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 © 2010 American Chemical Society
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 $f^{GB} = [r_{ij}^2 + R_iR_j \exp(-r_{ij}^2/4R_iR_j)]^{0.5}$, 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

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

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:

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

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:

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

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

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

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:

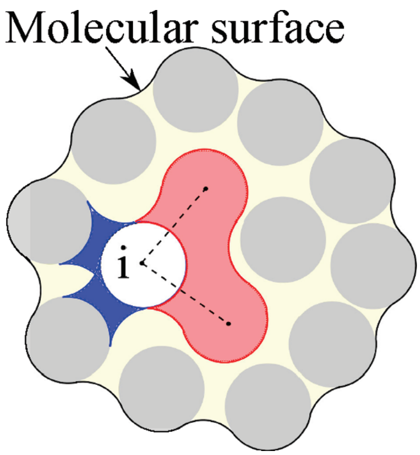

Figure 1. Illustration of the three regions of integration in eq 7 that are combined to approximate the molecular volume: VDW volume (light gray spheres), neck regions (dark blue), and "chunk" molecule (red). The open sphere represents atom *i*, and the dashed lines represent covalent bonds used to define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

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

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*: $r_{ij} \geq \rho_i + \rho_j$

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

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

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

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

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

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

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

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

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

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 $\rho_i + \rho_j + 2\rho_{\text{w}}$. To approximate the integral of $|\mathbf{r}|^{-6}$ over the "neck" region, we use the following analytical and empirical function:

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

for interatomic distances (r_{ij}) less than $\rho_i + \rho_j + 2\rho_{\text{w}}$ 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

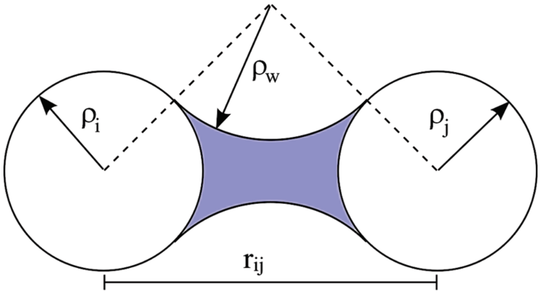

Figure 2. Neck region (blue) between two atoms with radii ρ_i and ρ_j and a water probe radius ρ_w. r_{ij} represents the distance between atoms *i* and *j*.

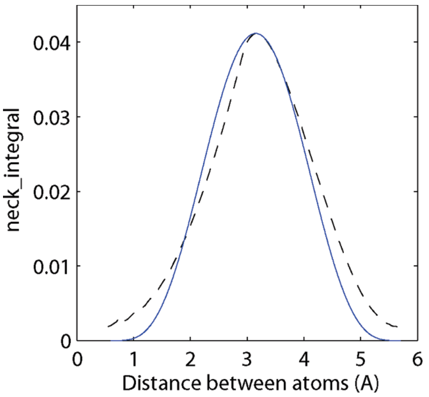

Figure 3. The numerical integration over the "neck" region (dashed black) compared with the analytical approximation (solid blue) used here. In this example, we have used ρ_i = 1.7, ρ_j = 1.2, and probe radius ρ_w = 1.4 Å.

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

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

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 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*, $\mathbf{I}_i^{\text{chunk}}$, can be computed essentially exactly by the NSR6 procedure at the setup stage and then subsequently reused at all other steps.

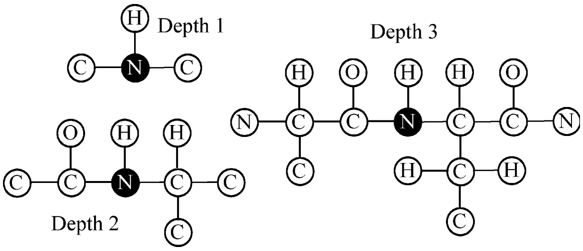

Figure 4. Examples of "chunk" molecules of depths 1, 2, and 3 used for the computation of the effective Born radius of a nitrogen atom (black circles).

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:

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

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

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 $\Psi \rightarrow \infty$, then $R_i \rightarrow A$. 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

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

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, $\{S_{\mathrm{vdw}},S_{\mathrm{neck}}\}$ 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.

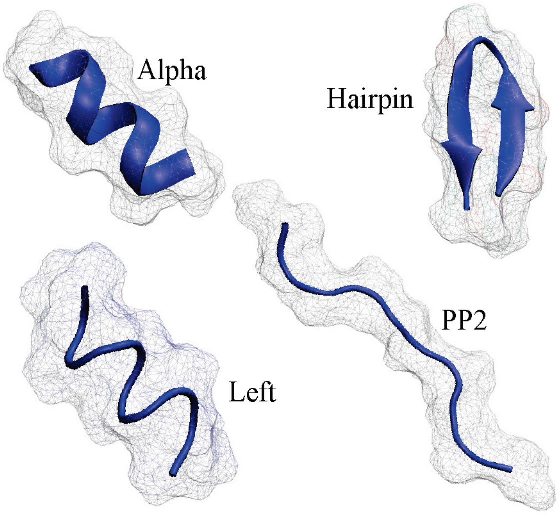

Figure 5. Cartoon representation of the four conformational states of alanine decapeptide, Ala10, used in this work.

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 $\Delta\Delta G_{\mathrm{el}}$ 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.

Table 2. Optimized Parameters

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

"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.

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.

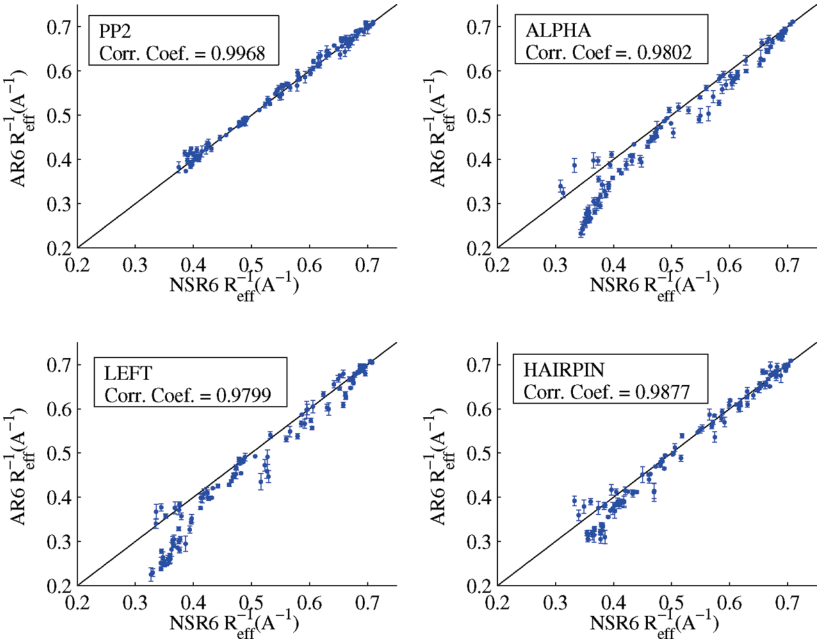

Figure 6. Comparison of the inverse of the approximated R6 effective Born radii (AR6) with the exact R6 effective Born radii (NSR6) for the four conformational states of Ala10. Every point represents the average Born radius over four possible "chunks", with the error bars representing standard deviations.

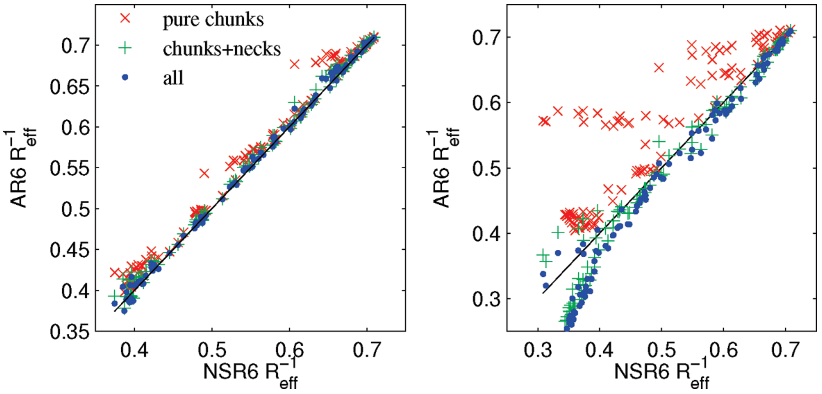

Figure 7. Comparison of the inverse of the approximated R6 effective Born radii (AR6) with the exact R6 effective Born radii (NSR6) for the pp2 (left) and alpha (right) conformational states of Ala10. Red × marks, AR6 with only the chunks contribution ($S_{vdw} = S_{neck} = 0$). Green + marks, AR6 with 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.

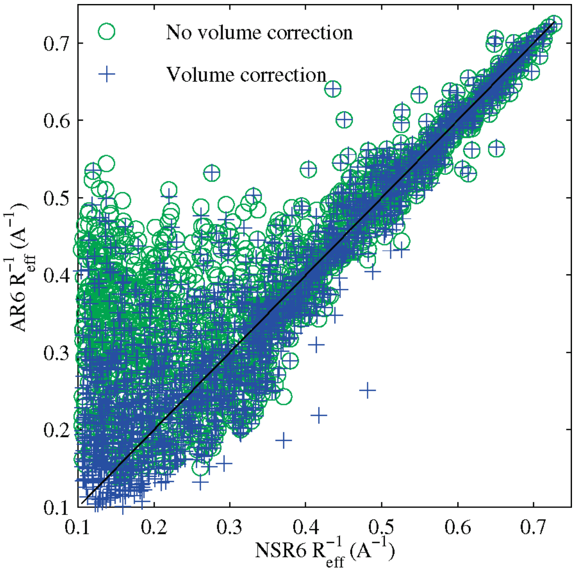

Figure 8. Comparison of the inverse of the approximated R6 effective Born radii (AR6) with the "exact" R6 effective Born radii (NSR6) for thieredoxin (2TRX). Green circles, AR6 with no additional volume correction ($\beta_1 = \beta_2 = 0$). Blue plus marks, AR6 with optimized parameters from Table 2.

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},

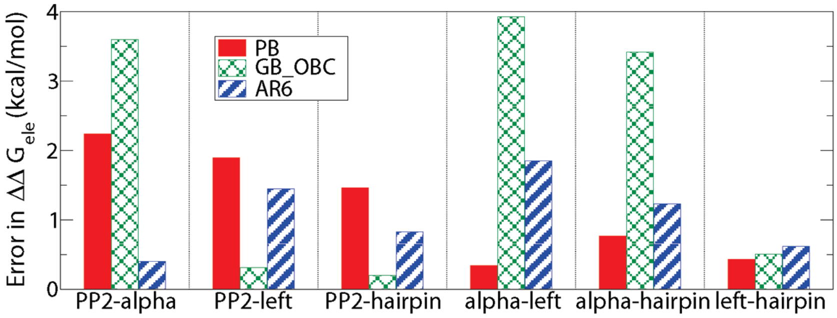

Figure 9. Absolute error in $\Delta\Delta G_{el}$, relative to the explicit solvent model, between four different conformational states of Ala10 (alpha, PP2, left, and hairpin). The energies were obtained using PB (solid red bars), GB_OBC (cross-hatched green bars), and AR6 (striped blue bars). The $\Delta\Delta G_{el}$ 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 $\Delta\Delta G_{\text{el}}$ 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 $100(\Delta G_{\text{el}}(\text{GB}) - \Delta G_{\text{el}}(\text{PB}))/|\Delta G_{\text{el}}(\text{PB})|$ 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 $\Delta\Delta G_{\text{el}}$ between the PB and AR6 for the denaturation trajectories of apo-myoglobin and protein A. The results are summarized

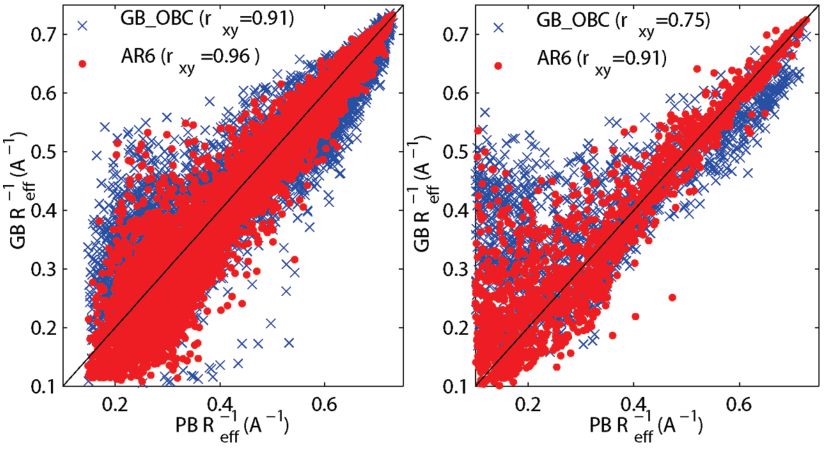

Figure 10. Comparison of the inverse of the approximated effective Born radii (GB R_{eff}^{-1}) with the "perfect" effective Born radii (PB R_{eff}^{-1}) for 19 small proteins (left) and thioredoxin (right). Approximated effective radii were computed by AR6 (red) and GB_OBC (blue). Correlation coefficients 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 $\varepsilon_{\mathrm{out}} = 1000$, $\varepsilon_{\mathrm{in}} = 1$, 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, $\Delta\Delta G = \Delta G_{\mathrm{el}}(\mathrm{N}) - \Delta G_{\mathrm{el}}(\mathrm{U})$ [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, $\Delta\Delta G_{\mathrm{el}}$ 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 $\Delta\Delta G_{\mathrm{el}}$ 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 $\Delta\Delta G_{\mathrm{el}}$.

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 $\Delta\Delta G_{\mathrm{el}}$ 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

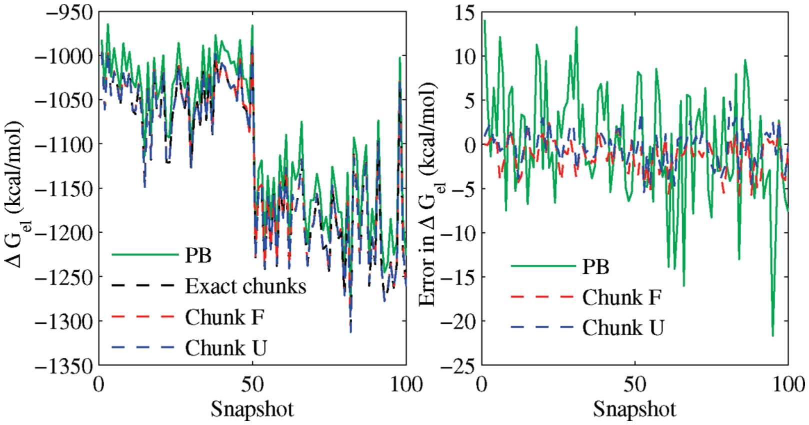

Figure 11. Left: solvation energy along the MD unfolding trajectory of protein-A (PDB ID: 1BDD) obtained by PB (solid lines) and AR6 (dashed lines). The AR6-based energies were obtained using “chunks” from one snapshot of the folded (chunk F, red) and unfolded (chunk U, blue) states, and from “chunks” 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 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.

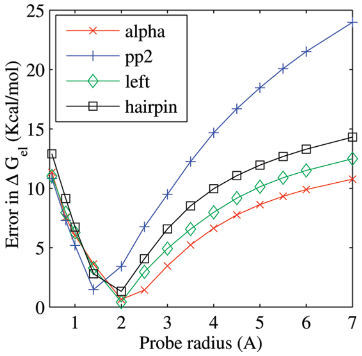

Figure 12. Absolute error of the numerical PB ΔG_{el}, relative to the explicit solvent (TIP3P) reference, as a function of the probe radius used to set the dielectric boundary in the PB calculations. The computations are performed for the four conformational states of alanine decapeptide shown in Figure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

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

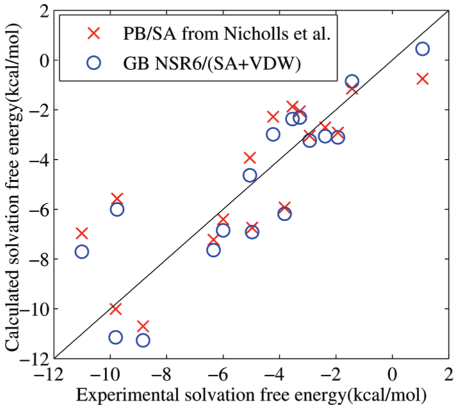

Figure 13. Comparison of computed solvation energies with the experimentally determined solvation energies for a "challenging" data set of small drug-like molecules. Blue: PB plus the cavity term (PB/SA) approach from Nicholls et al.[63] Red: GB NSR6 plus cavity and van der Waals terms (GB NSR6/(SA+VDW)).

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:

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

We compute ΔG_{cav} by applying an atom-independent surface tension γ to the solvent-accessible surface area (SASA): $\Delta G_{\text{cav}} = \gamma(\text{SASA})$, 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]

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

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 $\varepsilon_{\text{out}} = 1000$ to mimic the conductor limit $\varepsilon_{\text{out}} \rightarrow \infty$ 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 $\varepsilon_{\text{out}} = 80$, $\varepsilon_{\text{in}} = 1$, 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]

$$a_i = -\frac{16}{3}\pi d_{\text{w}} \varepsilon_{i\text{w}} \sigma_{i\text{w}}^6 \tag{28}$$

where $d_w = 0.033428$ Å^{-3} is the number density of water at standard conditions; $\varepsilon_{i\text{w}}$ and $\sigma_{i\text{w}}$ are computed by

$$\sigma_{i\text{w}} = \sqrt{\sigma_i \sigma_{\text{w}}} \tag{29}$$

$$\varepsilon_{i\text{w}} = \sqrt{\varepsilon_i \varepsilon_{\text{w}}} \tag{30}$$

where $\sigma_{\text{w}} = 1.7683$ Å and $\varepsilon_{\text{w}} = 0.1520$ 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