C4′(n+1)−C3′(n+1). The 3′-end (i.e., n+1) and 5′-end sugars are on the left and right, respectively, because the 5′−3′ direction is from right to left in this figure (opposite of Figure 1).

Group I (Figure 2) consists of two organophosphate models with only one sugar residue. The first system, SPM, (the abbreviation stands for sugar−phosphate−methyl; see Figure 2a), consists of a sugar residue and a methyl group linked via a phosphodiester bond as suggested by Orozco et al.[29] The second model compound, MOSPM, (i.e., methyl−oxygen−sugar−phosphate−methyl; see Figure 2b), represents an extended version of the SPM model with the H3′(n+1) atom replaced by a methoxy functional group (−O−CH3). In this system, the sugar moiety is situated in a more natural chemical environment because of the longer backbone fragment. To prevent formation of artificial intramolecular hydrogen bonds, the 3′ and 5′ ends of MOSPM, as well as the 5′ end of SPM, were terminated with methyl groups.

Group II (Figure 3) contains three models with the sugar−phosphate−sugar unit mimicking a dinucleotide step. The simplest system containing no additional groups was designed and used by MacKerell[13] and is further referred to as T3PS (tetrahydrofuran with 3′ phosphate with a capping sugar). Another model compound, SPSOM (sugar−phosphate−sugar−oxygen−methyl), used in our previous study[12] retains all the features characteristic for the DNA dinucleotide building blocks. The last, and the most complex, model[12] referred to as SPSOM-NCH2 was constructed out of SPSOM model by replacing H1′ of both sugar residues with a methylene-imino functional group ($-N{=}CH_2$). This extension was proposed to compensate, at least partially, for the neglect of the nucleobases, leading to a better description of the hyperconjugation effects along the sugar−phosphate backbone.[12]

Starting Geometries of the Conformers. All canonical (BI) DNA backbone geometries (geometries at and around the canonical structure) were labeled with the symbol “a”; that is, the particular conformers in the canonical series are referred to as a1, a2, ..., a11 (Table 1). The typical torsion angles for this most populated substate of free B-DNA duplex (α = 299° (g−), β = 179° (t), γ = 48° (g+), δ = 133°, ε = 182° (t), and ζ = 263°) were obtained by the means of analysis of crystal structures of 1531 dinucleotide steps in DNA.[7] The representative of the “average” BI-DNA and simultaneously the reference structure in this study is labeled a1; the remaining 10 geometries (a2-a11) were prepared to characterize PES in the vicinity of the above “average” BI structure. The a1 structure is the best representative of the BI cluster of geometries.

The first group of noncanonical DNA backbone geometries occupying the less populated α/γ = g+/t conformational

Table 1. Torsion Angle Values for the Canonical "a" and Noncanonical "b" Geometries

structure label	$\alpha+1^c/\gamma+1^a$	$\gamma+1^a$	$\beta+1^b$	$\alpha+1^c$	ζ^d	ε^e	δ^f	targeted perturbation
a1	g−/g+	45	180	300	260	180	136	BI; no perturbation
a2		50	180	280	270	190	136	$(\alpha+1)-20$
a3		40	180	320	260	180	136	$(\alpha+1)+20$
a4		45	170	300	260	190	136	$(\beta+1)-10$
a5		45	190	300	260	180	136	$(\beta+1)+10$
a6		35	180	310	260	185	136	$(\gamma+1)-10$
a7		55	180	295	260	180	136	$(\gamma+1)+10$
a8		50	180	300	260	190	116	δ: C1′-exo
a9		45	180	295	260	190	150	δ: C3′-exo
a10		50	190	310	270	160	136	$\varepsilon-20$
a11		45	170	300	260	200	130	$\varepsilon+20$
b1	g+/t	195	225	65	190	250	145	no perturbation
b2		200	220	50	190	265	150	$(\alpha+1)-15$
b3		190	220	80	190	225	140	$(\alpha+1)+15$
b4		190	215	60	190	260	150	$(\beta+1)-10$
b5		195	235	70	190	240	140	$(\beta+1)+10$
b6		205	225	60	190	250	145	$(\gamma+1)+10$
b7		195	220	75	210	235	145	$\zeta+20$
b8		190	220	65	180	255	145	$\zeta-10$

[a] $\gamma+1$: O5′(i+1)−C5′(i+1)−C4′(i+1)−C3′(i+1). [b] $\beta+1$: P(i+1)−O5′(i+1)−C5′(i+1)−C4′(i+1). [c] $\alpha+1$: O3′(i)−P(i+1)−O5′(i+1)−C5′(i+1). [d] ζ: C3′(i)−O3′(i)−P(i+1)−O5′(i+1). [e] ε: C4′(i)−C3′(i)−O3′(i)−P(i+1). [f] δ: C5′(i)−C4′(i)−C3′(i)−O3′(i).

Table 2. Torsion Angle Values for the Quadruplex Loop "q" Geometries

structure label	$\alpha+1^c/\gamma+1^a$	$\gamma+1^a$	$\beta+1^b$	$\alpha+1^c$	ζ^d	ε^e	δ^f	PDB code/NDB code	residue ID
q1	g+/t	176	189	77	73	230	143	1KF1/UD0017	DT11
q2		183	190	79	61	224	152	1KF1/UD0017	DT05
q3		195	184	63	64	220	147	1KF1/UD0017	DT17

[a] $\gamma+1$: O5′(i+1)−C5′(i+1)−C4′(i+1)−C3′(i+1). [b] $\beta+1$: P(i+1)−O5′(i+1)−C5′(i+1)−C4′(i+1). [c] $\alpha+1$: O3′(i)−P(i+1)−O5′(i+1)−C5′(i+1). [d] ζ: C3′(i)−O3′(i)−P(i+1)−O5′(i+1). [e] ε: C4′(i)−C3′(i)−O3′(i)−P(i+1). [f] δ: C5′(i)−C4′(i)−C3′(i)−O3′(i).

region is analogously labeled with symbol "b", that is, b1, b2, ..., b8 (Table 1). Note that labeling the two series of structures as "a" or "b" has nothing to do with the helix form designation. The second subset of α/γ = g+/t structures from human telomeric quadruplex loops[33,34] is labeled with symbol "q" and consists of three members denoted as q1, q2, and q3 (Table 2).

The geometries in the present Article were derived in the following way. At the beginning, we have taken two "parent" structures of the SPSOM model optimized at the B3LYP/6-31+G(d) level in our earlier paper (Supporting Information of ref 12, g−/g+ (p S13) and g+/t (p S15) structures). These structures (named spsom_a and spsom_b) served as the initial structures for the derivation of the "a" and "b" subsets of structures of the SPSOM system in the present Article. Next, we have modified (using modredundant route section keyword of the Gaussian 03 software[35]) all dihedral angles to the desired values (Table 1) to obtain structures a1−a11 and b1−b8, which were subsequently optimized at the respective theoretical levels (see below). Note that the actual final geometries derived in this Article are not affected by details of the two initial "parent" geometries, as we set up and constrained the dihedral angles upon optimizations.

However, due to the correlations involving pairs of backbone angles (known from the X-ray database study), the shift of each torsion angle from its canonical value introduces also the changes in values of other torsional angles. To cover these changes, each time when the target torsion angle was shifted from the canonical a1 structure, the remaining torsion angle values were adjusted accordingly to reflect correlation of torsion angles suggested by the X-ray database study (Table 1).[7] As this procedure takes the backbone torsion angles correlation into account, we suppose it is better for sampling the PES of the real DNA molecule than just keeping the remaining torsions fixed at their canonical values.

The geometries of the three q-conformers were prepared from the crystal structure of human telomeric quadruplex loops (pdb code: 1KF1, resolved at the 2.10 Å resolution) by extracting the corresponding SPSOM segments (Table 2) from the three independent loop structures. The addition of hydrogen atoms was carried out manually using Accelrys ViewerPro molecular modeling software. Their initial positions were adjusted according to the hybridization state of the linked heavy atom. The structures were then optimized at the respective theoretical levels (see below) with frozen dihedrals.

The other models, T3PS, SPM, MOSPM, and SPSOM-NCH2, were derived in the following manner. We have taken the a1, b1, and qx (x = 1, 2, and 3) MP2-optimized SPSOM geometries (Table 3). These were appropriately chemically modified to get the other model systems. These geometries served as "parent" structures for the T3PS, SPM, MOSPM, and SPSOM-NCH2 models (Table 3; parent structure name). These parent structures are not reoptimized after modification. The corresponding a1−a11 and b1−b8 structures were then derived from these parent structures by setting up the required combination of dihedral angles (Table 1) and subsequent constrained optimizations. The starting geometries, that is, all the above-noted parent structures of all model systems for the canonical "a", noncanonical "b", and quadruplex "q" variant (Table 3), are given in the Supporting

Table 3. List of Parent Geometries with Their Names and Origin[a]

model label	α/γ family	parent structure name	origin of the geometry
SPSOM	g−/g+	spsom_a	g−/g+ SPSOM[b]
	g+/t	spsom_b	g+/t SPSOM[b]
	g+/t G-DNA	spsom_q*x*[d]	quadruplex loop, NDB, UD0017; PDB, 1KF1[c]
T3PS	g−/g+	t3 ps_a	a1 SPSOM||MP2/6-31+G(d), with modifications
	g+/t	t3 ps_b	b1 SPSOM||MP2/6-31+G(d), with modifications
	g+/t G-DNA	t3 ps_q*x*	q*x* SPSOM||MP2/6-31+G(d), with modifications
SPSOM-NCH2	g−/g+	spsom_nch2_a	a1 SPSOM||MP2/6-31+G(d), with modifications
	g+/t	spsom_nch2_b	b1 SPSOM||MP2/6-31+G(d), with modifications
	g+/t G-DNA	spsom_nch2_q*x*	q*x* SPSOM||MP2/6-31+G(d), with modifications
SPM	g−/g+	spm_a	a1 SPSOM||MP2/6-31+G(d), with modifications
	g+/t	spm_b	b1 SPSOM||MP2/6-31+G(d), with modifications
	g+/t G-DNA	spm_q*x*	q*x* SPSOM||MP2/6-31+G(d), with modifications
MOSPM	g−/g+	mospm_a	a1 SPSOM||MP2/6-31+G(d), with modifications
	g+/t	mospm_b	b1 SPSOM||MP2/6-31+G(d), with modifications
	g+/t G-DNA	mospm_q*x*	q*x* SPSOM||MP2/6-31+G(d), with modifications

[a] All parent starting geometries are available in *xyz* format in the Supporting information. The notation "X||Y" denotes system "X" optimized at "Y" level of theory. The corresponding a1−a11 and b1−b8 structures were derived from these parent structures by modifying the dihedrals according to the Table 1 and subsequently using constrained gradient optimization. [b] See the Supporting Information of ref 12, g−/g+ (p S13) and g+/t (p S15) structures for spsom_a and spsom_b, respectively. [c] X-ray structure of the human telomeric quadruplex sequence. [d] "q*x*" represents q1, residue, DT11; q2, residue, DT05; and q3, residue, DT17.

Information. The Supporting Information further includes all SPSOM MP2-optimized geometries along with their reference CBS(T) (see below) energies.

QM Calculations. Every molecular electronic calculation with a finite basis set is susceptible to BSSE due to the fact that molecular orbitals are approximated by an expansion in terms of analytical basis functions centered on different points in real space (usually the nuclei) that are dependent on the geometry of the studied system.[36] The magnitude of the BSSE decreases as the size of the basis set increases and becomes zero at the limit of an infinite (also called complete) basis set (CBS).[37] For intermolecular noncovalent interactions, the BSSE can be efficiently corrected by the counterpoise procedure.[38] However, BSSE also affects potential energy surfaces of monomers.[39−44] Although several methods were proposed to eliminate the intramolecular BSSE,[39−41] the most reliable approach is to perform calculations with large basis sets,[40] ideally at the CBS limit. As the intramolecular BSSE is very significant in structures with conjugated aromatic systems,[45] the omission of nucleobases in our model compounds reduces this artifact substantially.

Geometry Optimizations. The gradient geometry optimization of the SPSOM structures using internally determined redundant coordinates was carried out at three levels of theory: (i) B3LYP/6-31+G(d),[46,47] (ii) MP2/6-31+G(d), and (iii) the resolution of identity (RI)[48,49] DFT TPSS meta-GGA functional[50] augmented with the empirical dispersion D-0.96-27 type term[51] (TPSS-D) with the large 6-311++G-(3df,3pd) (LP) basis set.

As we are primarily interested in biologically meaningful conformations rather than in pure and often unnatural gas-phase minima, we have fixed the highly flexible torsion degrees of freedom during the geometry optimization at their specified values (Table 1) via application of constraints to all backbone dihedrals, ranging from $\gamma+1$ to δ (in 3′→5′ direction). Obviously, the actual backbone conformation in real physiological environment is determined by numerous factors we do not take into account in this study, like interactions with solvent and counterions, base−base stacking, and edge-to-edge interactions, etc. It is thus unlikely that the experimentally observed and occupied conformational regions would be obtained by unconstrained (or insufficiently constrained) gas-phase optimizations of small model systems. To preserve the connection to biology, it is imperative to fix the backbone during optimization process and restrict minimization to conformational domains we are interested in. Otherwise, we would end up with intrinsically relaxed but irrelevant geometries, not occurring in real DNA structures. To illustrate the necessity of constrains, we performed several full optimizations, that is, without constraints, the results of which can be found in the Supporting Information (Table S12). The unconstrained optimizations drive the structure energetically downhill and away from initial (and relevant) conformational region.

While the MP2 and B3LYP optimizations were carried out using Gaussian 03 software, which is capable of fixing coordinates at values that differ from those of the input structure (i.e., it is possible to change torsions to desired values prior to constrained optimization within single optimization input), the Turbomole code is able to fix torsions at input values only. Because the module for the dispersion calculation of the TPSS-D optimization run was accessible for Turbomole package only, we used the B3LYP optimized geometries with the backbone torsion angles already set to target values as input structures for the subsequent TPSS-D reoptimization.

The minimum energy geometries of the remaining models (except for SPSOM-NCH2, where the optimization turned out to be problematic, see Results and Discussion) were obtained at the B3LYP/6-31+G(d) and MP2/6-31+G(d) levels of theory. All relevant backbone dihedrals starting from $\gamma+1$ (in 3′→5′ direction) were fixed during the optimization procedure in studied model systems.

Single-Point Calculations. The SPSOM model was chosen as the reference compound, which served to compare number of wave function-based and DFT-based methods with

the most accurate CBS(T) approximation (this abbreviation stands for estimated CCSD(T)/CBS calculations, see below).

Because of computational demand of the CCSD(T) calculations, the method with the highest consensus to CBS(T) was selected to benchmark conformers of the remaining models.

All relative energies are computed with respect to the conformer a1, which is considered as the best representative of the canonical BI-DNA.

Wave Function-Based Single-Point Calculations. The RIMP2/CBS energies were estimated using the extrapolation scheme suggested by Halkier and co-workers[52,53] and Dunning's augmented correlation-consistent basis sets of double-ζ and triple-ζ quality (aug-cc-pVDZ and aug-cc-pVTZ).[54,55] The extrapolation to the CBS effectively eliminates both BSSE and basis set incompleteness errors. Nevertheless, our preceding experience indicates this extrapolation scheme with aug-cc-pVDZ and aug-cc-pVTZ basis sets provides results approaching more likely the MP2/aug-cc-pVQZ calculations than the true MP2/CBS limit. Therefore, some residual BSSE and basis set incompleteness errors are likely to remain.[10,11,51] The Hartree−Fock (HF) energy and the correlation MP2 energy contribution are evaluated independently according to eqs 1 and 2:

$$E_{\mathrm{CBS}}^{\mathrm{HF}} = E_{\text{aug-cc-pVDZ}}^{\mathrm{HF}} + \frac{E_{\text{aug-cc-pVTZ}}^{\mathrm{HF}} - E_{\text{aug-cc-pVDZ}}^{\mathrm{HF}}}{0.760691} \quad (1)$$

$$E_{\mathrm{CBS}}^{\mathrm{MP2}} = E_{\text{aug-cc-pVDZ}}^{\mathrm{MP2}} + \frac{E_{\text{aug-cc-pVTZ}}^{\mathrm{MP2}} - E_{\text{aug-cc-pVDZ}}^{\mathrm{MP2}}}{0.703704} \quad (2)$$

Note that this equation notation for both the HF (eq 1) and the MP2 correlation component (eq 2) has been algebraically derived from the standard Helgaker's formulation (see refs 52 and 53) and is more suitable for practical evaluation of CBS extrapolated energies. The original extrapolation scheme and our formulation are thus equivalent.

To speed up the regular MP2 procedure, the RI-approximation[56−58] was utilized.

To account for higher order correlation effects, coupled cluster (CC) calculations with single, double, and perturbative, noniterative triple excitations utilizing 6-31+G(d) basis set (CCSD(T)/6-31+G(d)) were performed. Providing that the difference between the MP2 and CCSD(T) energies, often abbreviated as ΔCCSD(T), shows only small basis set dependence,[59−63] the CCSD(T)/CBS energies could be estimated as eq 3:

$$E_{\mathrm{CBS}}^{\mathrm{CCSD(T)}} \approx \mathrm{CBS(T)} = E_{\mathrm{CBS}}^{\mathrm{MP2}} + (E_{\text{6-31+G(d)}}^{\mathrm{CCSD(T)}} - E_{\text{6-31+G(d)}}^{\mathrm{MP2}}) = E_{\mathrm{CBS}}^{\mathrm{MP2}} + \Delta\mathrm{CCSD(T)} \quad (3)$$

The CCSD(T) calculations were carried out with the MOLPRO 2006.1 package.[64] The CBS(T) energies represent the reference values with which all other methods were compared. The CBS(T) abbreviation is used to indicate that the CCSD(T) CBS extrapolation is approximated.[10,11]

DFT-Based Single-Point Calculations. The DFT-based methods are computationally less demanding and less affected by the BSSE than conventional wave function-based methods. The older functionals are known to be deficient in the description of the dispersion interaction.[65,66] Thus, much recent effort has been spent on including the dispersion interaction either by adjusting current functionals,[67,68] by developing new functionals,[69−71] or by augmenting the existing functionals with empirical correction dispersion energy term.[51,72−74] In the present work, the performance of several traditional and recent functionals was compared to the reference CBS(T) energies of the SPSOM model system. The density functionals considered in this work fall into the following categories:

(i) The first is pure generalized gradient approximations (GGA) functionals. In the present work, the PBE functional[74] augmented with the Grimme's empirical correction term for long-range dispersion effects,[75] PBE-D, in conjunction with the TZVPP[76] triple-ζ basis set was employed.

(ii) The second is hybrid nonlocal GGA functionals containing a portion of the exact exchange interaction from the HF calculation. The only hybrid GGA functional used in the present study was the B3LYP[46,47] functional with the Pople's 6-31+G(d) basis set.

(iii) The third is hybrid meta-GGA type functionals that include also terms dependent on the kinetic energy density. The nonlocal meta-GGAs used in this work include M06, M06-HF, M06-2X, M08-HX, and M08-SO.[77−79] All listed Minnesota functionals belong to the widely used M06 and M08 suites of functionals and were applied with the 6-31+G(d) basis set.

(iv) The fourth is meta, HF-exchange excluded, GGA functionals with entirely local exchange-correlation description. The local treatment makes these functionals computationally very efficient; they are an order of magnitude faster than methods including HF exchange.[80] The two local meta-GGA functionals employed in the present study are M06-L[81,82] and TPSS[50] in combination with Jurečka's empirical dispersion B-0.96-27 type term (TPSS-D).[51] The 6-31+G(d) and 6-311++G(3df,3pd) "LP" basis sets were used for M06-L and TPSS-D calculations, respectively. The RI approximation was available and used for the TPSS functional calculation only.

(v) The fifth is double-hybrid density functionals, which improve the hybrid GGAs by adding a fraction of MP2 correlation energy on top of the HF exchange interaction. The mPW2-PLYP[83] functional with the Ahlrich's TZVPP[76] triple-ζ quality basis set was used.

All RI-approximated calculations were performed with Turbomole 5.10.[84] The mPW2-PLYP and PBE-D energies were calculated with Orca2.6.[85] All remaining ab initio calculations were carried out using Gaussian 03, revision E.01.[35]

Electronic Structure Analysis. The potential energy surface calculations were complemented by atoms-in-molecules (AIM)[86−88] and natural bond orbital (NBO)[89−95] analyses. The aim of the AIM calculations is to analyze the local electron density curvatures and reveal critical points, which can improve the understanding of the physical principles driving the stabilization of the individual conformers. The charge density topologies were computed from the nonfrozen core approximated MP2/6-31+G(d) converged

wave functions. The basis set contained 6d functions rather than the standard 5d ones. These calculations were performed with the AIMPAC code.[96,97]

NBO analysis was applied to identify orbital interactions leading to electron delocalization in the studied systems. As we have shown elsewhere,[12] both B3LYP and HF methods assign approximately the same fraction of electrons to non-Lewis delocalized orbitals and thus provide similar results. In this study, we performed the analysis for the HF/6-31+G(d) orbitals using MP2-optimized structures and the NBO 3.0 program[91,94] implemented in the Gaussian 03 code.[35]

Force Field Calculations. The force field energies were computed using the nonpolarizable ff99 force field,[25] as well as its reparametrized variant parmbsc0.[29] The poor description of the α/γ energetics in ff99 leading to the serious helix unwinding during long DNA simulations[26−28] is substantially improved in its parmbsc0 reparametrization. Prior to evaluation of both the ff99 and the parmbsc0 energies, the MP2-optimized model system geometries have been relaxed using the respective force field except for the fixed backbone dihedrals. That means that force field energies were derived using force field geometries. The relaxation was carried out to the default tolerances using the steepest descent technique for the first 250 iterations, followed by the conjugate gradient method; no cutoff was applied. To keep the backbone dihedrals at the values given in Tables 1 and 2, tight restraints have been imposed. The penalty function that pushes a given term toward the desired value was set at 3000 kcal mol^{-1}.

The electrostatic contribution to the internal energy is computed as a pairwise interaction between the atom-centered partial charges. Hence, the partial charges must be somehow derived and assigned to each atom with a constraint that the sum of partial charges equals the total charge of the system. QM offers numerous established schemes how to derive atomic charges, such as Mulliken and Voronoi population analysis, NBO analysis, AIM analysis, etc. However, AMBER force field calculations are based on the so-called ESP (electrostatic potential) or RESP (restrained ESP) charges.[98] The (R)ESP charges are used in MM calculations because they allow realistic estimates of molecular interactions and conformational preferences. The (R)ESP charges are determined in the following way: (i) the molecular geometry is optimized to a stable minimum conformation using a convenient QM method, (ii) then the electrostatic potential of the optimized molecule is calculated on a three-dimensional real-space grid, (iii) which is subsequently used to fit the atom-centered charges. So the (R)ESP charges are actually effective charges fitted solely to reproduce the QM-determined electrostatic potential of the system created by the electronic and nuclei distribution. Note that (R)ESP charges have no physical meaning as there is an infinite number of solutions of how to allocate charges among the atoms to reproduce the electrostatic potential to a desired precision. As our model systems do not belong to the standard residues for which the AMBER library contains precomputed partial charges, we used the RESP fitting procedure to obtain new charges for all our model systems. That means that the charges we used in this study are not exactly the same as in the original force field but have been derived with the same conception. This allows a consistent comparison between the QM and MM computations. The charges have been fitted at the HF/6-31G(d) level of theory for the most stable MP2/6-31+G(d) optimized geometry, which is the a2 structure. Note that the HF/6-31G(d) level is intentionally used to derive the Cornell et al. force field charges because the modestly overpolarized HF charge distributions are more compatible with the water models typically used for condensed-phase simulations. The charges are given in the Supporting Information. All force field calculations were performed with the sander module of Amber 10.0 suite of programs.[99]

Results and Discussion

SPSOM Geometries. The conformations corresponding to canonical (a-conformers), noncanonical g+/t (b-conformers), and G-DNA loop g+/t (q-conformers) values of α/γ angles were optimized at B3LYP/6-31+G(d), MP2/6-31+G(d), TPSS-D/LP, parm99, and parmbsc0 levels of theory with the backbone torsion angles kept constant at the values defined in Tables 1 and 2.

The C1′···C1′ distances within different conformational types increase in the order of a < b < q. The differences in C1′···C1′ distances between MP2 and TPSS-D structures are nearly negligible for a- and b-subsets of conformers. Regarding q-geometries, TPSS-D C1′···C1′ distances are about 0.14−0.28 Å longer than their MP2 equivalents.

The B3LYP geometries are generally more extended than the MP2 and TPSS-D ones with C1′···C1′ distances being on average 0.30 Å longer for all three structure sets. Further relaxation of the MP2-optimized geometries using empirical force fields leads to additional increase in C1′···C1′ distances. The parmbsc0 and parm99 geometries are nearly identical with predicted C1′···C1′ distances being 0.2−0.5 Å longer as compared to MP2 data for a- and q-conformers, respectively. The force field minimized geometries mutually differ more in the noncanonical b-region of the conformational space because parmbsc0 extends the C1′···C1′ distance by 0.2−0.3 Å more than parm99.

The fact that the MP2 and TPSS-D optimized conformations tend to be slightly more packed than the B3LYP and force field ones might be attributed to the formation of weak CH···O hydrogen bonds and van der Waals contacts (see later discussion on AIM analysis). The propensity to form close contacts between C and O atoms follows the subsequent order of methods: MP2 ∼ TPSS-D > B3LYP > force field. The interrelation between the C1′···C1′ distance and formation of CH···O interaction is illustrated in the Supporting Information, Figure S1, which shows that changing the method of calculations is accompanied by variation of the C2′···O5′ distance, while the 3′-sugar is somewhat repuckered. This ultimately affects also the C1′···C1′ distance. Although the majority of detected CH···O interactions are artifacts of optimization process in the absence of solvent environment and DNA context, they do affect both geometries and intrinsic energetics. Note that separation of the specific impact of such interactions on the optimal geometries

from the general propensities of the applied computational levels is far from being straightforward. This underlines the complexity of reference computations on flexible biomolecular fragments. However, we have estimated energy overstabilization contribution in case of significant CH···O interactions (see the AIM analysis below). The capability to establish such contacts also depends on the size of the model system. For a list of potential CH···O contacts, see Table S1.

The C1′···C1′ distances for SPSOM model system are listed in the Supporting Information, Table S2. Note that in real DNA duplex in condensed phase, C1′···C1′ distances represent variables strongly coupled to backbone torsions, glycosidic χ torsion angle, and base pairs stacking parameters (mainly slide, roll, and twist).[3] Although intrastrand C1′···C1′ distances depend on several structural parameters, in B-DNA, the majority of them range between 4.5 and 6.0 Å, which coincides with the present calculations.[3]

For the q-structures, the experimentally determined C1′···C1′ distances are longer than the computed ones with the discrepancy being up to 0.7 Å (Table S2), which is mainly caused by some modest sugar ring adjustments, as illustrated in Figure S1.

Another structural feature worth investigating is the change of the sugar pucker depending on the optimization method employed. While the sugar at the 5′-end has the pucker fixed during the optimization (the δ angle is kept constant; see Tables 1 and 2), the pucker of the 3′-sugar differs for geometries from diverse conformational space regions. While experimentally determined puckers of the a, b, and q-conformations are in the C2′-endo region, 3′-sugar puckers of the optimized a, b, and q-structures are close to C2′-endo, O4′-endo, and C1′-exo, respectively. In both canonical and noncanonical conformational space regions, MP2, TPSS-D, and B3LYP predict similar puckers with only marginal phase angle (P) differences. Sugar puckers predicted by parm99 and parmbsc0 are nearly identical in case of the a- and q-structures with P values below those of QM methods. As in the case of C1′···C1′ distances, parm99 and parmbsc0 3′-sugar puckers differ more significantly in the b-conformational region, for which parmbsc0 pushes the pucker into the O4′-endo domain, while parm99 optimization drives the pucker to the C1′-exo (see the Supporting Information, Table S3). The reason why parm99 and parmbsc0 differ in the 3′-sugar pucker description within the b-conformational space region while they provide virtually the same results for q-conformers is not clear. It may be related to lower ζ values in case of q-conformers leading to relaxation of a strain in the backbone.

Unlike our previous work,[12] the corresponding QM-minimized geometries are qualitatively equal with both B3LYP versus MP2 and TPSS-D versus MP2 RMSD values not exceeding 0.30 Å (q1). This is due to fixation of the torsions during optimizations. The RMSDs between equivalent force field and MP2 optimized geometries are a little bit larger with the maximum value of 0.36 Å for the q1 structure. Because of the extensive backbone fixation, the optimal geometries obtained at different levels of theory differ only marginally (for the largest difference observed between the force field and MP2 geometry, i.e., the q1 structure, see Figure S2).

The maximum difference between the RIMP2/CBS relative energies (all relative energies are calculated with respect to the a1 structure) calculated at the B3LYP and MP2 geometries equals 0.28 kcal mol^{-1} (for the b2 structure), which is ∼3.4% of the relative energy of the respective structure. This value can be considered as the upper limit of uncertainty of the RIMP2/CBS energies introduced by the difference between MP2 and B3LYP geometries. Although the MP2 method accounts better for dispersion interaction than B3LYP, it is more influenced by BSSE. On the other hand, TPSS-D functional combining better description of the dispersion interaction with the small susceptibility to the BSSE (supported by the fact that the dispersion correction is fitted to CBS data[51]) yields geometries similar to those of MP2. Thus, the MP2 geometries optimized using the moderate size 6-31+G(d) basis set are accurate enough, and they are utilized as reference geometries in the present study. Note, however, that the similarity of MP2/6-31+G(d) geometries to the TPSS-D/LP ones still does not guarantee insignificant influence of the intramolecular BSSE. It may also reflect compensation of errors, mainly BSSE versus underestimation of the dispersion energy due to the limited basis set size. Nevertheless, the data suggest that the calculated energetics is not dramatically sensitive to the geometry optimization method.

SPSOM Energies. *Wave Function-Based Methods.* The electronic energies of the SPSOM system conformers were calculated at the MP2/6-31+G(d), CCSD(T)/6-31+G(d), and RIMP2/CBS levels of theory using the MP2/6-31+G(d) minimized geometries. The latter two methods were used to construct the CBS(T) energies, that is, the estimated CCSD(T) energies extrapolated to the CBS limit (see Methods and Model Systems, eq 3). The CBS(T) relative energies represent reference values with which all other methods were compared. The relative energies of the wave function-based methods are given in Table 4 (ordering according to the increasing relative energies of the conformers is listed in the Supporting Information, Table S4).

In the canonical a-region, the MP2/6-31+G(d) results are in a good agreement with the CBS(T) energies with the maximum deviation of 0.21 kcal mol^{-1} for the a10 conformer (Table 4). On the other hand, b-conformers are systematically destabilized at the MP2/6-31+G(d) level of theory by ∼1.50 kcal mol^{-1}. As a result of medium size basis set, this discrepancy might be caused by the intramolecular BSSE, for its magnitude is structure-dependent. Because of slightly more packed geometry of a-conformers as compared to b- and q-structures, it can be anticipated that canonical conformers are more affected by artificially stabilizing BSSE than are b- and q-conformers. This should lead to a systematic destabilization of b- and q-structures with respect to a1, while the relative energetics of a-conformers should remain unaffected (provided that compactness variability within the canonical a-region is marginal). This assumption is supported by the fact that CCSD(T)/6-31+G(d) and MP2/6-31+G(d) relative single-point energies are almost identical to the values of ΔCCSD(T) energy differences being at most

Table 4. Relative Wave Function-Based Energies of the SPSOM Model Conformers Related to the Structure a1[a]

system/ method	MP2/ 6-31+G(d)	CCSD(T)/ 6-31+G(d)	RIMP2/ CBS[b]	CBS(T)[c]
a1	0.00	0.00	0.00	0.00
a2	−0.85	−0.84	−0.82	−0.76
a3	2.71	2.64	2.67	2.61
a4	−0.40	−0.41	−0.34	−0.34
a5	0.18	0.19	0.12	0.14
a6	1.17	1.14	1.18	1.16
a7	0.26	0.24	0.18	0.17
a8	0.69	0.60	0.72	0.66
a9	−0.09	−0.04	−0.10	−0.03
a10	2.40	2.35	2.26	2.19
a11	−0.15	−0.17	0.00	0.02
b1	9.07	8.77	7.83	7.58
b2	9.72	9.48	8.42	8.25
b3	9.49	9.17	8.09	7.82
b4	8.65	8.45	7.38	7.23
b5	9.56	9.19	8.28	7.93
b6	9.71	9.38	8.53	8.25
b7	9.41	9.11	7.98	7.74
b8	8.60	8.34	7.32	7.10
q1	2.34	2.21	1.95	1.87
q2	2.50	2.37	2.01	1.94
q3	2.50	2.34	1.91	1.81

[a] The energies were calculated using the MP2/6-31+G(d) optimized geometries. All energies are given in kcal mol^{-1}. [b] Estimated CBS energies using the extrapolation scheme suggested by Halkier and co-workers[52,53] via aug-cc-pVDZ and aug-cc-pVTZ basis sets (eqs 1 and 2). [c] Estimated CCSD(T) energies extrapolated to CBS according to eq 3.

−0.07, −0.34, and −0.08 kcal mol^{-1} for a, b, and q-conformers, respectively. The effect of inclusion of higher-order CCSD(T) correlation contributions can thus be regarded as negligible.

Among the wave function-based methods, RIMP2/CBS shows the best correlation with the CBS(T) results. The destabilization of b-conformers using the RIMP2/CBS method is increased by 0.1−0.3 kcal mol^{-1} as compared to the CBS(T) reference values. It is of the same order of magnitude as the uncertainty introduced by the choice of level of geometry optimization. This makes RIMP2/CBS a convenient alternative benchmark method, which is much faster than the complete CBS(T) calculation.

DFT-Based Methods. The ability of 10 different functionals to describe energetics of the SPSOM model system was assessed by comparison with the benchmark CBS(T) calculations. Relative energies related to a1 are given in Tables 5, 6, and S7. Ordering of conformers according to their increasing relative CBS(T) energies is listed in the Supporting Information, Tables S5 and S6.

The PBE functional with the triple-ζ quality TZVPP basis set augmented with the Grimme's empirical correction term (PBE-D) gives very good agreement with the CBS(T) calculations. For the a-set of canonical structures, the largest absolute value deviation observed between PBE-D/TZVPP and CBS(T) energies equals 0.68 kcal mol^{-1} for the a10 geometry. Noncanonical b-systems are, as compared to CBS(T), consistently overstabilized on average by 0.18 kcal mol^{-1}. For the q-systems, no significant difference from CBS(T) energies was detected.

The agreement between B3LYP/6-31+G(d) and CBS(T) is significantly worse. Although the largest difference between the B3LYP and CBS(T) relative energies of the a- and q-conformers is −1.09 kcal mol^{-1} (a8), the ordering of these conformers is rather diverse (Tables 5 and S5). The energy separation between a1 and b-conformers on the B3LYP PES is reduced by 1.22 kcal mol^{-1} (b5) to 2.09 kcal mol^{-1} (b3) with respect to the CBS(T) energies.

Table 5. Relative Energies of the SPSOM System Conformers Related to a1[a]

	functional:	PBE-D	B3LYP	mPW2-PLYP	TPSS-D	CBS(T)[b]
system	basis set:	TZVPP	6-31+G(d)	TZVPP	LP[c]	
a1		0.00	0.00	0.00	0.00	0.00
a2		−0.56	−0.99	−0.85	−0.81	−0.76
a3		2.08	2.30	2.53	2.19	2.61
a4		−0.18	−0.07	−0.18	−0.32	−0.34
a5		0.04	−0.32	−0.09	0.07	0.14
a6		0.95	1.03	1.19	0.87	1.16
a7		0.17	0.00	0.01	0.23	0.17
a8		0.63	−0.43	0.25	0.67	0.66
a9		0.13	0.07	0.00	0.17	−0.03
a10		1.51	2.58	2.27	1.86	2.19
a11		0.15	−0.49	−0.15	−0.11	0.02
b1		7.52	6.00	6.78	7.15	7.58
b2		7.88	6.25	7.29	7.80	8.25
b3		7.68	5.73	6.76	7.10	7.82
b4		6.98	5.35	6.27	6.91	7.23
b5		7.90	6.72	7.29	7.45	7.93
b6		8.17	6.46	7.45	7.70	8.25
b7		7.42	6.08	6.92	7.11	7.74
b8		7.12	5.40	6.23	6.77	7.10
q1		1.84	1.80	1.98	1.98	1.87
q2		2.10	2.36	2.37	2.30	1.94
q3		1.49	2.03	2.29	1.81	1.81

[a] Geometries were optimized at MP2/6-31+G(d) level of theory. All energies are given in kcal mol^{-1}. [b] Estimated CCSD(T) energies extrapolated to CBS according to eq 3. [c] LP stands for the 6-311++G(3df,3pd) basis set.

Table 6. Relative Energies of the SPSOM System Conformers Related to a1[a]

	functional:	M06-L	M06	M06-HF	M06-2X	
system	basis set:		6-31+G(d)			CBS(T)[b]
a1		0.00	0.00	0.00	0.00	0.00
a2		−0.79	−1.08	−1.16	−0.79	−0.76
a3		2.40	2.56	2.83	2.60	2.61
a4		−0.17	−0.68	−0.42	−0.25	−0.34
a5		0.28	0.15	0.13	0.14	0.14
a6		1.15	1.10	1.23	1.27	1.16
a7		0.23	0.06	0.14	0.16	0.17
a8		0.87	0.51	0.62	0.95	0.66
a9		0.21	−0.17	−0.21	−0.01	−0.03
a10		2.31	2.24	2.45	2.19	2.19
a11		0.18	−0.30	−0.12	0.24	0.02
b1		8.64	7.86	8.45	8.62	7.58
b2		9.46	8.85	8.97	9.15	8.25
b3		8.88	7.99	8.81	8.84	7.82
b4		8.24	7.55	7.92	8.02	7.23
b5		9.15	8.37	8.98	9.15	7.93
b6		9.33	8.61	9.01	9.28	8.25
b7		8.75	8.24	8.95	8.86	7.74
b8		8.21	7.34	7.97	8.01	7.10
q1		2.28	1.24	2.24	2.44	1.87
q2		2.68	1.58	2.53	2.63	1.94
q3		2.72	1.98	2.95	2.77	1.81

[a] Geometries were optimized at the MP2/6-31+G(d) level of theory. All energies are given in kcal mol^{-1}. Data for the M08 set of functionals are given in the Supporting information, Table S7. [b] Estimated CCSD(T) energies extrapolated to CBS according to eq 3.

The mPW2-PLYP/TZVPP method decreases the relative energy separation between b-conformers and canonical a1 conformation within the range of 0.64 kcal mol^{-1} (b5) to 1.07 kcal mol^{-1} (b3) (Table 5 and Table S5). mPW2-PLYP functional, as compared to CBS(T) results, subtly destabilizes also q-conformers with respect to a1 by 0.11 (q1), 0.44 (q2), and 0.48 (q3) kcal mol^{-1}. The ordering of the a-conformers according to their mPW2-PLYP relative energies nearly coincides with that of CBS(T).

The TPSS-D/LP energies agree well with the CBS(T) ones. The energetic ordering of a-conformers is identical to the reference CBS(T) data with a9 being the only exception. The energies of q1 and q2 relative to the a1 conformer are shifted upward by 0.11 and 0.36 kcal mol^{-1}, while the relative energy of the q3 system is exactly the same as at the CBS(T) level of theory (Table 5). The b-conformers are all overstabilized with respect to CBS(T) by 0.32 kcal mol^{-1} (b4) up to 0.72 kcal mol^{-1} (b3). Separate consideration of the empirical dispersion term (D) indicates lower degree of dispersion stabilization of b-conformations ($D \approx -17.4$ kcal mol^{-1}) when compared to a-conformations ($D \approx -18.8$ kcal mol^{-1}). Thus, neglect of the dispersion term would result in an artificial overstabilization of b-conformers on average by $\sim$1.4 kcal mol^{-1}. The performance of the TPSS functional alone (i.e., without empirical dispersion correction) along with the LP basis set is comparable to hybrid B3LYP/6-31+G(d).

The common feature of the foregoing functionals is the overstabilization of the b-conformers with respect to a1 as compared to CBS(T) results. Analysis of the PBE and TPSS single-point energies, both with and without dispersion correction, suggests that the overstabilization of b-region conformational subspace is, at least partially, due to the insufficient description of the dispersion interactions within the pure exchange-correlation functionals. Although both PBE-D and TPSS-D also slightly overstabilize b-conformers with respect to CBS(T), the overstabilization is markedly smaller in contrast to B3LYP, mPW2-PLYP, as well as PBE and TPSS without dispersion correction. Thus, for correct description of the potential energy landscape of this model system, functionals including dispersion energy contribution (mPW2-PLYP, PBE-D, and TPSS-D) must be employed. However, the high computational costs of the mPW2-PLYP prohibit its practical use, and PBE-D and TPSS-D functionals, being of at least comparable quality, are generally recommended. To quantitatively assess performance of the foregoing functionals, we present basic statistics in Table 7.

Five hybrid meta-GGA (M06, M06-HF, M06-2X, M08-HX, and M08-SO) and one local meta-GGA (M06-L) Truhlar's functionals yield consistent results. Unlike the previously studied functionals (Table 5), all quoted functionals of M06 and M08 suites destabilize b-systems (with respect to the reference CBS(T) calculations) and, with the exception of M06 functional, also the quadruplex q-conformers. The largest contribution to the destabilization probably comes from the HF repulsion. The higher energetic separation between a/b- and a/q-conformers may be explained in case of M06-2X and M06-HF by the overestimation of nonlocal exchange as M06-2X and M06-HF include 54% and 100% of the HF exchange energy, respectively. The a-systems are very well described by the M06 suite of functionals (Tables 6 and S6). While the M06 functional slightly falls behind the other M06-functionals as far as the a-region is concerned, it is markedly superior for a/b and a/q energy difference estimation and thus also for the overall performance. Regarding M08 functionals, the destabilization of b- and q-conformers with respect to a1 is larger as compared to M06 set of functionals. Moreover, the energetic description of a-conformers diverges from the CBS(T) more as compared to the M06 functionals (Table S7). M06 functionals thus appear to be more appropriate for energetic analyses of this kind of compounds than the M08 set of functionals. For statistical evaluation of the DFT methods, see Table 7.

Table 7. Correlation between the Reference CBS(T) and DFT Energies Computed Using MP2/6-31+G(d) Optimized Geometries[a]

				RSoS·n^{-1} [d]			
functional	basis set	r[b]	q[c]	a	b	q	all[e]
PBE-D	TZVPP	0.9979	**0.9736**	0.082	**0.041**	**0.044**	**0.102**
TPSS with D	LP[f]	0.9986	0.9372	0.039	0.256	0.048	0.263
TPSS without D		0.9864	0.7570	0.145	3.971	0.086	3.975
mPW2-PLYP	TZVPP	0.9963	0.8974	0.029	0.763	0.144	0.777
B3LYP	6-31+G(d)	0.9867	0.7872	0.189	3.086	0.076	3.093
M06-L	6-31+G(d)	0.9982	1.1418	0.021	1.204	0.515	1.310
M06		**0.9989**	1.0423	0.034	0.150	0.184	0.240
M06-HF		0.9977	1.1205	0.030	0.827	0.593	1.018
M06-2X		0.9982	1.1322	**0.014**	1.023	0.571	1.172
M08-HX		0.9965	1.1829	0.076	1.996	0.948	2.211
M08-SO		0.9961	1.2086	0.046	2.558	1.366	2.900

[a] a1 conformer is not included in the statistics as its relative energy equals by definition 0.0 kcal mol^{-1} for all methods. The "best" entry in the given column is highlighted. [b] Pearson's product−moment correlation coefficient detecting linear dependencies between CBS(T) and respective DFT energies defined as

$$\frac{N\sum_i E_i^{\mathrm{CBS(T)}}E_i^{\mathrm{DFT}} - \sum_i E_i^{\mathrm{CBS(T)}}\sum_i E_i^{\mathrm{DFT}}}{\sqrt{N\sum_i E_i^{\mathrm{CBS(T)}^2} - \left(\sum_i E_i^{\mathrm{CBS(T)}}\right)^2}\sqrt{N\sum_i E_i^{\mathrm{DFT}^2} - \left(\sum_i E_i^{\mathrm{DFT}}\right)^2}}$$

where N is the number of conformers (21); the reference a1 conformer is excluded. The summations are over all conformers, a1 excluded. [c] Slope of the linear regression line passing through the origin. The least-squares estimate of the slope (q) is defined as:

$$\sum_i E_i^{\mathrm{CBS(T)}}E_i^{\mathrm{DFT}} \cdot \left(\sum_i E_i^{\mathrm{CBS(T)}^2}\right)^{-1}$$

[d] Residual sum of squares (RSoS) divided by the number of conformers (n; $n_a = 10$, $n_b = 8$, $n_q = 3$) belonging to the respective conformational region (a, b, q):

$$\frac{1}{n}\sum_{i,\mathrm{region}}\left(E_i^{\mathrm{DFT}} - E_i^{\mathrm{CBS(T)}}\right)^2$$

[e] $[(\mathrm{RSoS}\cdot n_a^{-1})^2 + (\mathrm{RSoS}\cdot n_b^{-1})^2 + (\mathrm{RSoS}\cdot n_q^{-1})^2]^{1/2}$. [f] LP is short for 6-311++G(3df,3pd) basis set.

Force Field Energies. The correlation between force field (ff) relative energies calculated at the ff-minimized geometries and the CBS(T) energies is shown in Figure 4. In the canonical a-region, both force fields give similar results, which, with the exception of a8, a9, and a10 conformers (Figure 4), agree well with the CBS(T) energies (Table S8). Destabilization of the a10 structure in both force fields may

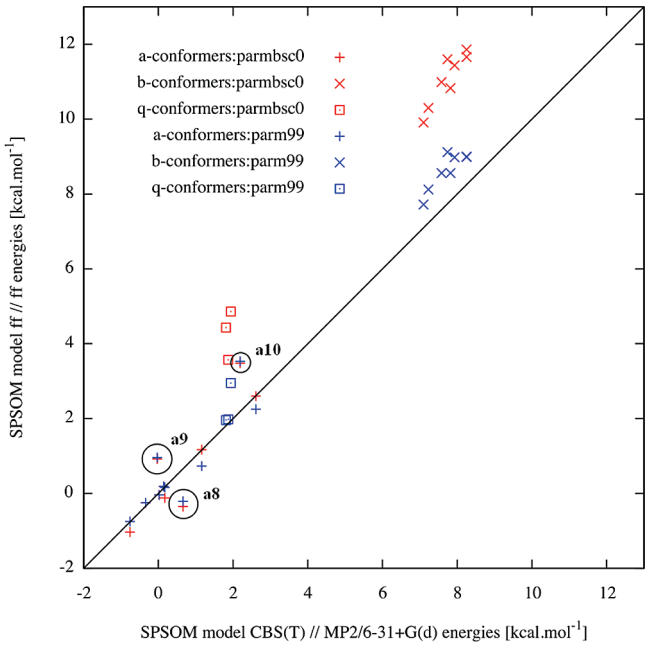


Figure 4. Plot of the correlation between ff//ff and CBS(T)//MP2/6-31+G(d) relative energies (for the corresponding values, see the Supporting Information, Table S8). The RESP charges were fitted to the HF/6-31G(d) electrostatic potential of the most stable conformer a2 (Table S9). The ideal correlation is represented by the black line with the unit slope. Three a-conformation outliers (a8, a9, and a10) are marked with a black circle.

be due to incorrect description of the short-range repulsion by the force field. The structure exhibits contact between H2′ and OP(n+1) atoms, which are as close as 2.4 Å. The sum of O2 and HC atomic types radii is ∼3.15 Å. Another contact occurs between H1′···H5′(n+1) whose distance is about 2.1 Å, while twice a H1 atomic type radius is ∼2.77 Å. Note that the 6-12 Lennard-Jones potential is known to severely exaggerate the short-range interatomic repulsion,[100] while it also has been noticed that the H(C) hydrogens of the Cornell et al. force field are too large.[101] The cause of the energy difference (with respect to CBS(T) results) for the a8 and a9 conformers is not clear. Although both force fields destabilize α/γ = g+/t structures (i.e., b and q-conformers), the parmbsc0 gives at first sight less accurate results. The relative energies of b and q-conformers are shifted upward (i.e., away from the CBS(T) reference values) by 1.6−2.8 kcal mol^{-1} (b-structures) and 1.6−2.5 kcal mol^{-1} (q-structures) with respect to parm99, respectively. It is to be noted, however, that parmbsc0 intentionally penalizes γ-trans geometries as compared to parm99 to prevent their formation in condensed phase molecular simulations.

One of the notorious problems in MM studies is the fact that constant (conformation-independent) atomic charges are used. As the electrostatic potential the charges are fitted on is conformation-dependent, different sets of charges are obtained when derived using different conformers. Thus, to get more insight into the force field performance, we have carried out yet another set of force field calculations, where the charges used for evaluation of the b-conformers were fitted on the HF/6-31G(d) potential of the most stable b-conformer (b8). The charges used to evaluate the a-conformers were kept as before (fitted on the HF/6-31G(d) potential of the a2 system, Table S9). This leads to the a1/b energy separation below ∼6.0 kcal mol^{-1} for parmbsc0 and below ∼3.5 kcal mol^{-1} for parm99, while the reference CBS(T) a1/b separation is 7.1−8.3 kcal mol^{-1}. Fitting charges of b-conformers on the "b-type" electrostatic potential renders parmbsc0 force field superior and supports the basic correctness of the γ-trans penalty of parmbsc0. It also illustrates how sensitive are the force field calculations to the choice of geometry for the derivation of their fixed atomic charges. Our evaluation with different sets of charges roughly corresponds to computations with conformation-dependent charges.

AIM and NBO Analysis. The importance of the weak CH···O hydrogen bonds was assessed by the means of atoms-in-molecules (AIM) Bader analysis. The 6d converged electron density (3;−1) critical points were determined by the AIM analysis of the MP2/6-31+G(d) wave function (see the Supporting Information, Table S10). The electron density and the Laplacian of the electron density threshold for a weak CH···O hydrogen bond was set to 0.01 au. Because of rather small structural differences among structures within the same conformational region (i.e., a, b, and q), we analyzed the first representative out of each region only (a1, b1, and q1). The analysis was done for SPSOM, T3PS, MOSPM, and SPM models (see below for structures of the later three systems).

All identified weak interactions are the so-called CH···O contacts, in which the van der Waals interaction is known to be relatively more important than in standard hydrogen bonds.[102] Because weak CH···O hydrogen bonds were found in canonical a1 system only, we expect their impact exclusively on energetics of the a-conformers. The parameters of critical points found in the remaining b and q-conformers are below the threshold and can thus be regarded as energetically insignificant. To get a basic idea about the extent of stabilization by CH···O hydrogen bonds, interaction energy of a single CH···O contact was also estimated using the AIM analysis. Two different energetic minima for the pucker conformation were localized at the B3LYP/6-31+G(d) and MP2/6-31+G(d) levels of theory in a4, a5, a9, and a11 conformers of the SPM model system (for details, see SPM model system results). The C4′-endo pucker predicted by the MP2/6-31+G(d) calculation in all a-structures allows one to form two almost equally strong (based on AIM) CH···O hydrogen bonds (C2′H···O5′ and C1′H···O5′). The C2′-endo sugar conformation obtained in a4, a5, a9, and a11 structures using B3LYP/6-31+G(d) optimization is stabilized only by the C2′H···O5′ interaction. Calculation of the RIMP2/CBS energies of both optimized geometries allowed one to estimate the energetic contribution (at the RIMP2/CBS level of theory) of one CH···O weak hydrogen bond to be ∼0.6−0.8 kcal mol^{-1}. This difference was indirectly estimated by comparing energies calculated on DFT geometry with C2′-endo pucker and MP2 geometry with C4′-endo. While the former geometry has one CH···O contact, the latter has two.

The only potentially biologically relevant CH···O interaction detected in our model systems is the C2′H···O5′ contact.

Table 8. Off-Diagonal Fock Matrix Elements (*F*, au) Characterizing the Delocalization Effects along the O5′−C5′−C4′−O4′ Bonds in a1 and b1 Structures[a]

	off-diagonal Fockian value			
	a1		b1	
direction	SPSOM	SPM	SPSOM	SPM
n(O4′)→σ*(C4′−C5′)	0.070	0.087	0.033, 0.030	0.035
σ(C4′−C5′)→Ryd(O4′)	0.049	0.046	0.043	0.039
n(O5′)→σ*(C4′−C5′)	0.050	0.049	0.033, 0.061	0.032, 0.062
σ(C4′−C5′)→Ryd(O5′)	0.034	0.039	0.034, 0.032	0.034, 0.031

[a] The results were obtained using the HF-wave functions at the MP2/6-31+G(d) optimized geometries. Two values listed in the same entry refer to two acceptor orbitals centered on the same atom (pair).

The C2′H···O5′ distance is quite frequently around 3.3 Å or even shorter in the B-DNA X-ray structures, including some nucleotides in ultrahigh resolution structures (e.g., ref 103; X-ray structure of a single chain of B double helix resolved at 0.74 Å resolution; the C2′···O5′ distances of DG-4 and DG-9 residues are 3.2 and 3.0 Å, respectively). Occurrence of this interaction in gas-phase computations has been noticed several times.[17,18,21] We nevertheless suppose its rather small impact on conformational preferences of the sugar−phosphate backbone in real environment as the experimental B-DNA C2′···O5′ distances are generally longer than 2.9−3.0 Å seen in gas-phase computations (Table S10). Its effect on the gas-phase energetics should be taken into account.

NBO analysis was carried out to obtain additional insight into destabilization of the b1 structure relative to a1 using the SPSOM and SPM model systems. The unconstrained sugar (i.e., the 3′-sugar of the SPSOM model and the only sugar residue of the SPM model system) in the b1 structure adopts the noncanonical O4′-endo conformation. The same sugar in the a1 structure remains in canonical C2′-endo (SPSOM) or flips to C4′-endo (SPM) conformation. The strongly stabilizing n(O4′)→σ*(C4′−C5′) hyperconjugation in the a1 structure is made impossible in the b1 conformation by O4′ atom pushed out of the C1′,C2′,C3′,C4′ plane. This conformational change is driven by the orbital interactions between n(O5′) and σ*(C4′−C5′), which induces a minor twist of the 5-membered ring along the O4′−C4′ bond due to the repulsion between the electron-rich C4′−C5′ bond and the lone pairs at O4′. The characteristic orbital delocalizations acting in a1 and b1 systems are listed in Table 8. Note that while the introduced stereoelectronic orbital interactions certainly contribute to stabilization of the a1 structure versus b1, we do not suggest that we can in this manner explain the whole energy difference between a- and b-conformers (7.1−8.3 kcal mol^{-1} for the SPSOM model at the CBS(T) level of theory). Note that the empirical force fields, inherently incapable of capturing QM effects, also destabilize b-conformers. In case of the force field, however, we should take into consideration the uncertainty introduced be the fixed atomic charges. When the charges are fitted to reproduce the electrostatic potential in the canonical a-region, they will necessarily introduce error in the b-region description, which could incidentally compensate for the missing electronic structure effects. The results reported above with the charges derived for the b-region geometry indirectly support this possibility. Because of the complex stereoelectronic effects, sugar conformations represent a major challenge for force field derivation. We have recently substantially reparametrized the Cornell et al. force field χ-torsion[104] to prevent ladder-like degradation in long RNA simulations.[105] However, we were still not capable to obtain a fully balanced simultaneous description of pucker and the χ-torsion.

Table 9. List of the Constrained Backbone Torsion Angles[a]

model system	fixed torsion angles
SPSOM-NCH2	γ+1,[b] β+1,[c] α+1,[d] ζ,[e] ε,[f] δ[g]
T3PS	γ+1, β+1, α+1, ζ, ε
MOSPM	γ+1, β+1, α+1, ζ
SPM	γ+1, β+1, α+1, ζ

[a] Values of the fixed torsions are given in Tables 1 and 2. [b] γ+1: O5′(*i*+1)−C5′(*i*+1)−C4′(*i*+1)−C3′(*i*+1). [c] β+1: P(*i*+1)−O5′(*i*+1)−C5′(*i*+1)−C4′(*i*+1). [d] α+1: O3′(*i*)−P(*i*+1)−O5′(*i*+1)−C5′(*i*+1). [e] ζ: C3′(*i*)−O3′(*i*)−P(*i*+1)−O5′(*i*+1). [f] ε: C4′(*i*)−C3′(*i*)−O3′(*i*)−P(*i*+1). [g] δ: C5′(*i*)−C4′(*i*)−C3′(*i*)−O3′(*i*).

Other Model Systems. Geometry optimizations of the remaining model systems (except of SPSOM-NCH2, see below) were carried out at the MP2/6-31+G(d) and B3LYP/6-31+G(d) levels of theory with the backbone torsions (Table 9) fixed at values listed in Tables 1 and 2. Their energies were compared at the RIMP2/CBS//MP2/6-31+G(d) level of theory (Supporting Information, Table S11).

The decision whether to fix the sugar pucker throughout geometry optimization or not is important. Restriction of the pucker to a defined region is advantageous for model comparison and for reduction of the number of potential CH···O interactions. On the other hand, relaxation of the sugar allows one to avoid possible steric conflicts, which may bias PES. To estimate the energetic bias induced by different pucker types, the a1, b1, and q1 conformers of SPSOM, T3PS, MOSPM, and SPM model systems were reoptimized with the sugar(s) kept at the C2′-endo conformation. As the largest change in the relative energies due to the pucker fixation was found to be ∼1.0 kcal mol^{-1} at the MP2/6-31+G(d) level of theory, the bias introduced by not imposing pucker constraints can be considered as acceptable. In the present work, we decided to relax the sugar pucker(s) of the T3PS, MOSPM, SPM, and SPSOM-NCH2 models during the optimization process. We do not claim that letting sugar pucker relax during minimization process is the correct practice as both options have their pros and cons. The estimated error in the relative energies introduced by our decision not to fix the pucker(s) is ∼1.0 kcal mol^{-1} and below.

T3PS Model System. The 3′-sugar (Figure 3) in the T3PS model adopts the C4′-endo conformation in a1, a7, a8, and a10 structures. These four conformers are probably stabilized by C1′H···O5′ contact, in addition to the C2′H···O5′ interaction, which is typical of all a-conformers. The a3 and a6 systems do not adopt C4′-endo pucker due to low value of γ+1 torsion angle (40° for a3 and 35° for a6), preventing formation of the C1′H···O5′ contact (Figure 5). The C4′-endo conformation adoption is also precluded in a2, a4, and a11 conformers because it would likely lead to H1′(*n*+1)/

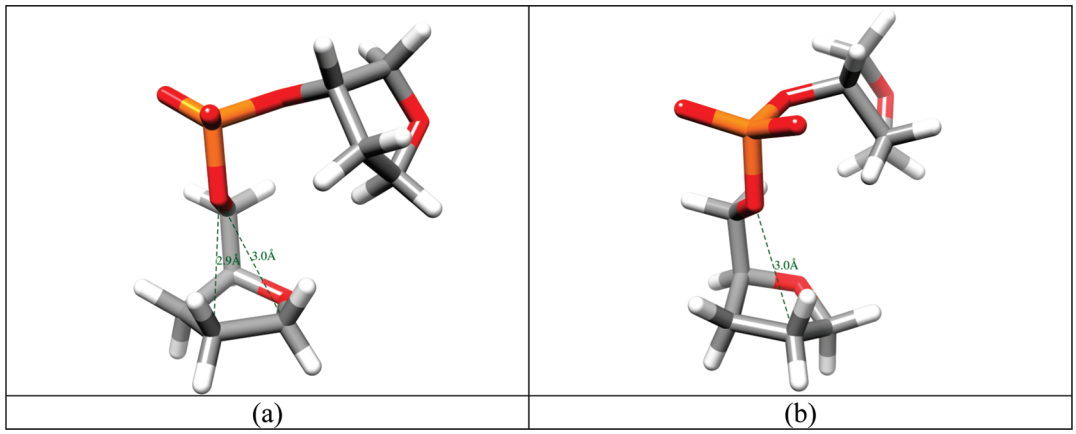


Figure 5. Two 3′-sugar pucker conformations observed in canonical structures of the T3PS model system: (a) C4′-endo pucker of the a1 conformer enabling simultaneous formation of C2′H···O5′ (C···O distance 2.9 Å) and C1′H···O5′ (C···O distance 3.0 Å) contacts. (b) C2′-endo pucker of the a6 conformer enabling a single C2′H···O5′ interaction (C···O distance 3.0 Å).

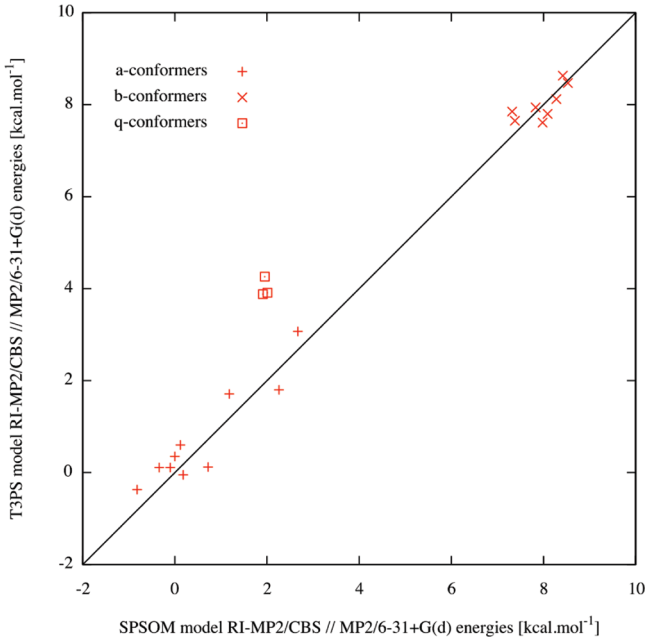


Figure 6. The correlation plot of the RIMP2/CBS//MP2/6-31+G(d) energies between T3PS and SPSOM models (Table S11). The ideal correlation is represented by the black line with the unit slope.

H2′ steric clash. The reason why the 3′-sugar residue of the remaining a-conformers (a5 and a9) do not adopt C4′-endo pucker is not obvious. (Note that the optimizations start from parent structures with C2′-endo arrangement.) The lack of the C1′H···O5′ attractive interaction likely destabilizes (relatively to a1) the respective T3PS a-conformers not adopting the C4′-endo pucker by ∼0.5 kcal mol^{-1} when compared to the SPSOM model (Table S11 and Figure 6). The only overstabilized T3PS conformers with respect to SPSOM (i.e., below diagonal in Figure 6) are those with C4′-endo 3′-sugar conformation, which is clear evidence of the C1′H···O5′ interaction. Moreover, the a1 conformer is affected by C2′H···OP(n+1) interaction whose biological relevancy is arguable. We presume that the alteration of the 3′-sugar pucker within the canonical (i.e., a-conformer) conformational region of T3PS is caused by replacing methoxy group on C3′ present in SPSOM and MOSPM models with a hydrogen atom as no such pucker variations have been observed in SPSOM and MOSPM (see below) models. This observation indicates that the T3PS model system is electronically incomplete and stresses the necessity of a longer backbone fragment at the 3′-end. Because nucleobases are attached via C1′ to the sugar ring, the C1′H···O5′ interaction is a consequence of the lack of nucleobases in our model systems and cannot occur in real DNA.

The increase of energy difference between a1 and q-conformations by ∼1.9−2.3 kcal mol^{-1} in T3PS with respect to SPSOM data (Figure 6 and Table S11) can be, at least partially, explained by the difference in the number and strength of CH···O hydrogen bonds detected in a- and q-conformers.

Although correlation of RIMP2/CBS energies between T3PS and SPSOM models depicted in Figure 6 is rather good, the above discussion clearly shows that the two models are not equivalent.

MOSPM Model System. For the MOSPM system, the B3LYP/6-31+G(d) and MP2/6-31+G(d) geometry optimizations give very similar geometries. The removal of the 5′-sugar moiety in MOSPM model eliminates some CH···O interactions (e.g., C2′H···OP(n+1) observed in a1 of the T3PS model) that are described differently by B3LYP or MP2 methods. Unlike the T3PS (and also SPM, see below) model, the optimization of the parent mospm_x (x = a, b, q1, q2 and q3) structures retains the pucker of the 3′-sugar residue. This is probably due to the methoxy group on C3′, which significantly alters the chemical environment and, as far as a-conformers are concerned, also stiffens the sugar ring while prohibiting the C4′-endo pucker formation.

The approximately same difference in the number and quality of the CH···O interactions in b- and q-conformers with respect to the a1 structure for both MOSPM and SPSOM model systems is responsible for high correlation of RIMP2/CBS MOSPM energies with the SPSOM model for all conformational domains (Figure 7). The a10 conformer constitutes the only outlier and is overstabilized in the MOSPM model (Table S11). The a10 conformer is ∼1.8−2.3 kcal mol^{-1} above a1 in the double-sugar residue model

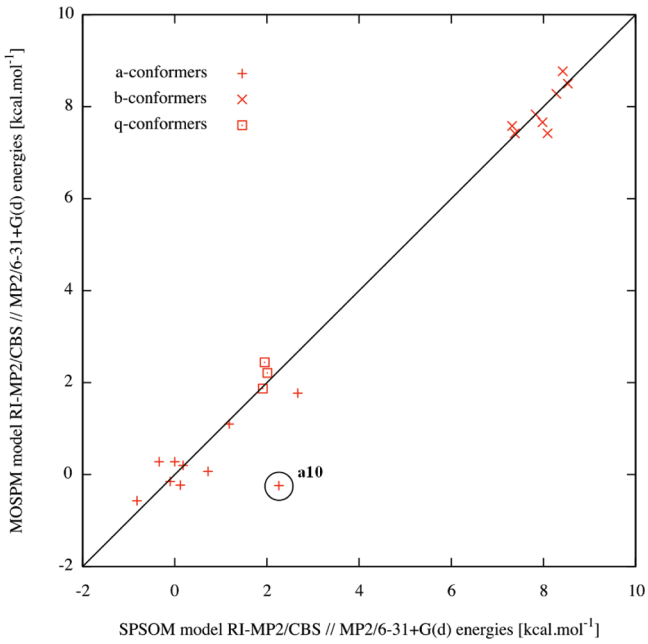


Figure 7. The correlation plot of the RIMP2/CBS//MP2/6-31+G(d) energies between MOSPM and SPSOM models (Table S11). The ideal correlation is represented by the black line with the unit slope. The a10 outlier is marked with a black circle and is discussed in the text.

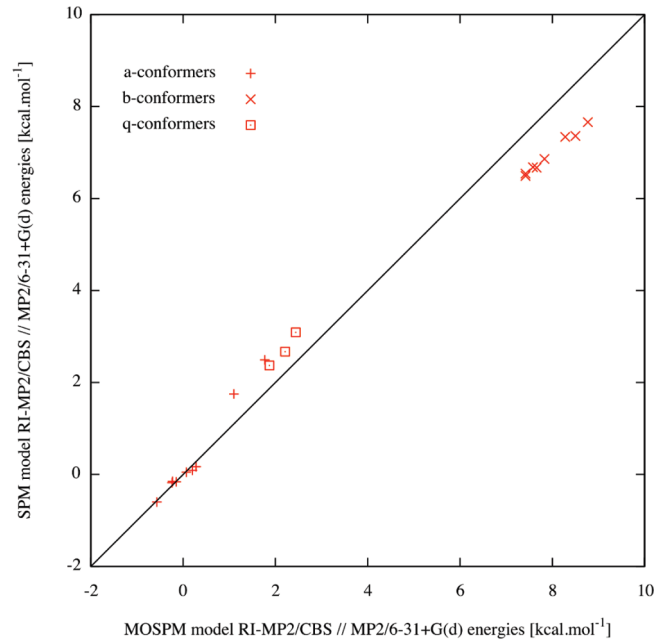


Figure 8. The correlation plot of the RIMP2/CBS//MP2/6-31+G(d) energies between SPM and MOSPM models (Table S11). The ideal correlation is represented by the black line with the unit slope.

systems (SPSOM and T3PS), while in single-sugar models (MOSPM and SPM) it is ∼0.2 kcal mol^{-1} below the a1 conformer. The reason for the destabilization of a10 in SPSOM and T3PS models is the low value of the ε torsion angle (160°, while the canonical value is ∼180°), which puts H2′ and OP(n+1) as close as 2.4 Å and H1′ and H5′(n+1) into a distance of 2.1 Å. As both the H1′ and the H2′ are attached to the 5′-sugar moiety, its substitution by a methyl group releases the repulsion. Thus, the only outlier does not reflect inconsistency of MOSPM and SPSOM model systems because the ε torsion is not defined in the MOSPM model and thus cannot be taken into account in the comparison. We suggest that the SPSOM and MOSPM model systems are equivalent as far as the γ+1, β+1, and α+1 torsions are concerned.

SPM Model System. B3LYP and MP2 optimizations of the SPM model system provide two groups of conformers. While minimum structures of the first group are independent of the used level of theory, B3LYP and MP2 methods yield different conformations differing in the sugar pucker in the second group consisting of a4, a5, a9, a11, and q1 conformers. In the first group, both B3LYP and MP2 minimizations led consistently to either C4′-endo, C2′-endo, or O4′-endo sugar pucker. The C4′-endo pucker (in a1, a2, a7, a8, and a10 systems) arises due to the presence of stabilizing C1′H···O5′ interaction. The low value of γ+1 torsion prevents the formation of this interaction in a3 (γ+1 = 40°) and a6 (γ+1 = 35°) conformers, in which the C2′-endo remains preserved. The O4′-endo pucker of the b- and q-conformers is induced by minimizing the overlap of the lone pair of O4′ with electron-rich σ*(C4′−C5′) bond (see NBO analysis).

The second group consists of a4, a5, a9, a11, and q1 structures. In the case of a-conformers, MP2 yields C4′-endo sugar pucker, while B3LYP optimization preserves the pucker in the C2′-endo region. This can be explained by the change in the fixed β+1 torsion from canonical 180° (a1) to 170° (a4), 190° (a5), and 170° (a11) values, which is probably accompanied by weakening of the C1′H···O5′ interaction that is sensed by the MP2 method but not by the DFT. The reason why B3LYP-optimized a9 structure does not adopt C4′-endo pucker is not clear as the β+1 torsion is fixed at the canonical value of 180°. The sugar pucker of the q1 conformer, which is in O4′-endo conformation at the B3LYP minima, is shifted toward the C1′-exo region at the MP2 PES. However, the reason for this behavior is not obvious.

The effect of the methoxy group was studied by comparing the RIMP2/CBS relative energies of the MP2 optimized conformers of SPM and MOSPM models (Figure 8). The energy ordering relative to the canonical conformation a1 is the same in both models (Table S11). The destabilization of a3, a6, and q-conformers (Table S11) with respect to a1 in the SPM system by ∼0.4−0.7 kcal mol^{-1} is due to the lack of the biologically irrelevant C1′H···O5′ interaction present in a1 SPM but absent in the five specified SPM structures. No such contact has been detected in the a1 MOSPM. Although relative energies of the b-conformers were expected to be shifted upward in the SPM model for the same reason as the a3, a6, and q-conformers, they are overstabilized with respect to MOSPM by ∼0.5 kcal mol^{-1} (Table S11, Figure 8). We ascribe this to the methoxy group, as it is the only difference between these two model systems. Note that apart from the methoxy group, both parent geometries (mospm_x and spm_x) from which optimization was initiated are identical. Both have the sugar pucker in the C2′-endo region. This indicates that the methoxy group effectively prevents the sugar pucker to adopt C4′-endo conformation, and it

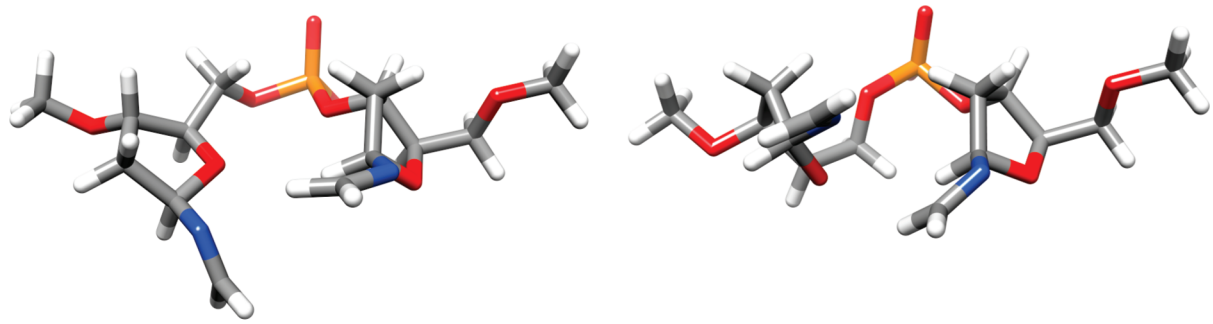


Figure 9. Two interaction modes of the $-N{=}CH_2$ groups: (a) T-shape-like interaction mode, and (b) stack-like interaction mode.

should be an inherent component of any model system for backbone computations.

The correlation of the RIMP2/CBS relative energies of the b- and q-structures between SPM versus MOSPM and SPM versus SPSOM model systems is similar. The q-conformers are destabilized, and the b-conformers are systematically overstabilized in case of SPM model (Table S11). Correlation between MOSPM and SPSOM models is worse. Hence, SPM and SPSOM model systems are not equivalent.

SPSOM-NCH2 Model System. The structures of this model system were optimized at the MP2/6-31+G(d) level of theory only as this model turned out to be inconvenient. The substitution of the C1′ hydrogen atom by the $-N{=}CH_2$ group, which partially mimics the effect of an aromatic nucleobase, leads to the extension of the hyperconjugation network over both sugar residues.[12] However, this system is unsuitable for studying the energetics of the sugar−phosphate backbone. The results of the geometry optimization depend on the initial orientation of the $-N{=}CH_2$ groups. The existence of various interaction modes of the $-N{=}CH_2$ groups (Figure 9) does not allow one to separate the energetic contribution of the methylene-imino groups from that of the sugar−phosphate backbone. It thus illustrates why model systems containing nucleobases (or even their simpler analogues) are not recommended for studying the energetics of the NA backbone. They do not allow to separate the intrinsic backbone preferences from other factors determining their PES.

Conclusions

To gain insight into the intrinsic energetics and electronic structure of the sugar−phosphate backbone, several model systems of 22 relevant DNA backbone conformations from three distinct conformational families were studied in the gas phase by the means of high level ab initio methods. The present study provides a set of accurate structure−energy data for DNA backbone model systems, which can be used as a benchmark database for assessment of other theoretical methods. The most accurate data are obtained at the MP2/CBS level corrected for CCSD(T) term using smaller basis set, that is, using the CBS(T) method. The study leads to the following conclusions:

To maintain the sugar−phosphate backbone to sample relevant conformations and combinations of dihedral angles found in crystal structures of DNA, multiple constraints on the backbone torsion angles have to be imposed. Essentially, it is necessary to fix all dihedral angles at their target values. Albeit fixation of the sugar pucker prevents formation of unnatural CH···O contacts, it may, on the other hand, lead to unnatural tensions biasing energetic analysis. For this reason, fixation of the pucker should be considered and examined from case to case.

The ΔCCSD(T) correction is virtually constant in all studied conformers. Thus, RIMP2 method with sufficiently large basis set (preferably extrapolated to CBS) is adequate for accurate description of nucleic acids backbone.

From the 10 tested DFT approaches, the best results close to the reference CBS(T) calculations are provided by the PBE and TPSS functionals augmented with an empirical dispersion term (PBE-D and TPSS-D), thus stressing the importance of including the dispersion interaction. Very good results were also obtained using the nonlocal meta-GGA M06 functional from the Minnesota M06 suite. The mPW2-PLYP functional also yields reasonable results in accord with CBS(T) reference calculations. Its applicability is, however, limited due to the computational requirements. The remaining M06-type functionals, that is, M06-L, M06-HF, and M06-2X, are of comparable performance. They provide results coinciding with CBS(T) in the canonical a-region but are less accurate in the evaluation of a1 versus b and a1 versus q energy difference. Functionals of the M08 set (M08-HX and M08-SO) are generally inferior to M06 functionals for the energetic analysis of this kind of compounds. The popular B3LYP performs rather unsatisfactorily.

The common attribute of the PBE-D, TPSS-D, and mPW2-PLYP functionals is a slight underestimation of destabilization of b-conformers with respect to the canonical a1 conformation as compared to the CBS(T) data. Neglect of the empirical dispersion terms would result in further stabilization of b-conformers with respect to a1 and thus to a bigger deflection from CBS(T) trend. Performance of pure TPSS/LP (i.e., without dispersion correction) is thus of B3LYP quality.

M06 and M08 functionals, on the other hand, follow the opposite trend as they further destabilize b- and q-conformers with respect to a1 as compared to CBS(T) data.

The intrinsic stability of the noncanonical α/γ = g+/t b-conformers is lower as compared to the canonical a-structures partially due to the deformation of the sugar

conformation, which leads to weakening of the strongly stabilizing n(O4′)→σ*(C4′−C5′) hyperconjugation effect. The conformational change in the 3′-sugar is driven by the orbital interactions between n(O5′) and σ*(C4′−C5′) inducing a minor twist of the sugar ring along the O4′−C4′ bond.

The energetics of the studied model systems is biased by the presence of the network of weak CH···O hydrogen bonds, the majority of which can be considered as the gas-phase artifact. Even though their stabilizing effect balances out to a large degree when structures taken from the same conformational region are confronted, their impact should be considered for comparison of structures from different regions of PES. The number of CH···O interactions and their strength are structure-dependent. Description of these interactions is also method-dependent. These usually undesired interactions greatly complicate reference calculations on fragments of DNA backbone. The only such contact that is occasionally seen in high-resolution B-DNA structures is C2′H···O5′.

The simplification of the SPSOM model system to the MOSPM one has only a marginal impact on the relative energies. Thus, we propose MOSPM as the potentially most appropriate model system for the QM studies of the sugar−phosphate backbone preferences in nucleic acids as a function of backbone torsion angles, excluding ε and ζ torsions. It has several advantages over the other studied model systems (SPSOM, T3PS, SPM, and SPSOM-NCH2): (i) It is smaller than SPSOM, T3PS, and SPSOM-NCH2 systems. (ii) The replacement of the 5′-sugar residue by the methyl group significantly reduces the number of CH···O interactions. (iii) In contrast to SPM, it offers (due to the presence of methoxy group) a more complete description of the electronic structure along the backbone. (iv) The addition of the $-N{=}CH_2$ groups in the SPSOM−NCH2 system does not introduce any advantage, as their presence significantly alters (and complicates) the shape of the potential energy surface and also increases the BSSE artifact. Although inherently incapable to model ε and ζ torsion profiles, the MOSPM system could replace the SPSOM model in future reference studies of δ, γ, β, and α torsions of the DNA backbone.

In future work, we plan to extend the present computations in two directions. One of them is inclusion of the solvent effects, and the other is consideration of other dihedral angle combinations.

Acknowledgment. This work was supported by the Grant Agency of the Academy of Sciences of the Czech Republic grant IAA400040802, Grant Agency of the Czech Republic grants 203/09/1476, P208/10/2302, and 203/09/H046, Ministry of Education of the Czech Republic LC06030, MSM0021622413, LC512, MSM6198959216, and MSM6046-137302, and Academy of Sciences of the Czech Republic, grant nos. AV0Z50040507, AV0Z50040702, and Z40550506. The present study was also financially supported by the Masaryk University, project ID: MUNI/A/0134/2009. We would like to thank Prof. D.G. Truhlar for giving us access to the recent DFT functionals developed by his group. Finally, we would like to thank the Brno MetaCentrum staff for the generous allotment of computer time.

Supporting Information Available: Figure of the C2′···O5′ and C1′···C1′ distances dependency on the method of calculation. Table of potential CH···O interactions. Table of C1′−C1′ distances and 3′-sugar pucker phase angles of the SPSOM model system. Figure illustrating the largest difference observed between the respective optimized geometries of SPSOM. Tables with relative energies of wave function-based, DFT-based, and force fields methods. Tables of RESP charges. List of the (3;−1) critical bond points. Energies and resulting backbone torsion angles of unconstrained optimizations of selected SPSOM conformers. Relative energies of conformers represented by T3PS, MOSPM, and SPM model systems. *xyz* coordinates of five parent structures (i.e., starting geometries, see Table 3) labeled as model_x (x = a, b, q1, q2, and q3) for each single model system (model = SPSOM, T3PS, MOSPM, SPM, and SPSOM-NCH2). *xyz* coordinates of all (22) MP2/6-31+G(d) optimized SPSOM conformers along with their reference CBS(T) energies. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Dickerson, R. E. *J. Mol. Biol.* **1983**, *166*, 419–441.

(2) Calladine, C. R.; Drew, H. R. *J. Mol. Biol.* **1984**, *178*, 773–781.

(3) Packer, M. J.; Hunter, C. A. *J. Mol. Biol.* **1998**, *280*, 407–420.

(4) Packer, M. J.; Dauncey, M. P.; Hunter, C. A. *J. Mol. Biol.* **2000**, *295*, 85–103.

(5) Hartmann, B.; Piazzola, D.; Lavery, R. *Nucleic Acid Res.* **1993**, *21*, 561–568.

(6) Sundaralingam, M. *Biopolymers* **1969**, *7*, 821–860.

(7) Svozil, D.; Kalina, J.; Omelka, M.; Schneider, B. *Nucleic Acid Res.* **2008**, *36*, 3690–3706.

(8) Neidle, S. *Nucleic Acid Structure and Recognition*; Oxford University Press: Oxford, 2002.

(9) Schwartz, T.; Rould, M. A.; Lowenhaupt, K.; Herbert, A.; Rich, A. *Science* **1999**, *284*, 1841–1845.

(10) Šponer, J.; Riley, K. E.; Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2595–2610.

(11) Šponer, J.; Jurečka, P.; Hobza, P. *J. Am. Chem. Soc.* **2004**, *126*, 10142–10151.

(12) Svozil, D.; Šponer, J. E.; Marchan, I.; Perez, A.; Cheatham, T. E.; Forti, F.; Luque, F. J.; Orozco, M.; Šponer, J. *J. Phys. Chem. B* **2008**, *112*, 8188–8197.

(13) MacKerell, A. D. *J. Phys. Chem. B* **2009**, *113*, 3235–3244.

(14) Foloppe, N.; MacKerell, A. D. *J. Phys. Chem. B* **1999**, *109*, 10955–10964.

(15) Wang, F. F.; Gong, L.-D.; Zhao, D. *J. Mol. Struct. (THEOCHEM)* **2009**, *909*, 49–56.

(16) Leulliot, N.; Ghomi, M.; Scalmani, G.; Berthier, G. *J. Phys. Chem. A* **1999**, *103*, 8716–8724.

(17) Louit, G.; Hocquet, A.; Ghomi, M. *Phys. Chem. Chem. Phys.* **2002**, *4*, 3843–3848.

(18) Shishkin, O. V.; Gorb, L.; Zhikol, O. A.; Leszczynski, J. *J. Biomol. Struct. Dyn.* **2004**, *21*, 537–553.

(19) Palamarchuk, G. V.; Shishkin, O. V.; Gorb, L.; Leszczynski, J. *J. Biomol. Struct. Dyn.* **2009**, *26*, 653–661.

(20) Millen, A. L.; Manderville, R. A.; Wetmore, S. D. *J. Phys. Chem. B* **2010**, *114*, 4373–4382.

(21) Hocquet, A. *Phys. Chem. Chem. Phys.* **2001**, *3*, 3192–3199.

(22) Poltev, V. I.; Anisimov, V. M.; Danilov, V. I.; Deriabina, A.; Gonzalez, E.; Garcia, D.; Rivas, F.; Jurkiewicz, A.; Les, A.; Polteva, N. *J. Mol. Struct. (THEOCHEM)* **2009**, *912*, 53–59.

(23) 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.

(24) Varnai, P.; Djuranovic, D.; Lavery, R.; Hartmann, B. *Nucleic Acids Res.* **2002**, *30*, 5398–5406.

(25) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(26) Beveridge, D. L.; Barreiro, G.; Byun, K. S.; Case, D. A.; Cheatham, T. E.; Dixit, S. B.; Giudice, E.; Lankaš, F.; Lavery, R.; Maddocks, J. H.; Osman, R.; Seibert, E.; Sklenar, H.; Stoll, G.; Thayer, K. M.; Varnai, P.; Young, M. A. *Biophys. J.* **2004**, *87*, 3799–3813.

(27) Varnai, P.; Zakrzewska, K. *Nucleic Acids Res.* **2004**, *32*, 4269–4280.

(28) Barone, F.; Lankaš, F.; Špačková, N.; Šponer, J.; Karran, P.; Bignami, M.; Mazzei, F. *Biophys. Chem.* **2005**, *118*, 31–41.

(29) Perez, A.; Marchan, I.; Svozil, D.; Šponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817–3829.

(30) Heddi, B.; Foloppe, N.; Oguey, Ch.; Hartmann, B. *J. Mol. Biol.* **2008**, *382*, 956–970.

(31) Fadrná, E.; Špačková, N.; Sarzynska, J.; Koča, J.; Orozco, M.; Cheatham, T. E., III; Kulinski, T.; Šponer, J. *J. Chem. Theory Comput.* **2009**, *5*, 2514–2530.

(32) Beššeová, I.; Otyepka, M.; Réblová, K.; Šponer, J. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10701–10711.

(33) Burge, S.; Parkinson, G. N.; Hazel, P.; Todd, A. K.; Neidle, S. *Nucleic Acids Res.* **2006**, *34*, 5402–5415.

(34) Parkinson, G. N.; Lee, M. P. H.; Neidle, S. *Nature* **2002**, *417*, 876–880.

(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.; Pople, J. A. *Gaussian 03*, revision E.01; Gaussian, Inc.: Wallingford, CT, 2004.

(36) Jensen, F. *Introduction to Computational Chemistry*; Wiley: New York, 2006; Chapter 5, pp 225−227.

(37) Dunning, T. H. *J. Phys. Chem. A* **2000**, *104*, 9062–9080.

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

(39) Reiling, S.; Brickmann, J.; Schlenkrich, M.; Bopp, P. A. *J. Comput. Chem.* **1996**, *17*, 133–147.

(40) Jensen, F. *Chem. Phys. Lett.* **1996**, *261*, 633–636.

(41) Senent, M. L.; Wilson, S. *Int. J. Quantum Chem.* **2001**, *82*, 282–292.

(42) Balabin, R. M. *J. Chem. Phys.* **2008**, *129*, 164101.

(43) Asturiol, D.; Duran, M.; Salvador, P. *Chem. Phys.* **2008**, *128*, 144108.

(44) Asturiol, D.; Duran, M.; Salvador, P. *J. Chem. Theory Comput.* **2009**, *5*, 2574–2581.

(45) Valdes, H.; Klusak, V.; Pitonak, M.; Exner, O.; Stary, I.; Hobza, P.; Rulisek, L. *J. Comput. Chem.* **2008**, *29*, 861–870.

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

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

(48) Weigend, F.; Haser, M. *Theor. Chim. Acta* **1997**, *97*, 331–340.

(49) Weigend, F.; Haser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143–152.

(50) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401–146405.

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

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

(53) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639–9646.

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

(55) Dunning, T. H., Jr. *J. Phys. Chem. A* **2000**, *104*, 9062–9080.

(56) Kendall, R. A.; Fruüchtl, H. A. *Theor. Chim. Acta* **1997**, *97*, 158–163.

(57) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. *Chem. Phys. Lett.* **1993**, *208*, 359–363.

(58) Vahtras, O.; Almlof, J.; Feyereisen, M. W. *Chem. Phys. Lett.* **1993**, *213*, 514–518.

(59) Hobza, P.; Šponer, J. *J. Mol. Struct. (THEOCHEM)* **1996**, *388*, 115–120.

(60) Jurečka, P.; Hobza, P. *Chem. Phys. Lett.* **2002**, *365*, 89–94.

(61) Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. *J. Am. Chem. Soc.* **2002**, *124*, 10887–10893.

(62) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2004**, *108*, 10200–10207.

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

(64) *MOLPRO, version 2006.1: A package of Ab initio Programs*; Cardiff University: Cardiff, U.K., 2006.

(65) Kristyan, S.; Pulay, P. *Chem. Phys. Lett.* **1994**, *229*, 175–180.

(66) Hobza, P.; Šponer, J.; Reschel, T. *J. Comput. Chem.* **1995**, *16*, 1315–1325.

(67) Lacks, D. J.; Gordon, R. G. *Phys. Rev. A* **1993**, *47*, 4681–4690.

(68) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664–675.

(69) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2003**, *367*, 778–784.

(70) Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. *Phys. Rev. Lett.* **2003**, *91*, 033201.

(71) Zhao, Y.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2701–2705.

(72) Hepburn, J.; Scoles, G. *Chem. Phys. Lett.* **1975**, *36*, 451–456.

(73) Ahlrichs, R.; Penco, R.; Scoles, G. *Chem. Phys.* **1977**, *19*, 119–130.

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

(75) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.

(76) Weigend, F.; Haser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143–152.

(77) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(78) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126–13130.

(79) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1849–1868.

(80) Trindle, C.; Shillady, D. *Electronic Structure Modeling*; CRC Press, Taylor & Francis Group: Boca Raton, FL, 2008.

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

(82) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157–167.

(83) Grimme, S.; Schwabe, T. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4398–4401.

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

(85) Nesse, F. Orca 2.6: An ab initio, DFT and semiempirical SCF-MO package.

(86) Bader, R. F. W. *Atoms in Molecules. A Quantum Theory*; Oxford University Press: Oxford, U.K., 1990.

(87) Bader, R. F. W. *Chem. Rev.* **1991**, *91*, 893–928.

(88) Bader, R. F. W. *J. Phys. Chem. A* **1999**, *103*, 304–314.

(89) Foster, J. P.; Weinhold, F. *J. Am. Chem. Soc.* **1980**, *102*, 7211–7218.

(90) Reed, A. E.; Weinhold, F. *J. Chem. Phys.* **1983**, *78*, 4066–4073.

(91) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735–746.

(92) Reed, A. E.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 1736–1740.

(93) Carpenter, J. E.; Weinhold, F. *J. Mol. Struct. (THEOCHEM)* **1988**, *46*, 41–62.

(94) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899–926.

(95) Weinhold, F.; Carpenter, J. E. In *The Structure of Small Molecules and Ions*; Naaman, R., Vager, Z., Eds.; Plenum: New York, 1988; pp 227−36.

(96) Biegler-Konig, F.; Schonbohm, J.; Bayles, D. *J. Comput. Chem.* **2001**, *22*, 545–559.

(97) Biegler-Konig, F.; Schonbohm, J. *J. Comput. Chem.* **2002**, *23*, 1489–1494.

(98) Bayly, C. I.; Ciepak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(99) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; 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.

(100) Morgado, C. A.; Jurečka, P.; Svozil, D.; Hobza, P.; Šponer, J. *J. Chem. Theory Comput.* **2009**, *5*, 1524–1544.

(101) Warmlander, S.; Šponer, J. E.; Šponer, J.; Leijon, M. *J. Biol. Chem.* **2002**, *277*, 28491–28497.

(102) Desiraju, G. R. *Acc. Chem. Res.* **2002**, *35*, 565–573.

(103) Kielkopf, C. L.; Ding, S.; Kuhn, P.; Rees, D. C. *J. Mol. Biol.* **2000**, *296*, 787–801.

(104) Banáš, P.; Hollas, D.; Zgarbová, M.; Jurečka, P.; Orozco, M.; Cheatham, T. E., III; Šponer, J.; Otyepka, M. *J. Chem. Theory Comput.*, in press.

(105) Mlýnský, V.; Banáš, P.; Hollas, D.; Réblová, K.; Walter, N. G.; Šponer, J.; Otyepka, M. *J. Phys. Chem. B* **2010**, *114*, 6642–6652.

CT1004593

JCTC Journal of Chemical Theory and Computation

Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins

Pavel Banáš,[†,‡] Daniel Hollas,[†] Marie Zgarbová,[†] Petr Jurečka,[†] Modesto Orozco,[§] Thomas E. Cheatham III,[||] Jiří Šponer,*[,†,‡] and Michal Otyepka*[,†,‡]

Regional Centre of Advanced Technologies and Materials, Department of Physical Chemistry, Faculty of Science, Palacky University Olomouc, tr. 17. listopadu 12, 771 46 Olomouc, Czech Republic, Institute of Biophysics, Academy of Sciences of the Czech Republic, Kralovopolska 135, 612 65 Brno, Czech Republic, Joint Research Program in Computational Biology, Institut de Recerca Biomédica and Barcelona Superocomputing Center, Baldiri i Reixac 10, Barcelona 08028, Spain, Jordi Girona 31, Barcelona 08028, Spain, and Departament de Bioquímica i Biología Molecular, Facultat de Biología, Avga Diagonal 645 Universitat de Barcelona, Barcelona 08028, Spain, and Departments of Medicinal Chemistry, Pharmaceutical Chemistry, and Pharmaceutics and Bioengineering, University of Utah, Salt Lake City, Utah 84112, United States

Received August 25, 2010

Abstract: The RNA hairpin loops represent important RNA topologies with indispensable biological functions in RNA folding and tertiary interactions. 5′-UNCG-3′ and 5′-GNRA-3′ RNA tetraloops are the most important classes of RNA hairpin loops. Both tetraloops are highly structured with characteristic signature three-dimensional features and are recurrently seen in functional RNAs and ribonucleoprotein particles. Explicit solvent molecular dynamics (MD) simulation is a computational technique which can efficiently complement the experimental data and provide unique structural dynamics information on the atomic scale. Nevertheless, the outcome of simulations is often compromised by imperfections in the parametrization of simplified pairwise additive empirical potentials referred to also as force fields. We have pointed out in several recent studies that a force field description of single-stranded hairpin segments of nucleic acids may be particularly challenging for the force fields. In this paper, we report a critical assessment of a broad set of MD simulations of UUCG, GAGA, and GAAA tetraloops using various force fields. First, we utilized the three widely used variants of Cornell et al. (AMBER) force fields known as *ff*94, *ff*99, and *ff*99bsc0. Some simulations were also carried out with CHARMM27. The simulations reveal several problems which show that these force fields are not able to retain all characteristic structural features (structural signature) of the studied tetraloops. Then we tested four recent reparameterizations of glycosidic torsion of the Cornell et al. force field (two of them being currently parametrized in our laboratories). We show that at least some of the new versions show an improved description of the tetraloops, mainly in the *syn* glycosidic torsion region of the UNCG tetraloop. The best performance is achieved in combination with the bsc0 parametrization of the α/γ angles. Another critically important region to properly describe RNA molecules is the *anti*/high-*anti* region of the glycosidic torsion, where there are significant differences among the tested force fields. The tetraloop simulations are complemented by simulations of short A-RNA stems, which are especially sensitive to an appropriate description of the *anti*/high-*anti* region. While excessive accessibility of the high-*anti* region converts the A-RNA into a senseless "ladder-like" geometry, excessive penalization of the high-*anti* region shifts the simulated structures away from typical A-RNA geometry to structures with a visibly underestimated inclination of base pairs with respect to the helical axis.

10.1021/ct100481h
Published on Web 11/09/2010

Introduction

RNA is an unbranched, linear polymer composed of four nucleotide units, A, C, G, and U. RNA molecules are usually single-stranded and fold back upon themselves. The 2′-OH group of ribose, absent in DNA, is a powerful donor and acceptor of hydrogen bonds (H-bonds) that is involved in an astonishing repertoire of non-Watson−Crick (noncanonical) interactions. The noncanonical interactions are essential features of RNA three-dimensional structure, dynamics, function, and evolution. Folded RNA molecules typically form short antiparallel double helices by aligning Watson−Crick-complementary stretches of a sequence. These canonical RNA double helices alternate with regions of nucleotides not forming canonical base pairs, i.e., formally unpaired regions. The secondary (2D) structure depicts canonical regions of the folded RNA molecule through the display of parallel lines representing canonical duplex RNA. All of the remaining nucleotides are shown as unpaired loops in such 2D plots. Although these are called loops, these nominally unpaired regions are usually precisely structured via noncanonical interactions and are of the utmost importance for RNA structure and function. The 2D structures of loops can be formally classified as hairpin loops formed by a single-strand segment folded on itself to terminate a helix, internal loops having two strand segments that occur between two helices, and multihelix junctions consisting of multiple-strand segments.[1,2]

The most frequently observed and functionally important hairpin loops are tetraloops (TLs), which cap canonical helices with four loop bases, abbreviated as L1−L4 in this paper. TLs facilitate the backbone inversion required for the formation of secondary and tertiary structures.[3−7] Among all possible combinations,[8] YNMG and GNRA (Y stands for pyrimidine, N for any nucleotide, M for adenine or cytosine, and R for purine), TL families are the most abundant.[5] These TL families are exceptionally thermodynamically stable (namely, when the TL is closed by CG base pairs in the stem),[9] have well-defined structures, and are involved in many biologically relevant processes. In general, TLs initiate folding of RNA structures[3,4,10] and are important interaction sites for tertiary contacts.[11−13]

UNCG Tetraloop. The UNCG TLs (a subfamily of the YNMG family) nucleate RNA global folding.[3] This tetraloop displays poor binding to natural ligands except cations and is not involved in RNA/RNA interactions. Experimental structures of this loop display very limited structural variability.[14−17] The most stable of UNCG TLs (UUCG, see Figure 1) has been extensively studied by several authors. Sakata showed that the 2′-OH groups of U_{L1}, C_{L3}, and G_{L4} and the amino group of G_{L4} are responsible for the thermodynamic stability of the UUCG motif.[18] Later Williams and Hall studied the role of 2′-OH groups of all nucleobases through ribose to 2′-deoxyribose mutations. They concluded that the most significant effect was observed for U_{L1}(2′-OH) deletion.[19,20]

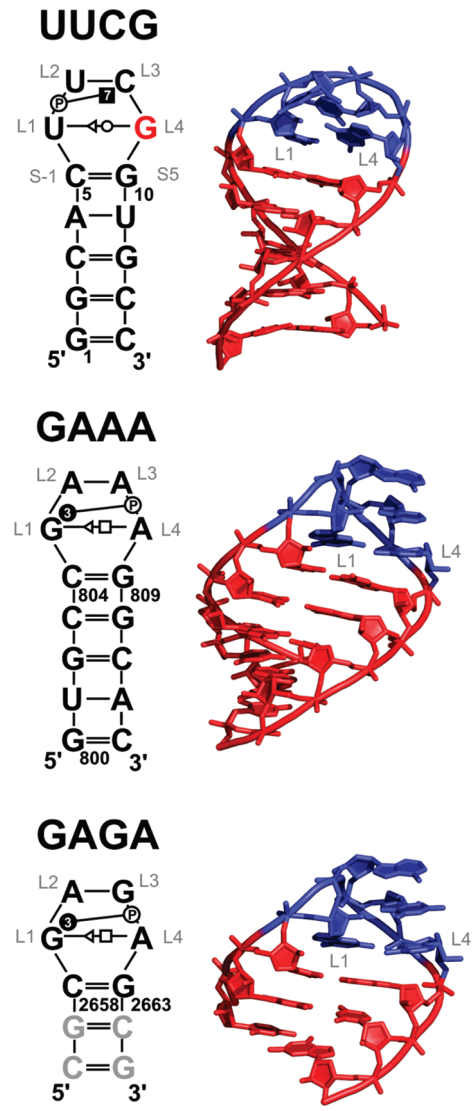


Figure 1. (Left) Secondary structures of the studied systems with base pairing and base−phosphate interactions annotated according to the standard classifications.[21,23] G_{L4} of the UUCG tetraloop having *syn* orientation is highlighted in red. The modeled GC pairs in the GAGA system are shown in gray. The loop residues are labeled as L1−L4 to avoid context numbering. For instance, U6 of UUCG is labeled as U_{L1}. (Right) Three-dimensional structures of studied systems. The A-RNA stem part is shown in red, while the tetraloop nucleotides are in blue.

The first NMR experiments identified the *trans*-Watson−Crick/sugar-edge (*t*WS)[21] G_{L4}/U_{L1} base pair with the U_{L1}(O2′)···G_{L4}(O6) H-bond as a signature interaction of the UUCG TL.[22] The NMR structure further revealed extensive stacking interactions and C_{L3}(N4)···U_{L2}(*pro*-R_p) base phosphate interaction type 7 (7BPh),[23] which are considered as the main source of the high thermodynamic stability.[17,22] The X-ray structures agreed well in the overall topology of the UUCG TL and unraveled two additional U_{L2}(O2′)···G_{L4}(N7) and C_{L3}(O2′)···C_{L3}(O2) H-bonds.[14] The

* Corresponding author. Tel.: +420 585 634 756 (M.O.). Fax: +420 585 634 761 (M.O.). E-mail: michal.otyepka@upol.cz (M.O.). Tel.: +420 541 517 133. E-mail: sponer@ncbr.chemi.muni.cz.
† Department of Physical Chemistry, Palacky University Olomouc.
‡ Academy of Sciences of the Czech Republic.
§ Joint Research Program in Computational Biology.
|| University of Utah.

latest NMR experiments of Schwalbe et al. gained an ultrahigh resolution of 0.25 Å for the loop region (0.3 Å for the stem region).[17] The structure confirmed the *t*WS G_{L4}/U_{L1} base pair with its characteristic $U_{L1}(O2')\cdots G_{L4}(O6)$ H-bond and 7BPh interaction between $C_{L3}(N4)$ and $U_{L2}(pro\text{-}R_p)$. Sugars of U_{L2} and C_{L3} adopted the C2′-endo puckering in agreement with the X-ray structure.

The stability of UUCG TL was also extensively studied by molecular dynamics (MD). Miller and Kollman[24] observed the destabilization of the $U_{L1}(O2')\cdots G_{L4}(O6)$ H-bond in explicit solvent MD simulations with the AMBER Cornell et al. *ff*94 force field and argued that the $U_{L1}(O2')\cdots G_{L4}(O6)$ interaction cannot be considered as the main source of the exceptional thermodynamic stability of the UUCG TL.[25] However, as we will demonstrate below, the loss of the signature $U_{L1}(O2')\cdots G_{L4}(O6)$ H-bond in simulations was in fact due to the imperfectness of the force field. The exceptional thermodynamic stability and structural features of UUCG TL were also addressed in many recent theoretical studies including replica exchange molecular dynamics and umbrella sampling PMF calculation.[26−30]

GNRA Tetraloops. Contrary to UNCG, GNRA TLs primarily mediate RNA tertiary interactions. An analysis of X-ray structures shows striking geometrical conservation also for the GNRA TLs.[8] Structural adaptations in GNRA TL−TL receptor complexes typically includes changes of the TL receptor while the TL is stiff.[31] Williams and coworkers identified 21 examples of standard TLs with the GNRA-like topology in the 2.4 Å resolution X-ray structure of *Haloarcula marismortui* (*H.m.*) large ribosomal subunit.[8] Although they occur in variable contexts within the ribosomal subunit, they adopt virtually identical geometries. The study further identified many hairpin loops with nucleotide insertions, deletions, switches, or strand clips which also adopt very similar 3D structures. Nevertheless, for the GNRA TLs, other experimental methods furnish evidence supporting their conformational dynamics, although in some cases the flexibility can also reflect error margins and inaccuracies in the experiments (see below). Note that even in the lower-resolution ribosomal X-ray structure data, refinement and noise inaccuracies cannot be ruled out as sources of error. For example, this may introduce *syn*/*anti* bias[32,33] and perhaps obscure the exact hairpin loop structures in some cases.

The first information about structural features of GNRA TL came from NMR[34] and lower resolution (∼3 Å) X-ray structures.[35,36] The high-resolution structures of the sarcin/ricin loop (SRL) domain of the large ribosomal subunit[37−40] together with a structural analysis[8] of the large ribosomal subunit[41] and NMR experiments[42,43] furnished in-depth insight into common features of the native fold of GNRA TLs. They include the *trans* Hoogsteen/sugar-edge (*t*HS) A_{L4}/G_{L1}[21] base pair; three signature H-bonds, namely, the $G_{L1}(N1/N2)\cdots A_{L4}(pro\text{-}R_p)$ 3BPh interaction, $G_{L1}(N2)\cdots A_{L4}(N7)$, and $G_{L1}(O2')\cdots R_{L3}(N7)$ (Figure 2); and stacked N_{L2}, R_{L3}, and A_{L4} bases. The backbone of GNRA TLs adopts classic U-turn topology. Contrary to UNCG TL, some structural variability of GNRA TLs is anticipated because, for instance,

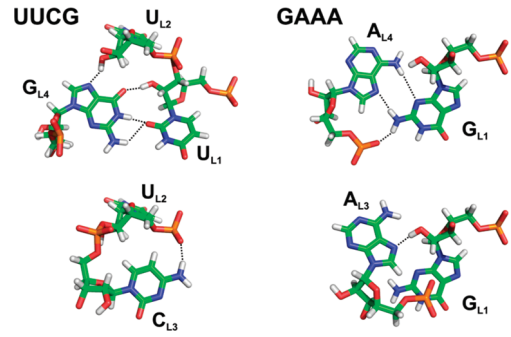


Figure 2. Signature H-bonds (black dashed lines) of UUCG and GAAA tetraloops on the left and right sides, respectively.

protein ribotoxin restrictocin binds an unfolded GNRA TL.[44] However, the majority of GNRA TLs (∼80%) adopt the canonical structure.[37] The question whether ribotoxins induce the conformation change or capture a temporarily unstructured GNRA TL remains open. The dynamics of GNRA TL are the subject of intensive experimental[43,45−49] and theoretical studies.[50−52]

Due to their small size, TLs have been a genuine target for simulation studies.[19,26−30,50,53,54] The simulation studies in general indicate rather substantial flexibility of the TLs, which exceeds variability that is inferred from atomic resolution experiments (see above), suggesting that simulations can be affected by the quality of force fields, typically parametrized keeping in mind the representation of regular helices, not compact irregular structures.[55−57]

In the present study, we investigate the structural dynamics of three representatives (UUCG, GAGA, and GAAA) of the UNCG and GNRA TL families. The aim of the paper is twofold: first, to get insights into the balance of forces in the TLs and, second, to better understand the performance of molecular mechanics force fields for these difficult systems. The TL simulations are supplemented by simulations of short A-RNA stems. We selected four widely used force fields for nucleic acids: three AMBER (Cornell et al.) force fields, *ff*94,[25] *ff*99,[58] and *ff*99bsc0,[59] and CHARMM27.[60] The *ff*99 and *ff*99bsc0 simulations were also performed at higher ionic strength using excess KCl salt to check the impact of ionic strength on the TL structure and dynamics.[61] Besides using the above established force fields, four recent reparameterizations (two of them from our laboratories) of χ glycosidic torsion of the AMBER force field are tested. They are combined with *ff*99 and *ff*99bsc0 force fields (see the Methods for details). These χ modifications were derived recently primarily on the basis of quantum-chemical (QM) computations but were not extensively tested in real simulations. Despite centering on high-level QM calculations, results from the reparameterizations differ considerably as different models and very different levels of QM computations were applied. Thus, in total, the performances of 12 RNA force field variants and combinations were considered

(CHARMM27, *ff*94, *ff*99, *ff*99bsc0, and the last two in combination with four χ modifications).

Methods

Starting Structures. The starting structure of UUCG TL was taken from the high-resolution NMR structure (PDB ID: 2KOC).[17] The GAAA TL was taken from the X-ray structure of a large ribosomal subunit of *Haloarcula maristumortui* (PDB ID, 1JJ2; mean resolution, 2.40 Å; residues 800−813).[62] The GAGA TL was derived from the high-resolution X-ray structure determined at 1.04 Å resolution of the sarcin−ricin loop (SRL; PDB ID, 1Q9A; residues 2658−2663)[39] and capped by two additional C/G base pairs. Short A-RNA stems were built using NAB available from the AmberTools package.[63]

AMBER Simulation Protocol. We performed classical MD simulations using well established simulation protocols.[57,61,64] Missing hydrogen atoms were added by the LeaP module of the AMBER package on the basis of standard residue templates. Each system was neutralized by Na^+ counterions (radius = 1.868 Å and well depth = 0.00277 kcal/mol) and immersed for the MD simulation in a rectangular water box (TIP3P)[65] with a 10-Å-thick layer of water molecules (60 × 50 × 45 Å^3 for UUCG and GAAA and 40 × 45 × 50 Å^3 for GAGA systems). The RNA−solvent system was minimized prior to the AMBER simulation as follows. Minimization of the solute hydrogen atoms was followed by minimization of counterions and water molecules. Subsequently, the hairpin was frozen, and solvent molecules with counterions were allowed to move during a 10-ps-long MD run, the purpose of which is to relax the density of the system. After that, the nucleobases were allowed to relax in several minimization runs with decreasing force constants applied to the backbone phosphate atoms. After full relaxation, the system was slowly heated to 298.15 K over 100 ps using 2 fs time steps and NpT conditions using a weak-coupling scheme with a coupling time of 1 ps.[66] The simulations were carried out under periodic boundary conditions (PBC) in the NpT ensemble (298.15 K, 1 atm) with 2 fs time steps. The particle-mesh Ewald (PME) method[67,68] was used to calculate electrostatic interactions with a cubic spline interpolation and ∼1 Å grid spacing, and a 10.0 Å cutoff was applied for Lennard-Jones interactions with automatic rebuilding of the buffered pair list when atoms moved more than 0.5 Å. The SHAKE algorithm was applied to fix all bonds containing hydrogen atoms. The SANDER module of AMBER 10.0[63] was used for simulations.

AMBER Force Fields. Standard AMBER force fields *ff*94,[25] *ff*99,[58] and *ff*99bsc0[59] were used for simulations. In addition, simulations were performed also with four variants of alternative profiles of the glycosidic χ torsion that were suggested recently as modifications of the *ff*99 force field:

(i) The Ode et al.[69] χ parameters are based on quantum chemical profiles obtained with high-accuracy *in vacuo* MP2/aug-cc-pVTZ//HF/6-31+G(d,p) energy calculations on small model compounds. The force field has been suggested to be compatible with both *ff*99 and *ff*99bsc0 basic parametrizations, and the respective simulations are henceforth labeled as *ff*99χ_{ODE} and *ff*99bsc0χ_{ODE} in the present paper. Note that this force field has not been tested in production runs so far except in our recent study on G-DNA quadruplexes, where it was shown to bring no advantage over the *ff*99 and *ff*99bsc0 force fields.

(ii) Reparameterization against the lower-quality in vacuo QM profile (MP2/6-31G(d)//HF/6-31G(d) level) of ribonucleosides of Yildirim et al.[70] was performed. The force field has not been tested for RNA simulations so far, but it was shown to improve the *syn* vs *anti* balance in nucleoside simulations. Although the original paper does not acknowledge the latest *ff*99bsc0 parametrization and considers the χ parameters exclusively in the context of *ff*99, we decided to test its performance with both *ff*99 and *ff*99bsc0. The respective simulations are marked as *ff*99χ_{YIL} and *ff*99bsc0χ_{YIL}.

(iii) Reparameterization based on a high-quality dispersion-corrected[71] DFT QM profile (PBE/6-311++G(3df,3pd)/D-1.06−23//PBE/6-311++G(3df,3pd)/COSMO method) of deoxyribonucleosides in a continuum water environment (this work and Zgarbova et al., manuscript in preparation) labeled as *ff*99$\chi_{OL\text{-}DFT}$ and *ff*99bsc0$\chi_{OL\text{-}DFT}$ was performed. (The label OL stands for Olomouc, see affiliations.)

(iv) Reparameterization based on the high-level QM profile (MP2/CBS//PBE/6-311++G(3df,3pd)/COSMO method) in continuum water considering weighted parameters for C2′-endo deoxyribose and C3′-endo ribose was performed; this variant is labeled *ff*99χ_{OL} and *ff*99bsc0χ_{OL} (this work and Zgarbova et al., manuscript in preparation).

The OL-DFT and OL parameter files are provided in the Supporting Information, while a full account of the parametrizations including extensive testing will be given separately (Zgarbova et al., manuscript in preparation). The OL force field should be considered as the final version; nevertheless, we also provide some results obtained with the preliminary OL-DFT version, as it provides important insights into the sensitivity of the results to the parametrization.

Note that the modified χ profiles are entirely independent of the recent *ff*99bsc0 reparameterization of the α/γ torsional profile, and therefore the *ff*99bsc0 force field is to be independently cited if applied together with any of the χ terms. The *ff*99bsc0 is essential, particularly for DNA, in modification of the preceding versions of the AMBER Cornell et al. force fields.

To assess effect of salt concentration on the stability of TLs, reference simulations under KCl salt excess ($c(K^+) \sim 0.45$ mol/L, $c(Cl^-) \sim 0.22$ mol/L) conditions and using the SPC/E water model[72] were carried out. Parameters for K^+ (radius, 1.593 Å; well depth, 0.4297 kcal/mol) and Cl^- (radius, 2.711 Å; well depth, 0.012 kcal/mol)[73] were used.

CHARMM Simulations. MD simulations of selected systems were also carried out with the CHARMM all27 force field[60] with the NAMD[74] package (ver. 2.6) using the following protocol. To avoid any differences in starting geometries, the neutralized and solvated system prepared for

AMBER simulations was used as a starting structure to prepare CHARMM27 topologies and coordinates in the CHARMM[75] software package (ver. 34b2). The waters and counterions were minimized in 2500 steps and shaken by short NpT dynamics (100 ps) at 300 K and 1 atm. The system was minimized prior to simulation in 3000 steps and then slowly heated to 300 K over 100 ps using 1 fs time steps and NpT conditions using Langevin dynamics.[76,77] The simulation was produced under periodic boundary conditions in the NpT ensemble (300 K, 1 atm) with 1 fs time steps, because the 2 fs integration step produced considerably less stable trajectories for A-RNA stems. The particle-mesh Ewald method was applied to calculate electrostatic interactions (PME tolerance 10^{-6}), and a 12.0 Å cutoff with an 8.0 Å switching distance was applied for Lennard-Jones interactions. The protocol applied performed well in test simulations on the B-DNA structure, in agreement with literature data.[78]

Table 1 summarizes all simulations analyzed in this study. The simulations were initially intended to be extended to 100 ns. However, some simulations were terminated earlier because of a major degradation of the TLs, i.e., an unfolding event in UUCG_charmm and the formation of a "ladder-like" structure in GAGA_bcs0, GAGA_99χ_{ODE}, and GAGA_bsc0χ_{ODE} simulations. On the other hand, the simulations carried out with reasonably performing force fields were extended to 300 ns (*ff*99bsc0χ_{YIL} and *ff*99bsc0$\chi_{OL\text{-}DFT}$) or to 0.8−1.0 μs (*ff*99bsc0χ_{OL}) to get better insight into the simulation behavior.

Analyses were performed using ptraj (from AmberTools package) and X3DNA.[79] H-bonds were analyzed using in-house software H-bonds (P. Banáš, http://fch.upol.cz/en/software/) using a 3.1 Å cutoff for the H-bond distance and 40° for the hydrogen−H-bond donor···H-bond acceptor angle.

Results

Signature Interactions in the Tetraloops. As explained in the Introduction, the UNCG and GNRA TLs are very precisely structured recurrent RNA motifs that adopt their native structure independently of their contexts. They therefore possess several characteristic (signature) structural features. For the UUCG TL, these include a *t*WS G_{L4}/U_{L1} base pair, *syn* conformation of G_{L4}, and C2′-endo sugar puckers for U_{L2} and C_{L3} (Figure 1). There are four UUCG signature H-bonds (Figure 2): $U_{L1}(O2')\cdots G_{L4}(O6)$, $G_{L4}(N1)\cdots U_{L1}(O2)$, $C_{L3}(N4)\cdots U_{L2}(pro\text{-}R_p)$, and $U_{L2}(O2')\cdots G_{L4}(N7)$. The latter one is seen only in approximately one-third of the high-resolution NMR structurally derived ensembles.[17] $U_{L1}(O2)$ tends to form a bifurcated H-bond to $G_{L4}(N1)$ and $G_{L4}(N2)$ in some X-ray structures.[14−16] On the other hand, the distance between $G_{L4}(N2)$ and $U_{L1}(O2)$ is always larger than 3.3 Å in the high-resolution NMR structure[17] (see also Table 2). The $C_{L3}(N4)\cdots U_{L2}(pro\text{-}R_p)$ H-bond corresponds to a type 7 base−phosphate interaction (7BPh[23]) between the C_{L3} base and U_{L2} phosphate.

The GNRA TLs include the *t*HS A_{L4}/G_{L1} ("sheared") base pair[37] complemented by three H-bonds (Figure 2): $G_{L1}(N2)\cdots A_{L4}(pro\text{-}R_p)$ (3BPh interaction, which is altered with the $G_{L1}(N1/N2)\cdots A_{L4}(pro\text{-}R_p)$ 4BPh interaction in MD or some X-ray structures), $G_{L1}(N2)\cdots A_{L4}(N7)$, and $G_{L1}(O2')\cdots R_{L3}(N7)$.[8] The N_{L2}, R_{L3}, and A_{L4} bases form a purine triple stack.

Table 1. Overview of MD Simulations of TL Systems Carried Out

label	force field	duration (ns)	the first appearance of "ladder-like" structure[80] (ns)[a]
UUCG_94	*ff*94	50	NO
UUCG_99	*ff*99	100	NO
UUCG_bsc0	*ff*99bsc0	100	NO
UUCG_99χ_{ODE}	*ff*99χ_{ODE}	100	NO
UUCG_bsc0χ_{ODE}	*ff*99bsc0χ_{ODE}	100	NO
UUCG_99χ_{YIL}	*ff*99χ_{YIL}	100	NO
UUCG_bsc0χ_{YIL}	*ff*99bsc0χ_{YIL}	300	NO
UUCG_99$\chi_{OL\text{-}DFT}$	*ff*99$\chi_{OL\text{-}DFT}$	100	NO
UUCG_bsc0$\chi_{OL\text{-}DFT}$	*ff*99bsc0$\chi_{OL\text{-}DFT}$	300	NO
UUCG_99χ_{OL}	*ff*99χ_{OL}	100	NO
UUCG_bsc0χ_{OL}	*ff*99bsc0χ_{OL}	800	NO
UUCG_charmm	CHARMM27	50	NO
UUCG_99SE[b]	*ff*99, KCl SE	100	NO
UUCG_bsc0SE[b]	*ff*99bsc0, KCl SE	100	NO
GAAA_99	*ff*99	50	NO
GAAA_bsc0	*ff*99bsc0	100	NO
GAAA_99χ_{ODE}	*ff*99χ_{ODE}	100	85
GAAA_bsc0χ_{ODE}	*ff*99bsc0χ_{ODE}	100	NO
GAAA_99χ_{YIL}	*ff*99χ_{YIL}	100	NO
GAAA_bsc0χ_{YIL}	*ff*99bsc0χ_{YIL}	300	NO
GAAA_99$\chi_{OL\text{-}DFT}$	*ff*99$\chi_{OL\text{-}DFT}$	100	NO
GAAA_bsc0$\chi_{OL\text{-}DFT}$	*ff*99bsc0$\chi_{OL\text{-}DFT}$	300	NO
GAAA_99χ_{OL}	*ff*99χ_{OL}	100	NO
GAAA_bsc0χ_{OL}	*ff*99bsc0χ_{OL}	800	NO
GAAA_charmm27	CHARMM27	100	NO
GAAA_99K^{+}[c]	*ff*99, K^+	100	95
GAAA_99SE[b]	*ff*99, KCl SE	100	NO
GAGA_99	*ff*99	100	36
GAGA_bcs0[d]	*ff*99bsc0	25	20
GAGA_99χ_{ODE}[d]	*ff*99χ_{ODE}	15	5
GAGA_bsc0χ_{ODE}[d]	*ff*99bsc0χ_{ODE}	25	21
GAGA_99χ_{YIL}	*ff*99χ_{YIL}	100	NO
GAGA_bsc0χ_{YIL}	*ff*99bsc0χ_{YIL}	300	NO
GAGA_99$\chi_{OL\text{-}DFT}$	*ff*99$\chi_{OL\text{-}DFT}$	100	NO
GAGA_bsc0$\chi_{OL\text{-}DFT}$	*ff*99bsc0$\chi_{OL\text{-}DFT}$	300	NO
GAGA_99χ_{OL}	*ff*99χ_{OL}	100	NO
GAGA_bsc0χ_{OL}	*ff*99bsc0χ_{OL}	1000	NO
GAGA_charmm27	CHARMM27	100	NO
GAGA_99SE[b]	*ff*99, KCl SE	100	50
GAGA_bsc0SE[b]	*ff*99bsc0, KCl SE	100	NO

[a] "NO" means not observed. [b] Simulations in excess of KCl salt. [c] Simulation under minimal salt conditions with Na^+ ions replaced by K^+. [d] Simulations were terminated because a "ladder-like" structure was irreversibly formed.

UUCG Tetraloop Dynamics. *AMBER Simulations: χ Reparameterizations Maintain Important Signature H-Bonds and Overall Integrity of the Tetraloop.* In all UUCG TL simulations with the standard AMBER force fields (with standard χ torsion, i.e, UUCG_94, UUCG_99, UUCG_bsc0, UUCG_99SE, and UUCG_bsc0SE simulations), we observed a loss of the signature $U_{L1}(O2')\cdots G_{L4}(O6)$ H-bond immediately after the simulation started. This H-bond was replaced by the $U_{L1}(O2')\cdots U_{L2}(O5')$ H-bond (Figure 3A). Despite the fact that the loops still stay, at first sight, locked close to the starting structure, these changes are clear signs of some force field imbalance. Considering the unambiguous structural data, this simulation development is not satisfactory.

Table 2. Basic Structural Characteristics of UUCG TL and H-Bond Populations Calculated from MD Simulations[a]

Structures or simulations	G_{L4}(N1)···U_{L1}(O2) (Å)	C_{L3}(N4)···U_{L2}(*pro*-R_P) (Å)	U_{L2}(O2′)···G_{L4}(N7) (Å)	U_{L1}(O2′)···G_{L4}(O6) (Å)	U_{L1} (O2′)···U_{L2}(O5′) (Å)	G_{L4} χ (deg)	*t*SW U_{L1}/G_{L4} propeller (deg)
NMR	2.7 ± 0.1	2.9 ± 0.1	2.9 ± 0.1	2.6 ± 0.1	3.4 ± 0.1	58 ± 4	−4.3 ± 4.5
X-ray	3.0 ± 0.1	2.9 ± 0.2	4.0 ± 0.5	2.7 ± 0.3	3.8 ± 0.2	60 ± 1	−7.8 ± 7.0
UUCG_94	88%	8%	0%	68%	8%	45 ± 11	−21 ± 11
UUCG_99	87%	9%	2%	65%	8%	48 ± 13	−22 ± 11
UUCG_bsc0	55%	68%	10%	14%	41%	42 ± 18	−32 ± 32
UUCG_99χ_{ODE}	90%	72%	54%	94%	4%	83 ± 13	−1 ± 13
UUCG_bsc0χ_{ODE}	74%	66%	50%	70%	18%	75 ± 17	−14 ± 22
UUCG_99χ_{YIL}	95%	76%	41%	87%	6%	65 ± 15	1 ± 11
UUCG_bsc0χ_{YIL}	92%	77%	46%	89%	6%	69 ± 17	2 ± 11
UUCG_99$\chi_{OL\text{-}DFT}$	92%	36%	28%	83%	7%	67 ± 14	−8 ± 13
UUCG_bsc0$\chi_{OL\text{-}DFT}$	93%	70%	50%	88%	9%	76 ± 14	−1 ± 11
UUCG_99χ_{OL}	93%	68%	42%	80%	15%	64 ± 15	−5 ± 12
UUCG_bsc0χ_{OL}	92%	71%	49%	85%	10%	72 ± 17	−3 ± 12
UUCG_99SE	82%	55%	8%	15%	53%	38 ± 15	−25 ± 12
UUCG_bsc0SE	79%	59%	6%	7%	57%	34 ± 13	−29 ± 11

[a] NMR values were averaged from a set of 20 structures taken from PDB 2KOC.[17] X-ray values were averaged from X-ray structures 1F7Y (res. 2.8 Å),[14] 1I6U (res. 2.6 Å),[15] and 1FJG (res. 3.0 Å).[16] Some of the values are presented as average ± standard deviation. H-bond populations are calculated from respective MD simulations of UUCG TL (see Methods).

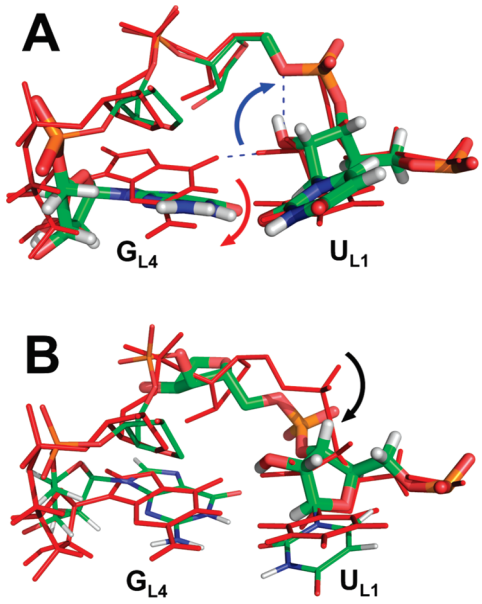


Figure 3. The MD snapshots (colored by atom types) compared with high-resolution NMR structure (in red) showing structural problems seen in simulations of the UUCG tetraloop (some atoms are not shown for clarity, and important parts are shown as sticks). (A) The disruption of the U_{L1}(O2′)···G_{L4}(O6) H-bond and formation of a new U_{L1}(O2′)···U_{L2}(O5′) H-bond observed in all MD simulations with standard χ profiles are highlighted by the blue arrow, while the simultaneous decrease of χ of G_{L4} leading to a change in the U_{L1}/G_{L4} propeller is shown by the red arrow. (B) The U_{L2} phosphate α/γ flip is depicted by the black arrow. See the text for full details.

Simultaneously with the disruption of the U_{L1}(O2′)··· G_{L4}(O6) H-bond, we observed significant propeller twisting (changing from −2° to ∼−25°) of the G_{L4}/U_{L1} *t*WS base pair and mainly a shift of *syn* G_{L4} χ torsion from 60° to 40° in all above-mentioned simulations (Table 2). The same shifts of χ torsion toward lower values in the *syn* region were also observed for *syn* G+1 and A38H$^+$ nucleobases in *ff*99 MD simulations of the hairpin ribozyme (Supporting Information, Table S1).[80] This structural shift comes from the strain in G_{L4} χ torsion, as the energy profile of guanosine χ torsion in AMBER *ff*99 (Supporting Information, Figure S1) shows that the minimum in the *syn* region equals 40°. On the basis of a structural analysis of the above-mentioned MD simulations, we suggest that the shift of G_{L4} χ torsion is the primary source of perturbation of the signature interaction.

All simulations with reparameterized χ torsion prevented the shift of G_{L4} χ torsion, and the signature U_{L1}-(O2′)···G_{L4}(O6) H-bond was stable, except in some cases where it was disrupted due to some unrelated perturbations elsewhere in the structure (Table 2). Thus, UUCG_99χ_{ODE}, UUCG_99χ_{OL}, UUCG_99χ_{YIL}, UUCG_bsc0$\chi_{OL\text{-}DFT}$, UUCG_bsc0χ_{OL}, and UUCG_bsc0χ_{YIL} were the most stable trajectories keeping all signature H-bonds (Table 2). In other words, stable UUCG TL was observed in all simulations with modified χ torsion parameters, except for UUCG_99$\chi_{OL\text{-}DFT}$, showing an undesired α/γ flip of the U_{L1} phosphate (see below) and UUCG_bsc0χ_{ODE} where a "ladder-like" artifact of $C_{S\text{-}1}$ and U_{L1} χ torsions, described below, occurred. $C_{S\text{-}1}$ denotes stem cytosine at the 5′ side of the TL, i.e., C5 in the presented model of UUCG (Figure 1). This indicates that the modified χ torsion profiles locally improve sampling within the *syn* region.

The formation of the U_{L1}(O2′)···U_{L2}(O5′) H-bond in simulations with an unmodified χ profile was likely partially facilitated by modest shifts of ε and ζ torsions of $C_{S\text{-}1}$ (ε from −126° to ∼−150° and ζ from −80° to ∼−60°) and U_{L1} (ε from −160° to ∼−175° and ζ from −100° to ∼−90°). This backbone adaptation occurred in entirely all AMBER UUCG simulations in the initial minimization and was irreversible. This shift of ε and ζ torsions moved the U_{L2}(O5′) oxygen closer to the U_{L1}(2′-OH) hydroxyl (from 3.4 Å in the NMR structure to ∼3.0 Å in MD simulations), which supported the formation of the new U_{L1}(O2′)···U_{L2}(O5′) H-bond. However, we do not consider this backbone adaptation to be the most crucial force field problem, since in the simulations with reparametrized χ torsions the original U_{L1}(O2′)···G_{L4}(O6) H-bond remains stable despite the ε/ζ shift.

ff99bsc0 Clearly Improves the Stability of γ Angle Distribution. In the advanced stages of simulations (on the

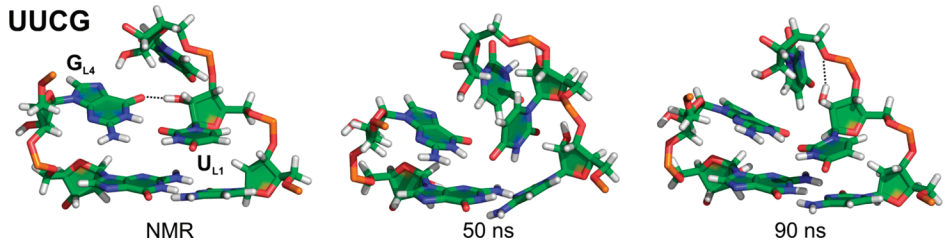


Figure 4. Structures of UUCG TL at the beginning of the UUCG_bsc0 simulation, at 50 ns and at 90 ns, showing the distortion of the UUCG tetraloop. C_{L3} is not shown, for clarity.

tens of nanoseconds time scale), we evidenced further problems due to a disruption of the $C_{L3}(N4)\cdots U_{L2}(pro\text{-}R_p)$ 7BPh interaction in UUCG_94, UUCG_99, and UUCG_99$\chi_{OL\text{-}DFT}$ trajectories. This was caused by α/γ flip of the U_{L2} phosphate (Figure 3B). The flip involved a shift of $\alpha(U_{L2})$ from $\sim-160°$ to $\sim-50°$, $\varepsilon(U_{L1})$ from $\sim-170°$ to $\sim-100°$, and $\gamma(U_{L2})$ from $\sim50°$ to $\sim-170°$. It further correlated with C3′-endo to C2′-endo U_{L2} sugar repuckering. In all cases, the U_{L2} phosphate remained distorted until the simulation ended. The γ torsion of the U_{L2} phosphate sampled the gauche(+) region in all *ff*99bsc0 simulations except for UUCG_bsc0χ_{ODE} and UUCG_bsc0χ_{OL}, where we also observed weakly populated (population about ~5%) and fully reversible γ-*trans* substates. Thus, in contrast to *ff*99 simlations, the bsc0 correction prevents an irreversible α/γ flip of the U_{L2} phosphate to γ-*trans*. The *ff*99bsc0 force field has been designed to prevent pathological γ-*trans* substates in B-DNA MD simulations.[59] It is worth noting that the native position of $\gamma(G_{L4})$ is *trans* due to a sharp bend of the RNA strand at the tip of UUCG TL. Interestingly, $\gamma(G_{L4})$ kept its native γ-*trans* orientation in all simulations with *ff*99bsc0, despite some expectations that *ff*99bsc0 may occasionally overcorrect the γ-*trans* substates.[61,64] Clearly, at least when starting simulations from the native structure, *ff*99bsc0 is superior to *ff*99 for the UUCG TL, as it prevents one undesired irreversible γ-trans flip while keeping the native γ-*trans* nucleotide stable.

The Occurrence of High-anti Substates in Correlation with the Force Field Artifact of Forming a "Ladder-Like" Structure. An almost reversible disruption of the TL signature accompanied by a shift of χ torsions of $C_{S\text{-}1}$ and U_{L1} from the *anti* to the high-*anti* region (from $\sim-150°$ to $\sim-90°$) and breaking of the $G_{L4}(N1)\cdots U_{L1}(O2)$ and $U_{L1}(O2')\cdots G_{L4}(O6)$ H-bonds was observed in UUCG_bsc0 and UUCG_bsc0χ_{ODE} simulations (Figure 4). The shift of χ torsions to the high-*anti* region corresponds to a recently discovered common force field artifact named a "ladder-like" structure of RNA stems, because the most characteristic feature of the "ladder-like" structure is a transition of χ torsion to the high-*anti* region with a value $\sim-85°$ (the exact value slightly depends on the system and force field).[80] In a fully developed "ladder-like" structure of a duplex, besides the shift of the χ torsion, the sugar puckering, ε and ζ torsions, slide, twist, and peaks in the P−P radial distribution function are also modestly affected by the transition, which in addition is not reversible (see ref 80 for more details). In the present simulations, although the $C_{S\text{-}1}$ and U_{L1} χ torsions later returned to the *anti* region, the signature $U_{L1}(O2')\cdots G_{L4}(O6)$ H-bond was not fully stabilized and experienced fluctuations. As will be discussed below, the unmodified *ff*99 and χ_{ODE}[69] parametrizations support the formation of the "ladder-like" artifact, while the remaining three χ reparameterizations appear to prevent its formation. It is entirely consistent with the behavior of UUCG simulations. The independent *ff*99bsc0 modification of α/γ dihedrals is neutral with respect to the "ladder-like" structure formation.

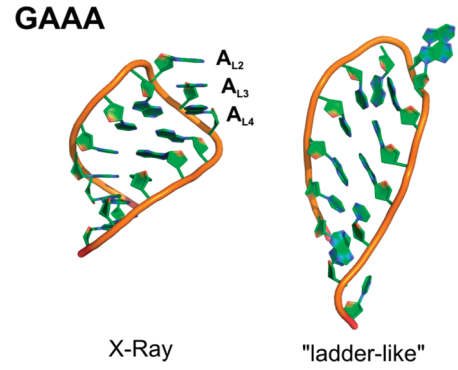


Figure 5. Ladder-like conformer as observed in a simulation of the GNRA tetraloop with AMBER force fields, unless the χ torsion profile is appropriately modified. Initial geometry is on the left, and "ladder-like" conformer is on the right.

CHARMM Simulations. The MD simulation of UUCG TL carried out with a CHARMM27 force field showed complete melting during the first 10 ns. This is in full agreement with recently published simulation data from Deng and Cieplak.[28] The signature $U_{L1}(O2')\cdots G_{L4}(O6)$ and $G_{L4}(N1)\cdots U_{L1}(O2)$ H-bonds were broken at ~0.5 ns, and G_{L4} departed from its initial position. The C_{L3} nucleobase unstacked from U_{L1} at 9.5 ns, breaking its 7BPh $C_{L3}(N4)\cdots U_{L2}(pro\text{-}R_p)$ H-bond. The ribose pucker of C_{L3} switched from C2′-endo to C3′-endo at 20 ns, and G_{L4} switched from a *syn* to an *anti* orientation at ~40 ns. Stem base pairs were also not stable and exhibited frequent breathing.

GNRA Tetraloop Dynamics. In simulations of GNRA TLs with the standard χ AMBER force fields, we typically observed a transition of the A-RNA stem bearing the GNRA TL to the underwound "ladder-like" structure (Figure 5). The "ladder-like" structure is a force field artifact occurring on the tens of nanoseconds time scale, which we first described in our study on hairpin ribozymes.[80] This transition was always preceded by an irreversible disruption of the GNRA signature. The "ladder-like" structure of the stem bearing GNRA TL occurred with standard AMBER force fields as well as with the χ torsion reparameterization of Ode et al.[69] (see Table 1). On the other hand, it has been prevented by

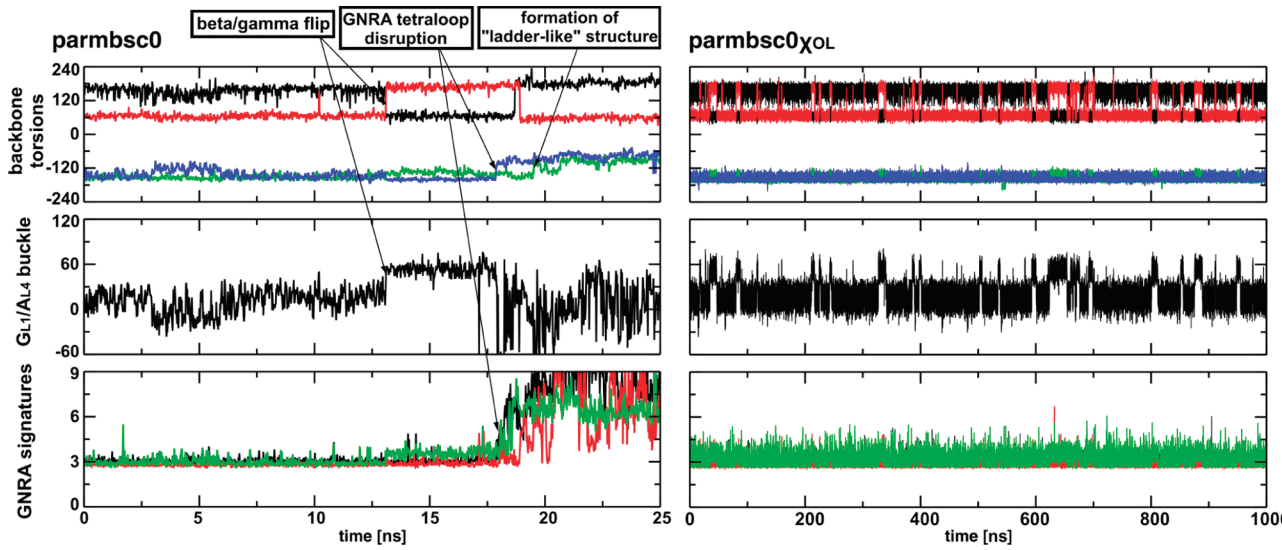


Figure 6. Typical progression of GNRA tetraloop AMBER simulations. Left: simulations without the χ correction or with Ode et al.'s correction illustrated by a 25 ns GAGA_bsc0 simulation. Right: simulations with $\chi_{OL\text{-}DFT}$, χ_{OL}, and χ_{YIL} variants illustrated by a 1 μs GAGA_bsc0χ_{OL} simulation. The upper graphs present a time evolution of the $\beta(G_{L3})$ torsion (black line), $\gamma(G_{L3})$ torsion (red line), and mean χ torsion averaged over either stem nucleobases (green line) or the GNRA tetraloop (blue line). The middle graph shows the G_{L1}/A_{L4} buckle, and the lower graph presents the GNRA tetraloop signature H-bonds: $G_{L1}(N2)\cdots A_{L4}(N7)$, $G_{L1}(N2)\cdots A_{L4}(pro\text{-}R_p)$, and $G_{L1}(O2')\cdots G_{L3}(N7)$ in black, red, and green, respectively.

the other three χ torsion reparameterizations ($\chi_{OL\text{-}DFT}$, χ_{OL}, and χ_{YIL}) and was also not observed in CHARMM simulations.

Although the formation of the "ladder-like" structure was shown primarily to be the force field artifact of A-RNA stems[80] (also see below), a detailed structural analysis, mainly the monitoring of χ torsions, revealed that in our present case the formation of a "ladder-like" structure in the stem was typically preceded by the structural degradation of GNRA TL. Thus, it appears that the loss of the TL integrity facilitates the "ladder-like" transition. Both issues are, however, most likely interconnected. More specifically, the loss of structural integrity of GNRA TL was mainly facilitated by a reversible flip of the β/γ torsions of the R_{L3} phosphate from the *trans*/*gauche*(+) to the *gauche*(+)/*trans* conformation (Figure 6). This β/γ flip increased the buckle of the *t*HS A_{L4}/G_{L1} base pair from ~5° to ~60° (Figure 6). Although the β/γ flips are reversible, we suggest that the increased A_{L4}/G_{L1} buckle causes some steric strain of the stem−loop junction and accelerates the structural degradations of the system including a transition of the stem to a "ladder-like" structure (when the force field does not prevent this "ladder-like" artifact).

As noted above, we did not observe any formation of a "ladder-like" structure in the simulations with $\chi_{OL\text{-}DFT}$, χ_{OL}, and χ_{YIL} reparameterizations. However, simulations with *ff*99 (combining *ff*99 with either the $\chi_{OL\text{-}DFT}$, χ_{OL}, or χ_{YIL} parametrization) exhibited serious distortion of GNRA TL caused by either the R_{L3} β/γ flip or less often by a flip of the $\alpha(A_{L2})$ torsion from *trans* to *gauche*, usually accompanied by the shift of $\gamma(A_{L2})$ torsion from *gauche*(+) to *trans*. Furthermore, these flips caused disruption of GNRA signatures in four of the six *ff*99 simulations with $\chi_{OL\text{-}DFT}$, χ_{OL}, and χ_{YIL} (GAGA_99χ_{YIL}, GAGA_99χ_{OL}, GAAA_99χ_{YIL}, and GAAA_99$\chi_{OL\text{-}DFT}$). On the other hand, the simulations combining *ff*99bsc0 with $\chi_{OL\text{-}DFT}$, χ_{OL}, or χ_{YIL} parametrizations exhibited stable behavior of GNRA TL on the hundreds of nanoseconds time scale. It should be noted that both R_{L3} β/γ and A_{L2} α/γ flips were still present in simulations with the bsc0 correction; however, these flips were reversible and short-lived and thus did not result in distortion of GNRA TL.

The R_{L3} β/γ flip, which seems to be the main source of GNRA TL destabilization in AMBER simulations, might be a consequence of imperfect force field parameters of β and γ torsions. Nonetheless, the R_{L3} phosphate undergoing the β/γ flip is positioned in proximity to the A_{L2} phosphate (P−P distance in the X-ray structure is 5.9 and 5.8 Å in GAGA and GAAA, respectively) because of sharp inversion of the sugar−phosphate backbone path. Thus, the R_{L3} β/γ flip might also be alternatively caused by insufficiently compensated electrostatic repulsion between these two phosphates or some other imbalance of the intermolecular terms. This is in agreement with the fact that structural degradation is initiated by the flip of either R_{L3} or A_{L2} phosphate. However, the involvement of *ff*99bsc0 correction significantly attenuates these phosphate flips, although it is not able to completely eliminate them. Thus, it seems that both imperfect $\alpha/\beta/\gamma$ torsion parameters and an imbalance of the intermolecular terms can contribute to structural degradation of GNRA TL in the *ff*99 force field. Nevertheless, the present GNRA TL simulations are substantially improved when combining the bsc0 correction together with the $\chi_{OL\text{-}DFT}$, χ_{OL}, or χ_{YIL} modification.

We extended the *ff*99bsc0χ_{OL} force field simulations of GAAA and GAGA systems to 0.8 and 1.0 μs, respectively, to test the performance of this force field on the microsecond time scale. We found that the GAGA_bsc0χ_{OL} simulation was entirely stable on the microsecond time scale. However,

we observed conformational changes of the TL region in the GAAA_bsc0χ_{OL} simulation after 0.56 μs. The deformation of GAAA TL is related neither to the flip of the R_{L3} and A_{L2} phosphates nor to the formation of a "ladder-like" structure. It may be caused by some other force field imbalances which become visible on the microsecond time scale (see Supporting Information, Figure S2).

Neither the formation of a "ladder-like" conformation nor a R_{L3} β/γ flip was observed in simulations with CHARMM27. However, we observed local switches of A_{L2} and R_{L3} phosphates including a rapid fluctuation of $\alpha(A_{L2})$ between native *trans* and *gauche*(+), $\alpha(R_{L3})$ between *trans* and native *gauche*(+), and rapid switches of ε and ζ in all four TL nucleobases that were accompanied by structural distortion of the GNRA TL in both CHARMM27 simulations. Nonetheless, the most distinctive feature of CHARMM27 simulations was the instability of the stem bearing the GNRA TL that exhibited extensive terminal base pair breathing followed by the disruption of base-pairing in the stem and subsequent unfolding of the structure, similar to what has been reported, for example, for stems in simulations of kissing-loop complexes.[81,82]

Simulations of A-RNA Stems. The Supporting Information describes a set of simulations of short canonical A-RNA stems. These simulations illustrate rather visible differences between the modified χ parametrizations in the description of the canonical A-RNA structure, mainly a different inclination of base pairs with respect to the helical axis. Although not related directly to the main topic of this paper, these A-RNA simulations provide further insight into the sensitivity of A-RNA simulations to force field parameters and help understanding of the simulation behavior of the TLs.

Discussion

The 5′-UNCG-3′ and 5′-GNRA-3′ RNA tetraloops (TL) are the two most important classes of RNA hairpin loops. These thermodynamically very stable TLs belong to the most prominent RNA motifs, i.e., recurrent RNA building blocks with a precisely defined context-independent 3D structure. While the UNCG TLs play a key role in RNA folding, the GNRA TLs are involved in tertiary interactions and recognition processes.

As with each RNA motif, the UNCG and GNRA TLs are characterized by signature molecular interactions which define their native structure and, subsequently, following the isostericity principle, their consensus sequences.[33] The 3D signature of the studied UUCG TL includes the *t*WS G_{L4}/U_{L1} base pair, *syn* conformation of G_{L4}, south C2′-endo pucker of U_{L2} and C_{L3}, and four UUCG signature H-bonds: U_{L1}(O2′)···G_{L4}(O6), G_{L4}(N1)···U_{L1}(O2), C_{L3}(N4)···U_{L2}(*pro*-R_p), and U_{L2}(O2′)···G_{L4}(N7). The U_{L1}(O2′)···G_{L4}(O6) and C_{L3}(N4)···U_{L2}(*pro*-R_p) H-bonds are unambiguous. The U_{L2}(O2′)···G_{L4}(N7) H-bond is seen only in approximately one-third of the high-resolution NMR structure ensemble.[17] The GNRA TLs are structured with a *t*HS A_{L4}/G_{L1} ("sheared") base pair[37] complemented by three H-bonds: G_{L1}(N1/N2)···A_{L4}(*pro*-R_p), G_{L1}(N2)···A_{L4}(N7), and G_{L1}(O2′)···R_{L3}(N7).[8] The GNRA signature further includes a N_{L2}, R_{L3}, and A_{L4} triple base stack.

Although it cannot be ruled out that the TLs (especially the GNRA one) exhibit some structural dynamics or can be remodeled in some structural contexts, structural biology data as well structural bioinformatics convincingly show that the above-described signature interactions define the genuine native structures of these RNA TL classes.[8,37] Therefore, correct computational methods should be capable of reproducing the characteristic structures of UNCG and GNRA RNA TLs, identifying them as global minima, and dominantly sampling them. However, as noted in the literature, a correct force field description of nucleic acid hairpin loops may be a considerable challenge for the contemporary molecular mechanical force fields.[57] Hairpin loops are characterized by a complex mixture of different molecular interactions and noncanonical backbone conformations and are substantially exposed to the solvent.

The RNA TLs, due to their small size and biochemical importance, became a favorable target for simulation studies in the past several years. These studies were primarily concentrated on the folding of the TLs, using sophisticated enhanced sampling methods and massive large-scale parallel computations.[28,29,50] These impressive studies clearly demonstrated the basic capability of the simulation technique to correctly identify the stem base pairing and subsequently fold the structure. However, less attention has been paid to the exactness of the final or most stable structures identified as the native states. Closer inspection of the published data reveals that at least in some cases the predicted topology is not fully consistent with the native topology as known from structural biology.

In the present paper, we have considered a less ambitious but perhaps no less important task. We investigate the capability of the established force fields to keep the native topology of the UNCG and GNRA TLs. We analyze typical structural rearrangements seen on the ~100+ ns time scale and their force field dependence. Simultaneously, we use the TLs as model systems to test four recent attempts (two of them from our laboratories) to adjust the χ glycosidic torsion profile of the Cornell et al. force field, in addition to the basic *ff*99 and *ff*99bsc0 force field variants. The *ff*99bsc0 is the only viable AMBER force field for DNA simulations, while RNA was until now considered to be almost equivalently well described by all basic Cornell et al. force field variants.[57,59] The χ modifications were derived on the basis of QM computations using different model systems, different QM levels, and different overall philosophies of parametrization (see the Introduction and Methods and the Supporting Information for parameters). There are considerable differences among the four suggested χ glycosidic torsion profiles, while they also substantially differ from the original parametrization (supplementary Figure S1B, Supporting Information). We also performed a set of simulations on short A-RNA stems (Supporting Information). Although the A-RNA simulations are not directly related to the main topic of this paper, these A-RNA simulations provide insights into the sensitivity of A-RNA simulations to force field parameters and help with understanding the simulation behavior of the TLs. The A-RNA simulations indicate that adjusting the χ torsion has a visible effect on the calculated inclination

and base pair roll of A-RNA helices (see Supporting Information, Table S3). The global (helical) structure parameter inclination and local (wedge) parameter roll are mathematically interconnected. They characterize the degree to which the base pairs in the helix adopt the A-RNA geometry having the base pairs inclined with respect to the global helical axis. Due to helical twisting, the inclination then leads to base pair roll in the local base pair step coordination frames.[83,84] The $\chi_{OL\text{-}DFT}$ and χ_{OL} variants of χ adjustment modestly reduce the inclination/roll values compared with simulations using the unmodified force field (see Supporting Information, Table S3). The χ_{YIL} adjustment leads to a qualitative reduction (and underestimation) of inclination/roll (see Supporting Information, Table S3). The impact of these effects on RNA simulations is under further investigation. The χ_{ODE} parametrization destabilizes the A-RNA by promoting the "ladder-like" structure. This behavior reflects the balance of *anti* and high-*anti* regions of the respective parametrizations.

For the tetraloops, we have obtained the following results. In all UUCG TL simulations with the standard AMBER force fields (with unmodified χ torsion), we observed a loss of the signature $U_{L1}(O2')\cdots G_{L4}(O6)$ H-bond immediately after the simulation start, i.e., within few picoseconds. This H-bond is replaced by the $U_{L1}(O2')\cdots U_{L2}(O5')$ H-bond (Figure 3A). Despite the fact that the loop subsequently remains close to the starting structure, the loss of the signature interaction is not in agreement with structural data. Simultaneously with the disruption of the $U_{L1}(O2')\cdots G_{L4}(O6)$ H-bond, we observed significant propeller twisting of the G_{L4}/U_{L1} *t*WS base pair and mainly a shift of G_{L4} χ *syn* torsion from 60° to 40°. We argued that the G_{L4} χ torsion shift is the primary source of perturbation of the signature interaction. All studied modifications of χ torsions (i.e., χ_{ODE}, χ_{YIL}, $\chi_{OL\text{-}DFT}$, and χ_{OL}) improve the behavior most likely because they provide a more realistic description of the *syn* region of G (Supporting Information, Figure S1B).

Further analysis indicates that the use of bsc0 correction improves the simulation behavior by stabilizing the observed distribution of the γ backbone angles, mainly by preventing the undesired and irreversible γ-*trans* flip of the U_{L2} phosphate. Interestingly, the native γ-*trans* flip of G_{L4} is kept. Therefore, this TL is best described when using the *ff*99bsc0 basic parametrization with some of the χ torsion adjustments.

The most significant feature of GNRA simulations with the standard variants of the AMBER force field is a loss of the GNRA integrity on a scale of dozens of nanoseconds followed by a subsequent "ladder-like" conversion of the whole helical stem (Figure 6). Adding the suggested χ corrections (except of χ_{ODE}) improves the behavior of the GNRA TL simulations and prevents larger degradations on the ~100 ns time scale, which is the typical time scale for presently published RNA simulations. The most likely reason why the three successful χ corrections improve the GNRA simulation behavior is the change of the profile in the high-*anti* region compared to that in the *anti* region (Supporting Information, Figure S1B). Compared with the basic *ff*99/*ff*99bsc0 parametrizations, $\chi_{OL\text{-}DFT}$ and χ_{OL} bring a modest penalty to the high-*anti* region, which however seems to be enough to prevent the forming of a "ladder-like" structure. The χ_{YIL} works in the same direction, but the high-*anti* penalty is much more vigorous. The χ_{ODE} rather supports the high-*anti* χ region, and that is why it does not prevent the ladder-like artifact. Note that despite the overall improvement the GNRA simulations exhibit some local dynamics which may indicate some more subtle imbalances. This will require further studies. We noticed reversible flips of β/γ torsions of the R_{L3} phosphate from the *trans/gauche*(+) to the *gauche*(+)/*trans* conformation (Figure 6) which are associated with a dramatically increased buckle of the G_{L1}/A_{L4} base pair from ~5° to ~60°. Such base pair distortion may accelerate further undesired rearrangements of the TLs. The use of *ff*99bsc0 correction improves the simulation behavior by stabilizing the native conformation of TLs phosphates but does not completely prevent these flips. We need to keep in mind that some of the observed dynamical effects may be related to imbalances of the intermolecular terms of the force field. In such a case, the ability of the torsion angle adjustments to improve the simulations may be limited. As noted above, we suspect that the lack of a fully balanced description of the interphosphate repulsion may contribute to the observed backbone dynamics. A balanced description of some such effects may thus require the development of polarization force fields or at least some reparameterization of the solvation terms. Taken together, the GNRA TL is best described by the *ff*99bsc0 basic parametrization with χ_{YIL}, $\chi_{OL\text{-}DFT}$, or χ_{OL} adjustments.

In summary, our data show that three of the χ glycosidic torsion profiles, namely, $\chi_{OL\text{-}DFT}$, χ_{OL}, and χ_{YIL}, improve the description of the TLs, especially when combined with the *ff*99bsc0 basic parametrization.[59] Mainly, they prevent the formation of the degrading "ladder-like" structures of RNA stems, which break down the simulated GNRA tetraloops. The "ladder-like" conformation is associated with an excessive high-*anti* shift of the χ torsion.[80] On the other hand, the χ glycosidic torsion reparameterization of Ode et al.[69] is much less suitable for RNA simulations, as it accelerates the formation of the ladder-like structures. All four χ modifications locally improve the description of the *syn* region, which stabilizes the UUCG TL simulations. Again, bsc0 is to be used as the basic force field for the UUCG simulation. The present results, however, should be taken as preliminary, and considerably more extensive tests on numerous different RNA and DNA systems are under way. We would like to stress that although some of the χ torsion adjustments show significant potential for improving RNA simulations, mainly by preventing the "ladder-like" structure, it cannot be ruled out that we will in the future identify also a worsening of some other properties of the simulated molecules.

When assessing the significance of the results, we have to make a few cautionary notes. First, the χ torsion reparameterizations are applicable exclusively to RNA. We have tested them (not shown) also for B-DNA and DNA quadruplex loops, and they do not improve DNA simulations. In fact, it appears that fine-tuning of the χ torsion simultaneously for DNA and RNA is not possible, unless some other

parameters are modified too. This is related to the second cautionary comment. When using simple analytical force fields, the description of the simulated system is unavoidably only approximate, despite careful parametrization procedures. Thus, the force field is inescapably physically inexact and incomplete. Therefore, the impact of adjustments of the individual torsions, although potentially improving the simulation performance, should not be overrated. The physically inexact torsional potentials are used to approximate the effective overall sum of many diverse physical contributions. The QM reparameterization of torsions does not per se guarantee that the simulations are subsequently improved, as the force field performance depends on the overall balance of all of the force field terms. Therefore, before any application of a modified force field, it must be carefully tested for relevant nucleic acid systems. It is therefore a rather unusual practice that the χ_{ODE} and χ_{YIL} parametrizations were made available without any testing.[69,70] In addition, it is well-known that improving one feature of the simulated systems may have undesired side effects elsewhere. This explains why the χ torsion adjustments tested here do not improve the behavior of DNA simulations.

For the sake of completeness, we also performed limited simulations using the CHARMM27 force field. The main dynamics that we noticed in GNRA TLs are phosphate flips and fluctuations, similar to those reported above for the AMBER simulations. Nevertheless, the simulated structures were later destabilized in their stem regions (base pair fluctuations and fraying), which ultimately also affected the TLs. Note that the simulations were carried out with quite short stems. Such reduced stability of the short stems is consistent with literature data.[81] The UNCG TL trajectory was unstable. This simulation result is identical to more extensive data reported and in more detail described by Deng and Cieplak.[28]

The 100+ ns simulations are sufficient for many useful applications, such as the very basic MD characterization of existing RNA structures. Note that perturbation of the TLs may cause bias in the overall assessment of the data even when the TL is not the primary focus of a given simulation study. Therefore, stabilization of the RNA TLs on this time scale is important. Work is in progress to investigate the TLs using much longer simulations and also using substantially larger RNA systems to prevent eventual end effects.

Acknowledgment. This study was supported by Grants CZ.1.05/2.1.00/03.0058, LC512, LC06030, and MSM6198959216 from the Ministry of Education of the Czech Republic; Grants 203/09/1476 and 203/09/H046 from the Grant Agency of the Czech Republic; Grant IAA400040802 from the Grant Agency of the Academy of Sciences of the Czech Republic; Grants AV0Z50040507 and AV0Z50040702 from the Academy of Sciences of the Czech Republic; and Student Project PrF_2010_025 of Palacký University.

Supporting Information Available: Library files with *ff*99$\chi_{OL\text{-}DFT}$, *ff*99χ_{OL}, *ff*99bsc0$\chi_{OL\text{-}DFT}$, and *ff*99bsc0χ_{OL} force fields, a detailed description of how to include these force fields in the AMBER standard preparation tool *leap*, detailed information about shifted energy minima for *syn* region of glycosidic χ torsion in standard AMBER force fields, detailed description of the structural degradation of GAAA TL that appeared on a very large time scale, and a description of the simulations of short A-RNA stems including tables with helical parameters. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Leontis, N. B.; Westhof, E. Analysis of RNA motifs. *Cur. Opin. Struct. Biol.* **2003**, *13*, 300–308.

(2) Mathews, D. H.; Turner, D. H. Prediction of RNA secondary structure by free energy minimization. *Cur. Opin. Struct. Biol.* **2006**, *16*, 270–278.

(3) Tuerk, C.; Gauss, P.; Thermes, C.; Groebe, D. R.; Gayle, M.; Guild, N.; Stormo, G.; Daubentoncarafa, Y.; Uhlenbeck, O. C.; Tinoco, I.; Brody, E. N.; Gold, L. CUUCGG Hairpins - Extraordinarily Stable RNA Secondary Structures Associated with Various Biochemical Processes. *Proc. Natl. Acad. Sci.* **1988**, *85*, 1364–1368.

(4) Uhlenbeck, O. C. Nucleic-Acid Structure - Tetraloops and RNA Folding. *Nature* **1990**, *346*, 613–614.

(5) Woese, C. R.; Winker, S.; Gutell, R. R. Architecture of Ribosomal-RNA - Constraints on the Sequence of Tetra-Loops. *Proc. Natl. Acad. Sci.* **1990**, *87*, 8467–8471.

(6) Wolters, J. The Nature of Preferred Hairpin Structures in 16s-Like Ribosomal-RNA Variable Regions. *Nucleic Acids Res.* **1992**, *20*, 1843–1850.

(7) Bevilacqua, P. C.; Blose, J. M. Structures, kinetics, thermodynamics, and biological functions of RNA hairpins. *Annu. Rev. Phys. Chem.* **2008**, *59*, 79–103.

(8) Hsiao, C.; Mohan, S.; Hershkovitz, E.; Tannenbaum, A.; Williams, L. D. Single nucleotide RNA choreography. *Nucleic Acids Res.* **2006**, *34*, 1481–1491.

(9) Blose, J. M.; Proctor, D. J.; Veeraraghavan, N.; Misra, V. K.; Bevilacqua, P. C. Contribution of the Closing Base Pair to Exceptional Stability in RNA Tetraloops: Roles for Molecular Mimicry and Electrostatic Factors. *J. Am. Chem. Soc.* **2009**, *131*, 8474–8484.

(10) Varani, G. Exceptionally Stable Nucleic-Acid Hairpins. *Annu. Rev. Biophys. Biomed.* **1995**, *24*, 379–404.

(11) Marino, J. P.; Gregorian, R. S.; Csankovszki, G.; Crothers, D. M. Bent Helix Formation between RNA Hairpins with Complementary Loops. *Science* **1995**, *268*, 1448–1454.

(12) Chauhan, S.; Woodson, S. A. Tertiary interactions determine the accuracy of RNA folding. *J. Am. Chem. Soc.* **2008**, *130*, 1296–1303.

(13) Jaeger, L.; Michel, F.; Westhof, E. Involvement of a GNRA Tetraloop in Long-Range Tertiary Interactions. *J. Mol. Biol.* **1994**, *236*, 1271–1276.

(14) Ennifar, E.; Nikulin, A.; Tishchenko, S.; Serganov, A.; Nevskaya, N.; Garber, M.; Ehresmann, B.; Ehresmann, C.; Nikonov, S.; Dumas, P. The crystal structure of UUCG tetraloop. *J. Mol. Biol.* **2000**, *304*, 35–42.

(15) Tishchenko, S.; Nikulin, A.; Fomenkova, N.; Nevskaya, N.; Nikonov, O.; Dumas, P.; Moine, H.; Ehresmann, B.; Ehresmann, C.; Piendl, W.; Lamzin, V.; Garber, M.; Nikonov, S. Detailed analysis of RNA-protein interactions within the ribosomal protein S8-rRNA complex from the archaeon Methanococcus jannaschii. *J. Mol. Biol.* **2001**, *311*, 311–324.

(16) Carter, A. P.; Clemons, W. M.; Brodersen, D. E.; Morgan-Warren, R. J.; Wimberly, B. T.; Ramakrishnan, V. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature* **2000**, *407*, 340–348.

(17) Nozinovic, S.; Furtig, B.; Jonker, H. R. A.; Richter, C.; Schwalbe, H. High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. *Nucleic Acids Res.* **2010**, *38*, 683–694.

(18) Sakata, T.; Hiroaki, H.; Oda, Y.; Tanaka, T.; Ikehara, M.; Uesugi, S. Studies on the Structure and Stabilizing Factor of the CUUCGG Hairpin Rna Using Chemically Synthesized Oligonucleotides. *Nucleic Acids Res.* **1990**, *18*, 3831–3839.

(19) Williams, D. J.; Hall, K. B. Unrestrained stochastic dynamics simulations of the UUCG tetraloop using an implicit solvation model. *Biophys. J.* **1999**, *76*, 3192–3205.

(20) Williams, D. J.; Boots, J. L.; Hall, K. B. Thermodynamics of 2′-ribose substitutions in UUCG tetraloops. *RNA* **2001**, *7*, 44–53.

(21) Leontis, N. B.; Westhof, E. Geometric nomenclature and classification of RNA base pairs. *RNA* **2001**, *7*, 499–512.

(22) Allain, F. H. T.; Varani, G. Structure of the P1 Helix from Group-I Self-Splicing Introns. *J. Mol. Biol.* **1995**, *250*, 333–353.

(23) Zirbel, C. L.; Sponer, J. E.; Sponer, J.; Stombaugh, J.; Leontis, N. B. Classification and energetics of the base-phosphate interactions in RNA. *Nucleic Acids Res.* **2009**, *37*, 4898–4918.

(24) Miller, J. L.; Kollman, P. A. Theoretical studies of an exceptionally stable RNA tetraloop: Observation of convergence from an incorrect NMR structure to the correct one using unrestrained molecular dynamics. *J. Mol. Biol.* **1997**, *270*, 436–450.

(25) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(26) Villa, A.; Widjajakusuma, E.; Stock, G. Molecular dynamics simulation of the structure, dynamics, and thermostability of the RNA hairpins uCACGg and cUUCGg. *J. Phys. Chem. B* **2008**, *112*, 134–142.

(27) Riccardi, L.; Nguyen, P. H.; Stock, G. Free-Energy Landscape of RNA Hairpins Constructed via Dihedral Angle Principal Component Analysis. *J. Phys. Chem. B* **2009**, *113*, 16660–16668.

(28) Deng, N. J.; Cieplak, P. Free Energy Profile of RNA Hairpins: A Molecular Dynamics Simulation Study. *Biophys. J.* **2010**, *98*, 627–636.

(29) Garcia, A. E.; Paschek, D. Simulation of the pressure and temperature folding/unfolding equilibrium of a small RNA hairpin. *J. Am. Chem. Soc.* **2008**, *130*, 815–817.

(30) Zuo, G. H.; Li, W. F.; Zhang, J.; Wang, J.; Wang, W. Folding of a Small RNA Hairpin Based on Simulation with Replica Exchange Molecular Dynamics. *J. Phys. Chem. B* **2010**, *114*, 5835–5839.

(31) Qin, P. Z.; Feigon, J.; Hubbell, W. L. Site-directed spin labeling studies reveal solution conformational changes in a GAAA tetraloop receptor upon Mg2+-dependent docking of a GAAA tetraloop. *J. Mol. Biol.* **2005**, *351*, 1–8.

(32) Reblova, K.; Razga, F.; Li, W.; Gao, H. X.; Frank, J.; Sponer, J. Dynamics of the base of ribosomal A-site finger revealed by molecular dynamics simulations and Cryo-EM. *Nucleic Acids Res.* **2010**, *38*, 1325–1340.

(33) Stombaugh, J.; Zirbel, C. L.; Westhof, E.; Leontis, N. B. Frequency and isostericity of RNA base pairs. *Nucleic Acids Res.* **2009**, *37*, 2294–2312.

(34) Heus, H. A.; Pardi, A. Structural Features That Give Rise to the Unusual Stability of RNA Hairpins Containing GNRA Loops. *Science* **1991**, *253*, 191–194.

(35) Pley, H. W.; Flaherty, K. M.; Mckay, D. B. 3-Dimensional Structure of a Hammerhead Ribozyme. *Nature* **1994**, *372*, 68–74.

(36) Scott, W. G.; Finch, J. T.; Klug, A. The Crystal-Structure of an all-RNA Hammerhead Ribozyme - a Proposed Mechanism for RNA Catalytic Cleavage. *Cell* **1995**, *81*, 991–1002.

(37) Correll, C. C.; Swinger, K. Common and distinctive features of GNRA tetraloops based on a GUAA tetraloop structure at 1.4 angstrom resolution. *RNA* **2003**, *9*, 355–363.

(38) Correll, C. C.; Wool, I. G.; Munishkin, A. The two faces of the Escherichia coli 23 S rRNA sarcin/ricin domain: The structure at 1.11 angstrom resolution. *J. Mol. Biol.* **1999**, *292*, 275–287.

(39) Correll, C. C.; Beneken, J.; Plantinga, M. J.; Lubbers, M.; Chan, Y. L. The common and the distinctive features of the bulged-G motif based on a 1.04 angstrom resolution RNA structure. *Nucleic Acids Res.* **2003**, *31*, 6806–6818.

(40) Correll, C. C.; Munishkin, A.; Chan, Y. L.; Ren, Z.; Wool, I. G.; Steitz, T. A. Crystal structure of the ribosomal RNA domain essential for binding elongation factors. *Proc. Natl. Acad. Sci.* **1998**, *95*, 13436–13441.

(41) Ban, N.; Nissen, P.; Hansen, J.; Moore, P. B.; Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 angstrom resolution. *Science* **2000**, *289*, 905–920.

(42) Szewczak, A. A.; Moore, P. B.; Chan, Y. L.; Wool, I. G. The Conformation of the Sarcin Ricin Loop from 28s Ribosomal-RNA. *Proc. Natl. Acad. Sci.* **1993**, *90*, 9581–9585.

(43) Szewczak, A. A.; Moore, P. B. The Sarcin Ricin Loop, a Modular RNA. *J. Mol. Biol.* **1995**, *247*, 81–98.

(44) Yang, X. J.; Gerczei, T.; Glover, L.; Correll, C. C. Crystal structures of restrictocin-inhibitor complexes with implications for RNA recognition and base flipping. *Nat. Struct. Biol.* **2001**, *8*, 968–973.

(45) Jucker, F. M.; Heus, H. A.; Yip, P. F.; Moors, E. H. M.; Pardi, A. A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J. Mol. Biol.* **1996**, *264*, 968–980.

(46) Menger, M.; Eckstein, F.; Porschke, D. Dynamics of the RNA hairpin GNRA tetraloop. *Biochemistry* **2000**, *39*, 4500–4507.

(47) Xia, T. B. Taking femtosecond snapshots of RNA conformational dynamics and complexity. *Curr. Opin. Chem. Biol.* **2008**, *12*, 604–611.

(48) Johnson, J. E.; Hoogstraten, C. G. Extensive Backbone Dynamics in the GCAA RNA Tetraloop Analyzed Using C-13 NMR Spin Relaxation and Specific Isotope Labeling. *J. Am. Chem. Soc.* **2008**, *130*, 16757–16769.

(49) Zhao, L.; Xia, T. B. Direct revelation of multiple conformations in RNA by femtosecond dynamics. *J. Am. Chem. Soc.* **2007**, *129*, 4118–4119.

(50) Sorin, E. J.; Engelhardt, M. A.; Herschlag, D.; Pande, V. S. RNA simulations: Probing hairpin unfolding and the dynamics of a GNRA tetraloop. *J. Mol. Biol.* **2002**, *317*, 493–506.

(51) Spackova, N.; Sponer, J. Molecular dynamics simulations of sarcin-ricin rRNA motif. *Nucleic Acids Res.* **2006**, *34*, 697–708.

(52) Depaul, A. J.; Thompson, E. J.; Patel, S. S.; Haldeman, K.; Sorin, E. J. Equilibrium conformational dynamics in an RNA tetraloop from massively parallel molecular dynamics. *Nucleic Acids Res.* **2010**, *38*, 4856–4867.

(53) Bowman, G. R.; Huang, X. H.; Yao, Y.; Sun, J.; Carlsson, G.; Guibas, L. J.; Pande, V. S. Structural insight into RNA hairpin folding intermediates. *J. Am. Chem. Soc.* **2008**, *130*, 9676–9678.

(54) Ferner, J.; Villa, A.; Duchardt, E.; Widjajakusuma, E.; Wohnert, J.; Stock, G.; Schwalbe, H. NMR and MD studies of the temperature-dependent dynamics of RNA YNMG-tetraloops. *Nucleic Acids Res.* **2008**, *36*, 1928–1940.

(55) Fadrna, E.; Spackova, N.; Stefl, R.; Koca, J.; Cheatham, T. E.; Sponer, J. Molecular dynamics simulations of guanine quadruplex loops: Advances and force field limitations. *Biophys. J.* **2004**, *87*, 227–242.

(56) Fadrna, E.; Spackova, N.; Sarzynska, J.; Koca, J.; Orozco, M.; Cheatham, T. E.; Kulinski, T.; Sponer, J. Single Stranded Loops of Quadruplex DNA As Key Benchmark for Testing Nucleic Acids Force Fields. *J. Chem. Theory Comput.* **2009**, *5*, 2514–2530.

(57) Ditzler, M. A.; Otyepka, M.; Sponer, J.; Walter, N. G. Molecular Dynamics and Quantum Mechanics of RNA: Conformational and Chemical Change We Can Believe In. *Acc. Chem. Res.* **2010**, *43*, 40–47.

(58) Wang, J. M.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(59) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. Refinenement of the AMBER force field for nucleic acids: Improving the description of alpha/gamma conformers. *Biophys. J.* **2007**, *92*, 3817–3829.

(60) MacKerell, A. D.; Banavali, N. K. All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. *J. Comput. Chem.* **2000**, *21*, 105–120.

(61) Besseova, I.; Otyepka, M.; Reblova, K.; Sponer, J. Dependence of A-RNA simulations on the choice of the force field and salt strength. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10701–10711.

(62) Klein, D. J.; Schmeing, T. M.; Moore, P. B.; Steitz, T. A. The kink-turn: a new RNA secondary structure motif. *EMBO J.* **2001**, *20*, 4214–4221.

(63) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B. P.; 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.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 10*; University of California: San Francisco, 2009.

(64) Banas, P.; Jurecka, P.; Walter, N. G.; Sponer, J.; Otyepka, M. Theoretical studies of RNA catalysis: Hybrid QM/MM methods and their comparison with MD and QM. *Methods* **2009**, *49*, 202–216.

(65) 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.

(66) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular-Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(67) 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.

(68) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(69) Ode, H.; Matsuo, Y.; Neya, S.; Hoshino, T. Force Field Parameters for Rotation Around chi Torsion Axis in Nucleic Acids. *J. Comput. Chem.* **2008**, *29*, 2531–2542.

(70) Yildirim, I.; Stern, H. A.; Kennedy, S. D.; Tubbs, J. D.; Turner, D. H. Reparameterization of RNA chi Torsion Parameters for the AMBER Force Field and Comparison to NMR Spectra for Cytidine and Uridine. *J. Chem. Theory Comput.* **2010**, *6*, 1520–1531.

(71) Riley, K. M.; Pitonak, M.; Jurecka, P.; Hobza, P., Stabilization and Structure Calculations for Noncovalent Interactions in Extended Molecular Systems Based on Wave Function and Density Functional Theories. *Chem. Rev.* **2010**, in press.

(72) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(73) Joung, I. S.; Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.

(74) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(75) Brooks, B. R.; Bruccoleri, R. E.; Olafson, D. J.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

(76) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant-Pressure Molecular-Dynamics Algorithms. *J. Chem. Phys.* **1994**, *101*, 4177–4189.

(77) Feller, S. E.; Zhang, Y. H.; Pastor, R. W.; Brooks, B. R. Constant-Pressure Molecular-Dynamics Simulation - the Langevin Piston Method. *J. Chem. Phys.* **1995**, *103*, 4613–4621.

(78) Perez, A.; Lankas, F.; Luque, F. J.; Orozco, M. Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res.* **2008**, *36*, 2379–2394.

(79) Lu, X. J.; Olson, W. K. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.* **2008**, *3*, 1213–1227.

(80) Mlynsky, V.; Banas, P.; Hollas, D.; Reblova, K.; Walter, N. G.; Sponer, J.; Otyepka, M. Extensive Molecular Dynamics Simulations Showing That Canonical G8 and Protonated A38H+ Forms Are Most Consistent with Crystal Structures of Hairpin Ribozyme. *J. Phys. Chem. B* **2010**, *114*, 6642–6652.

(81) Reblova, K.; Fadrna, E.; Sarzynska, J.; Kulinski, T.; Kulhanek, P.; Ennifar, E.; Koca, J.; Sponer, J. Conformations of flanking bases in HIV-1 RNA DIS kissing complexes studied by molecular dynamics. *Biophys. J.* **2007**, *93*, 3932–3949.

(82) Sarzynska, J.; Reblova, K.; Sponer, J.; Kulinski, T. Conformational transitions of flanking purines in HIV-1 RNA dimerization initiation site kissing complexes studied by Charmm explicit solvent molecular dynamics. *Biopolymers* **2008**, *89*, 732–746.

(83) Sponer, J.; Kypr, J. Different Intrastrand and Interstrand Contributions to Stacking Account for Roll Variations at the Alternating Purine-Pyrimidine Sequences in a-DNA and a-Rna. *J. Mol. Biol.* **1991**, *221*, 761–764.

(84) Bhattacharyya, D.; Bansal, M. A Self-Consistent Formulation for Analysis and Generation of Non-Uniform DNA Structures. *J. Biomol. Struct. Dyn.* **1989**, *6*, 635–653.

CT100481H

JCTC Journal of Chemical Theory and Computation

Effects of Water Placement on Predictions of Binding Affinities for p38α MAP Kinase Inhibitors

James Luccarelli, Julien Michel,† Julian Tirado-Rives, and William L. Jorgensen*

Department of Chemistry, Yale University, New Haven, Connecticut 06520-8107, United States

Received September 3, 2010

Abstract: Monte Carlo free energy perturbation (MC/FEP) calculations have been applied to compute the relative binding affinities of 17 congeneric pyridazo-pyrimidinone inhibitors of the protein p38α MAP kinase. Overall correlation with experimental data was found to be modest when the complexes were hydrated using a traditional procedure with a stored solvent box. Significant improvements in accuracy were obtained when the MC/FEP calculations were repeated using initial solvent distributions optimized by the water placement algorithm JAWS. The results underscore the importance of accurate placement of water molecules in a ligand binding site for the reliable prediction of relative free energies of binding.

Introduction

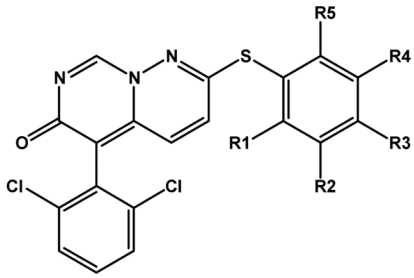
The accurate computation of free energies of binding of ligands is an important goal for computational chemistry that has the potential to improve the efficiency of drug discovery.[1,2] Numerous computational methodologies exist, but among these, free energy simulations are particularly attractive because they provide a formally rigorous way to compute free energies of binding. Nevertheless, progress is hindered by the limitations of classical force fields, difficulties in adequate sampling of protein and ligand flexibility with Monte Carlo (MC) or molecular dynamics methods (MD), and challenges in accurately taking into account changes in hydration.[3−7]

This report focuses on the impact of the initial placement of water molecules in a protein binding site on computed binding affinities of ligands. There is significant evidence in the literature that the computed free energies of binding can be strongly affected by the number and positions of water molecules present in a protein binding site.[8−11] For instance, information about the locations and thermodynamic properties of water molecules has been shown to substantially improve the scoring of protein−ligand interactions.[12] However, depending on the nature of the system under study, impossibly lengthy MC or MD simulations may be required before an equilibrium distribution of water molecules in the protein−ligand binding site is obtained. To efficiently address this issue, the water placement algorithm JAWS was recently developed.[13] The procedure has been shown to accurately detect hydration sites in protein−ligand complexes and has been used in conjunction with Monte Carlo free energy perturbation (MC/FEP) simulations to rationalize changes in free energies of binding for analogs that expel ordered water molecules from a protein binding site.[6] The preceding study, however, was concerned with a small number of protein−ligand complexes, where clear crystallographic evidence supporting a change in hydration between different analogs was available. Further investigation is desirable, especially in the context of lead optimization, where a large number of structurally related compounds may be considered, and for which subtle changes in hydration could affect the outcome of the free energy calculations. For this purpose, a series of 17 inhibitors of the protein p38α MAP kinase, previously reported by Pearlman and Charifson,[14] was chosen for detailed analyses (Figure 1). It is an attractive data set since it reflects a classic problem in medicinal chemistry, the optimal choice of substituents on a benzene ring. Furthermore, the series spans 2−3 orders of magnitude in activity, it involves typical small changes in substituents for a lead optimization exercise,[2] and it has been used as

* Corresponding author e-mail: william.jorgensen@yale.edu.

† Current address: Institute of Structural and Molecular Biology, The University of Edinburgh, Edinburgh, EH9 3JR, U.K.

10.1021/ct100504h
Published on Web 11/23/2010



Cmpd	R1	R2	R3	pIC_{50}
1	H	H	H	6.602
2	H	H	F	7.000
3	H	H	CH_3	5.854
4	H	Cl	Cl	6.097
5	H	CH_3	H	5.854
6	H	CH_3	CH_3	5.721
7	H	F	H	6.347
8	CH_3	H	H	6.699
9	H	Cl	F	6.301
10	H	Cl	H	6.553
11	CH_3	H	Cl	6.745
12	Br	H	H	6.602
13	CH_3	H	CH_3	6.577
14	OH	H	H	6.444
15	NH_2	H	F	6.658
16	Cl	H	F	7.444
17	F	H	F	8.046
18	F	H	H	N/A

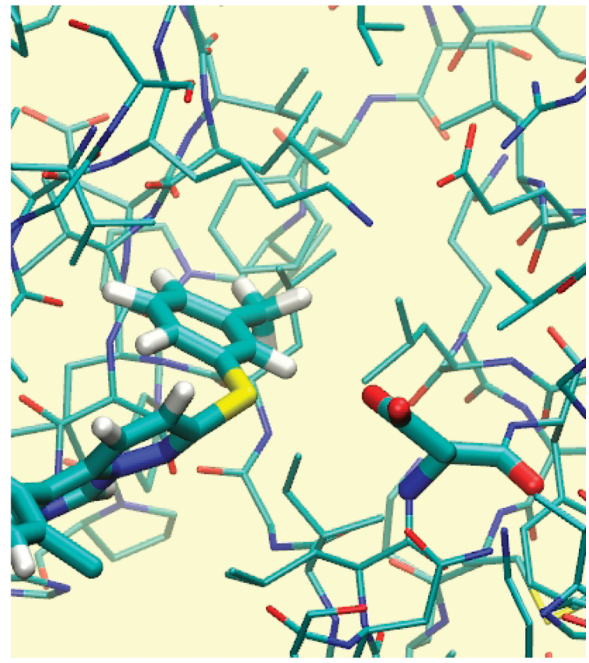


Figure 1. Left: Investigated inhibitors and measured activities for the inhibition of kinase activity with IC_{50} values in M (ref 14). Right: Computed image of ligand **5** bound in the "R1, R2" pose to p38α MAP kinase. As 180° flips of the thiophenyl group are not observed in the simulations, the R1 and R2 positions are considered distinct from the R4 and R5 positions. In the "R1, R2" pose, ligands **14** and **15** are able to hydrogen bond to nearby Asp168, drawn in thicker sticks. For clarity, some protein residues and all protein hydrogen atoms have been omitted. Compound **18**, which was not in the experimental study, was used as a convenient intermediate for the MC/FEP calculations. Image prepared using the software VMD.[16]

a benchmark to test alternative computational approaches for activity predictions.[14,15]

Methods

Protein Setup. An X-ray crystal structure of p38α MAP kinase in complex with a pyrido-pyrimidinone inhibitor (PDB ID: 1OUY),[17] which is very similar to **17**, provided the structural starting point. The ligand from the crystal structure was replaced with the parent ligand **1**, which was constructed with the program *BOSS*,[18] and protein and ligand Z matrixes were prepared using the programs *chop* and *pepz*.[18] Protein residues with any atom within 17.5 Å of a ligand atom were retained. The degrees of freedom of the side chains of protein residues with any atom within 12.5 Å of a ligand atom were sampled during the MC simulations. Backbone degrees of freedom and side chain bond lengths were kept frozen following a short conjugate-gradient relaxation. The net charge of the systems was set to zero by neutralizing protein residues distant from the ligand. The protonation states of histidine side chains were assigned with the assistance of the software PROPKA 2.0.[19] The OPLS-AA force field was used for the protein.[20]

Ligand Setup. Initial structures were generated using the molecule growing program *BOMB*.[21] The unsubstituted inhibitor **1** provided the core to grow the desired analogs. For consistency, the inhibitors were numbered as originally reported,[14] with the addition of the 1,3-difluoro compound as **17**. As the aromatic ring to which the substituents are attached is capable of rotating relative to the rest of the molecule, it is possible for the ligand to bind in two alternative binding modes, related by a 180° flip around the thioether bond (Figure 1). As these rotamers are not expected to interconvert during the MC simulations, structures were generated for both binding modes. For the MC simulations, the ligands were treated as fully flexible, and their energetics were represented with the OPLS/CM1A force field.[22] The CM1A atomic charges were scaled by 1.14.[23]

Solvent Setup. For the ligands alone in water, a 25 Å water cap was used containing ca. 2000 TIP4P water molecules. Each protein−ligand complex was solvated by ca. 1250 TIP4P water molecules in a ca. 25 Å radius water cap. A half-harmonic potential with a force constant of 1.5 kcal/mol/Å^2 was applied to water molecules at distances greater than 25 Å from the center of the system to prevent evaporation. The initial solvent distribution was derived from a stored solvent box using the default procedure with the *MCPRO* 2.1 program.[18] Specifically, a protein or ligand atom near the center of the binding site is taken as the origin of the system, and a cube containing 27 images of an equilibrated (298 K, 1 atm) cube of 512 TIP4P water molecules is centered on it. Each of the 13 824 water molecule is considered, and one is deleted if its oxygen atom is found to be within 2.5 Å of any non-hydrogen atom of a solute, or if it is outside the system boundary defined by the cap radius.

Though this straightforward procedure is typical of MC and MD programs, the number of retained water molecules and their initial coordinates depend on the choice of center atom. In principal, any associated artifacts should be removed if the MC or MD sampling is complete. However, it is easy to imagine that, for example, water molecules could be absent from or trapped in solute pockets and not be able to diffuse in or out of the pocket in the course of a simulation.

Alternatively, the initial water placement for the protein−ligand complexes was determined using the JAWS algorithm. The details of JAWS are described elsewhere.[13] Briefly, a 3-D cubic grid with 1 Å spacing is positioned to envelop the binding site. The grid region is defined by overlapping spheres of 4 Å radius, centered on user-selected solute atoms in the binding site. MC simulations are then performed to first find potential hydration sites and then to determine their occupancies. The putative hydration sites are detected by allowing "θ" water molecules to sample the grid volume while simultaneously scaling their intermolecular interactions between "on" and "off". The full system that is simulated consists of the protein, ligand, θ-water molecules, and regular water molecules. After the most probable sites are identified, a new MC simulation is run with θ water molecules constrained near the sites and with θ sampling. The absolute binding affinity of a water molecule at a given site is estimated from the ratio of probabilities that the water molecule is "on" or "off". The locations of hydration sites were determined using 5 million (5 M) MC configurations with sampling of just the water molecules, followed by 10 M configurations that sampled the water, protein, and ligand degrees of freedom. Then, the second phase covered 50 M configurations to estimate the occupancy of the sites.

Free Energy Calculations. Relative binding free energies for the ligands were computed from the standard thermodynamic cycle evaluating the free energy change in solution and in complex with the protein.[1−3] The free energy changes were computed with the *MCPRO* 2.1 program,[18] using Metropolis MC simulations to sample configurations of the system,[24] the single-topology technique for the structural perturbations,[25] and 11 windows of simple overlap sampling to compute the free energy change between the initial and final ligand structures.[26,27] For the ligands alone in water, each FEP window consisted of 10 M configurations of equilibration and 20 M configurations of averaging. For the protein−ligand complexes using the default water setup, the equilibration period was 12.5 M configurations for the first window and 10 M for the subsequent 10 windows. The windows were run serially; the initial configuration for windows 2−11 was based on the last configuration of the previous window and was well equilibrated. The simulations where the initial solvent coordinates came from the JAWS calculations were run at a later date using a higher-throughput protocol, whereby the 11 windows for each FEP calculation were run in parallel on 11 processors. In this case, equilibration for each window entailed 5 M configurations of water-only sampling, followed by 10 M configurations of full equilibration. For both protocols, the averaging period for each window was 10 M configurations. The averaging for the JAWS-based calculations was then extended to 20 M configurations for further checking of convergence. In all cases, evaluation of the potential energy employed 9 Å residue-based cutoffs, and the MC simulations were run at 298 K. The JAWS procedure takes about the same amount of computer time as running 2−3 FEP windows, so it adds ca. 25% to the overall computational effort.

A set of perturbations was devised to compute free energies of binding for all analogs relative to ligand **1**. In order to minimize the steric and electrostatic changes in each perturbation, consistent with past FEP studies, larger analogues were perturbed in multiple steps, e.g., OH → F → H or Cl → F → H.[11,28] Full details of all perturbations are provided in the Supporting Information. To account for the two possible "R1, R2" or "R4, R5" poses for the unsymmetrical ligands **4**−**17**, the relative free energies of binding of each pose were combined to produce an overall free energy of binding $\Delta\Delta G$ using eq 1

$$\Delta\Delta G = -RT \ln[\exp(-\Delta\Delta G_{R1,R2}/RT) + \exp(-\Delta\Delta G_{R4,R5}/RT)] + RT \ln 2 \quad (1)$$

where R is the ideal gas constant, T is 298 K, and $\Delta\Delta G_{R1,R2}$ and $\Delta\Delta G_{R4,R5}$ are the relative free energies of binding of the two poses. The second term in eq 1 penalizes the computed free energies of binding of the unsymmetrical ligands **4**−**17** by RT ln 2 because they are relative to the symmetrical ligand **1**. Thus, when the relative free energies of binding of the two poses differ by greater than ca. 2 kcal/mol, the free energy of binding is essentially that of the more favorable pose plus RT ln 2. Alternatively, if the relative free energies of binding of the two poses are the same, the RT ln 2 penalty is removed.

Though in this study the JAWS calculations were only applied to the initial state, the FEP calculations were run from the larger to smaller ligand to minimize the possibility of trapping water molecules by growing in the opposite manner. Another use for JAWS-like protocols would be to evaluate the preferred hydration pattern for the initial and final states of a proposed free-energy calculation. If significant differences were detected that would likely not be overcome by normal sampling, then alternative perturbation pathways could be considered.

Analysis. The agreement between predicted and measured free energies of binding was assessed by computing root-mean square deviations (RMSDs), mean unsigned errors (MUEs), and predictive indices (PIs). The latter has been proposed by Pearlman and Charifson to measure the quality of a rank-ordering by the potency of a series of ligands and is computed according to eq 2,[14]

$$\begin{aligned} \mathrm{PI} &= \frac{\sum_{j>i}\sum_{i} w_{ij}C_{ij}}{\sum_{j>i}\sum_{i} w_{ij}} \\ w_j &= |E(j) - E(i)| \\ C_{ij} &= -1 \text{ if } \frac{E(j) - E(i)}{P(j) - P(i)} < 0 \\ &= +1 \text{ if } \frac{E(j) - E(i)}{P(j) - P(i)} > 0 \\ &= 0 \text{ if } P(j) - P(i) = 0 \end{aligned} \quad (2)$$

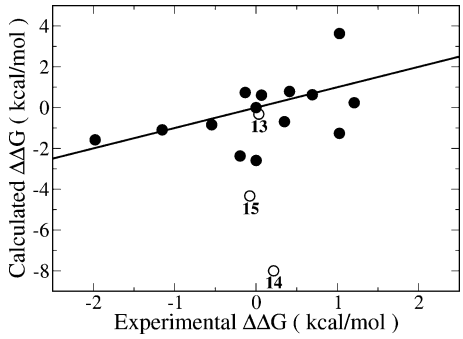


Figure 2. Calculated vs experimental relative free energies of binding for the p38 inhibitors using the conventional protocol with the stored water box. Free energy differences (kcal/mol) are relative to inhibitor **1**. Open circles indicate the computed relative free energy of binding for ligands **13**, **14**, and **15,** which are discussed in the main text.

where $E(i)$ and $P(i)$ are the experimental and predicted binding free energies of compound i. The PI index ranges from −1 to +1, depending on how well the predicted ranking matches the experimental ordering. A value of +1 indicates perfect predictions, a value of −1 indicates predictions that are perfectly anticorrelated, and a value of 0 arises from random results. In essence, the method considers each pair of compounds i and j in turn. Large differences in binding free energies have a large weight, w_{ij}, which provides a large positive contribution to the final PI, if the rank-ordering of the pair is correct. Conversely, if i and j have a small difference in measured binding affinity, an incorrect prediction of the most potent binder has a minor impact on the final PI.

Results and Discussion

Approximate free energies of binding for the inhibitors were obtained from the experimental pIC_{50} values.[14] While the relationship between K_i and $\Delta\Delta G$ is linear, the correlation between IC_{50} and K_i is not exact; thus pIC_{50} and $\Delta\Delta G$ cannot be expected to be perfectly linearly related.

The MC/FEP results using the default hydration protocol are compared with the experimental data in Figure 2. The overall RMSD is 2.65 kcal/mol, while the MUE is 1.69 kcal/mol, and the PI is 0.41, representing modest predictive power.[14] Four of the ligands (**8**, **9**, **10**, and **17**) were found to bind most favorably in the "R4, R5" mode, while the rest bound in the "R1, R2" mode. The most significant outliers from these calculations were the 2-hydroxyl and 2-amino-substituted ligands **14** and **15**, which are capable of hydrogen bonding to the carboxylate group of Asp168. No obvious features stand out for the errors for the remaining ligands. An inspection of snapshots from the calculations, however, revealed substantial inconsistencies between different ligands in the number and positioning of water molecules within the binding site.

For instance, as illustrated in Figure 3, two water molecules were placed in the phenyl substituent pocket for the *meta*-fluoro analog **7**, but they were absent for the *meta*-chloro analog **10**. One water molecule is in a fairly hydrophobic environment and can donate only a single

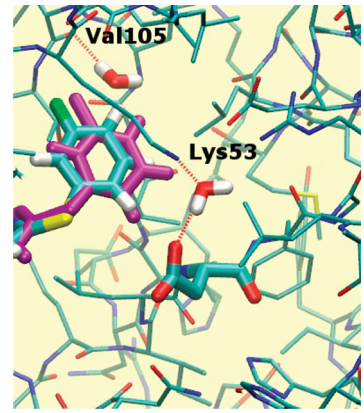


Figure 3. Hydration sites in the vicinity of the p38 inhibitor **7** (colored sticks) obtained using the default hydration protocol. These hydration sites are not observed for inhibitor **10** (purple sticks). Ligand and Asp168 atoms are drawn in thicker sticks. Other water molecules, selected protein residues, and all protein hydrogen atoms have been omitted for clarity. Hydrogen bonding interactions between protein atoms and water molecules are depicted by dotted red lines.

hydrogen bond to the backbone carbonyl of Val105. Thus, though there is sufficient space to insert a water molecule in this region of the binding site when **7** is bound, it is unclear whether this would be thermodynamically favorable. The other water molecule is involved in strong hydrogen bonding interactions with Lys53 and Asp168, and it is doubtful that it should be absent when **10** is bound.

Clarification of the water distributions in the binding site was sought using the water placement algorithm JAWS for each ligand. These calculations revealed the presence of several hydration sites within the binding pocket that were inconsistently found when the stored solvent box was used. To test whether consistency in solvent distribution alone was sufficient to improve accuracy, the MC/FEP calculations were repeated starting with the solvent distribution computed using the JAWS protocol for the complex of **1** for all complexes. This resulted in only marginal improvement over the original results, with an RMSD of 2.52 kcal/mol, a MUE of 1.95 kcal/mol, and a PI of 0.55 (see the Supporting Information). While using the same initial solvent distribution eliminated errors resulting from varying numbers of water molecules, occasional large errors were introduced in instances where the water distribution derived for the smallest analog (**1**) was used for simulations of much larger analogs, for example, the 3,4-dimethyl one (**6**). In these cases, some water molecules were observed to be trapped in high-energy configurations between the protein and ligand. The extensive sampling of ligand, protein, and solvent degrees of freedom required to resolve such steric problems is not systematically achieved with the standard MC simulation protocol.

To eliminate this source of error, the MC/FEP calculations were repeated with the JAWS-derived water distributions for each starting ligand state. This resulted in much improved accuracy, with the errors roughly halved. For 10 M configurations of averaging, the RMSD for all ligands is reduced to 1.35 kcal/mol, the MUE to 0.95 kcal/mol, and the PI improved to 0.62 (Figure 4). These results changed little upon extension of the averaging period to 20 M configurations;

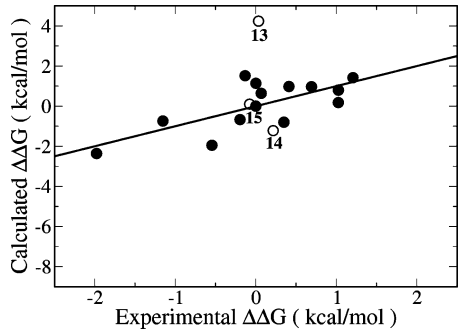


Figure 4. Calculated vs experimental relative free energies of binding for the p38 inhibitors using the JAWS protocol to determine initial water coordinates. Details are the same as in Figure 2. To aid in comparison with Figure 2, the axis scales and data symbols are the same.

the RMSD, MUE, and PI became 1.44, 0.92, and 0.61 (Supporting Information). Only ligand **17** was found to prefer binding in the "R4, R5" pose, which is observed in the 3FC1 crystal structure.[15] However, the relative free energies of binding for ligand **5** in both poses were the same within 0.2 kcal/mol; thus, it does not have a preference. The "R1, R2" pose is calculated to be significantly more favorable for the remaining unsymmetrical ligands.

A detailed analysis of the output of the MC/FEP simulations was undertaken to elucidate the improvements in binding affinity predictions using the JAWS-derived water distributions. An analysis of the hydrogen-bonding ligands **14** and **15**, which were previously predicted to bind overly favorably, revealed that the JAWS calculations located two additional hydration sites in the vicinity of Asp168, as illustrated in Figure 5 (bottom) for the hydroxyl analog **14**. These hydration sites are located in a cavity partially shielded from bulk solvent and were not populated using the default solvation protocol (Figure 5, top). The first water molecule receives two hydrogen bonds from the ligand's hydroxyl group and Lys53 and donates two hydrogen bonds to the other water molecule and Asp168. The second water molecule also donates two hydrogen bonds to Glu71 and the backbone carbonyl of Phe169. The occupancy of the additional hydration sites can be expected to affect the outcome of the FEP calculations. In the MC/FEP simulations equilibrated following the default solvent-box protocol, the hydroxyl group of **14** is donating a hydrogen bond to Asp168 (Figure 5, top). In the MC/FEP simulations equilibrated after the JAWS setup, the carboxylate group of Asp168 ends up rotated away from its initial position to accommodate better the additional water molecules. The interaction between the hydroxyl group and Asp168 is no longer direct, but it is water-mediated. Consequently, the addition of the *meta* hydroxyl or amino group onto the phenyl ring is less favorable, in agreement with the experimental activity measurements. Specifically, the relative binding affinity of **14** was computed in two steps, **14**→**18** and **18**→**1**. Compared with the original simulations, the relative free energy of binding of **14**→**18** is 3.3 kcal/mol less favorable for **14**, and the relative free energy of binding of **18**→**1** is 3.5 kcal/mol less favorable for **18** with the JAWS setup. Overall, the error in the computed free energy of binding for **14** is reduced

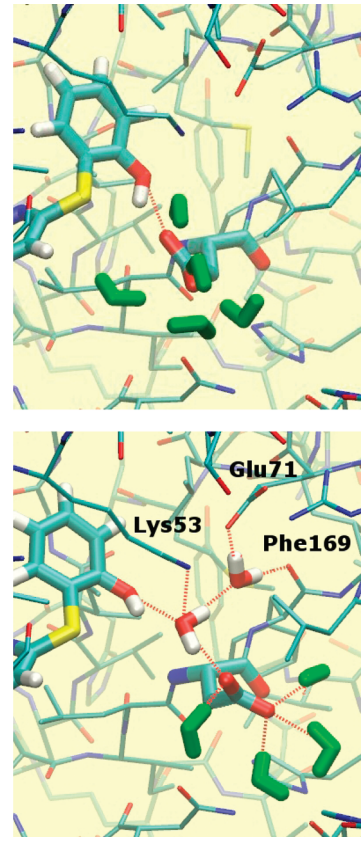


Figure 5. Representative snapshots from MC/FEP simulations for inhibitor **14**. Top: The solvent distribution that originated from the default procedure. Bottom: The solvent distribution obtained after equilibration using JAWS. Ligand and Asp168 atoms are drawn in thicker sticks. Hydrogen bonding interactions between the ligand hydroxyl group or buried water molecules are depicted by dotted red lines. Solvent exposed water molecules solvating Asp168 are shown in green sticks. Other water molecules, selected protein residues, and all protein hydrogen atoms have been omitted for clarity.

from 8.2 to 1.4 kcal/mol. Similarly, the relative binding affinity of **15** was computed in two steps, **15**→**2** and **2**→**1**. Compared with the original simulations, the relative free energy of binding of **15**→**2** is 5.5 kcal/mol less favorable for **15**, and the relative free energy of binding of **2**→**1** is 1.1 kcal/mol more favorable for **1**. Overall, the error in the computed free energies of binding for **15** is reduced from 4.2 to 0.2 kcal/mol.

The most significant remaining error is for the 2,4-dimethyl compound **13**, whose relative binding affinity is too unfavorable by 4.2 kcal/mol with the JAWS simulation protocol, whereas the initial simulations yielded an error of only 0.4 kcal/mol. The relative binding free energy of **13** was computed in two steps, **13**→**3** and **3**→**1**. The small error for **13** using the traditional solvent-box protocol is fortuitous because it represents a cancellation of errors for the two steps, +1.9 and −2.3 kcal/mol, respectively. With the JAWS protocol, the corresponding errors are +5.0 and −0.8 kcal/mol, so the problem was predominantly in the **13**→**3** step. One possibility is that the starting water distribution for **13** was not appropriate for or did not evolve properly for **3**. As pointed out previously, substantial errors can be expected if

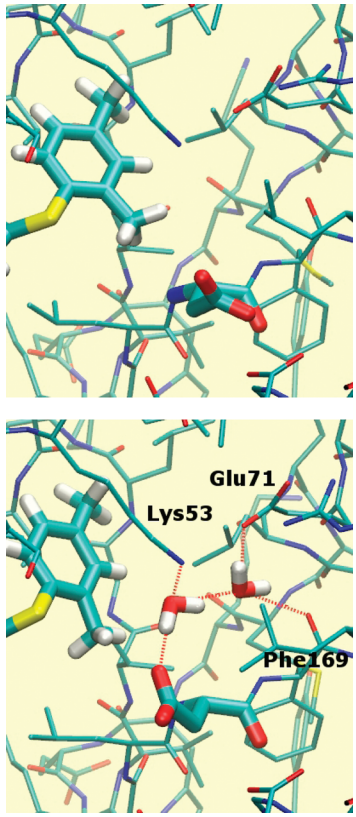


Figure 6. Representative snapshots from MC/FEP simulations of inhibitor **13** bound to p38α MAP kinase. Top: The configuration near Asp168 that arose in the simulation starting from the default hydration procedure. Bottom: The configuration near Asp168 after equilibration using JAWS. Ligand and Asp168 atoms are drawn in thicker sticks.

a perturbation induces changes in hydration in the binding site, unless the computed relative free energies of binding are corrected by computing the absolute free energy of binding of the displaced water molecules.[6] However, this situation does not seem to occur in the present case since the JAWS-computed hydration patterns for the starting and ending states, **13** and **3**, are identical.

The problem with the **13**→**3** perturbation appears to be more complex and associated with the conformation of the Asp168 side chain. From the free energy changes for the individual FEP windows, it is apparent that the ca. 3 kcal/mol difference in free energy changes between the two simulations arises at the beginning of the perturbation, when the 2-methyl group of **13** starts to be shrunk into a hydrogen atom (see the Supporting Information). A visualization of snapshots saved during the simulations with the traditional hydration protocol reveals that the carboxylate group of Asp168, which was initially pointed toward the ligand, rotates away from the 2-methyl group (Figure 6, top). This did not occur in the JAWS equilibrated simulations, presumably because rotation of the carboxylate group would break the hydrogen bond with one of the two nearby water molecules placed using JAWS (Figure 6, bottom). The MC/FEP calculations were repeated using the same JAWS-derived solvent distribution, but with Asp168 rotated to adopt the conformation observed with the solvent-box protocol. The computed change in free energy of binding for **13**→**3** became −2.0 ± 0.2 kcal/mol, which is intermediate between the results obtained with the solvent box and JAWS protocols. In turn, this reduces the error for **13**→**1** to 2.5 kcal/mol. Thus, the conformations of Asp168, which is part of the flexible DFG motif, are likely not adequately sampled in the present MC/FEP simulations. This highlights complexities associated with the fact that hydration of the system and the conformation of the ligand and protein are all coupled. The hydration may be set up properly for one conformation, but it is possible that the system relaxes away from this conformation to one that would prefer a different population of water molecules that cannot be achieved with computationally reasonable sampling periods.

Finally, it is worth reflecting on the degree of agreement that can actually be achieved between the computations and the experiment, given the uncertainties in the measured activity data. As noted previously, the conversion of ΔIC_{50}s into $\Delta\Delta G$'s of binding is approximate, but another source of error is the variability of the IC_{50} measurements themselves. Although uncertainties were not reported for the experimental data used here,[14,15] a study of a large corporate database found a median standard deviation for activity measurements of approximately 0.3 log unit.[29] This corresponds to a factor of 2 in IC_{50} or ±0.41 kcal/mol. Our own experiences with repeated measurements for compounds used as standards in multiple biological assays are similar.[2,21,28] Drawing samples from a Gaussian distribution centered around the reported IC_{50}'s for each ligand according to this standard deviation, the predictive index (eq 2) can be computed between two independent simulated activity measurements for the entire data set. Following a procedure similar to the one reported by Brown et al.,[29] the sensitivity of the PI to uncertainties in the measured IC_{50}'s is derived by repeating the calculation 1 million times. Assuming the above-mentioned errors, the median achievable PI is 0.76 for this data set. Though the distribution of PI values is not Gaussian (see the Supporting Information), approximately 67% of PI measurements would fall within the range 0.67−0.84. The median achievable PI for this data set is thus below unity because the error bars on the measured IC_{50}'s are large enough to qualitatively change the rankings of some of the ligands. Given these considerations, the improvement of the PI from 0.41 to 0.62 upon using a JAWS-optimized water distribution for the MC/FEP simulations is reinforced as being significant. The PI of 0.62 is also greater than the PI achieved for this data set by various scoring function and MM-PBSA approaches that were previously tested.[14,15] The only higher PI, an impressive 0.85, was achieved using thermodynamic integration and molecular dynamics with the Amber program.[14]

Conclusion

The results presented here illustrate that the initial placement of water molecules can significantly affect the outcome of computations of protein−ligand binding affinities. Details such as this need to be considered to allow current computational methods to evolve to the accuracy required for routine, reliable guidance of lead-optimization programs in drug discovery and of molecular design in general. It was

found that optimization of the distribution of water molecules in the protein−ligand binding site using the water placement algorithm JAWS substantially improved the quality of subsequent MC/FEP results for a data set of 17 inhibitors of p38α MAP kinase. Use of the JAWS-derived water distributions reduced the RMSD for relative free energies of binding from 2.65 to 1.35 kcal/mol and improved the predictive index (eq 2) from 0.41 to 0.62. Though further optimization of JAWS and other water-placement procedures is possible,[12,13] additional issues affecting the outcome of free-energy calculations also continue to warrant concerted attention.[1−3] Force-field and sampling problems remain, and it is sometimes necessary to consider more complex perturbation cycles where binding-site water molecules must be forced to disappear.[6,8−10] As pointed out here in the context of Figure 6, the complexity of sampling issues can be great, as it simultaneously involves all components of the modeled systems. Initial choices for the placement of water molecules, every dihedral angle in the ligand and the protein, and the protonation state of each ionizable residue can all have ramifications that are not removed by standard sampling procedures. Nevertheless, the present results have demonstrated that significant gains in accuracy can be realized by more thorough consideration of initial water placement in calculations of the free energetics of protein−ligand binding.

Acknowledgment. Gratitude is expressed to the National Institutes of Health (GM32136) for support of this research and to Dr. Paul S. Charifson for helpful discussions. J.M. also acknowledges support from a Marie Curie International Outgoing Fellowship from the European Commission (FP7-PEOPLE-2008-4-1-IOF, 234796-PPIdesign).

Supporting Information Available: Results from MC calculations using a single JAWS-derived solvent distribution. Results from each perturbation calculation and from extending the MC/FEP calculations to 20 M configurations of averaging. Plot of the distribution of achievable PI values given the uncertainties in assay results (11 pages). This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Chipot, C.; Pohorille, A. In *Springer Series in Chemical Physics*; Chipot, C., Pohorille, A., Eds.; Springer-Verlag: Berlin, 2007; Vol 86: Free Energy Calculations: Theory and Applications in Chemistry and Biology, pp 33−75.

(2) Jorgensen, W. L. *Acc. Chem. Res.* **2009**, *42*, 724–733.

(3) Michel, J.; Foloppe, N.; Essex, J. W. *Mol. Inf.* **2010**, *29*, 570–578.

(4) Williams, D. H.; Stephens, E.; O'Brien, D. P.; Zhou, M. *Angew. Chem., Int. Ed.* **2004**, *43*, 6596–6616.

(5) Olsson, T. S.; Williams, M. A.; Pitt, W. R.; Ladbury, J. E. *J. Mol. Biol.* **2008**, *384*, 1002–1017.

(6) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2009**, *131*, 15403–15411.

(7) Michel, J.; Essex, J. W. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 639–658.

(8) Helms, V.; Wade, R. C. *J. Am. Chem. Soc.* **1998**, *120*, 2710–2713.

(9) Price, M. L. P.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2000**, *122*, 9455–9466.

(10) Deng, Y.; Roux, B. *J. Chem. Phys.* **2008**, *128*, 115103.

(11) Leung, C. S.; Zeevaart, J. G.; Domaoal, R. A.; Bollini, M.; Thakur, V. V.; Spasov, K. A.; Anderson, K. S.; Jorgensen, W. L. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 2485–2488.

(12) (a) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 808–813. (b) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. *J. Am. Chem. Soc.* **2008**, *130*, 2817–2831.

(13) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2009**, *113*, 13337–13346.

(14) Pearlman, D. A.; Charifson, P. S. *J. Med. Chem.* **2001**, *44*, 3417–3423.

(15) Pearlman, D. A. *J. Med. Chem.* **2005**, *48*, 7796–7807.

(16) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–45.

(17) Fitzgerald, C. E.; Patel, S. B.; Becker, J. W.; Cameron, P. M.; Zaller, D.; Pikounis, V. B.; O'Keefe, S. J.; Scapin, G. *Nat. Struct. Mol. Biol.* **2003**, *10*, 764–769.

(18) Jorgensen, W. L.; Tirado-Rives, J. *J. Comput. Chem.* **2005**, *26*, 1689–1700.

(19) Bas, D. C.; Rogers, D. M.; Jensen, J. H. *Proteins* **2008**, *73*, 765–783.

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

(21) Barreiro, G.; Kim, J. T.; Guimarães, C. R. W.; Bailey, C. M.; Domaoal, R. A.; Wang, L.; Anderson, K. S.; Jorgensen, W. L. *J. Med. Chem.* **2007**, *50*, 5324–5329.

(22) Jorgensen, W. L.; Tirado-Rives, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665–6670.

(23) Udier-Blagovic, M.; Morales De Tirado, P.; Pearlman, S. A.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 1322–1332.

(24) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

(25) Jorgensen, W. L.; Ravimohan, C. *J. Chem. Phys.* **1985**, *83*, 3050.

(26) Lu, N.; Kofke, D. A.; Woolf, T. B. *J. Comput. Chem.* **2004**, *25*, 28–40.

(27) Jorgensen, W. L.; Thomas, L. L. *J. Chem. Theory Comput.* **2008**, *4*, 869–876.

(28) Leung, S. S. F.; Tirado-Rives, J.; Jorgensen, W. L. *Bioorg. Med. Chem.* **2009**, *17*, 5874–5886.

(29) Brown, S. P.; Muchmore, S. W.; Hajduk, P. J. *Drug Discovery Today* **2009**, *14*, 420–427.

CT100504H