Improving of the accuracy and efficiency of implicit solvent models in Biomolecular Modeling

Boris Aguilar Huacan

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Alexey Onufriev, Chair
Chandrajit Bajaj
David Bevan
Yang Cao
Adrian Sandu

January 23, 2013
Blacksburg, Virginia

Keywords: Molecular Modeling, Implicit solvents , Generalized Born Model

Improving of the accuracy and efficiency of implicit solvent models in Biomolecular Modeling

Boris Aguilar Huacan

(ABSTRACT)

Biomolecular Modeling is playing an important role in many practical applications such as biotechnology and structure-based drug design. One of the essential requirements of Biomolecular modeling is an accurate description of the solvent (water). The challenge is to make this description computationally facile –that is reasonably fast, simple, robust and easy to incorporate into existing software packages. The most rigorous procedure to model the effect of aqueous solvent is to explicitly model every water molecule in the system. For many practical applications, this approach is computationally too intense, as the number of required water atoms is on average one order of magnitude larger than the number of atoms of the molecule of interest.

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 computationally more efficient. The Generalized Born (GB) implicit solvent has become quite popular due to its relative simplicity and computational efficiency. However, recent studies showed serious deficiencies of many GB variants when applied to Biomolecular Modeling such as an over-stabilization of alpha helical secondary structures and salt bridges.

In this dissertation we present two new GB models aimed at computing solvation properties with a reasonable compromise between accuracy and speed. The first GB model, called NSR6, is based on a numerically surface integration over the standard molecular surface. When applied to a set of small drug-like molecules, NSR6 produced an accuracy, with respect to experiments, that is essentially at the same level as that of the expensive explicit solvent treatment. Furthermore, we developed an analytic GB model, called AR6, based on an approximation of the volume integral over the standard molecular volume. The accuracy of the AR6 model is tested relative to the numerically exact NSR6. Overall AR6 produces a good accuracy and is suitable for Molecular Dynamics simulations which is the main intended application.

Dedication

To my father Pablo Cesar Aguilar and my girlfriend and best friend Grecia Heredia, for their unconditional support.

Acknowledgments

I would like to thank first and foremost Dr. Alexey Onufriev for his guidance, support, and advice. In addition, many of my fellow students and colleagues have contributed directly and indirectly to the succes of my research. In particular, I would like to thank Reinhard Laubenbacher, Ramu Anandakrishnan, Igor Tolokh, Abhishek Mukhopadhyay, Saeed Izadi, Andrew Fenley, Lida Uribe, Bartleby Ordonez, David Murrugarra, Edgar Saenz, Cristina Letona, Moises Guerra, and Alan Veliz-Cuba.

Contents

List of Figures

List of Tables

Glossary

Before entering into the main part of this dissertation, we introduce the definition of the following key terms used throughout this document.

Poisson Boltzmann Model A commonly used model to describe the electrostatic properties of a molecule in an aqueous medium. A molecule is represented as a set of spheres whose internal dielectric properties differs from the external solvent. It is based on the Poisson equation that is solved by a variety of standard numerical methods.

Generalized Born model An analytical approximation to the Poisson-Boltzmann model, see equation 3.1 of chapter 3.

Solvation Energy Represents the gain in energy for transferring a molecule from vacuum to an aqueous medium. The solvation energy plays a key role in Implicit Solvent Models because it represents the interaction between the water and the molecule of interest.

Intrinsic Radii The radii of each atom of a molecule. In Biomolecular Modeling atoms are typically represented by spheres.

Water Probe A sphere that represent a water molecule. A radius of 1.4Å is commonly used for the water probe.

Van der Waals (VDW) surface Surface of the union of atomic spheres defined by the Intrinsic Radii.

Solvent Accessible Surface Defined as the area traced out by the center of the water probe after rolling the probe over the spherical atoms.

Molecular Surface The surface of a molecule defined as the boundary of of the union of all possible probes which do not overlap with the atomic spheres of a molecule. This surface is sometimes referred as the Solvent Excluded Surface and is the most commonly used surface definition in PB and GB models.

Molecular Volume Volume enclosed by the Molecular Surface.

Molecular Dynamics A computational method to simulate the physical movement of atoms and molecules. Successive configurations (snapshots) of the system are generated by integrating the Newtons's law of motion.

R6 Integration $|r|^6$ integration over the Molecular Volume.

Chapter 1

Specific Aims

The main goal of the this dissertation is to develop new GB models capable of computing solvation energies with a good balance of accuracy and computational efficiency. The approach is based on the so called “R6 integration” which is an integral over the molecular volume, see equation 2.4 below. The hypothesis is that, based on the “R6 integration”, it is possible to develop computationally efficient GB models suitable for computing various properties of large sets of biomolecules and for Molecules Dynamic simulations. This overall goal is achieved through the following more specific objectives:

1. Develop and implement a numerical GB model (NSR6), based on the “R6 integration”, to accurately compute solvation free energies. Compare the efficiency and accuracy of NSR6 with those of the PB model.

2. Develop an analytic GB model that approximate the “R6 integration”, taking the Numerical GB NSR6 as reference. Compare the efficiency and accuracy of AR6 with previously developed and popular GB models.

3. Evaluate the accuracy of the solvation free energies produced by the new GB models on common benchmarking data sets of small molecules –applicable to rational drug design–, amino acids, and small proteins.

4. Implement the analytical GB method in a Molecular Dynamics package and perform extensive testing to evaluate its efficiency and accuracy.

Chapter 2

Introduction

Accurate computation of the solvation free energy (ΔG_{solv}) of a molecule, which represents the gain in energy for transferring a molecule from vacuum to an aqueous medium, is central to numerous areas of biomedical and industrial research. Alchemical free energy calculations, in which all water molecules are explicitly incorporated in the model, is arguably the most realistic, accurate, and practically available procedure to compute ΔG_{solv}. Recent studies have shown that the approach is able to reproduce experimental solvation energies with a good degree of accuracy, approximately 1.2kcal/mol of average error to experiment for a large set of small molecules[83, 57]. One of the main problems of the explicit solvent treatment is, however, that the approach is computationally too expensive for many practical applications. An adequate representation of solvated biomolecules typically requires a number of water atoms that is one order of magnitude larger than the number of atoms of the biomolecule. As a result, this framework spends most of the computational resources in modeling the movement of the water atoms rather than focusing on the molecule of interest.

The implicit solvent model is a popular alternative framework used to compute solvation free energies. In this framework, discrete water molecules are replaced by an infinite continuum medium with the average dielectric properties of water, thus considerably reducing the number of degrees of freedom of the system. The computational cost associated with the use of these models is therefore significantly smaller than the cost of representing water atoms explicitly.

Within the implicit solvent framework the solvation free energy is typically divided into electrostatic (ΔG_{el}) and non-polar (ΔG_{nopol}) components:

$$\Delta G_{\text{solv}} = \Delta G_{\text{el}} + \Delta G_{\text{nopol}} \tag{2.1}$$

ΔG_{el} takes into account electrostatic interactions and ΔG_{nopol} other interactions such as cavity formation and attractive Van der Waals interactions. The main topic of this dissertation focuses on developing efficient methods to compute ΔG_{el}.

Solving the Poisson-Boltzmann equation (PB) is theoretically the most rigorous implicit solvent model for computing ΔG_{el} [34, 41, 55, 12, 80, 18, 9, 6]. This method computes the electrostatic potential $\phi(\mathbf{r})$ produced by a set of "fixed" atomic charges q_i located at positions $\mathbf{r}_i$ (the center of the atomic spheres), through the following equation:

$$\nabla[\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})] = -4\pi \sum_i q_i \delta(\mathbf{r} - \mathbf{r}_i). \tag{2.2}$$

where $\epsilon(\mathbf{r})$ represents the position-dependent dielectric constant. Typically, two values of dielectric constants are used to model a solvated molecule, 80 for regions outside the molecular volume (water region) and 1 for regions inside (molecule region) see figure 2.1. The sharp transition of dielectric constants defines the so called Dielectric Boundary (DB). The electrostatic component of the solvation free energy is then computed by $\Delta G_{\text{el}} = \frac{1}{2}\sum_i q_i[\phi(\mathbf{r}_i) - \phi_{vac}(\mathbf{r}_i)]$, where $\phi_{vac}(\mathbf{r}_i)$ is computed by equation 2.2 using a constant value of $\epsilon(\mathbf{r}) = 1$ inside and outside the dielectric boundary.

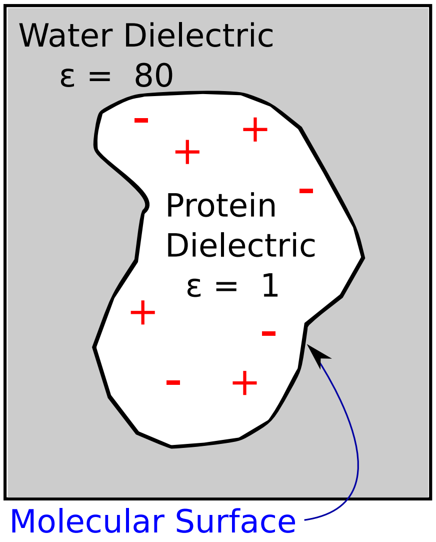

Figure 2.1: Two-dimensional representation of a solvated protein system typically used by the PB model.

One of the most important steps in implicit solvent models is defining the solute/solvent Dielectric Boundary – a region of space over which the dielectric constant $\epsilon(\mathbf{r})$ changes from the value characteristic of the molecular interior ($\epsilon = 1$) to that of water ($\epsilon = 80$). Although many Dielectric boundary definitions has been proposed [6, 20], in molecular modeling, three basic surface definitions are commonly used as dielectric boundary. The schematics of these surfaces are shown in figure 2.2. Among these, the Van der Waals (vdW) representation of the molecule – the union of hard atomic spheres – is the simplest and most computationally facile. However, it is the Lee-Richard Molecular Surface, figure 2.2, that has been utilized most often in numerical PB and GB calculations[8, 76, 61], because it produces solvation energies that are considerably more accurate than those based on the vdW surface[68].

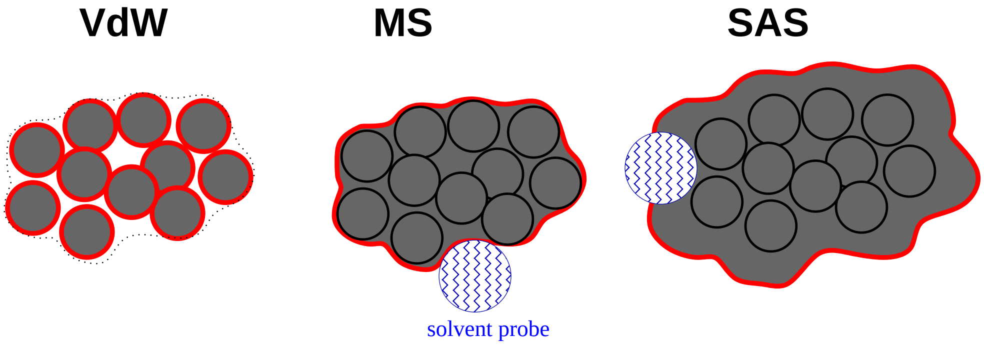

Figure 2.2: The three representations of solute–solvent dielectric boundary (thick red line) commonly used in implicit solvent framework. **Left panel** The van-der-Waals (vdW) boundary coincides with the surface of atomic spheres, the inter-atomic interstitial space is treated as high dielectric solvent (white). **Middle panel:** The Lee-Richards molecular surface (MS): all interstitial space, small voids and invaginations inside the surface are treated as low dielectric solute (Grey). (3) **Right panel:** The solvent accessible surface (SAS) defines the boundary: all interstitial space and small voids inside the SAS are treated as low dielectric solute.

Analytical solutions of the equation 2.2 can be derived only for few specific cases with highly symmetric geometries. For realistic Biomolecular Modeling applications, only numerical solutions of the PB equation are possible. Some of the methods to solve equation 2.2 include simple and multi-grid finite differences[10, 8], and finite elements [76, 63, 41]. These methods may become quite time-consuming, especially if applied to a large set of conformations of a macromolecule or if they are incorporated into Molecular Dynamics (MD). One of the most challenging problems for implementing the PB model in Molecular Dynamics is the computation of solvation forces, as they require derivatives of the solvation free energy. Numerical instabilities are produced when the molecular surface changes abruptly due to small variations in atomic positions. While several attempts to implement PB in MD were published none of them fulfilled the requirements of efficiency and simplicity to become mainstream.

The computational complexity of solving equation 2.2, combined with technical difficulties associated with computing forces due to changes in the molecular surface motivated the search for alternative analytical methods to be used in Molecular Dynamics. The Generalized Born model (GB) has become popular as an alternative to the PB model especially in Molecular Dynamics as it offers a mathematically simple, closed-form formula to compute ΔG_{solv}. Importantly, the GB approximation is also aimed at working for arbitrary shapes. The algorithmic simplicity and computational efficiency of the GB model, combined with accuracy improvements, have made it the method of choice in implicit solvent MD. [26, 89, 39, 40, 81, 75, 24, 43, 32, 11, 49, 28, 78, 21, 19, 95, 14, 88, 86, 97, 67, 71, 30, 50].

Altough GB models are routinely used in molecular modeling, recent studies[45] have shown that the accuracy of commonly used GB flavors in computing ΔG_{solv} is considerable worse than that of the PB model, when explicit solvent models are taken as reference. The problems in accuracy are manifested when the GB model is applied to Biomolecular Modeling, for instance it is know that the GB model produce an undesirable bias towards helical secondary structures and over-stabilization of salt bridges[77, 31]. More accurate GB variants are needed that keep the computational efficiency of the original GB model.

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

$$R_i^{-1} = \rho_i^{-1} - \frac{1}{4\pi}\int_{|\mathbf{r}-\mathbf{r}_i|>\rho_i}^{solute} |\mathbf{r}-\mathbf{r}_i|^{-4} d\mathbf{V}, \tag{2.3}$$

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[71] 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. [77]) 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. [77] 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 recently proposed by Svrcek-Seiler[90] and independently by Grycuk[37] as an alternative to the CFA:

$$R_i^{-1} = \left(\frac{3}{4\pi}\int_{ext}\frac{d\mathbf{V}}{|\mathbf{r}-\mathbf{r}_i|^6}\right)^{1/3} = \left(\rho_i^{-3} - \frac{3}{4\pi}\int_{r>\rho_i}^{solute}|\mathbf{r}|^{-6}d\mathbf{V}\right)^{1/3} = \left(\rho_i^{-3} - \mathbf{I}_i^{tot}\right)^{1/3}, \tag{2.4}$$

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. Recently, Mongan et al. [61] have shown that when the "R6 integration" are computed by essentially exact numerical integration of equation 2.4, 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 integration" can potentially eliminate some of the deficiencies of the methods based on CFA.

However, few efficient algorithms that compute equation 2.4 over physically correct molecular surface/volume exists [7]. Due to the $|r|^6$ term, equation 2.4 is very sensitive to small integration inaccuracies in the vicinity of the atom in question, i. Thus, a methodology to compute the R6 integration requires a high resolution, at least in the local region of the atom i making an standard numerical volume integration of equation 2.4 computationally too expensive for practical applications in biomolecular Modeling.

We developed a numerical methodology to compute "R6 integration", NSR6, which is based on a direct surface integration over a numerically triangulated molecular surface. While NSR6 is mathematically equivalent to the a molecular volume integration approach[61] the surface-based routine is much faster. The applications of NSR6 are currently limited to energy evaluation over static structures. Examples of these applications are pKa calculation and the computation of binding affinities. Since NSR6 is based on the sharp molecular surface (solvent excluded), the method is susceptible to discontinuities due to small atom displacements, which is inappropriate for Molecular Dynamic Simulations.

In order to perform MD simulation, we developed a new analytical method (AR6) to compute the effective Born radii based on an approximation of the "R6 integration". Although the method starts with a computationally efficient pair-wise approximation over the volume occupied by the atomic spheres, 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 sharp molecular surface (solvent excluded). We show that the proposed method is suitable for MD simulations and keeps the computational efficiency and stability of the previous GB models currently implemented in the AMBER package.

Chapter 3

Description of the two proposed methods: NSR6 and AR6

This chaper appeared as the reference [4].

In this chapeter we provide a detailed description of a numerical method (NSR6) and an approximate analytical (AR6) method for computing the effective Born radii of biomolecules. Both methods are based on the so called “R6 integration” (see equation 2.4 of the Introduction).

3.1 Introduction

An accurate description of solvent is essential for modeling and simulation of biological macromolecules. Currently, the most rigorous procedure to model 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 [18, 41, 12, 56, 35, 80, 54, 1, 5]. Within the framework of implicit solvent models, macromolecules are treated as a low dielectric 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 non polar (ΔG_{nopol}) 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} [34, 41, 55, 12, 80, 18, 9]. 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} [26, 89, 39, 40, 81, 75, 24, 43, 32, 11, 49, 28, 78, 21, 19, 95, 14, 88, 86, 97, 67, 71, 30, 50], especially in MD.

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

$$\Delta G_{\mathrm{el}} \approx \Delta G_{\mathrm{GB}} = -\frac{1}{2}\sum_{ij} \frac{q_i q_j}{f^{\mathrm{GB}}(r_{ij}, R_i, R_j)} \left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right), \tag{3.1}$$

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[89] of f^{GB} is $f^{\mathrm{GB}} = \left[r_{ij}^2 + R_i R_j \exp(-r_{ij}^2/4R_i R_j)\right]^{\frac{1}{2}}$, although other similar expressions have been tried[43, 69].

Recently, it has been shown that equation 5.1 produces a systematic error (with respect to PB results) when applied to systems with finite values of ϵ_{in} and ϵ_{out} [85]. Sigalov et al.[84] have proposed a modified GB model (ALPB) that eliminates this systematic error while keeping the computational efficiency of the Still's original equation:

$$\Delta G_{\mathrm{el}} \approx -\frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\frac{1}{1+\beta\alpha}\sum_{ij} q_i q_j \left(\frac{1}{f^{\mathrm{GB}}} + \frac{\alpha\beta}{A}\right), \tag{3.2}$$

where $\beta = \epsilon_{in}/\epsilon_{out}$, $\alpha = 0.571412$, and A is the electrostatic size of the molecule, which is essentially the over-all size of the structure, that can be computed analytically [84]. The ALPB model is currently implemented in AMBER[16] 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 equation 5.1, the GB ΔG_{el} are in close agreement with those of the PB [69]. The computationally expensive "perfect" effective Born radii are commonly used for benchmarking and testing different GB "flavors"– approximations that compute R_i.

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

$$R_i^{-1} = \rho_i^{-1} - \frac{1}{4\pi}\int_{|\mathbf{r}-\mathbf{r}_i|>\rho_i}^{solute} |\mathbf{r}-\mathbf{r}_i|^{-4} d\mathbf{V}, \tag{3.3}$$

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[71] 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. 77) 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. 77 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[90] and independently by Grycuk[37] as an alternative to the CFA:

$$R_i^{-1} = \left(\frac{3}{4\pi} \int_{ext} \frac{d\mathbf{V}}{|\mathbf{r} - \mathbf{r}_i|^6} \right)^{1/3} = \left(\rho_i^{-3} - \frac{3}{4\pi} \int_{r>\rho_i}^{solute} |\mathbf{r}|^{-6} d\mathbf{V} \right)^{1/3} = \left(\rho_i^{-3} - \mathbf{I}_i^{tot} \right)^{1/3}, \qquad (3.4)$$

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 equation 3.3, the "R6 radii" are exact for any location of a charged atom within a perfect spherical solute in the $\epsilon_{out}/\epsilon_{in} \gg 1$ limit. Recently, Mongan et al. [61] have shown that when the "R6 radii" are computed by essentially exact numerical integration of equation 5.2, 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 equation 5.1 or 4.1 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 physically realistic molecular (Lee-Richards[48]) 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[94] and Labute [46] have reported analytical methods to compute "R6 radii" in which equation 5.2 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[52, 92]. 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.[51] 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[42] in CHARMM or AMBER GB variants[15]. 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 speed-up of conformational search[70] . 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[17].

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 pair-wise 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 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.

3.2 Theory

3.2.1 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.[61]. Within this formulation, R_i is calculated by the following equation:

$$R_i^{-1} = \left(-\frac{1}{4\pi} \oint_{\partial V} \frac{\mathbf{r} - \mathbf{r_i}}{|\mathbf{r} - \mathbf{r_i}|^6} \cdot d\mathbf{S} \right)^{1/3}, \tag{3.5}$$

which by Gauss-Ostrogradski theorem, is equivalent to equation 5.2. 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:

$$R_i^{-1} \approx \left(-\frac{1}{4\pi} \sum_k \frac{(\mathbf{c}_k - \mathbf{r_i}) \cdot \hat{\mathbf{n}}_k S_k}{|\mathbf{c}_k - \mathbf{r_i}|^6} \right)^{1/3}, \tag{3.6}$$

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 $\hat{\mathbf{n}}_k$ 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[79], 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[61] over the molecular volume in equation 5.2; the reason for the relative inefficiency of the numerical volume-based approach in the context of the R6 is mentioned below.

3.2.2 Approximate analytical computation of the R6 radii: AR6

In this section, we propose an analytic approach to approximate the R6 radii based on the integration of equation 5.2 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.

Overall Approach

In order to analytically compute R6 radii (equation 5.2), 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 equation 5.2, 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 pre-defined small group of covalently linked atoms, including atom i, over which the R6 integration is pre-computed numerically. This group of atoms will be referred 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.[60] in which "necks" were originally introduced in the context of $|r|^{-4}$ integrals. Finally, atoms outside the "chunk" region (arguably the region where equation 5.2 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 3.1: 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 equation 5.2 is approximated by:

$$\mathbf{I}_i^{tot} = \frac{3}{4\pi}\int_{r>\rho_i}^{solute} |\mathbf{r}|^{-6} dV \approx \mathbf{I}_i^{vdw} + \mathbf{I}_i^{neck} + \mathbf{I}_i^{chunk}, \tag{3.7}$$

where $\mathbf{I}_i^{vdw}$ represents the R6 integration over the van der Waals volume outside the "chunk" of atom i, $\mathbf{I}_i^{neck}$ represents the R6 integration over the "neck" regions (see Ref. 60 for details), and $\mathbf{I}_i^{chunk}$ is the R6 integration over the molecular volume of the "chunk". In figure 3.1 the regions of integration of $\mathbf{I}_i^{vdw}$, $\mathbf{I}_i^{neck}$, and $\mathbf{I}_i^{chunk}$ are represented by light Grey, Blue, and Red color respectively.

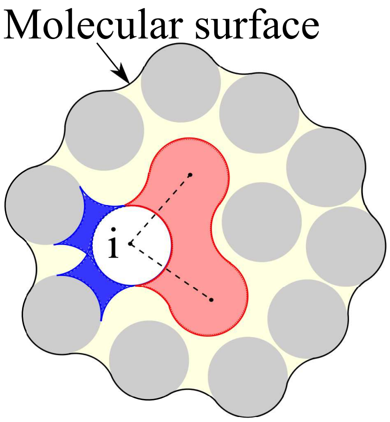

Figure 3.1: Illustration of the three regions of integration in equation 3.7 that are combined to approximate the molecular volume: VDW volume (light Grey spheres), Neck regions (dark Blue), and "chunk" molecule (Red). The open sphere represents atom i and the dashed lines represent covalent bonds used to define which atoms belong to the chunk molecule.

The above approximation will over-count overlapping regions between necks and atoms outside the "chunks". Therefore, the contribution of $\mathbf{I}_i^{vdw}$ and $\mathbf{I}_i^{neck}$ 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 of 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 3.1. To address this underestimation, we use an additional volume correction which requires the use of two additional parameters.

Integration over Van der Waals volume: $\mathbf{I}_i^{vdw}$

Here we compute the $\mathbf{I}_i^{vdw}$ integral in equation 3.7 over the individual VDW atomic spheres that make up the molecule; the $|\mathbf{r}|^{-6}$ integral contribution of VDW sphere of atom j to the effective Born radius of atom i was analytically calculated previously [91, 94]. 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 $\mathbf{I}_i^{vdw}$ 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: $r_{ij} \geq \rho_i + \rho_j$

$$\mathbf{F}_6(\rho_i, \rho_j, r_{ij}) = \frac{\rho_j^3}{(r_{ij}^2 - \rho_j^2)^3}. \tag{3.8}$$

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

$$\mathbf{F}_6(\rho_i, \rho_j, r_{ij}) = \frac{1}{16 r_{ij}} \left(\frac{r_{ij} + 3\rho_j}{(r_{ij} + \rho_j)^3} + \frac{3(\rho_j^2 - \rho_i^2 - (r_{ij} - \rho_i)^2) + 2 r_{ij}\rho_i}{\rho_i^4} \right). \tag{3.9}$$

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

$$\mathbf{F}_6(\rho_i, \rho_j, r_{ij}) = \frac{1}{\rho_i^3} + \frac{\rho_j^3}{(r_{ij}^2 - \rho_j^2)^3}. \tag{3.10}$$

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

$$\mathbf{F}_6(\rho_i, \rho_j, r_{ij}) = 0. \tag{3.11}$$

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 $S_{vdw}^j < 1$, to correct for over counting of the volume due to possible overlaps between VDW spheres of neighboring atoms. Then, the total contribution of VDW spheres is:

$$\mathbf{I}_i^{vdw} = \sum_{j \notin \text{"chunk"}\ i} \mathbf{F}_6(\rho_i, (S_{vdw}^j)\rho_j, r_{ij}), \tag{3.12}$$

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 $S_{vdw}^j = S_{vdw}$ is constant for all atoms of the molecule (We have found that $S_{vdw} = 0.6211$ gives the best results, see below).

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.[60] 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 3.2, is completely determined by their VDW radius ρ_i and ρ_j, the distance r_{ij} between them, and the water probe radius ρ_w. Moreover, the "neck" exists only if the distance between atoms i and j is less than $\rho_i + \rho_j + 2\rho_w$. To approximate the integral of $|\mathbf{r}|^{-6}$ over the "neck" region we use the following analytical and empirical function:

$$neck\_integral(r_{ij}, \rho_i, \rho_j) \approx A_{ij}(r_{ij} - B_{ij})^4(\rho_i + \rho_j + 2\rho_w - r_{ij})^4, \tag{3.13}$$

for inter atomic distances (r_{ij}) less than $\rho_i + \rho_j + 2\rho_w$ and greater than B_{ij}. Otherwise $neck\_integral(r_{ij}, \rho_i, \rho_j)$ 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 of atoms in the molecule. It is important to note that equation 3.13 is an empirical approximation to the exact R6 integration over the neck region. We have selected the lowest degree polynomial that satisfy the following properties required for molecular dynamic simulations. First, the polynomial approximation has continuous derivatives with respect to atomic distances. The polynomial also approximates the real integration with a good degree of accuracy, see figure 3.3. The $neck\_integral(r_{ij}, \rho_i, \rho_j)$ function is parametrized by A_{ij} and B_{ij} which depend on ρ_i, ρ_j and ρ_w. Following a similar procedure of Ref. 60, 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 3.2). 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 $neck\_integral(r_{ij}^{max}, \rho_i, \rho_j) = Neck_{max}$. 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 3.3 illustrates that equation 3.13 is a reasonable approximation of the R6 integration over the "neck" region. By construction, equation 3.13 is differentiable in the entire domain of r_{ij}.

Finally, the total integral over neck regions is approximated by:

$$\mathbf{I}_i^{neck} = \frac{3}{4\pi} S_{neck} \sum_{j \notin \text{"chunk" } i} neck\_integral(r_{ij}, \rho_i, \rho_j), \tag{3.14}$$

where S_{neck} is a free parameter used to correct for the volume over-counting 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).

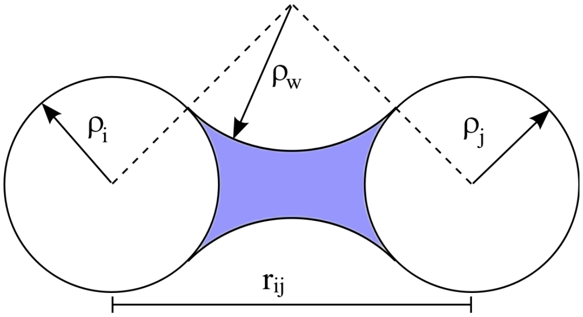

Figure 3.2: Neck region (blue) between two atoms with radius ρ_i and ρ_j and a water probe radius ρ_w. r_{ij} represents the distance between atoms i and j.

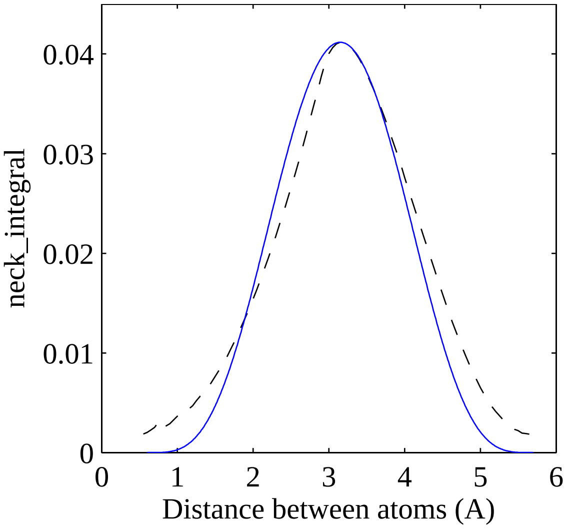

Figure 3.3: The numerical integration over the "neck" region (dashed black) compared with the analytical approximation (solid blue) used here. In this example, we have used $\rho_i = 1.7, \rho_j = 1.2$, and probe radius $\rho_w = 1.4$ Å.

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[61] rather than its practical implementation, the required accuracy was achieved by brute force via inefficient numerical volume integration over a very fine 3D mesh in the vicinity of i. Since here we are set to develop an efficient analytical model, we take a completely different approach. We isolate a small set of neighboring atoms covalently connected to the atom of interest i, see the exact definition below. The geometrical configuration of this small set of atoms, which will be called "chunk", is not expected to change substantially during dynamics. Thus, the contribution of the "chunk" to the effective Born radius of atom i, $\mathbf{I}_i^{chunk}$, can be computed essentially exactly by the NSR6 procedure at the set-up stage and then subsequently re-used at all other steps.

The neighbor atoms that form the "chunk" molecule for a given atom i are determined by setting the "chunk depth", which is defined as the maximum possible integer distance (in the graph-theoretic sense where atoms are the vertices and covalent bonds are edges) between atom i and any other atom in the "chunk". In figure 3.4 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 one. 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 3.4 is composed of all of the atoms of this molecule, which is the same as the "chunk" of Deph= 1 for the nitrogen atom.

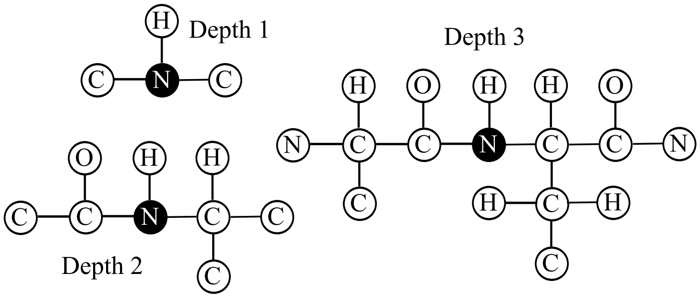

Figure 3.4: Examples of "chunk" molecules of depths 1, 2, and 3 used for the computation of the effective Born radius of a nitrogen atom (black circles).

Note that:

1. 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 over-all shape is maintained during dynamics.
2. 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 $\mathbf{I}_i^{chunk}$ as follows. The idea is to use a fast analytical expression for $\mathbf{I}_i^{chunk}$, 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 $\mathbf{I}_i^{chunk}$; the constant λ_i is estimated once at the set-up stage (e.g. at time = 0). For all other steps, $\mathbf{I}_i^{chunk}$ is computed analytically based on 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:

$$\lambda_i = \frac{\rho_i^{-3} - (\alpha_i^{chunk})^3}{\sum_{k \neq i}^{M} \mathbf{F}_6(\rho_i, \rho_k, r_{ik}^o)}, \tag{3.15}$$

$$\alpha_i^{chunk} = \left(-\frac{1}{4\pi} \oint_{\partial V_{chunk}} \frac{\mathbf{r} - \mathbf{r_i}}{|\mathbf{r} - \mathbf{r_i}|^6} \cdot d\mathbf{S} \right)^{1/3}, \tag{3.16}$$

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). ∂V_{chunk} represent the surface of the "chunk" molecule, and 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 equation 3.16, 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 set-up stage for each atom, the values of $\mathbf{I}_i^{chunk}$ for all of the following steps are computed by:

$$\mathbf{I}_i^{chunk} = \lambda_i \sum_{k \neq i}^{M} \mathbf{F}_6(\rho_i, \rho_k, r_{ik}). \tag{3.17}$$

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, equation 3.13, 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 re-capitulation of the first peak in the PFMs –signature of the use of true molecular volume– presented in Ref. 60 in which necks were originally defined, will still be maintained.

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}:

$$\tilde{\rho}_i^{-3} = \rho_i^{-3} - \mathbf{I}_i^{chunk}, \tag{3.18}$$

$$c_i = 1 - \frac{1}{A^3 \tilde{\rho}_i^{-3}}, \tag{3.19}$$

$$\Psi = (\mathbf{I}_i^{vdw} + \mathbf{I}_i^{neck}), \tag{3.20}$$

$$\beta_0 = 1/c_i, \tag{3.21}$$

$$R_i^{-1} \approx (\tilde{\rho}_i^{-3} - c_i \tilde{\rho}_i^{-3} tanh(\beta_0 \Psi \tilde{\rho}_i^3 - \beta_1 (\Psi \rho_i^3)^2 + \beta_2 (\Psi \rho_i^3)^3))^{1/3} + B. \tag{3.22}$$

Here, A is the electrostatic size of the molecule which is essentially its "global" size, see Ref. 84 for details. Simple and robust routines for computing this parameter are available; in practical MD simulations it can be approximated by a constant. The rescaling process in Eqs. (18)-(22) was built such that if $\Psi \to \infty$ then $R_i \to A$. Thus, the effective Born radius is upper bounded by the molecular size A. On the other hand, if $\Psi << 1$ then $R_i^{-1} \approx \left(\tilde{\rho}_i^{-3} - \mathbf{I}_i^{vdw} - \mathbf{I}_i^{neck}\right)^{1/3}$: the effective Born radius of surface atoms (with small effective radii) are not affected by the rescaling process.

The constant offset parameter B was defined in Ref. 61 and has a value of 0.028 Å^{-1}. This parameter was introduced to minimize the difference between the computed the 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.

Additional volume correction

When equation 3.22 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, equation 3.20, 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[38] analytical model of solvation:

$$V_i = \frac{\sum_{j=1, j \neq i}^{N} \rho_j^3 \Theta_{ij}}{R_s^3}. \tag{3.23}$$

where

$$\Theta_{ij} = \begin{cases} \left(1 - \left(\frac{r_{ij}}{R_s}\right)^2\right)^2 & r_{ij} \leq R_s \\ 0 & r_{ij} > R_s \end{cases} \tag{3.24}$$

The parameter R_s is set to 10 Å which is the same value used in the FACTS method [38]. The inverse of the effective Born radii are then computed by the following differentiable expression:

$$R_i^{-1} \approx (\tilde{\rho}_i^{-3} - c_i \tilde{\rho}_i^{-3} tanh(\beta_0 \Psi \tilde{\rho}_i^3 - \beta_1 (V_i \Psi \rho_i^3)^2 + \beta_2 (V_i \Psi \rho_i^3)^3))^{1/3} + B, \tag{3.25}$$

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

3.2.3 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) equation 5.1, 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. 61 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 non-uniform, 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 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[69]. 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[61]. We also test agreement with the corresponding perfect radii, see below. To reduce the possibility of overfitting further, we fit the two sets of parameters, $\{S_{vdw}, S_{neck}\}$ and $\{\beta_1, \beta_2\}$ independently.

The rescaling factors S_{vdw}, and S_{neck}, Eqs. (12) and (14), are optimized such that the total electrostatic solvation energies ΔG_{el} obtained by AR6 (through equation 4.1) match the ΔG_{el} of the NSR6 procedure for four conformational states of an Alanine decapeptide (Ala10) represented in figure 3.5. For the optimization, each of the four conformational states of Ala10 was represented by 10 MD snapshots[77]. 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 over-fitting. 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 over-fitting, and allows for an exhaustive exploration of the parameter domain.

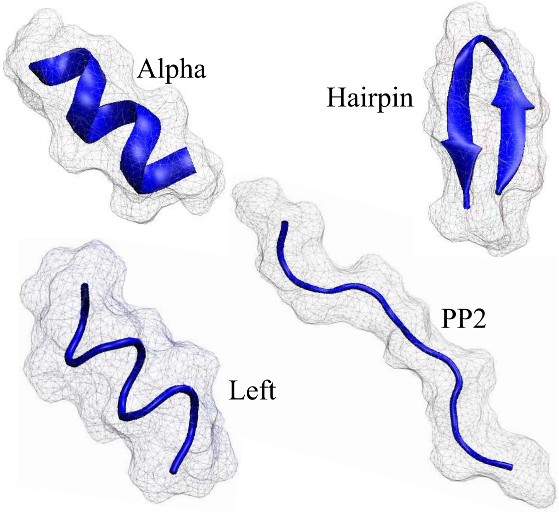

Figure 3.5: Cartoon representation of the four conformational states of alanine decapeptide, Ala10, used in this work.

We have used the Nelder-Mead[62] 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 4 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 4 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 3.1. Although the accuracy of the approximation (determined by the RMSD values of table 3.1) 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.

Table 3.1: Free energies of solvation for different conformations of Ala10 (kcal/mol) obtained with the AR6 and the NSR6 procedures. Solvation energies were computed using $\epsilon_{out} = 80$, $\epsilon_{in} = 1$, 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.

	NSR6	AR6			
		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) $\Delta\Delta G_{\text{el}}$		
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

The energies obtained by using AR6 with optimized parameters S_{vdw} and S_{nech} 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 (equation 5.1) and the errors due to the approximation of the effective Born Radii. Figure 3.6 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 "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 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 equation 5.1

Parameters β_1 and β_2 are meant to control the rescaling process for large radii in equation 3.25, such that the rescaling is large for deeply buried atoms and small for the exposed ones. 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

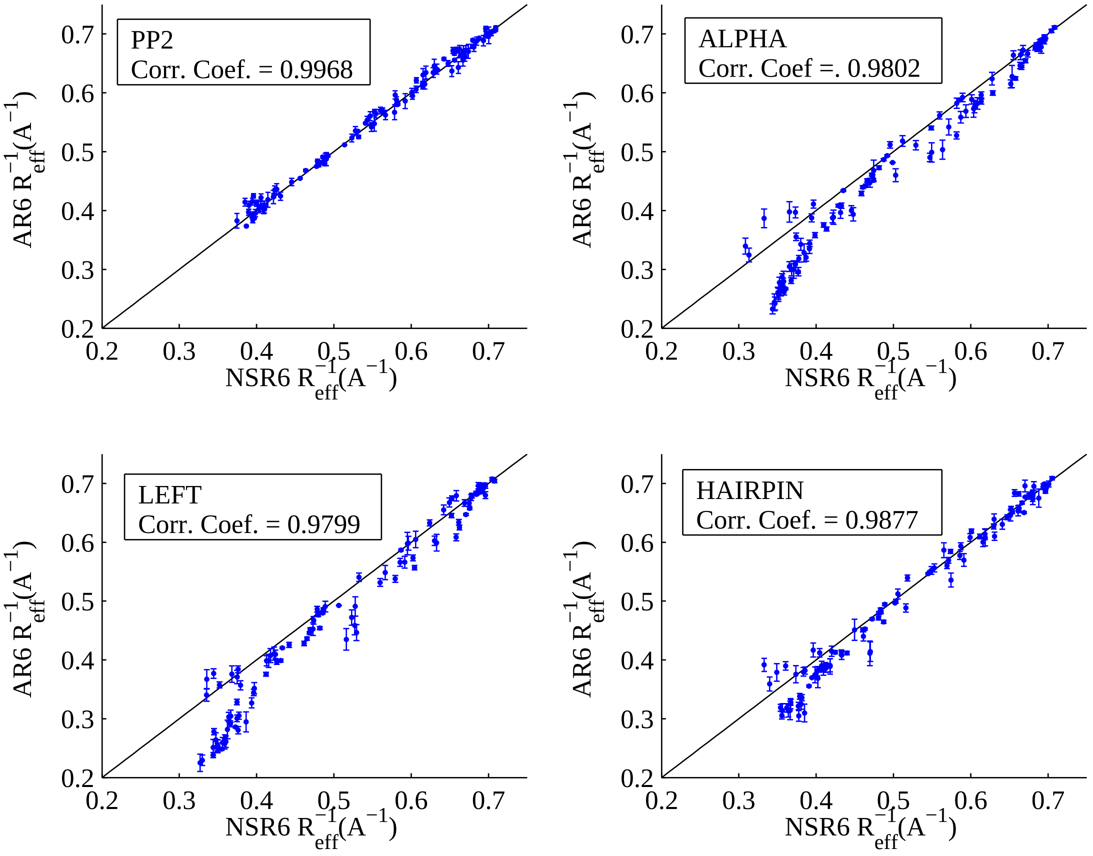

Figure 3.6: Comparison of the inverse of the approximated R6 effective Born radii (AR6) with the exact R6 effective Born radii (NSR6) for the four conformational states of Ala10. Every point represents the average Born radius over four possible "chunks", with the errors bars representing standard deviations.

in table 3.7 (bold letters). 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 5.2, these values were used for all of the calculations presented in the Results section.

3.2.4 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 3.7 shows the correlation between the effective Born radii com-

Table 3.2: Optimized parameters.

Parameter	Value
S_{vdw}	0.6211
S_{neck}	0.4058
β_1	18.4377
β_2	313.7171

puted 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 3.1,“switched off”. These results shows that it is the combination of necks’ contribution and the approximation of the R6 in the “chunk” regions that contribute most to the good approximation to the numerically exact R6 integration for small molecules such as Ala10. 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.

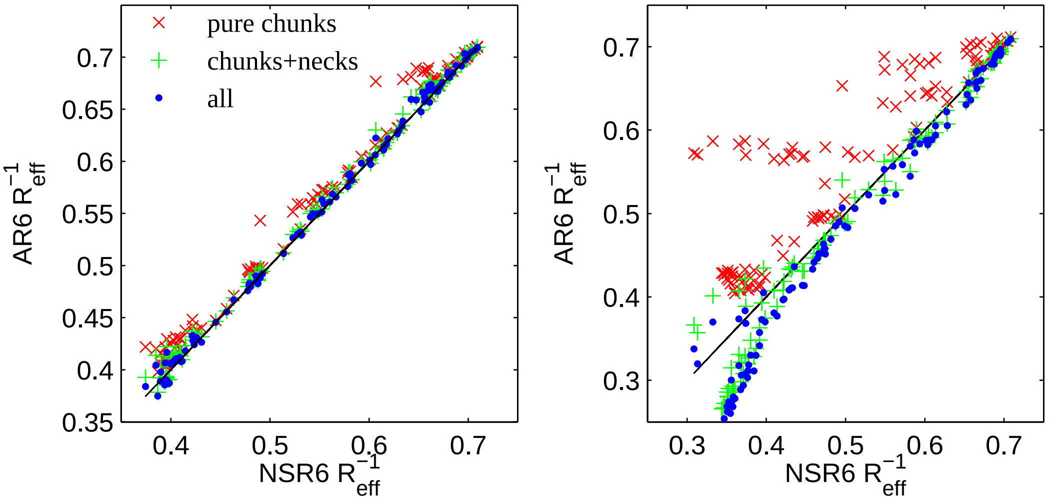

Figure 3.7: Comparison of the inverse of the approximated R6 effective Born radii (AR6) with the exact R6 effective Born radii (NSR6) for the pp2 (left) and alpha (right) conformational states of Ala10. Red × marks, AR6 with only the chunks contribution ($S_{vdw} = S_{neck} = 0$). Green + marks, AR6 with chunks and neck contribution ($S_{vdw} = 0, S_{neck} = 0.4058$). Blue circles, AR6 with all 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.

The contribution of the additional volume correctio, equation 3.23, 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 3.8

shows that when no volume correction is applied ($\beta_1 = \beta_2 = 0$), the effective Born radii of buried atoms (located in leftmost side) are systematically underestimated. When the additional volume correction is activated, the effective Born radius of buried atoms are substantially shifted down towards the correct values of NSR6. Notably, atoms with small effective Born radius (located in leftmost side of figure 3.8) are almost unaffected by the rescaling via $\beta_1, \beta_2 > 0$.

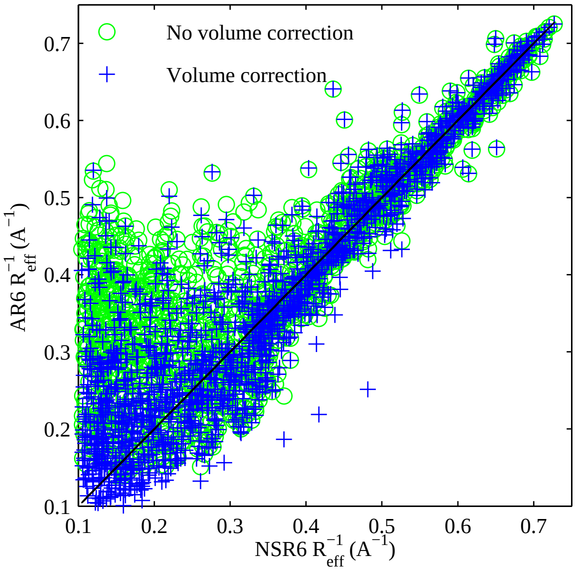

Figure 3.8: Comparison of the inverse of the approximated R6 effective Born radii (AR6) with the "exact" R6 effective Born radii (NSR6) for thieredoxin (2TRX). Green circles, AR6 with no additional volume correction ($\beta_1 = \beta_2 = 0$). Blue plus marks, AR6 with optimized parameters from table 5.2.

3.3 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.[77] 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 3.5. In figure 3.9 we show the error, with respect to explicit solvent, of the $\Delta\Delta G_{\text{el}}$ computed by numerical PB, GB_OBC, and AR6, between the four conformational states of Ala10. The $\Delta\Delta G_{\text{el}}$ 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.

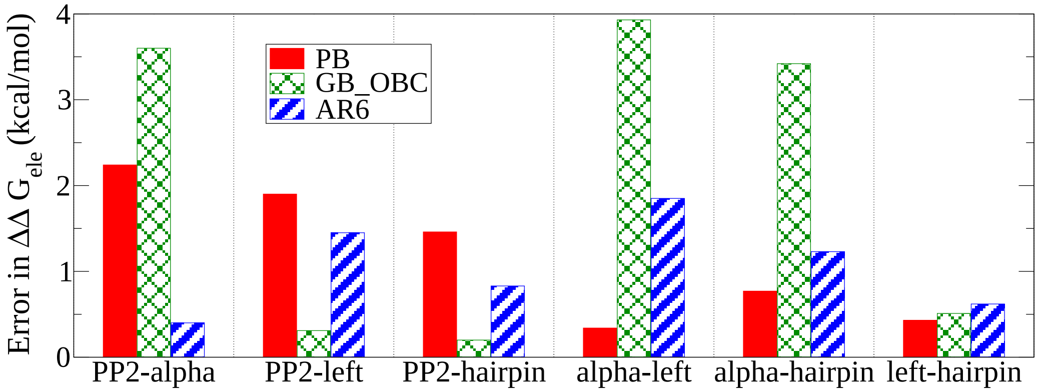

Figure 3.9: Absolute error in $\Delta\Delta G_{\mathrm{el}}$, relative to the explicit solvent model, between four different conformational states of Ala10 (alpha, PP2, left, and hairpin). The energies were obtained using PB (solid red bars), GB_OBC(cross-hatched green bars) and AR6 (striped blue bars). The $\Delta\Delta G_{\mathrm{el}}$ for conformational states A and B is defined as $\Delta\Delta G_{\mathrm{el}}(A - B) = \Delta G_{\mathrm{el}}(A) - \Delta G_{\mathrm{el}}(B)$.

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 3.3 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 3.3 for comparison.

Table 3.3: RMSD of the solvation energies (kcal/mol), relative to the PB reference of three GB flavors. The computation was carried out on a set of 22 structures using optimized parameters from table 5.2 and "chunk" Depth 3.

	NSR6	AR6	GB_OBC
RMS	9.98	16.72	50.49

Finally, the agreement in the computed $\Delta\Delta G_{\mathrm{el}}$ values between numerical PB and AR6 is also verified on the denaturation trajectories of Apo-myoglobin and protein-A, see the results in table 3.4. In the following subsections, these results are explored in more detail.

Table 3.4: The change in the electrostatic part of the solvation free energy, $\Delta G_{\mathrm{el}}(N) - \Delta G_{\mathrm{el}}(U)$ [kcal/mol], of Apo-myoglobin and Protein-A in 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

3.3.1 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.[1] Table 3.5 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[77]. 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 4 possible values of ΔG_{el} for each conformational state of Ala10, corresponding to the 4 possible conformational states used to set up “chunks”. The final ΔG_{el} presented in table 3.5 for each conformational state is obtained by averaging these 4 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 3.5 show that compared to the other analytical GB flavors tested, the $\Delta\Delta G_{\mathrm{el}}$’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 ($\Delta\Delta G_{\mathrm{el}}$). Table 3.5 shows that, relative to TIP3P, the values of $\Delta\Delta G_{\mathrm{el}}$ 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 $\Delta\Delta G_{\mathrm{el}}$ by only -0.4 kcal/mol relative to the TIP3P. This suggests that AR6 is not biased towards alpha conformation in contrast to GB_OBC. The AR6 model overestimates TIP3P values by only 1.45 kcal/mol for the $\Delta\Delta G_{\mathrm{el}}$ between PP2 and left, and by 0.8 kcal/mol for the $\Delta\Delta G_{\mathrm{el}}$ between PP2 and hairpin. Overall the $\Delta\Delta G_{\mathrm{el}}$ 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.[77], and essentially the same as the PB result.

When using the original Still’s equation instead of equation 4.1 used throughout this work, the overall RMSD of ΔG_{el} and $\Delta\Delta G_{\mathrm{el}}$ 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 3.5. Thus, the improvement showed in table 3.5 is mostly due to the use of AR6 for effective radii

[1] However, the latter may not be as clean since GB only computes ΔG_{el}, not the total solvation energy ΔG_{solv} available from the experiments.

Table 3.5: Free energies of solvation between different conformations of Ala10 (kcal/mol). The data of TIP3P, GB_HCT, GB_OBC, GBNeck, and PB were taken from Roe et al. [77]. Solvation energies were calculated using $\epsilon_{out} = 80$, $\epsilon_{in} = 1$, and $\kappa = 0$.

	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) $\Delta\Delta G_{\text{el}}$				
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) $\Delta\Delta G_{\text{el}}$ 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

computation rather than the use of equation 4.1 instead of the original Still's equation.

3.3.2 Accuracy of the effective Born radii

The "perfect" (obtained via numerical PB calculations) effective Born radii are often used as benchmark for the accuracy of different GB flavors as such comparisons can help identify sources of error in the computation of the approximate effective radii [61, 60, 68]. In table 3.6 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 equation 4.1. 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 3.10 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 equation 4.1. 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.

Table 3.6: RMSD (Å^{-1}) between the inverse of the 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

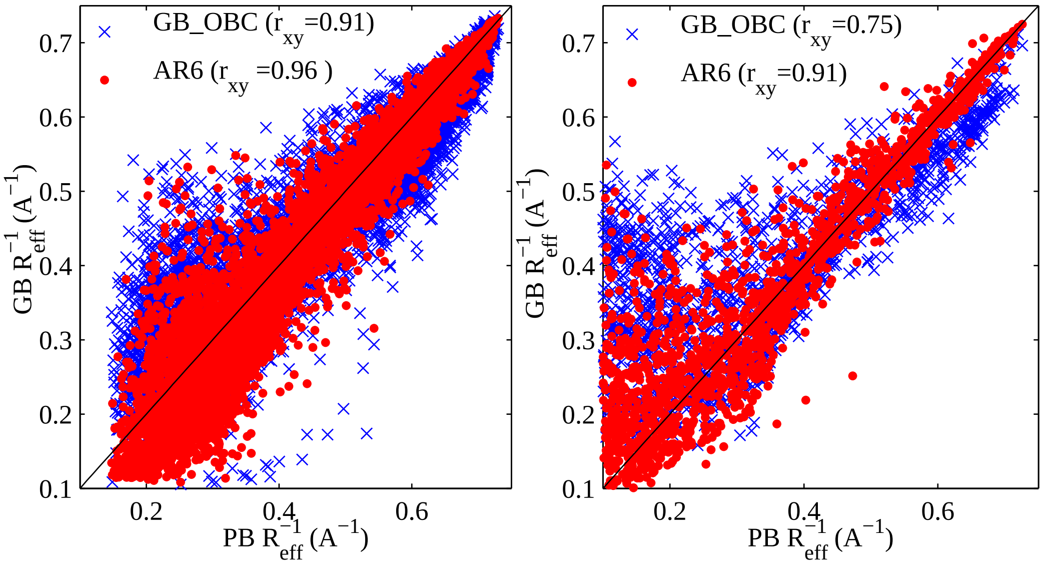

Figure 3.10: Comparison of the inverse of the approximated effective Born radii (GB R$_{\rm eff}^{-1}$) with the "perfect" effective Born radii (PB R$_{\rm eff}^{-1}$) for 19 small proteins (left) and thioredoxin (right). Approximated effective radii were computed by AR6 (Red) and GB_OBC (blue). Correlation coefficients r_{xy} are indicated in parentheses.

3.3.3 Accuracy of $\Delta G_{\rm el}$ 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 [61]. The results of this comparison are shown in table 3.7. 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 3.7 were calculated as the arithmetic mean of $100(\Delta G_{\mathrm{el}}(\mathrm{GB}) - \Delta G_{\mathrm{el}}(\mathrm{PB}))/|\Delta G_{\mathrm{el}}(\mathrm{PB})|$ 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.

Table 3.7: Electrostatic solvation energies (kcal/mol) for a set of 22 structures. The solvation energies were calculated using $\epsilon_{out} = 1000$, $\epsilon_{in} = 1$, and $\kappa = 0$. In all cases we used the optimized parameters on table 5.2 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.

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 Per. err.		1.07%	1.55%	2.67%
RMSD		9.69	16.72	50.49

3.3.4 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 $\Delta G_{\rm el}$. Here, we test the sensitivity of AR6 to the choice of the structure used to set up "chunks". The values of $\Delta G_{\rm el}$ for AR6 in table 3.5 (upper block) for each conformational state of Ala10 were obtained by averaging the 4 $\Delta G_{\rm el}$ values corresponding to each of the 4 conformational states of Ala10 used to set up "chunks". The corresponding standard deviations, considering the 4 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 $\Delta G_{\rm el}$ due to the choice of structure for setting up "chunks" is very small relative to the absolute values of $\Delta G_{\rm el}$. To further analyze this sensitivity in larger molecules, we compare the $\Delta\Delta G_{\rm el}$ between the PB and AR6 for the denaturation trajectories of apo-myoglobin and protein A. The results are summarized in table 3.8. Having two protein conformations (in the Native and Unfolded state), it is possible to compute "chunk" contributions from two completely different sources, one from a snapshot of the Native state (chunk "N" in table 3.8), and the other from a snapshot of the Unfolded state (chunk "U" in table 3.8). Ideally, $\Delta\Delta G_{\rm el}$ computed using such different "chunks" should be identical. According to our procedure, the "chunks" contribution (and thus $\Delta G_{\rm el}$) depends slightly on the initial configuration. The results in table 3.8 show that the change in $\Delta\Delta G_{\rm el}$ 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 $\Delta\Delta G_{\rm el}$.

Table 3.8: The change in the electrostatic part of solvation free energy, $\Delta\Delta G = \Delta G_{el}(N) - \Delta G_{el}(U)$ [kcal/mol], of Apo-Myoglobin and Protein-A in going from the Unfolded (U) to the native (N) state computed with the PB and AR6 models. 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)

	PB	AR6	
		Chunk N	Chunk U
(Apo)myoglobin , pH =2	-2087	-2088.2	-2083.8
Protein-A, pH = 7	143.37	144.02	144.27

In order to further extend the analysis of the sensitivity of energy to the use of different "chunks", we show in figure 3.11 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.

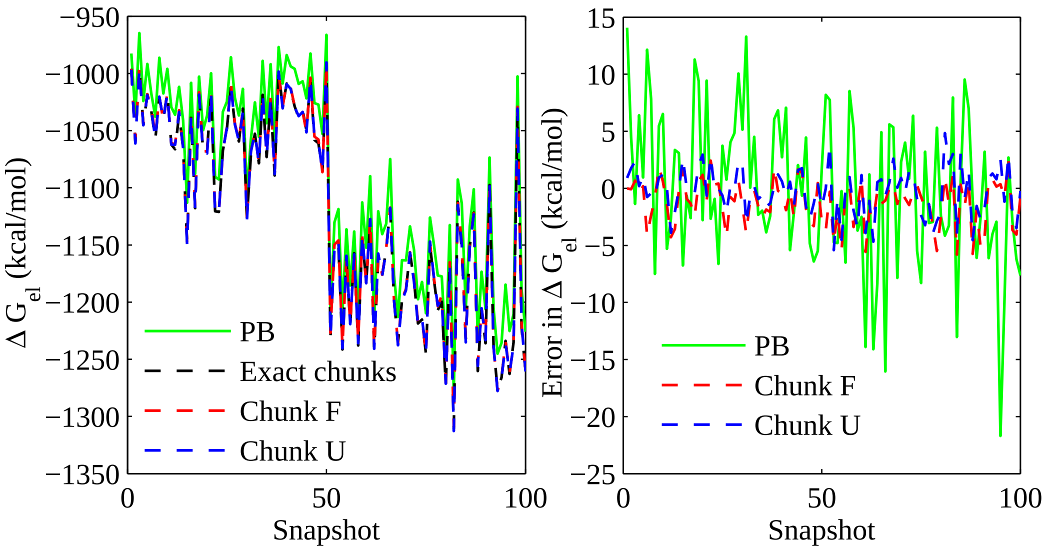

Figure 3.11: Left: solvation energy along the MD unfolding trajectory of Protein-A (PDB ID: 1BDD) obtained by PB (solid lines) and AR6 (dashed lines). The AR6 based energies were obtained using "chunks" from one snapshot of the Folded (Chunk F, red) and Unfolded (Chunk U, blue) states respectively, and 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".

3.3.5 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 set-up 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, equation 3.17, for every atom type within a specific amino acid or nucleotide, and save it in a look-up table for all future computations. Within this protocol, the set-up stage will consist only of reading "chunk" contributions from a look-up table, which is inexpensive. We test this strategy in table 3.9, where we present the values of ΔG_{el} and $\Delta\Delta G_{\text{el}}$ 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 pre-tabulated $\{\lambda_i\}$ 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 3.5: 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 set-up process, introducing little deviation from the original procedure in which λ_i is numerically computed for every atom of the molecule, at the set-up stage.

Table 3.9: Free energies of solvation between different conformations of Ala10 (kcal/mol). The data of TIP3P, and PB were taken from Roe et al. [77]. Solvation energies were calculated using $\epsilon_{out} = 80$, $\epsilon_{in} = 1$, and $\kappa = 0$.

			AR6	
	TIP3P	PB	Original chunks	Pre-tabulated 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) $\Delta\Delta G_{\text{el}}$				
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) $\Delta\Delta G_{\text{el}}$ 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

3.3.6 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[52, 92]. On the other hand, the use of the van der Waals surface in this context has also been advocated[74, 22], including some recent implementations of the R6 flavor[94, 46]. While the precise nature of the physically realistic dielectric boundary is still an open and complex issue[23] 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?

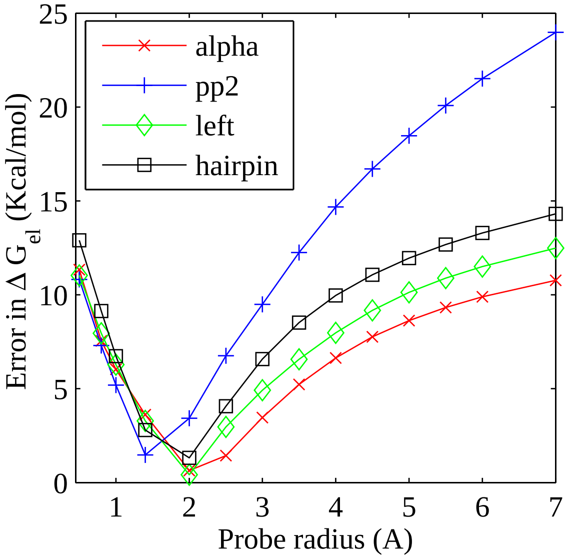

Figure 3.12: Absolute error of the numerical PB ΔG_{el}, relative to the explicit solvent (TIP3P) reference, as a function of the probe radius used to set the dielectric boundary in the PB calculations. The computations are performed for the 4 conformational states of alanine decapeptide shown in figure 3.5.

The unambiguous answer is presented in figure 3.12, 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 3.12 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 3.5, the use of the Lee-Richards molecular surface for the dielectric boundary in PB calculations results in consistently better agreement with TIP3P solvent model, than do the VDW-based definitions. Since the GB model is essentially an approximation of the PB model, these results suggest that in order to obtain more accurate electrostatic solvation free energies relative to the explicit solvent, the dielectric boundary used in the computation of the effective Born radii should strive to approximate the Lee-Richards molecular surface, not the VDW surface.

3.4 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 integrand – 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 pre-defined groups of atoms, "chunks", over which the integration is performed numerically exactly, at the set-up stage. The "chunk" contributions to the total integral are then re-used. 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 set-up 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 over-counting 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 over-fitting.

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. Sensitivity of ΔG_{el} to several key approximations have been tested as well. We have also explored a variant of the approach to eliminate the set-up costs via the use of pre-tabulated chunks. The accuracy of this variant, which carries no set-up costs, is virtually the same as that of the original. While a difference in the set-up efficiency is probably not critical in MD simulations, where the set-up time is only a tiny fraction of the whole simulation time, the pre-tabulated 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 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 non-polar part estimated via the cavity and VDW terms as proposed earlier by Gallicchio et al.[30]. 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 non-polar 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

3.5 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. [77]. 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 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 [63] 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, $\epsilon_{out} = 78.5$, $\epsilon_{in} = 1$, 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.[27], the selection criterion being that the compounds are small enough to allow for high-resolution grid computations. Their PDB IDs are presented in table 3.7 in which the PDB IDs in bold were used as training set. Chain “A” or “model 1” has been chosen when appropriate. The assignment of partial charges, protonation states, etc. are described in Ref. 27. In addition, a canonical B-DNA 10 base pair structure from Ref. 95 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: 7 mostly α helical, 4 mostly β sheet, 5 roughly equal mix of α/β, and 5 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.[8] 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 $\epsilon_{out} = 1000$ to mimic the conductor limit $\epsilon_{out} \to \infty$, and therefore, avoid masking the geometry-specific deficiencies of the standard GB model by its inaccuracies arising from finite ϵ_{out}[85]. 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 reference PB solver in table 3.7. 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[10].

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[73]. The native state is represented by 50 consecutive snapshots (2 ps apart from each other) with near-native radius of gyration, $\sim$ 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 $\sim$ 30 Å –as is experimentally observed in the unfolded state[25]. 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 [63] with a cubic box, and a grid spacing of 0.5 Å. The dielectric constant for protein interior is 1, and the ionic strength is zero.

The surface triangulation used in the NSR6 procedure and the computation of “chunks” contribution, were carried out using the MSMS package [79] using a probe radius of 1.4 and triangle density of 10.

The surface triangulation used in the NSR6 procedure and the computation of “chunks” contribution, were carried out using the MSMS package [79] using a probe radius of 1.4 and triangle density of 10.

Chapter 4

Evaluation of the accuracy of GB NSR6 in computing solvation energies.

This chapter appeared as the reference [2].

In this chapter we evaluate the performance of the accuracy and computational efficiency of the numerical method GB_NSR6. For accuracy testing, the computed solvation energies are compared to experiment and also to those computed by explicit solvent model (TIP3P), for a common benchmark set of 504 small drug-like molecules.

4.1 Introduction

Accurate computation of the solvation free energy (ΔG_{solv}) of a molecule is central to numerous areas of biomedical and industrial research. In particular, availability of computationally efficient and accurate methods of ΔG_{solv} calculation is important for protein ligand binding and rational drug design [44, 58, 82, 33], which involves screening potential drug molecules for optimal binding affinity with the target protein in the presence of aqueous solvent. Here, the solvation energy is one of the main components of the computed binding affinity. Accurate determination of solvation energy is also important in the study of many other physical properties relevant to drug discovery, such as ionization state changes, solubility, phase transfer and aggregation[53, 64].

Alchemical free energy calculations, in which all water molecules are explicitly incorporated in the model, is arguably the most realistic and accurate practically available procedure to compute ΔG_{solv}. Recent studies have shown that the approach is able to reproduce experimental solvation energies with a good degree of accuracy [83, 57]. For instance, using

Molecular Dynamic (MD) simulation with explicit TIP3P water model, Mobley et al. [57] have obtained an agreement with experimental solvation energies within 1.2 kcal/mol (RMS deviation). However, an adequate representation of a solvated small molecule (tens of atoms) typically requires hundreds of discrete water molecules; exploring all degrees of freedom in such a system is computationally too intense for many practical applications, especially to analyze large sets of molecules.

The implicit solvent model is a popular alternative framework used to compute solvation free energies of small drug-like molecules. In this model, discrete water molecules are replaced by an infinite continuum medium with the average dielectric properties of water, thus considerably reducing the number of degrees of freedom of the system. The computational cost associated with the use of these models is therefore significantly smaller than the cost of representing water explicitly. Recent studies using implicit solvent models show a reasonable compromise between accuracy and computational efficiency [64, 65, 66, 47, 59], although high quality explicit solvent simulations provide a more accurate representation of solvation effects.

Within the implicit solvent framework the solvation free energy is typically divided into polar (ΔG_{el}) and non-polar ($\Delta G_{\mathrm{nopol}}$) components: $\Delta G_{\mathrm{solv}} = \Delta G_{\mathrm{el}} + \Delta G_{\mathrm{nopol}}$. The Generalized Born (GB) model is often employed to compute ΔG_{el}. This model has become quite popular due to its relative simplicity and computational efficiency compared to more standard numerical methodologies, such as solving the Poisson Boltzmann (PB) equation numerically[27]. Many flavors of the GB model have been proposed recently; these flavors differ in the way they compute the so-called effective Born radii R_i which play a key role in the GB model, determining its accuracy and efficiency, see *e.g.* refs. 70, 11, and 26 for reviews of the model. Many of the GB flavors are commonly used in MD simulations, due to a reasonable compromise between accuracy and speed[70, 11, 27, 26]. However, a recent study[45] showed that the accuracy of commonly used GB flavors in computing ΔG_{solv} of small molecules is still worse than the $\sim$ 1kcal/mol "accuracy limit" expected from continuum models that share the same underlying physics with the more fundamental PB treatment[66].

A new family of GB flavors have been recently developed [4, 47, 94]; these are based on the so-called R6 effective Born radius which is calculated as a single $\vec{r}/|\vec{r}|^6$(R6) integral over the Lee-Richards[48] molecular surface (although some authors advocate the use of the Van der Waals surface (or volume) instead [94, 47]). Unlike the approximations that had been the basis of the majority of the previous generation flavors, the R6 model is exact for what is, perhaps, the single most important limiting case – the sphere. Previous work demonstrated that the values of ΔG_{el} obtained by the Lee-Richards based R6 models were in a very close agreement with the values of ΔG_{el} computed via the more fundamental PB model for small proteins and DNA[61]; in fact, as far as ΔG_{el} was concerned, the R6 effective radii appear to have reached the accuracy limit of the so-called perfect effective radii based directly on the PB model. Good agreement of the Lee-Richards based R6 with the TIP3P explicit solvent model for different conformations of alanine polypeptide was also reported[4]. Given the high promise of the R6 flavor, the natural question to ask at this point is whether the

R6 is also as good for small molecules relevant to drug design. Two key ingredients are needed to answer this question: a “pure”, parameter free R6 GB, and a common, diverse test set of small molecules. A recent implementation of the R6 flavor[4], called GB_NSR6, is a good fit for the task: its crux is the numerically exact computation of the R6 effective Born radii over the Lee-Richards[48] molecular surface. Importantly, once the geometrical properties of a molecule have been specified, GB_NSR6 requires no additional parameters to compute ΔG_{el}. Thus, GB_NSR6 can be used to test the accuracy of the different radii sets commonly used in solvation energy calculations. As an added bonus, GB_NSR6 inherits the expected efficiency of the GB models. At the same time, a fairly large set of small molecules has recently emerged that has already been used extensively in testing of various solvent models including several common GB flavors[45]. This is the set of 504 neutral small molecules for which experimental solvation free energies are available, conveniently along with the solvation energies from explicit solvent alchemical calculations reported by Mobley et al. [57].

The rest of this article is organized as follows. First, we present a brief description of the data set and the methodology employed here to compute total solvation energies. The “Results” section contains an evaluation of the accuracy and computational efficiency of the GB_NSR6 model, compared to explicit solvent models and experimental data. A brief comparison with several other common GB flavors is also included. The results are summarized in the “Conclusions” section.

4.2 Methods

In this work we compute solvation energies of a dataset of 504 neutral small molecules. The experimental solvation energy, the coordinates and topology files were obtained from the supporting information of Mobley et al. [57], details of these structures can be found in the same reference.

4.2.1 Polar component of ΔG_{solv}

The polar component of the solvation energy was calculated by the ALPB model[84], which re-introduces physically correct dependence on dielectric constant into the original GB model of Still et al.[89], while maintaining the efficiency of the original. The ALPB model approximates ΔG_{el} using the following formula:

$$\Delta G_{\mathrm{el}} \approx -\frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\frac{1}{1+\beta\alpha}\sum_{ij} q_i q_j \left(\frac{1}{f^{\mathrm{GB}}} + \frac{\alpha\beta}{A}\right), \tag{4.1}$$

where ϵ_{in}, and ϵ_{out} are the dielectric constants of the solute and the solvent respectively, $\beta = \epsilon_{in}/\epsilon_{out}$, $\alpha = 0.571412$, and A is the electrostatic size of the molecule, which is essentially the over-all size of the structure, that can be computed analytically [84]. Here, q_i is the partial charge of atom i. We employ the most widely used functional form[89] of f^{GB}: $f^{\mathrm{GB}} = \left[r_{ij}^2 + R_i R_j \exp(-r_{ij}^2/4R_i R_j)\right]^{\frac{1}{2}}$, where R_i is the so-called *effective Born radius* of atom i, and r_{ij} is the distance between atoms i and j. We set $\epsilon_{in} = 1$, and $\epsilon_{out} = 80$ in equation 4.1.

In this work, the effective Born radii R_i were calculated by the following equation:

$$R_i^{-1} = \left(\frac{3}{4\pi}\int_{ext}\frac{d\mathbf{V}}{|\mathbf{r}-\mathbf{r}_i|^6}\right)^{1/3}, \tag{4.2}$$

proposed by Svrcek-Seiler[90] and independently by Grycuk[37]. In equation 5.2 the integral (ext) is taken over the region outside the molecule, and r_i is the position of atom i. We use an equivalent formulation described in Mongan et al. [61]:

$$R_i^{-1} = \left(-\frac{1}{4\pi}\oint_{\partial V}\frac{\mathbf{r}-\mathbf{r_i}}{|\mathbf{r}-\mathbf{r_i}|^6}\cdot d\mathbf{S}\right)^{1/3}, \tag{4.3}$$

which by Gauss-Ostrogradski theorem, is equivalent to equation 5.2. Here, ∂V represents the molecular surface of the molecule, and $d\mathbf{S}$ is the infinitesimal surface vector.

At the first stage of the computation of R_i, the methodology uses MSMS package[79] to build a numerical triangulation of the Lee-Richards[48] molecular surface. There are no adjustable parameters in the GB_NSR6 model used here (the empirical constant offset B to the inverse radii used in Mongan et al. [61] was set to zero in this work), that is no fitting of the model to the data was performed. While technically the resolution of the MSMS molecular surface triangulation is controlled by a density parameter, it was not fitted – we chose a value (6 vertex/Å^2) high enough to insure that the molecular surface is accurate enough. Higher densities result in only marginal increases in accuracy, and are not justified by the associated increase in computational costs. Equation 4.3 is then numerically approximated using the triangles that form the molecular surface, and their respective unit vectors that are orthogonal to the triangles. The surface of the molecule is determined by the intrinsic radius of each atom type of the molecule as well as the solvent probe radius; these parameters obviously affect the ΔG_{solv} but are external to the GB model, and are used in the same manner as the equivalent PB treatment –to define molecular surface. Here we have tested three standard sets of intrinsic atomic radii, BONDI[13], PARSE[87], and ZAP9[64]. Partial atomic charges are also external to the GB model, the same set is used for the corresponding explicit solvent calculations, see below.

4.2.2 Non-polar component of ΔG_{solv}

The non-polar component of the solvation free energy is commonly modeled as being proportional to the Solvent Accessible Surface area (SASA). However it was shown that this strategy generates a poor correlation when compared to explicit and experimental results[93, 96, 98]. Recently, more sophisticated methodologies have been proposed [47, 93, 96, 98, 30]. Here, unless otherwise specified, the non-polar component of the solvation energy is computed in a manner similar to the approach proposed by Gallicchio et al. [30], in which the non-polar component of the solvation energy is decomposed into cavity and van der Waals (VDW) terms,

$$\Delta G_{\text{nopol}} = \Delta G_{\text{cav}} + \Delta G_{\text{vdW}}, \tag{4.4}$$

The cavity component is calculated by:

$$\Delta G_{\text{cav}} = \gamma \cdot \text{SASA}, \tag{4.5}$$

where γ is the surface tension parameter. In the original formulation of Gallicchio et al.[30], γ is an atom-dependent parameter. Here we use a constant atom-independent parameter instead. The parameter γ was optimized using the procedures described further in this section. SASA and surface triangulation were computed by the MSMS[79] package on the same footing.

The solute-solvent van der Waals interaction term is calculated by:

$$\Delta G_{\text{vdW}} = \sum_i \mu_i \frac{a_i}{(R_i + \rho_w)^3}, \tag{4.6}$$

where μ_i is a dimensionless adjustable parameter that depends on the atom type, ρ_w is the water probe radius, R_i is the effective Born radius of atom i previously computed by GB_NSR6 for computation of polar solvation energy. The values of a_i in equation 4.6 are computed as [30]:

$$a_i = -\frac{16}{3}\pi d_w \epsilon_{iw} \sigma_{iw}^6, \tag{4.7}$$

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

$$\sigma_{iw} = \sqrt{\sigma_i \sigma_w}, \tag{4.8}$$

$$\epsilon_{iw} = \sqrt{\epsilon_i \epsilon_w}, \tag{4.9}$$

where $\sigma_w = 1.7683$ Å and $\epsilon_w = 0.1520$ 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 4.1.

To find an optimum set of values for μ_i and γ parameters, we randomly selected 104 molecules (our training set) from the main dataset. Their names are specified in the Supporting

Material. The remaining 400 molecules make up the test set. We used the Nelder-Mead simplex algorithm[62] for optimization. The objective function employed in minimizations consists of the equally weighted sum of two terms: the RMS deviation of the computed total solvation energies and the experimental energies, and the RMS deviation of the ΔG_{nopol} values obtained by GB_NSR6 with respect to those obtained by TIP3P. This is because using only the RMS deviation with respect to experimental ΔG_{solv} would compensate not only for the error in ΔG_{nopol} but for errors in the polar part as well, likely resulting in over-fitting. Because the Nelder-Mead algorithm finds only local minima, we have tried different initial guesses and selected the one that produces the smallest value of the objective function.

Table 4.1 presents the optimum values of μ_i for each atom type, and the optimal γ for the ZAP9 radii set. A total of 10 atom types were considered. Optimum parameters for BONDI and PARSE radii sets are presented in the Supporting Material. In all of the calculations, we have used water probe radius of 1.4Å.

Table 4.1: Lennard-Jones parameters used for the computation of ΔG_{vdW}, for the ZAP9 radii set. The optimized value of γ is 0.01 kcal/mol/Å^2, and is the same for all atom types.

	σ_i (Å)	ϵ_i (kcal/mol)	μ_i
H	1.4870	0.0157	-1.5869
C	1.9080	0.1094	4.1621
N	1.8240	0.1700	5.983
O	1.6612	0.2100	4.8457
S	2.0000	0.2500	0.4412
P	2.1000	0.2000	-1.6941
F	1.75	0.061	1.3331
Cl	1.948	0.265	2.7095
Br	2.22	0.320	0.3188
I	2.35	0.40	-0.8997

The negative values of μ_i can potentially produce positive values of ΔG_{vdW}, see equation 4.6 and equation 4.7. This is unphysical because ΔG_{vdW} is an approximation of the attractive van der Waals part of the energy and should be negative. Moreover, the negative values of μ_i may indicate over-fitting of the polar component —the objective function includes both polar and nonpolar terms. To further investigate this issue, we have optimized the μ_i parameters using only the RMS deviation of the computed ΔG_{nopol} energies relative to those of TIP3P. First, we have used the unconstrained Nelder-Seilder optimization method. The resulting μ_i values are still negative, see Supporting Material. This excludes over-fitting of the polar component of the solvation energy, because the new objective function is based on the non-polar component only. The negative values of μ_i in this case are suggestive of inconsistencies in the functional form of equation 4.6 itself. The problems are unlikely to be severe though. When the μ_i parameters were constrained to the $0 < \mu_i < 2$ range, the resulting RMS deviation (1.25 kcal/mol) of ΔG_{solv} relative to experiment was only marginally larger than the RMSD (1.2 kcal/mol) based on the parameters of table 4.1, some of which negative. Since the set of parameters in table 4.1 still produces slightly more accurate results, and since the computed total van der Waals energy of each of the 504 molecules is still positive in this case, we decided to use the μ_i values of table 4.1 throughout this work. All the sets of optimized parameters (μ_i and γ) and their corresponding RMS values relative to experiment are available in the Supporting Material.

It is important to note that the μ_i and γ parameters are optimized for small molecules and are not necessarily transferable for proteins or peptides. The set of non-negative $\{mu_i\}$ may be interesting to explore in that respect.

4.2.3 Poisson Boltzmann calculations

All the Poisson Boltzmann calculations were performed using ZAP v2.0, the PB solver from OpenEye[36]. The polar component of the solvation energy was computed using the Gaussian dielectric boundary (solute/solvent interface) option, with the dielectric constant of 80 for water and unity for the internal dielectric. We used a grid resolution of 0.25 Å and a buffer region around the molecule of 4 Å from the surface to box boundary.

4.3 Results

4.3.1 Accuracy of computed total solvation free energies

We have used GB_NSR6 to calculate the ΔG_{solv} for the dataset of 504 neutral small molecules; this set was introduced in ref 59 and subsequently used by a number of authors[57, 29, 45]. The computed values of total ΔG_{solv} are compared with experimental solvation energies; the polar and non-polar components of the solvation energy are compared separately with those obtained previously by explicit solvent MD simulations[57]. We begin by comparing the accuracy of the computed ΔG_{solv} with experimental data; in the computation we have used three different intrinsic radii sets, see table 4.2. Clearly, ZAP9 produces more accurate values of ΔG_{solv} with respect to experimental data, compared to the other two radii sets.

Since both polar and non-polar components of ΔG_{solv} computed with TIP3P explicit solvent model are available, we can evaluate the relative contribution of the two components to the accuracy of the computed ΔG_{solv}. Figure 4.1 shows that among the different radii sets, the values of ΔG_{el} obtained using ZAP9 agree best with the TIP3P ΔG_{el} values, with the RMSD of 0.89 kcal/mol. It is important to point out that 124 out of the 504 molecules in the data set used here were also used for the original ZAP9 parametrization [64]. ZAP9 radii set was optimized for the PB solver ZAP[36] using experimental solvation energies.

The values of non-polar energy for typical small molecules (ranging from 0 to 3.5 kcal/mol for

Table 4.2: RMS deviation and correlation coefficients (r^2) of computed ΔG_{solv} (kcal/mol) with respect to experimental data.

Radii Set	RMS Deviation (kcal/mol)	Correlation Coefficient (r^2)
PARSE	2.34	0.74
BONDI	1.79	0.70
ZAP9	1.20	0.86

this particular data set) are much smaller than the corresponding polar solvation energies, and thus contribute much less to the accuracy of the total solvation energy than the polar component which ranges from 0 to -15 kcal/mol. Thus, at least for the current test set, the accuracy of the total solvation energy is determined mainly by the calculation of the polar component of the solvation energy, in which the radii set plays a key role. We have thus chosen ZAP9 to be our "benchmark" radii set throughout the paper.

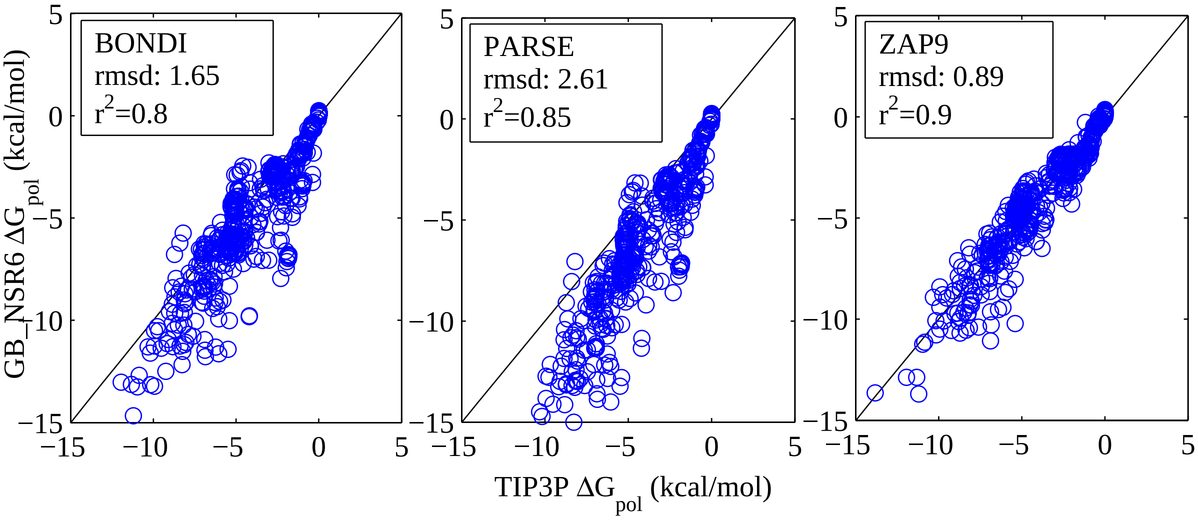

Figure 4.1: Correlation plots between the values of ΔG_{el} computed by GB_NSR6 and those included in ref 57 obtained by explicit(TIP3P) solvent model, for a set of 504 neutral small molecules. Three intrinsic atomic radii sets are used in the GB_NSR6 method: BONDI, PARSE, and ZAP9.

The total solvation energies obtained by our GB flavor agrees with experiment to the same extent as do the energies based on TIP3P explicit solvent calculations, figure 4.2. Namely, GB_NSR6 augmented by the non-polar term as described in "Methods" produces an RMS deviation of 1.2 kcal/mol and a correlation coefficient $r^2 = 0.86$ relative to experimental data, while TIP3P explicit solvent yields in an RMS deviation of 1.26 kcal/mol and a correlation coefficient of 0.89 relative to experiment.

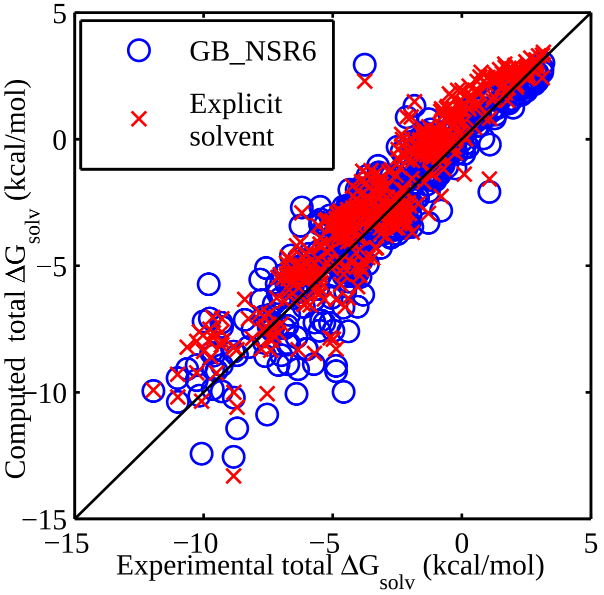

Figure 4.2: Correlation between the computed and experimental solvation free energies, ΔG_{solv} for the entire set of 504 small molecules. The computed ΔG_{solv} are obtained via GB_NSR6 (using SASA+VDW as nonpolar term), as described in the "Methods" section. The explicit solvent ΔG_{solv} were taken from ref 57. ZAP9 radii are used in the GB_NSR6 calculations.

Additional statistics of the GB_NSR6 model quality, evaluated by a direct comparison with experimental data, are presented in table 4.3 in which the results are divided into the training (104 molecules) and the test (400 molecules) sets. The RMS deviation, the average unsigned errors, and the correlation coefficients obtained by GB_NSR6 in both training and test sets are similar to those obtained by the TIP3P explicit solvent model. Moreover, the average error obtained by GB_NSR6 is slightly closer to zero than that obtained by the TIP3P model. About 25% of the molecules in the test set present gross errors, defined as the unsigned errors greater than $2k_BT \sim 1.2$ kcal/mol. This percentage is somewhat smaller than that of TIP3P for which 40% of the molecules present gross errors. Finally, GB_NSR6 produces an RMS deviation of 3.5 kcal/mol for the 5% of the molecules in the testing set with the largest unsigned error. This number is slightly lower, 3.0 kcal/mol, for the TIP3P solvation energies.

A rigorous comparison of accuracies of different GB flavors is a difficult task, as many of them were developed in different contexts and for different applications. For example, some of the flavors were parametrized for specific radii set(s) and for MD simulations of proteins; these radii sets may not perform optimally on small molecules. Thus, comparing between

Table 4.3: Accuracy in ΔG_{solv} calculation (kcal/mol), with respect to experimental data. The explicit(TIP3P) solvent results are from ref 57. The last two rows represent the percentage of molecules with gross errors (error $> 2k_BT \approx 1.2$kcal/mol at 300K), and the RMS deviation of the 5% of the molecules with the largest unsigned errors, respectively. ZAP9 radii are used in the GB_NSR6 calculations.

	Training Set		Test Set	
	TIP3P	GB_NSR6	TIP3P	GB_NSR6
RMS Deviation	1.21	1.16	1.27	1.21
Avg. Error	-0.49	0.01	0.72	-0.25
Avg. —Error—	1.01	0.88	1.04	0.86
Corr. coef. (r^2)	0.89	0.88	0.89	0.86
%—Error—¿ $2k_BT$	36.5%	28.9%	41%	25%
5% worst RMS	2.70	3.07	3.00	3.57

GB flavors with their respective default parameters may not be as benefitial for model developers who seek to understand the root cause of the differences. Likewise, an optimal radii set may result in suboptimal preformance within a "radii-dependent" GB flavor, which also complicates meaningful comparison. Nevertheless, it is important to put the results of GB_NSR6 in perspective with other GB flavors. Such a comparison may be of particular interest to practitioners who want to identify the most suitable "canned" method for the specific task at hand. A recent extensive survey of implicit solvent models by Knigth et. al.[45] offers an excellent reference point, as the set of molecules employed in that survey is the exact same set used in this work. The RMS deviation relative to experiment of the GB models benchmarked in Knight et. al. varies between 1.5 - 2.1 kcal/mol with correlations coefficients between r^2 =0.66 and 0.81. GB_NSR6 with ZAP9 radii set produces an RMS deviation of 1.2 kcal/mol with a correlation coeficient of r^2 =0.86, see table 4.3. These numbers essentially do not change if we use the SASA only model for the nonpolar term as in Knight et. al., see the next subsection for more details.

Overall, these results show that GB_NSR6 in combination with the ZAP9 radii set provides ΔG_{solv} values in reasonable agreement with experiment for the entire data set of 504 small molecules. The overall level of accuracy is similar to the that of the (orders of magnitude more computationally expensive) TIP3P explicit solvent approach. An analysis of the computational efficiency of the model is presented in the following subsections.

4.3.2 Non-polar solvation energy

The non-polar part of the total solvation free energy is calculated via the methodology developed by Gallicchio and Levy [30]. Basically, the non-polar term is divided into two components, the cavity formation and the solute-solvent van der Waals interaction (SASA+VDW), see "Methods". This approach requires 10 parameters which we have optimized against experimental total solvation energies and non-polar solvation energy calculated by TIP3P. Please refer to the "Methods" section for details of this strategy.

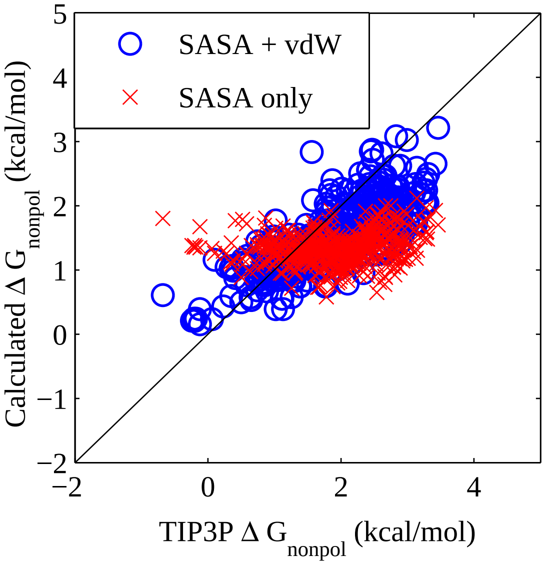

Figure 4.3: Correlation plots between the "SASA only" and "SASA + VDW" models of the non-polar solvation energies (ΔG_{nopol}) and the corresponding explicit solvent (TIP3P) results from ref 57.

The results are presented in figure 4.3: blue circles show a reasonable, though far from perfect, agreement (RMSD = 0.49 kcal/mol and a correlation coefficient of $r^2 = 0.6$) between the values of ΔG_{nopol} obtained by SASA+VDW and the explicit solvent (TIP3P) ΔG_{nopol}. To asses potential advantages of the relatively more complicated (Gallicchio and Levy [30]) form of the ΔG_{nopol} used here vs. a cruder model in which $\Delta G_{\text{nopol}} = \Delta G_{\text{cav}} = \gamma \cdot SASA$, we have investigated how the total solvation energy would agree with experiment if only the cavity term (SASA-only) of the non-polar component were considered ($\Delta G_{\text{nopol}} = \gamma SASA$). The surface tension parameter, γ, for this case has been re-optimized by minimizing the RMS deviation of the total solvation energy obtained by SASA-only and the experimental data available for the training set. The optimum value thus obtained for γ is 0.0051 kcal/mol/Å^2. The RMS deviations and the correlation coefficients relative to experiments obtained by the SASA-only model and the full SASA+VDW are included in table 4.4. The SASA-only model yields an RMS deviation of 1.16 kcal/mol between the computed and experimental total solvation energies for the 504 molecules, which is essentially the same as the RMSD obtained by the full SASA+VDW approach based on 10 additional parameters. Nevertheless, the red cross symbols in figure 4.3 show that the simpler SASA-only model produces a very poor correlation (RMS deviation of 0.83 and correlation coefficient of 0.04) with respect to

Table 4.4: RMS deviation (kcal/mol) and correlation coefficients (r^2) obtained by the SASA+VDW and (SASA-only) procedures, with respect to experimental ΔG_{solv} values (A), and TIP3P computed values of $\Delta G_{\mathrm{nopol}}$ (B).

	SASA+VDW		SASA-only	
	RMSD	r^2	RMSD	r^2
(A) ΔG_{solv}, relative to experiment				
Entire dataset	1.20	0.86	1.16	0.86
Alkanes only	0.41	0.66	0.79	0.50
(B) $\Delta G_{\mathrm{nopol}}$, relative to TIP3P				
Entire dataset	0.83	0.60	0.83	0.04
Alkanes only	0.78	0.37	1.23	0.19

the non-polar solvation obtained by TIP3P solvent model.

To further investigate this trend we have calculated solvation energies of 35 alkane molecules included in the dataset. In this type of molecules the non-polar contribution to the total solvation free energy is no longer considerably smaller than the polar component, so that any inaccuracy in $\Delta G_{\mathrm{nopol}}$ will produce a noticeable effect on the accuracy of the computed total solvation energy. These results, see table 4.4, show that for these very non-polar compounds SASA+VDW provides a distinctly better degree of correlation (relative to SASA-only term) with experimental and TIP3P based values of $\Delta G_{\mathrm{nopol}}$. The loss of correlation of the simpler SASA-only model reported here is consistent with previous observations[98], in which the SASA-only model was used for the calculation of solvation energies of alkane molecules.

Computational Performance Analysis

The computation of the total solvation free energy of the 504 molecules takes approximately 20s (on average 0.04 seconds per molecule), on a commodity PC with a Pentium(R) Dual-Core CPU T4200 2GHz Intel processor, and 2GB of RAM memory. Most of the computational time is consumed by the R6 effective Born radii calculation (50%), followed by the combined processes of surface triangulation and SASA calculation (40%) , which are performed by MSMS. The rest of the computational time is consumed by relatively faster processes such as evaluation of equation 4.1 and computation of the electrostatic size. Note that once the computation of polar free energy is finished, the van der Waals term is obtained at virtually no additional cost because the effective Born radius of that atom, R_i, in equation 4.1 have been already computed.

Any detailed and exhaustive comparison of currently available methods is outside the scope of this work. Moreover, our experience suggests that when comparing speeds of very differnt

Table 4.5: Average computational time per molecule and RMS deviation to experiment, for the calculation of total solvation energies of 504 neutral small molecules. The data for TIP3P and SEA performance are from ref 29. Computations performed on a single processor commodity PC.

Methodology	Computational time	RMSD to experiment
Fully explicit solvent (TIP3P)	Hours – days	1.2 kcal/mol
Semi explicit (SEA)	Around a second	1.3 kcal/mol
Numerical implicit solvent (GB_NSR6)	Few tens of milliseconds	1.2 kcal/mol
Analytical implicit solvent (GB_HCT)	Few milliseconds	2.5 kcal/mol

algorithms, seeking to achieve a distinction more accurate than about a factor of two may not be prudent. The speed of any particular implementation of a complex algorithm can always be improved by clever optimizations at various levels. Accordingly, here we use performance data from a recent study that provides a general performance comparison sketch of different methodologies to compute solvation free energies of small molecules, see table 4.5. Readers still interested in precise timings for some of the methods we tested in the course of this study are referred to the Supporting Material, table 4.5. We notice that, being fully implicit, GB_NSR6 is expectedly faster than fully, or semi explicit methods. Analytical GB flavors such as the widely used GB_HTC [40] from Amber package [16] are several times faster than the numerical GB_NSR6, but at the expense of a substantial reduction of the accuracy, see table 4.5.

Comparison between GB_NSR6 and the Poisson-Boltzmann method

Here we examine the accuracy of the GB_NSR6 and the numerical PB treatment on the same footing, relative to both experiment and explicit solvent model (TIP3P). The RMS deviation of the computed ΔG_{solv} and ΔG_{el} free energies relative to experiment and TIP3P respectively figure 4.4 are shown in figure 4.4. In all of these calculations the SASA only model was used to compute ΔG_{nopol}. Overall, these results show that for an optimum radii set, GB_NSR6 is at the same level of accuracy as the PB model.

Within the PB model, (squares, figure 4.4) ZAP9 radii also performs better than the other two radii sets. This is expected since ZAP9 was optimized for numerical PB using experimental solvation energies[64]. Considering the optimum ZAP9 radii set, GB_NSR6 and the PB produce almost the same RMS deviation relative to experiment, 1.2 kcal/mol and 1.21 kcal/mol respectively. Relative to explicit solvent calculations (TIP3P), GB_NSR6 appears to be slightly more accurate than the PB. However, the difference between GB_NSR6 and the PB (RMS error 0.9 kcal/mol) is smaller than the error of the TIP3P explicit solvent model relative to experiment (RMS error 1.26 kcal/mol), and so the GB_NSR6 vs. PB difference here should not be viewed as significant.

For suboptimal radii sets (BONDI and PARSE) the RMS errors of both the PB and GB_NSR6 solvation energies are substantially larger than those produced by ZAP9. A noticeable difference also exists between the GB_NSR6 and the PB results, see the green and blue symbols in figure 4.4. There may be many reasons for the systematic errors produced by the BONDI and PARSE radii sets; we discuss one concrete conjecture in the the Supporting material.

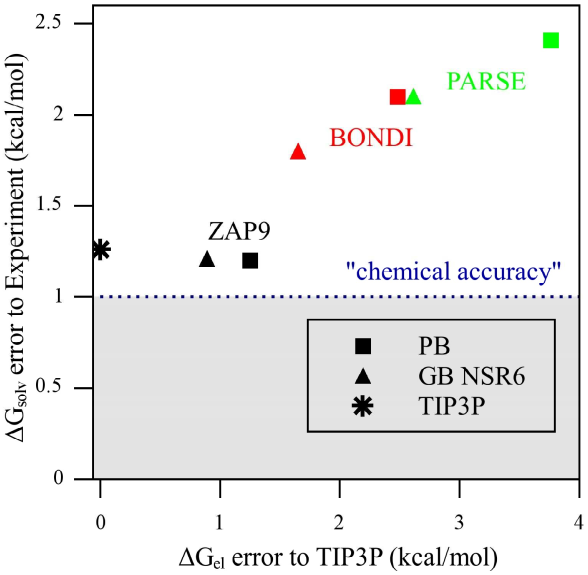

Figure 4.4: RMS errors of the computed total ΔG_{solv} and polar ΔG_{el} relative to experiment and explicit solvent models (TIP3P) respectively. The computations are by the PB (squares) and GB_NSR6 (triangles) methods with ZAP9 (black), BONDI (red), and PARSE (green) radii sets. The shaded region represent the desirable "chemical accuracy" of 1 kcal/mol relative to experiment.

4.4 Conclusions

In this work we have evaluated the performance of a recently developed model, GB_NSR6[4], which is based on the so-called "R6" flavor of the generalized Born implicit solvent approximation. Our main motivation was to test how well the parameter-free GB_NSR6 can perform on small molecules relevant to drug-design efforts where efficient and accurate computation of solvation energies is key. We have used a common dataset of 504 small molecules with available experimental and explicit solvent total solvation energies[57]. For this particular data set, GB_NSR6 produces an RMS deviation in ΔG_{solv} of 1.2 kcal/mol relative to experimental solvation energies. This level of accuracy is the same as that (1.26 kcal/mol) obtained by computationally much more expensive MD simulations with explicit (TIP3P) solvent model. The $\sim$ 1kcal/mol accuracy limit for small molecule ΔG_{solv} is the current expectation for models that share the same underlying physics with the Poisson treatment[66]. While higher accuracy may be achived for specific data sets at the expense of a large number of adjustable parameters[47], transferability will likely be affected. The proposed GB_NSR6 is free from this defect as the electrostatic part of ΔG_{solv} is computed with no parameters (and the non-polar part has little affect on the over-all accuracy). Thus, we are confident that the achieved accuracy will be relevant beyond the specific test set we used.

Comparing computational efficiencies between very different methods (e.g fully explicit solvent, semi-explicit[29] or fully implicit such as the GB) is always difficult. However, rough order-of-magnitude comparisons can still be made, especially if based on the same test set[29]. In this respect, GB_NSR6 – which is based on numerical integration to obtain effective Born radii – is secondary only to optimized analytical GB models such as those available in AMBER. The latter are still several times faster than GB_NSR6, but at the expense of substantial accuracy loss.

The polar component of the solvation energy is computed with no adjustable parameters by GB_NSR6, producing an RMS deviation of 0.89 kcal/mol relative to explicit solvent treatment. This degree of accuracy is similar to those obtained by other recent implicit or hybrid methods such as the one proposed by Fennell et al. [29]. Our results show that the intrinsic radii set, which define the molecular surface, plays a key role in the accuracy of the polar part of solvation energy calculations. The ZAP9 radii set recently proposed by Nicholls et al. [64] provides more accurate results than other standard radii set such as BONDI and PARSE, when compared to TIP3P explicit solvent model. For this optimal set, both the total and polar solvation energies obtained by GB_NSR6 and the Poisson Boltzmann treatment are very similar.

These observations warrant further investigation to evaluate performance of ZAP9 radii set for larger molecules such as nucleic acids or proteins, as its transferrability outside of small molecules is by no means guaranteed.

In this work, the non-polar component of the solvation free energy is calculated as the sum of two components, the cavity and VDW terms as proposed by Gallicchio and Levy[30]. Once ΔG_{el} is computed, the VDW term is calculated with almost no cost, since it depends on the R6 Born radii obtained previously in the computation of ΔG_{el}. Our results shows that the RMS deviation of this methodology (relative to experimental data and TIP3P results) is similar to that obtained when $\Delta G_{\mathrm{nopol}}$ is calculated as being proportional to Solvent Accessible Surface Area. However, the methodology provides a better correlation with TIP3P based $\Delta G_{\mathrm{nopol}}$, and experimental solvation free energies of alkane molecules, in which the non-polar component is predominant, see table 4.4.

Overall the GB_NSR6 model used in this work shows competitive performance when applied to a data set of 504 small molecules, providing a good balance between computational efficiency and accuracy. In our opinion, interested parties may benefit from these features especially for fast computation of large dataset of molecules. Another useful fature of this GB flavor is that it is not optimized for any specific radii set and thus can be used to evaluate relative preformance of different sets. All of the software developed in this work is freely available from http://people.cs.vt.edu/ onufriev/software.php

Chapter 5

Molecular Dynamic based on AR6

This chapter will appear as the reference [3].

In this chapter we modified the original formulation of AR6 and used for Molecular Dynamics simulations. More specifically, we added more parameters and a treat hydrogen atoms with an slightly different approximation. This version of AR6 is then used to fold an small protein of 10 amino acids, CLN025.

5.1 Introduction

An accurate description of solvent is essential for modeling and simulation of biological macromolecules. Currently, the most rigorous procedure to model the effect of aqueous solvent is to explicitly include the water molecules in the system. However, this method is computationally too intense for many practical applications. Implicit solvent models, in which solvent molecules are represented by a continuum function, have become a popular alternative to explicit solvent methods, especially in Molecular Dynamics simulations, as they are more computationally efficient [18, 41, 12, 56, 35, 80, 54]. Within the framework of implicit solvent models, macromolecules are treated as a low dielectric 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}, which is typically divided into polar (ΔG_{el}) and non polar ($\Delta G_{\mathrm{nopol}}$) terms. Here we focus on efficient methods to estimate ΔG_{el} values of biomolecules, because the computation of ΔG_{el} is currently the computational bottleneck of Molecular Dynamic simulations based on the GB model.

Within the linear response continuum implicit solvent framework, solving the Poisson-Boltzmann equation (PB) is theoretically the most rigorous way to compute ΔG_{el} [34, 41, 55, 12, 80, 18, 9]. However, the PB model may become quite time-consuming, especially when incorporated into molecular dynamics (MD) simulations, where its practical implementation faces several other challenges. The Generalized Born model (GB) has become a popular alternative to the PB model for the computation of ΔG_{el} [26, 89, 39, 40, 81, 75, 24, 43, 32, 11, 49, 28, 78, 21, 19, 95, 14, 88, 86, 97, 67, 71, 30, 50], especially in MD simulations, due to its closed and simple form expression for computing ΔG_{el}. Unfortunately, several examples exists that show a poor accuracy on ΔG_{el} estimation of many available GB variants. The problem has serious implications in Molecular Dynamic simulations, for instance, it is know that the GB model produce an undesirable bias towards helical secondary structures and over-stabilization of salt bridges[77, 31].

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

$$\Delta G_{\text{el}} \approx \Delta G_{\text{GB}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\sum_{ij}\frac{q_i q_j}{f^{\text{GB}}(r_{ij}, R_i, R_j)}, \tag{5.1}$$

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[89] of f^{GB} is $f^{\text{GB}} = \left[r_{ij}^2 + R_i R_j \exp(-r_{ij}^2/4R_i R_j)\right]^{\frac{1}{2}}$, although other similar expressions have been tried[43, 72].

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 equation 5.1, the ΔG_{el} computed by GB are in close agreement with those computed by PB [72]. The computationally expensive “perfect” effective Born radii are commonly used for benchmarking and testing different GB “flavors”–approximations that compute R_i.

An expression to compute the effective Born radii, which will be called here “R6 radii”, was proposed by Svrcek-Seiler[90] and independently by Grycuk[37]:

$$R_i^{-1} = \left(\frac{3}{4\pi}\int_{ext}\frac{d\mathbf{V}}{|\mathbf{r} - \mathbf{r}_i|^6}\right)^{1/3} = \left(\rho_i^{-3} - \frac{3}{4\pi}\int_{r>\rho_i}^{solute}|\mathbf{r}|^{-6}d\mathbf{V}\right)^{1/3} \tag{5.2}$$

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. The “R6 radii” are exact for any location of a charged atom within a perfect spherical solute in the $\epsilon_{out}/\epsilon_{in} \gg 1$ limit. Recently, it has been shown that when the “R6 radii” are computed by essentially exact numerical integration of equation 5.2, 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 equation 5.1 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 physically realistic molecular (Lee-Richards[48]) 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.

We have recently developed a method, named AR6, to approximate the effective Born radii based on equation 5.2. The approximation is composed of analytical functions of atomic positions whose derivatives are continuous with respect to atomic positions. These features make the approximation suitable for molecular dynamics simulations as the computation of atomic forces requires the computation of the derivatives of the potential energy of the system. Here we modified AR6 adding more parameters to the model. The modified AR6 is tested in the computation of ΔG_{el} on an extensive dataset of structures. We compare the results with those obtained by explicit explicit solvent models and by the PB model.

5.2 Methods

5.2.1 Computation of effective Born radii, AR6

Recently, we have proposed an approach to approximate effective Born radii based on the R6 integration, equation 5.2. The method is based on the analytical integration (equation 5.2) over several geometrical regions that together, locally approximate the molecular volume[4]. The integration over these regions are shown in the appendix. It is important to note the derivatives of the functions that represent the R6 integration over these different regions are continuous with respect to atomic positions. This is important for a molecular dynamic simulations as discontinuities in derivatives of the effective Born radii can produce unphysically large atomic forces. In this work we simplified this method for Molecular dynamics simulation. Moreover, we added additional parameters to improve the accuracy of ΔG_{el} estimation. Here we provided a brief description of the method. The molecular volume that surrounds atom i is approximated by the union of three distinct regions:

1) The atomic VDW volume corresponding to atomic spheres, excluding atoms inside 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 between close pairs of atom.

3) The essentially exact molecular volume of the "chunk" of atom i.

The Integration over the VDW volume represent the R6 integration (equation 3.7 over the atomic spheres not bonded to any atom of the "chunk" of atom i ("nochunk"). The contribution of the VDW sphere of atom j to the effective Born radius of atom i was analytically calculated previously. In this work it is represented by the function $F_6(\rho_i, \rho_j, r_{ij})$ defined in the Appendix. To reduce the error for not considering overlapping between atoms, we shrink the atoms by atom type factor f_j, $\mathbf{I}_i^{vdw}$ is then computed by:

$$\mathbf{I}_i^{vdw} = \sum_{j \in \text{``nochunk''}\ i} \mathbf{F}_6(\rho_i + \rho_w, S_j, r_{ij}), \tag{5.3}$$

In order to correct the underestimation of effective Born radii of buried atoms, we use 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 analytical model of solvation. V_i is computed by the following equation

$$\mathbf{V}_i = R_s^{-3} \sum_{j \neq i} \Theta_{ij}, \tag{5.4}$$

The Neck region is the intermediate volume region enclosed by two atomic spheres i and j, and a probe radius that rolls over them, please see chapter 3, section 3.2.2 for a more detailed definition. The integration over the neck region was originally developed by Mongan et al. [60] in the context of CFA ("R4 integration"), here we use an approximation of the integration over the "neck" of atoms i and j in the context of "R6 integration". The approximation of the R6 integration over the neck region is defined by N_{ij} (Appendix), the contribution of the necks close to atom i is defined by the equation 5.5 in which n_i is a fitting parameter that depend on the atom type of i. Only the atoms not bonded to any atom of "chunk" are considered ("nochunk") to compute $\mathbf{I}_i^{neck}$. It is important to note that the derivative of equation 5.5 with respect to atomic distances is continuous which is an important requirement for molecular dynamic simulations.

$$\mathbf{I}_i^{neck} = n_i \sum_{j \in \text{``nochunk''}\ i} \mathbf{N}_{ij}, \tag{5.5}$$

The "chunk" of atom i is defined as a small set of neighboring atoms such that the number of covalent bonds between the atoms of the "chunk" and atom i is at most 3. The R6 integration over the volume covered by the atoms of the "chunk" is computed by the equation 5.6, where λ_i is a parameter that was calculated with the procedure described in reference [4], $\mathbf{F_6}$ is a function defined in the Appendix.

$$\mathbf{I}_i^{chk} = \lambda_i \sum_{j \in \text{``chunk''}\ i} \mathbf{F}_6(\rho_i, \rho_j, r_{ij}), \tag{5.6}$$

The effective Born radii of atom i is then approximated by the following equations, which require the fitting of parameters n_i and s_i

$$\mathbf{I}_i = \mathbf{I}_i^{vdw} + (1 + s_i \mathbf{V}_i^3)\mathbf{I}_i^{neck}, \tag{5.7}$$

$$\tilde{\rho}^{-3} = \rho^{-3} - \mathbf{I}_i^{chk}, \tag{5.8}$$

$$R_i^{-1} = \left[\tilde{\rho}^{-3} - (\tilde{\rho}^{-3} - A_0)\tanh(\frac{\mathbf{I}_i}{\tilde{\rho}^{-3} - A_0})\right]^{1/3} + B \tag{5.9}$$

Where s_i and n_i are fitting parameters that depend on the atom type. We used a total of 8 different atom types for "necks" and a single parameter for s_i. In total we fitted 9 parameters. A_0 is an offset parameter used to limit the value of the effective Born Radii. We set up its value to 0.001.

5.2.2 Especial treatment of hydrogen atoms

Analyzing the accuracy of effective Born radii per atom type, we found that buried hydrogen atoms are the least accurate. This is true for AR6 and also for other GB models currently implemented in AMBER, igb5 and igb8. The issue is relevant because hydrogen are the most abundant atoms in proteins. In order to improve the accuracy for these atoms, we propose the following device. The effective Born radii of surface hydrogens atoms (with small effective radii) are computed with equation 5.9, while the effective radii of buried hydrogens (with large effective radii) are equal to the effective radii of their corresponding covalently bonded atom. To obtain an smooth transition between the two previous cases, we use the following sigmoid function that depends on the volume factor described in equation 5.4:

$$\mathrm{sig}(x) = \frac{1}{1 + \exp(-a_H(x - b_H))} \tag{5.10}$$

Where a_H and b_H are fitting parameters that modulate the sigmoid function. The final effective radii of Hydrogen atoms are then computed by the following equation.

$$R_{i_H}^{-1} = R_{i_H}^{-1}(1 - \mathrm{sig}(\mathbf{V}_{i_H})) + \mathrm{sig}(\mathbf{V}_{i_H})R_{i_B}^{-1} \tag{5.11}$$

Where i_H is the index of a hydrogen atom and i_B is the index of the atom covalently bonded to atom i_H. The optimum values of a_H and b_H are 70.333 and 0.2436 respectively. These vales were obtained by minimizing the RMS deviation of the computed effective Born radii of hydrogens atoms, using NSR6 as reference.

5.2.3 Training set for parameter fitting

We have used the NelderMead simplex algorithm to optimize the parameters of the AR6 model. The objective function to be minimized is the weighted sum of four terms. Three

Table 5.1: RMS error in $\Delta\Delta G_{\mathrm{el}}$ to NSR6 for Ala10 and protein-A, and RMS error in inverse effective Born radii relative to NSR6 for a set of 11 proteins. The objective function is obtained as a weighted sum of RMS errors.

	RMS error	weight
$\Delta\Delta G_{\mathrm{el}}$ relative to pp2	0.0007 kcal/mol	30
$\Delta\Delta G_{\mathrm{el}}$ non pp2	0.0008 kcal/mol	30
$\Delta\Delta G_{\mathrm{el}}$ protein A	0.0008 kcal/mol	30
Inverse effective radii	0.0564 $\mathring{A}^{-1}$	1000
Objective function	56.50	

values of RMS error of differences in solvation energies ($\Delta\Delta G_{\mathrm{el}}$), and the RMS error of the computed inverse effective Born radii, see table 5.1. The set of structures used to compute RMS errors consists of four conformations of Ala10, the folded and unfolded state of protein A, and set of 11 proteins. The details of the structures and the computation of the objective function are provided in the following paragraphs.

Roe et al.[77] used explicit solvent TI calculations to estimate the solvation energy of four conformations of Ala10 (pp2, alpha, left, and hairpin); we used this system in our training set. Each conformation of Ala10 is composed of 10 structures obtained from MD simulations of the TI calculation. The ΔG_{el} corresponding to each conformational state is computed by averaging the values of ΔG_{el} of each of the corresponding MD snapshots. This system contributes to the objective function with two terms, first the RMS error in $\Delta\Delta G_{\mathrm{el}}$s between the alpha, left, and hairpin conformation, relative of pp2. Second, the RMS error of the $\Delta\Delta G_{\mathrm{el}}$ values between left and alpha, between hairpin and alpha, and between hairpin and left. The values of $\Delta\Delta G_{\mathrm{el}}$ described above are computed by AR6 and compared to those computed by NSR6. The error in estimation of $\Delta\Delta G_{\mathrm{el}}$ between the folded and unfolded conformation of protein-A is also included in the objective function. The folded and unfolded state ensemble are represented by 50 consecutive snapshots obtained from the implicit solvent simulation, see reference [68] for more details. Finally, the RMSD of the computed inverse of the effective radii is included into the objective function. We used the inverse of the effective radii obtained by NSR6 as reference. A set of small structures ranging from 19 to 76 aminoacids are used for effective Born radii computation.

Table 5.2: Parameters of AR6

Atom type	n_i	s_i
C bonded to 3 Hs	0.1822	50.6587
C bonded to 2 Hs	0.4939	50.6587
C	0.4512	50.6587
N	0.2938	50.6587
O	0.4911	50.6587
H	0.4098	50.6587
H bonded to N	0.3056	50.6587
S	0.4500	50.6587

5.3 Results

5.3.1 Effective Born radii

In this section we test the accuracy of the effective Born radii computed by AR6 and compare it with the accuracy of other methods currently implemented in AMBER. The effective Born radii obtained via numerical PB calculations (the so called "perfect Born radii", see [69]) are often used as benchmarks to test the accuracy of different GB flavors. Here, we use a methodology called NSR6 as reference instead. NSR6 is based on a numerical surface integration of equation 5.2, and produces effective Born radii and energies that are in very good agreement with respect to those based on numerical PB. For instance, the correlation plot between the NSR6 effective radii and the perfect (PB based) effective radii shown in figure **??** for thioredoxin (1654 atoms) shows that NSR6 has an excellent accuracy when compared to the perfect Born radii.

Therefore we used NSR6 as reference to test AR6 in computing effective Born radii and solvation energies. Moreover, having NSR6 as reference can help identify some sources of error in methods that approximate effective Born radii, such as AR6. We have used a set of 11 proteins to test the accuracy of computed effective Born radii. Figure 5.2 shows a correlation plot between the approximated (AR6, igb5, and igb8) inverse of the effective Born radii and the numerically exact effective Born radii(NSR6). We observe that AR6 shows an improvement over igb5 and igb8 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 equation 5.1. AR6 is also, on average, more accurate in regions of large effective Born radii that correspond to atoms deeply buried inside the protein.

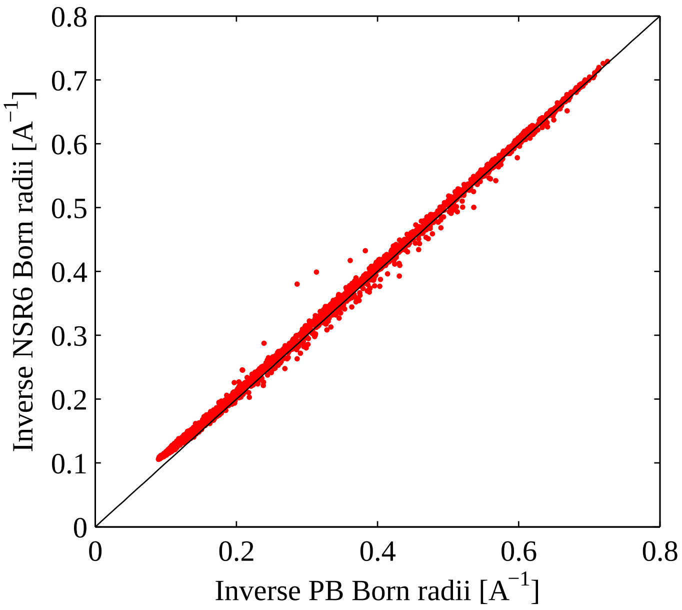

Figure 5.1: Scatter plot comparison of inverse effective radii calculated by the numerical surface integration method NSR6, (red) to inverse “perfect” PB radii for thioredoxin (PDB code 2TRX). Diagonal line indicates perfect agreement.

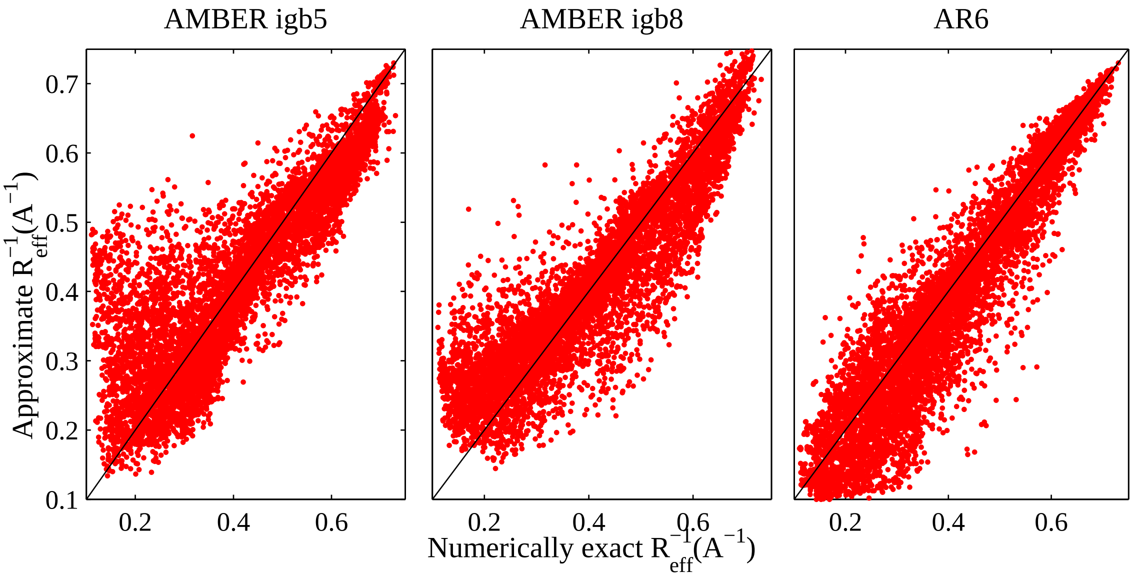

Figure 5.2: Correlation plots between the values of effective Born radii computed by igb5, igb8, and AR6, and those computed by the reference NSR6 (x axis).

5.3.2 Solvation energies

The electrostatic part of the solvation energy is calculated by the NSR6, AR6, igb5, and igb8 methodologies, on a data set of 200 small proteins, see the Methods section for details. These structures were used earlier for parametrization of GB models. The distributions of of errors in ΔG_{el} values for AR6, igb8, and igb5 are shown in Figure, where NSR6 is used as reference. AR6 has an average error of 5.5 kcal/mol, which is at the same level of error than igb8 (8.8 kcal/mol average error), and better than igb5 (30.8 kcal/more average error). AR6 also produces a better standard deviation of the error (20.0 kcal/mol) compare to that of igb8 (27.8 kcal/mol) and igb5 (31.5 kcal/mol). In summary, the results show that on average, AR6 provides a good accuracy compared to other available GB models.

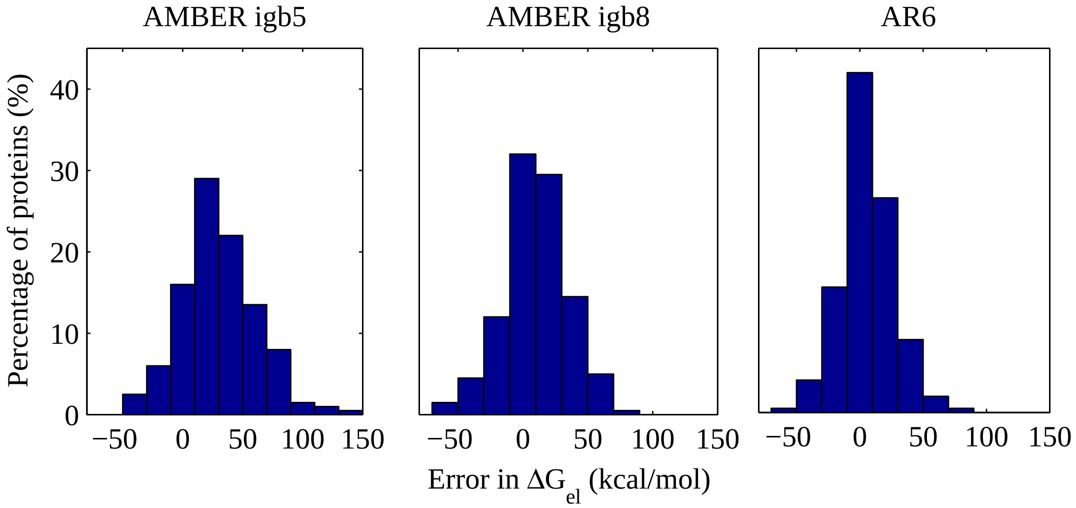

Figure 5.3: Distribution of errors in ΔG_{el} estimation for igb5, igb8, and AR6. The errs are computed for a set of 200 proteins using GB NSR6 as reference.

5.3.3 Solvation energies relative to explicit solvent

The best strategy to test the accuracy of the GB model is to compare the computed solvation energies directly with experiments. This strategy has been used previously in small molecules. However, experimental salvation energies for a statistically significant set of proteins are not yet available. Here, we test the accuracy of GB (and also PB) comparing salvation energies with those obtained by using explicit solvent (TIP3P). The comparison was carried out over a set of 19 small proteins of about 30 aminoacids, details of these structures are found in Methods. To avoid conformational sample issues conformations of all the test molecules are fixed. We choose only net neutral proteins mainly to mitigate the

Table 5.3: Electrostatic part of the solvation energy (kcal/mol) of 19 small proteins computed by using the TIP3P explicit solvent model (TI), PB, NSR6, AR6, and IGB8.

PDB_ID	TI	PB	NSR6	AR6	IGB8
1az6	-236.21	-253.11	-242.19	-247.23	-260.49
1byy	-372.59	-381.62	-381.05	-379.91	-373.93
1eds	-295.36	-315.50	-306.17	-301.02	-309.05
1g26	-280.56	-299.68	-294.53	-296.40	-304.35
1qfd	-347.10	-354.79	-351.81	-356.82	-368.82
1bh4	-239.80	-240.78	-231.61	-229.02	-230.15
1cmr	-374.18	-384.40	-378.23	-382.75	-392.51
1fct	-248.98	-265.73	-260.77	-263.03	-262.27
1ha9	-387.33	-383.31	-376.34	-371.81	-381.75
1qk7	-509.11	-518.94	-514.48	-513.06	-521.06
1bku	-396.43	-404.48	-399.10	-395.24	-404.12
1dfs	-469.62	-483.92	-476.74	-481.21	-483.00
1fmh	-321.26	-343.86	-339.54	-326.90	-346.56
1hzn	-439.23	-434.90	-425.76	-420.10	-427.96
1scy	-288.15	-294.94	-287.82	-279.10	-297.04
1brv	-206.74	-213.62	-212.78	-207.52	-217.80
1dmc	-360.81	-380.44	-372.47	-375.75	-381.68
1fwo	-462.71	-485.18	-473.52	-473.12	-482.76
1paa	-478.93	-479.71	-476.02	-465.91	-474.69
RMSD to TI		13.53	9.42	11.03	15.66
RMSD to PB			6.90	9.79	5.99

uncertainties associated with the absence of charge hydration asymmetry. Table 5.3 shows the electrostatic solvation energies for each of the 19 small proteins computed by explicit solvent (TIP3P), PB, NSR6, AR6, and igb8. Surprisingly, among the tested methods NSR6 is closer to the results obtained by explicit solvent model with a RMS deviation in polar solvation energies of 9.42 kcal/mol. PB is second with an RMS deviation of 13.53 kcal/mol. These results highlight the differences in GB and PB accuracies and suggest that using PB as reference can results in over fitting and error compensation problems in parameter based GB variants. Finally, AR6 shows a a RMS deviation 11.03 kcal/mol relative to TIP3P which is better than the other methods. This level of accuracy is the result of our strategy of fitting the parameters of AR6 using NSR6 as reference rather than PB.

5.3.4 MD simulations with the improved AR6 model

One of the most attractive features of the GB model is the possibility to use it in MD simulations where it can effectively mimic solvation effects without the need of including water molecules explicitly in the system. An essential component of Molecular Dynamic simulations based on implicit solvent models is the computation of solvation forces, which requires the effective radii and their corresponding derivatives with respect to atomic positions. The equations to compute the derivatives of the effective radii are included in the appendix. We implemented the computation of solvation forces in an local version of AMBER12 which were used to generate the results of this chapter.

We begin to test the performance AR6 model in MD, by running two 200ns MD simulations of the small (10 aminoacids) protein CLN025. One MD simulation is performed at 300K, a temperature in which CLN025 is mostly folded in a hairpin conformation. The second MD simulation is carried out at 340K, the experimental folding temperature of CLN025. At this temperature, we expect to observe many transitions from folded to unfolded conformation, and the opposite transition events. Figure 5.4 shows the backbone RMSD from the folded structure obtained experimentally by NMR. At 300K, CLN025 stays in the folded state after a transition period. As expected, at 340K we observe several folding-unfolding events. For comparison we also included a MD simulation obtained by using the popular OBC method (igb = 5 in AMBER). At 300K the OBC simulation does not stabilize in the expected folded conformation, but transitions from the folded to unfolded simulation. At 340K OBC is in the unfolded state most of the simulation time.

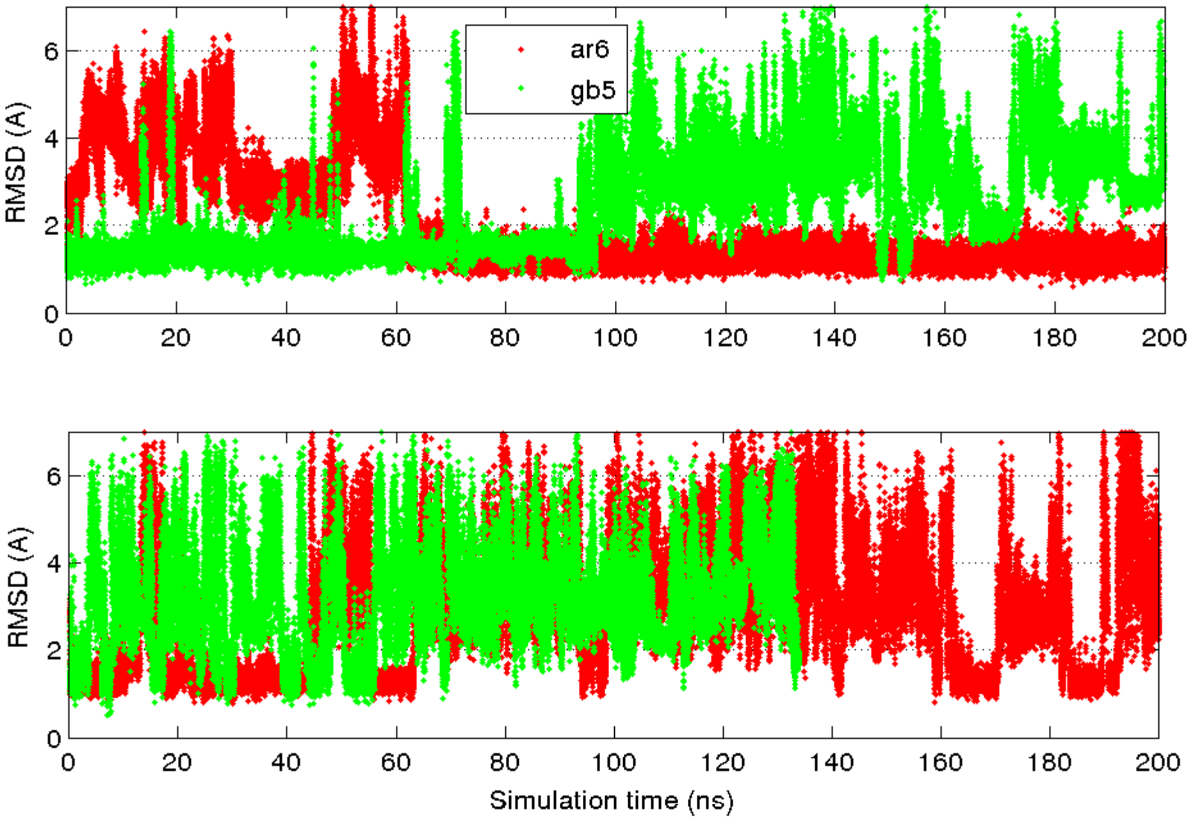

Figure 5.4: Backbone Root Mean Square deviation (RMSD) in Angstrom units(A) from experimental NMR structure during an MD simulation of CLN025 using AR6(Red) and gb5 available in AMBER12(Green).

Chapter 6

Conclusion

In this work we have developed two GB models aimed at computing solvation free energies with a good balance of efficiency and accuracy. The first GB model (NSR6) is based on a numerical integration over the molecular surface, which provides solvation energies comparable to those computed by the more computationally expensive PB model. The second GB model (AR6) is an analytical approximation of NSR6 suitable for Molecular Dynamic simulations.

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 small theoretical structures. The R6 approach advocated here is simple – based on a single integrand – 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 model.

We have performed a fairly extensive set of accuracy tests for the new models. These included comparing electrostatic solvation free energies (ΔG_{el}) against the numerical PB, explicit solvent simulations, and experimental solvation energies where available. In particular, we evaluated the performance of NSR6 in a common dataset of 504 small molecules with available experimental and explicit solvent total solvation energies[57], and a set of 19 small protein structures with available explicit solvent solvation energies. Moreover, we evaluated 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. The effective radii computed by AR6 is compared to NSR6 in a set of 11 small proteins, and the performance of AR6 in molecular Dynamic Simulations is evaluated by running MD simulations of the miniprotein CLN025.

Overall the NSR6 model developed in this work shows competitive performance when applied to our testing data set, providing a good balance between computational efficiency and accuracy. for instance, NSR6 produces an RMS deviation in ΔG_{solv} of 1.2 kcal/mol relative to experimental solvation energies on the data set of 504 small molecules. This level of accuracy is the same as that (1.26 kcal/mol) obtained by computationally much more expensive MD simulations with explicit (TIP3P) solvent model. In our opinion, interested parties may benefit from these features especially for fast computation of large dataset of molecules. The analytical AR6 method to compute the effective Born radii offers a clear improvement in accuracy over a set of popular pairwise methods implemented in the popular package AMBER, without apparent sacrifices in computational complexity. This makes the approach a promising candidate for applications that require repetitive computations of ΔG_{el}. AR6 was developed with MD in mind as robustness, stability and differentiability were strictly enforced. AR6 is implemented in an local version of the AMBER package.

Bibliography

[1] B. Aguilar, R. Anandakrishnan, J. Z. Ruscio, and A. V. Onufriev. Statistics and physical origins of pK and ionization state changes upon protein-ligand binding. *Biophys. J.*, 98:872–880, Mar 2010.

[2] Boris Aguilar and Alexey V. Onufriev. Efficient computation of the total solvation energy of small molecules via the r6 generalized born model. *Journal of Chemical Theory and Computation*, 8(7):2404–2411, 2012.

[3] Boris Aguilar and Alexey V. Onufriev. Protein folding with a new generalized born model. *In progess*, 2014.

[4] Boris Aguilar, Richard Shadrach, and Alexey V. Onufriev. Reducing the secondary structure bias in the generalized born model via r6 effective radii. *J. Chem. Theory Comput.*, 6(12):3613–3630, December 2010.

[5] R. Anandakrishnan, B. Aguilar, and A. V. Onufriev. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic acids research*, 40:W537–W541, 2012.

[6] C. Bajaj, S. Chen, and A. Rand. An efficient higher-order fast multipole boundary element solution for poissonboltzmann-based molecular electrostatics. *SIAM Journal on Scientific Computing*, 33(2):826–848, 2011.

[7] C. Bajaj and W. Zhao. Fast molecular solvation energetics and forces computation. *SIAM Journal on Scientific Computing*, 31(6):4524–4552, 2010.

[8] N A Baker, D Sept, S Joseph, M J Holst, and J A McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.*, 98(18):10037–10041, Aug 2001.

[9] Nathan A Baker. Improving implicit solvent simulations: a Poisson-centric view. *Curr. Opin. Struct. Biol.*, 15(2):137–143, Apr 2005.

[10] Donald Bashford. *An Object-Oriented Programming Suite for Electrostatic Effects in Biological Molecules*, volume 1343 of *Lecture Notes in Computer Science*, pages 233–240. Springer, Berlin, Germany, 1 edition, 1997.

[11] Donald Bashford and David A Case. Generalized Born Models of Macromolecular Solvation Effects. *Annu. Rev. Phys. Chem.*, 51:129–152, Jan 2000.

[12] P. Beroza and David A. Case. Calculation of Proton Binding Thermodynamics in Proteins. *Methods Enzymol.*, 295:170–189, 1998.

[13] A. Bondi. van der waals volumes and radii. *J. Phys. Chem.*, 68(3):441–451, 1964.

[14] N Calimet, M Schaefer, and T Simonson. Protein molecular dynamics with the generalized Born/ACE solvent model. *Proteins*, 45(2):144–158, Nov 2001.

[15] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The Amber biomolecular simulation programs. *J. Comput. Chem.*, 26(16):1668–1688, December 2005.

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

[17] Jana Chocholousová and Michael Feig. Balancing an accurate representation of the molecular surface in generalized born formalisms with integrator stability in molecular dynamics simulations. *J. Comp. Chem.*, 27(6):719–729, 2006.

[18] C. J. Cramer and D. G. Truhlar. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.*, 99:2161–2200, 1999.

[19] L David, R Luo, and Michael K Gilson. Comparison of generalized Born and Poisson models: Energetics and dynamics of HIV protease. *J. Comput. Chem.*, 21(4):295–309, Mar 2000.

[20] Sergio Decherchi, José Colmenares, Chiara Eva E. Catalano, Michela Spagnuolo, Emil Alexov, and Walter Rocchia. Between algorithm and model: different molecular surface definitions for the Poisson-Boltzmann based electrostatic characterization of biomolecules in solution. *Communications in computational physics*, 13:61–89, January 2013.

[21] B N Dominy and Charles L Brooks. Development of a generalized Born model parametrization for proteins and nucleic acids. *J. Phys. Chem. B*, 103(18):3765–3773, May 1999.

[22] Feng Dong and Huan-Xiang Zhou. Electrostatic contribution to the binding stability of protein-protein complexes. *Proteins*, 65(1):87–102, 2006.

[23] J. Dzubiella, J. M. Swanson, and J. A. McCammon. Coupling nonpolar and polar solvation free energies in implicit solvent models. *J. Chem. Phys.*, 124(8):084905, February 2006.

[24] S R Edinger, C Cortis, P S Shenkin, and R A Friesner. Solvation free energies of peptides: Comparison of approximate continuum solvation models with accurate solution of the Poisson-Boltzmann equation. *J. Phys. Chem. B*, 101(7):1190–1197, Feb 1997.

[25] D. Eliezer, J. Yao, H. J. Dyson, and P. E. Wright. Structural and dynamic characterization of partially folded states of apomyoglobin and implications for protein folding. *Nat. Struct. Biol.*, 5:148–155, 1998.

[26] Michael Feig and Charles L Brooks. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.*, 14(2):217–224, Apr 2004.

[27] Michael Feig, Alexey Onufriev, Michael S Lee, Wonpil Im, David A Case, and Charles L Brooks. Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem.*, 25(2):265–284, Jan 2004.

[28] A K Felts, Y Harano, E Gallicchio, and R M Levy. Free energy surfaces of beta-hairpin and alpha-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model. *Proteins*, 56(2):310–321, Aug 2004.

[29] Christopher J. Fennell, Charles W. Kehoe, and Ken A. Dill. Modeling aqueous solvation with semi-explicit assembly. *Proc. Natl. Acad. Sci. U. S.Ȧ.*, 108(8):3234–3239, February 2011.

[30] E Gallicchio and R M Levy. AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comput. Chem.*, 25(4):479–499, Mar 2004.

[31] Raphaël Geney, Melinda Layten, Roberto Gomperts, Viktor Hornak, and Carlos Simmerling. Investigation of salt bridge stability in a generalized born solvent model. *J. Chem. Theory Comput.*, 2(1):115–127, November 2006.

[32] A Ghosh, C S Rapp, and R A Friesner. Generalized Born model based on a surface integral formulation. *J. Phys. Chem. B*, 102(52):10983–10990, Dec 1998.

[33] M. K. Gilson and H. X. Zhou. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.*, 36:21–42, 2007.

[34] Michael K Gilson, M E Davis, B A Luty, and J Andrew McCammon. Computation of Electrostatic Forces On Solvated Molecules Using the Poisson-Boltzmann Equation. *J. Phys. Chem.*, 97(14):3591–3600, Apr 1993.

[35] Micheal K. Gilson. Theory of Electrostatic Interactions in Macromolecules. *Curr. Opin. Struct. Biol.*, 5:216–223, 1995.

[36] J. Andrew Grant, Barry T. Pickup, and Anthony Nicholls. A smooth permittivity function for Poisson-Boltzmann solvation methods. *J. Comp. Chem.*, 22(6):608–640, April 2001.

[37] T Grycuk. Deficiency of the Coulomb-field approximation in the generalized Born model: An improved formula for Born radii evaluation. *J. Chem. Phys.*, 119(9):4817–4826, Sep 2003.

[38] Urs Haberthür and Amedeo Caflisch. Facts: Fast analytical continuum treatment of solvation. *J. Comput. Chem.*, 29:701–715, October 2007.

[39] G D Hawkins, C J Cramer, and D G Truhlar. Pairwise Solute Descreening of Solute Charges From a Dielectric Medium. *Chem. Phys. Lett.*, 246(1-2):122–129, Nov 1995.

[40] G D Hawkins, C J Cramer, and D G Truhlar. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.*, 100(51):19824–19839, Dec 1996.

[41] Barry Honig and Anthony Nicholls. Classical Electrostatics in Biology and Chemistry. *Science*, 268:1144–1149, 1995.

[42] Wonpil Im, Michael S Lee, and Charles L Brooks. Generalized Born model with a simple smoothing function. *J Comput Chem*, 24(14):1691–702, Nov 2003.

[43] B Jayaram, Y Liu, and D L Beveridge. A modification of the generalized Born theory for improved estimates of solvation energies and pK shifts. *J. Chem. Phys.*, 109(4):1465–1471, Jul 1998.

[44] William L. Jorgensen. The many roles of computation in drug discovery. *Science*, 303(5665):1813–1818, March 2004.

[45] Jennifer L. Knight and Charles L. Brooks. Surveying implicit solvent models for estimating small molecule absolute hydration free energies. *J. Comput. Chem.*, 32(13):2909–2923, 2011.

[46] Paul Labute. The generalized born/volume integral implicit solvent model: Estimation of the free energy of hydration using london dispersion instead of atomic surface area. *J. Comp. Chem.*, 29:1693–1698, 2008.

[47] Paul Labute. The generalized Born/volume integral implicit solvent model: Estimation of the free energy of hydration using london dispersion instead of atomic surface area. *J. Comp. Chem.*, 29(10):1693–1698, 2008.

[48] B Lee and F M Richards. Interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.*, 55(3):379, Jan 1971.

[49] M S Lee, Freddie R Salsbury, and Charles L Brooks. Novel generalized Born methods. *J. Chem. Phys.*, 116(24):10606–10614, Jun 2002.

[50] Mathew C Lee and Yong Duan. Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized Born solvent model. *Proteins*, 55(3):620–634, May 2004.

[51] Michael S Lee, Michael Feig, Freddie R Salsbury, and Charles L Brooks. New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem.*, 24(11):1348–1356, Aug 2003.

[52] Michael S. Lee and Mark A. Olson. Evaluation of poisson solvation models using a hybrid explicit/implicit solvent method. *J. Phys. Chem. B*, 109(11):5223–5236, March 2005.

[53] Antonio Llinàs, Robert C. Glen, and Jonathan M. Goodman. Solubility challenge: Can you predict solubilities of 32 molecules using a database of 100 reliable measurements? *J. Chem. Inf. Model.*, 48(7):1289–1303, July 2008.

[54] R. Luo, L. David, and M. K. Gilson. Accelerated Poisson-Boltzmann calculations for static and dynamic systems. *J. Comp. Chem.*, 23:1244–1253, 2002.

[55] J D Madura, J M Briggs, R C Wade, M E Davis, B A Luty, A Ilin, Jan M Antosiewicz, Michael K Gilson, B Bagheri, L R Scott, and J Andrew McCammon. Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. *Comput. Phys. Commun.*, 91(1-3):57–95, Jan 1995.

[56] Jeffrey D. Madura, Malcolm E. Davis, Michael K. Gilson, Rebecca C. Wade, Brock A. Luty, and J. Andrew McCammon. Biological Applications of Electrostatic Calculations and Brownian Dynamics. *Rev. Comp. Chem.*, 5:229–267, 1994.

[57] David L. Mobley, Christopher I. Bayly, Matthew D. Cooper, Michael R. Shirts, and Ken A. Dill. Small molecule hydration free energies in explicit solvent: An extensive test of fixed-charge atomistic simulations. *J. Chem. Theory Comput.*, 5(2):350–358, February 2009.

[58] David L. Mobley and Ken A. Dill. Binding of Small-Molecule ligands to proteins: what you see is not always what you get. *Structure*, 17(4):489–498, April 2009.

[59] David L. Mobley, Ken A. Dill, and John D. Chodera. Treating entropy and conformational changes in implicit solvent simulations of small molecules. *J. Phys. Chem. B*, 112(3):938–946, January 2008.

[60] J. Mongan, C. Simmerling, J. A. Mccammon, D. A. Case, and A. Onufriev. Generalized born model with a simple, robust molecular volume correction. *J. Chem. Theory Comput.*, 3(1):156–169, January 2007.

[61] John Mongan, Andreas Svrcek-Seiler, and Alexey Onufriev. Analysis of integral expressions for effective born radii. *J. Chem. Phys.*, 127(18):185101–185101, 2007.

[62] J A Nelder and R Mead. A simplex method for function minimization. *Comput. J.*, 7:308–315, 1965.

[63] Anthony Nicholls and Barry Honig. A Rapid Finite Difference Algorithm, Utilizing Successive Over Relaxation to solve the Poisson-Botzmann Equation. *J. Comp. Chem.*, 12:435–445, 1991.

[64] Anthony Nicholls, David L. Mobley, J. Peter Guthrie, John D. Chodera, Christopher I. Bayly, Matthew D. Cooper, and Vijay S. Pande. Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. *J. Med. Chem.*, 51(4):769–779, February 2008.

[65] Anthony Nicholls, Stanislaw Wlodek, and J. Grant. SAMPL2 and continuum modeling. *J. Comput. Aided Mol. Des.*, 24(4):293–306, April 2010.

[66] Anthony Nicholls, Stanislaw Wlodek, and J. Andrew Grant. The SAMP1 solvation challenge: Further lessons regarding the pitfalls of parametrization†. *J. Phys. Chem. B*, 113(14):4521–4532, April 2009.

[67] H Nymeyer and A E Garcia. Simulation of the folding equilibrium of alpha-helical peptides: A comparison of the generalized Born approximation with explicit solvent. *Proc. Natl. Acad. Sci. U.S.A.*, 100(24):13934–13939, Nov 2003.

[68] A. Onufriev, D. Bashford, and D. A. Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins*, 55(2):383–394, May 2004.

[69] A. Onufriev, D. A. Case, and D. Bashford. Effective born radii in the generalized born approximation: the importance of being perfect. *J Comput Chem*, 23(14):1297–1304, November 2002.

[70] Alexey Onufriev. *Continuum Electrostatics Solvent Modeling with the Generalized Born Model*, pages 127–165. Wiley, USA, 1 edition, 2010.

[71] Alexey Onufriev, Donald Bashford, and David A Case. Exploring protein native states and large-scale conformational changes with a modified generalized Born model. *Proteins*, 55(2):383–394, May 2004.

[72] Alexey Onufriev, David A Case, and Donald Bashford. Effective Born radii in the generalized Born approximation: the importance of being perfect. *J. Comput. Chem.*, 23(14):1297–1304, Nov 2002.

[73] Alexey Onufriev, David A Case, and Donald Bashford. Structural details, pathways, and energetics of unfolding apomyoglobin. *J. Mol. Biol.*, 325(3):555–567, Jan 2003.

[74] Sanbo Qin and Huan-Xiang Zhou. Do electrostatic interactions destabilize protein-nucleic acid binding? *Biopolymers*, 86(2):112–118, 2007.

[75] D Qiu, P S Shenkin, F P Hollinger, and W C Still. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A*, 101(16):3005–3014, Apr 1997.

[76] W. Rocchia, E. Alexov, and B. Honig. Extending the applicability of the nonlinear poisson-boltzmann equation: Multiple dielectric constants and multivalent ions. *J. Phys. Chem. B*, 105(28):6507–6514, July 2001.

[77] D. R. Roe, A. Okur, L. Wickstrom, V. Hornak, and C. Simmerling. Secondary structure bias in generalized born solvent models: Comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation. *J. Phys. Chem. B*, 111(7):1846–1857, February 2007.

[78] A N Romanov, S N Jabin, Y B Martynov, A V Sulimov, F V Grigoriev, and V B Sulimov. Surface Generalized Born Method: A Simple, Fast, and Precise Implicit Solvent Model beyond the Coulomb Approximation. *J. Phys. Chem. A*, 108(43):9323–9327, Oct 2004.

[79] M. F. Sanner, A. J. Olson, and J. C. Spehner. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–320, March 1996.

[80] M. Scarsi, J. Apostolakis, and A. Caflisch. Continuum Electrostatic Energies of Macromolecules in Aqueous Solutions. *J. Phys. Chem. A*, 101:8098–8106, 1997.

[81] M Schaefer and M Karplus. A comprehensive analytical treatment of continuum electrostatics. *J. Phys. Chem.*, 100(5):1578–1599, Feb 1996.

[82] M. R. Shirts, Mobley D. L., and S. P. Brown. *Free energy calculations in structure-based drug design*, pages 61–85. Lecture Notes in Computer Science. Cambridge University Press, Cambridge, New York USA, 1 edition, 2010.

[83] Devleena Shivakumar, Yuqing Deng, and Benoit Roux. Computations of absolute solvation free energies of small molecules using explicit and implicit solvent model. *J. Chem. Theory Comput.*, 5(4):919–930, April 2009.

[84] Grigori Sigalov, Andrew Fenley, and Alexey Onufriev. Analytical electrostatics for biomolecules: Beyond the generalized born approximation. *J. Chem. Phys.*, 124(12):124902, 2006.

[85] Grigori Sigalov, Peter Scheffel, and Alexey Onufriev. Incorporating variable dielectric environments into the generalized Born model. *J. Chem. Phys.*, 122(9):094511, Mar 2005.

[86] Carlos Simmerling, B Strockbine, and A E Roitberg. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.*, 124(38):11258–11259, Sep 2002.

[87] Doree Sitkoff, Kim A. Sharp, and Barry Honig. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.*, 98(7):1978–1988, 1994.

[88] V Z Spassov, L Yan, and S Szalma. Introducing an implicit membrane in generalized Born/solvent accessibility continuum solvent models. *J. Phys. Chem. B*, 106(34):8726–8738, Aug 2002.

[89] W C Still, A Tempczyk, R C Hawley, and T Hendrickson. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.*, 112(16):6127–6129, Aug 1990.

[90] Andreas Svrcek-Seiler. Personal communication, 2001.

[91] Wolfgang A. Svrcek-Seiler. *Force Field based Investigations on Strucutre and Dynamics of RNA Molecules.* PhD thesis, University of Vienna, Vienna, Austria, 2003.

[92] Jessica M J Swanson, John Mongan, and J Andrew McCammon. Limitations of Atom-Centered Dielectric Functions in Implicit Solvent Models. *J. Phys. Chem. B*, 109(31):14769 –14772, Aug 2005.

[93] Chunhu Tan, Yu-Hong Tan, and Ray Luo. Implicit nonpolar solvent models. *J. Phys. Chem. B*, 111(42):12263–12274, October 2007.

[94] H. Tjong and H. X. Zhou. GBr6: A parameterization-free, accurate, analytical generalized born method. *J. Phys. Chem. B*, 111(11):3055–3061, March 2007.

[95] V Tsui and David A Case. Molecular Dynamics Simulations of Nucleic Acids with a Generalized Born Solvation Model. *J. Am. Chem. Soc.*, 122(11):2489–2498, Mar 2000.

[96] Jason A. Wagoner and Nathan A. Baker. Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms. *Proc. Natl. Acad. Sci. U. S. Ȧ.*, 103(22):8331–8336, May 2006.

[97] T Wang and R C Wade. Implicit solvent models for flexible protein-protein docking by molecular dynamics simulation. *Proteins*, 50(1):158–169, Jan 2003.

[98] Martin Zacharias. Continuum solvent modeling of nonpolar solvation: improvement by separating surface area dependent cavity and dispersion contributions. *J. Phys. Chem. A*, 107(16):3000–3004, April 2003.

Appendix A

Derivative of the effective Born radii with respect to atomic positions

Let's define:

$$\tilde{R_i}^{-1} = R_i^{-1} - B, \tag{A.1}$$

The derivative of the R6 Born radius of atom i is :

$$\frac{\partial R_i^{-1}}{\partial d_{ij}} = -\frac{1}{3}\frac{\tilde{R_i}^{-1}}{\tilde{\rho}^{-3}}\left((1+\frac{\mathbf{I}_i}{\tilde{\rho}^{-3}}(1+\tanh(\frac{\mathbf{I}_i}{\tilde{\rho}^{-3}})))\frac{\partial \mathbf{I}_i^{chk}}{\partial d_{ij}} + (1+\tanh(\frac{\mathbf{I}_i}{\tilde{\rho}^{-3}}))\frac{\partial \mathbf{I}_i}{\partial d_{ij}}\right) \tag{A.2}$$

$$\frac{\partial \mathbf{I}_i^{chk}}{\partial d_{ij}} = \lambda_i \frac{\partial \mathbf{F}_6}{\partial d_{ij}}(\rho_i, \rho_j, r_{ij}) \tag{A.3}$$

$$\frac{\partial \mathbf{I}_i}{\partial d_{ij}} = \left(\frac{\partial \mathbf{F}_6}{\partial d_{ij}}(\rho_i, S_j, r_{ij}) + n_i \frac{\partial \mathbf{N}_{ij}}{\partial r_{ij}}\right)(1 + s_i V_i^3) + 3(\mathbf{I}_i^{vdw} + \mathbf{I}_i^{neck}) s_i \frac{V_i^2}{R_s^3}\frac{\partial \Theta_{ij}}{\partial r_{ij}} \tag{A.4}$$

$\frac{\mathbf{I}_i^{chk}}{\partial d_{ij}} = 0$ for atoms outside the "chunk" region. The expression of of $\mathbf{F}_6$, $\mathbf{N}_{ij}$, Θ_{ij} and their corresponding derivatives with respect to atomic distances are presented in the following sections.

A.1 Grycuk Equations

$$\mathbf{F}_6(\rho_i, \rho_j, r_{ij}) = \begin{cases} \frac{\rho_j^3}{(r_{ij}^2-\rho_j^2)^3} & (r_{ij} \geq \rho_i + \rho_j) \\ \frac{1}{16r_{ij}}\left(\frac{r_{ij}+3\rho_j}{(r_{ij}+\rho_j)^3} + \frac{3(\rho_j^2-\rho_i^2-(r_{ij}-\rho_i)^2)+2r_{ij}\rho_i}{\rho_i^4}\right) & (\rho_i + \rho_j > r_{ij} \geq |\rho_i - \rho_j|) \\ \frac{1}{\rho_i^3} + \frac{\rho_j^3}{(r_{ij}^2-\rho_j^2)^3} & (r_{ij} < |\rho_i - \rho_j|, \rho_j \geq \rho_i) \\ 0 & \text{otherwise} \end{cases} \tag{A.5}$$

$$\frac{\partial \mathbf{F}_6}{\partial r_{ij}}(\rho_i, \rho_j, r_{ij}) = \begin{cases} \frac{-6r_{ij}\rho_j^3}{(r_{ij}^2-\rho_j^2)^4} & (r_{ij} \geq \rho_i + \rho_j) \\ \frac{-3}{16r_{ij}^2}\left(\frac{r_{ij}^2+\rho_j^2+4\rho_j r_{ij}}{(r_{ij}+\rho_j)^4} + \frac{r_{ij}^2+\rho_j^2-2\rho_i^2}{\rho_i^4}\right) & (\rho_i + \rho_j > r_{ij} \geq |\rho_i - \rho_j|) \\ \frac{-6r_{ij}\rho_j^3}{(r_{ij}^2-\rho_j^2)^4} & (r_{ij} < |\rho_i - \rho_j|, \rho_j \geq \rho_i) \\ 0 & \text{otherwise} \end{cases} \tag{A.6}$$

Where $\mathbf{F}_6$ is the R6 volume integral ($\frac{3}{4\pi}\int \frac{dV}{r^6}$) over a sphere j of radius ρ_j used to compute the effective Born radius of atom i with radius ρ_i. r_{ij} is the distance between atoms i and j.

A.2 Approximation of the R6 integration over the Neck region

Lets define $D_{ij}^w = \rho_i + \rho_j + 2\rho_w$, where ρ_w is the water probe radius. The approximation of the R6 integral over a neck between atoms i and j, and its derivative are:

$$\mathbf{N}_{ij} = \begin{cases} A_{ij}(r_{ij} - B_{ij})^4(D_{ij}^w - r_{ij})^4 & (B_{ij} < r_{ij} < D_{ij}^w) \\ 0 & \text{otherwise} \end{cases} \tag{A.7}$$

$$\frac{\partial \mathbf{N}_{ij}}{\partial r_{ij}} = \begin{cases} 4A_{ij}(D_{ij}^w + B_{ij} - 2r_{ij})(r_{ij} - B_{ij})^3(D_{ij}^w - r_{ij})^3 & (B_{ij} < r_{ij} < D_{ij}^w) \\ 0 & \text{otherwise} \end{cases} \tag{A.8}$$

A.3 Volume Measure

$$\Theta_{ij} = \begin{cases} \rho_j^3\left(1 - \left(\frac{r_{ij}}{R_s}\right)^2\right)^2 & r_{ij} \leq R_s \\ 0 & \text{otherwise} \end{cases} \tag{A.9}$$

$$\frac{\partial \Theta_{ij}}{\partial r_{ij}} = \begin{cases} -4\frac{\rho_j^3}{R_s^2}\left(1-\left(\frac{r_{ij}}{R_s}\right)^2\right) r_{ij} & r_{ij} \leq R_s \\ 0 & \text{otherwise} \end{cases} \tag{A.10}$$