A Generalized Kirkwood Implicit Solvent for the Polarizable AMOEBA Protein Model

Rae A. Corrigan[1], Andrew C. Thiel[1], Jack R. Lynn[1], Thomas L. Casavant[2], Pengyu Ren[4],

Jay W. Ponder[5] and Michael J. Schnieders[1,3]

[1]Roy J. Carver Department of Biomedical Engineering, U. of Iowa, Iowa City, IA, 52242, USA

[2]Department of Electrical and Computer Engineering, U. of Iowa, Iowa City, IA 52242, USA

[3]Department of Biochemistry and Molecular Biology, U. of Iowa, Iowa City, IA, 52242, USA

[4]Department of Biomedical Engineering, U. of Texas in Austin, Austin, TX

[5]Department of Chemistry, Washington U. in St. Louis, St. Louis, MO

Corresponding Author

michael-schnieders@uiowa.edu

Abstract

Computational simulation of biomolecules can provide important insights into protein design, protein-ligand binding calculations, and ab initio biomolecular folding, among other applications. Accurate treatment of the solvent environment is essential in such applications, but use of explicit solvent can add considerable cost. Implicit treatment of solvent effects using a dielectric continuum model is an attractive alternative to explicit solvation since it is able to describe solvation effects without the inclusion of solvent degrees of freedom. Previously, we described the development and parameterization of implicit solvent models for small molecules. Here, we extend the parameterization of the generalized Kirkwood (GK) implicit solvent model for use with biomolecules described by the AMOEBA force field via the addition of interstitial space corrections to account for biomolecular geometry. These corrections include updating pairwise descreening scale factors to be element-specific and adding neck and tanh corrections to the calculation of effective radii. We then apply the AMOEBA/GK implicit solvent to a set of nine proteins and achieve an average RMSD of 2.1 Å across 500 ns simulations. Overall, the continued development of implicit solvent models will help to facilitate simulation of arbitrary biomolecules on biologically relevant timescales.

Introduction

Biomolecular simulation is a powerful tool that can be used to understand important biological processes such as biomolecular folding[1,2] and binding[3,4]. Simulations can also aide in biomolecular design[5] and provide insights into molecular interactions. Often, simulations are limited by the use of accurate but costly explicit descriptions of solvent, making it difficult to achieve biologically relevant timescales. Implicit solvent models that represent solvent effects

using a dielectric continuum provide a complementary alternative that eliminates explicit representation of solvent molecules. The total implicit solvent potential of mean force can be divided into polar (electrostatic) and non-polar terms. The polar term can be calculated numerically using Poisson-Boltzmann (PB) solvers such as the adaptive Poisson-Boltzmann solver (APBS)[6] and PyGBe[7,8]. Alternatively, the popular generalized Born (GB)[9-11] model for fixed partial charges or the generalized Kirkwood (GK)[12] model for polarizable multipoles offer efficient analytic approximations.

A foundational component of biomolecular simulations is the selection of a force field. Various GB implicit solvent models for proteins and nucleic acids have been described for fixed charge force fields. An early GB implicit solvent model developed by Hawkins, Cramer, and Truhlar[13] (HCT) presented a pairwise descreening method to calculate effective radii (Figure 1) analytically based on a van der Waals solute volume. Alternatively, the GBSW (GB simple switching)[14,15] model, implemented in CHARMM[16,17], samples atomic density around individual atoms to determine contributions to effective radii and employs a switching function to smooth the dielectric boundary. Ideally, effective radii should be computed using a molecular volume (*i.e.*, Lee-Richards[18]). This motivates the GBMV (GB molecular volume)[19] and GBMV2[20] models, also implemented in CHARMM, that leverage a close approximation of molecular volume to calculate effective radii. Further work in AMBER[21] led to a model from Onufriev, Bashford, and Case[22] (OBC) that added a molecular volume correction in the form of tanh rescaling of effective radii. This correction to effective radii calculated based on pairwise descreening of van der Waals radii helps to account for high dielectric interstitial spaces. An additional molecular volume correction was introduced by Mongan *et al.*[23] for the Coulomb field approximation (CFA) and later by Aguilar *et al.*[24] for the Grycuk[25] approach, which both use an approximate "neck" contribution to

describe the solvent-excluded space between pairs of nearby atoms. Both tanh and CFA neck corrections were implemented in AMBER within the GB-neck2 models for proteins[26] and nucleic acids[27], and were shown to increase the accuracy of effective radii. A broader description of GB models is presented in a recent review by Onufriev and Case[11].

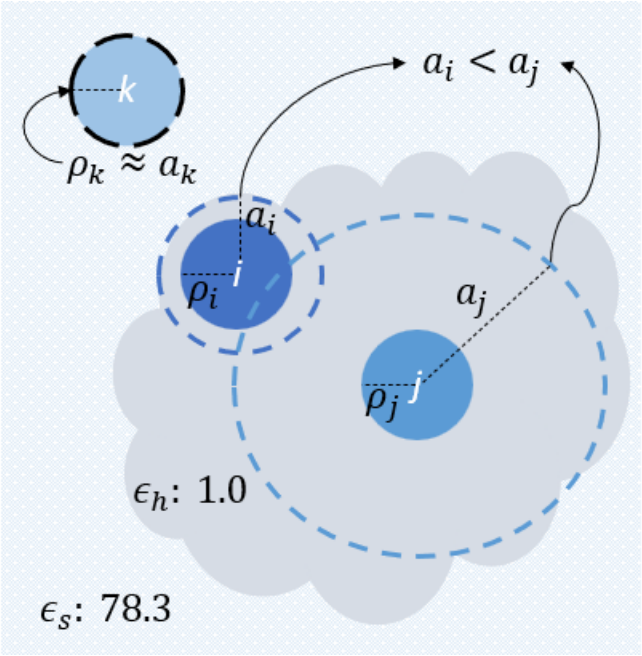

Figure 1 Pictorial representation of effective Born radii for an arbitrary globular molecule in implicit water. The effective Born radii, a_i and a_j, for atoms i and j, respectively are larger than the intrinsic radii for both atoms (ρ_i and ρ_j). Since atom j is more deeply buried within the molecule than atom i, a_j is larger than a_i. The effective Born radius for ion k is approximately equivalent to its intrinsic radius, ρ_k – the effective Born radius could be slightly larger due to descreening by a nearby molecule.

As an alternative to fixed charge force fields, polarizable force fields have been developed that in principle should provide a more transferable description of biomolecular electrostatics. Early work by Maple *et al.* combined a polarizable force field with numerical PB continuum solvation to study protein-ligand interactions[28]. Similarly, the AMOEBA force field for proteins[29] and nucleic acids[30] has been combined with PB[6,8], ddCOSMO[31], and GK[12,32] electrostatic

continuum models. More recently, the Drude oscillator polarizable force field[33-47] has been combined with PB electrostatics and used to study pK_a shifts[48,49].

Previously we described a polarizable implicit solvent model for the AMOEBA force field using Generalized Kirkwood (GK) electrostatics for small molecules[32]. This model was designed for small molecules and computed effective radii using an integral over a van der Waals volume rather than an integral over a molecular volume. Here we describe three modifications to extend the model to biomolecules. These include – using element-specific HCT overlap scale factors for pairwise descreening, the addition of a neck correction to account for interstitial space volumes between nearby atoms, and finally a *tanh* correction to account for more distal interstitial spaces. The non-polar model described in the previous work is also extended to biomolecules, which includes a cavitation term based on GaussVol[50] and a Weeks-Chandler-Anderson (WCA)[51] dispersion term. Protein simulations are presented to demonstrate the efficacy of the current implicit solvent model and future work to expand to nucleic acid and biomolecular complex simulation is discussed. This AMOEBA/GK model is implemented in both Force Field X[52] and OpenMM[53], and is being ported to Tinker[54,55].

Theory

The aqueous solvation free energy difference of a molecule (ΔG_{solv}) is the change in free energy between a molecule in vacuum and in water. To formulate an implicit solvent, ΔG_{solv} can be decomposed into three separate path dependent free energy differences[56] to give

$$\Delta G_{solv} = \Delta G_{cav} + \Delta G_{disp} + \Delta G_{elec}$$

Equation 1

where ΔG_{cav} is the unfavorable formation of a molecule-shaped cavity in water and ΔG_{disp} is the favorable addition of solute-water dispersion interactions in the previously formed cavity. Collectively, these first two terms combine to make up the non-polar portion of solvation free energy differences ($\Delta G_{non-polar} = \Delta G_{cav} + \Delta G_{disp}$). Overall, our non-polar term builds on the many advancements and insights contained in the AGBNP family of implicit solvents[57-60]. The final term ΔG_{elec} accounts for the interaction of solute charge density (*i.e.*, fixed partial atomic charges or polarizable atomic multipoles) with the continuum solvent. The implementation and parameterization of the non-polar term for the current AMOEBA implicit solvent model, as well as the implementation and parameterization of the polar term for small molecules has been described previously[32]. Here, updates to the AMOEBA GK implicit solvent model to facilitate its use with biomolecules are described.

Reference values for ΔG_{elec} in the specific case of the polarizable AMOEBA force field can be determined either by solving the Poisson-Boltzmann equation (PBE) numerically using the APBS multigrid finite-difference solver[6,61,62] or via a boundary integral approach implemented in PyGBe. While numerical solutions to the PBE can be systematically improved (*e.g.,* by using progressively finer grids or surface meshes), they are generally too expensive to be used for molecular dynamics simulations. For this reason, several approximations have been proposed, including the well-known generalized Born approximation. GB employs a summation over pairwise and self-interactions for fixed atomic partial charge force fields to yield ΔG_{elec} as

$$\Delta G_{GB} = -\frac{1}{2}\left(\frac{1}{\varepsilon_h} - \frac{1}{\varepsilon_s}\right)\sum_{i,j}\frac{q_i q_j}{f_{ij}}$$

Equation 2

where ε_s is the permittivity of solvent (78.3 for water), ε_h is the permittivity of a homogenous reference state (1.0 for vacuum), q_i and q_j are the partial charges of atoms *i* and *j*, respectively, and a commonly used form of the generalizing function *f* is given by

$$f_{ij} = \sqrt{r_{ij}^2 + a_i a_j \exp\left(-r_{ij}^2 / c a_i a_j\right)}$$

Equation 3

where r_{ij} is the atomic separation distance in Angstroms, a_i and a_j are the effective Born radii, and c controls the transition from the Born regime $(f = 1/a)$ to the screened Coulomb's law regime $(f = 1/r)$. For most GB implementations, $c = 4$, but here we treat c as a tunable parameter and fix its value to 2.455 as determined previously[12]. Generalized Kirkwood extends the GB approximation to arbitrary degree multipole moments, which facilitates the use of polarizable atomic multipole solute electrostatics. The GK monopole term $G_{GK}^{(0)}$ is equivalent to the GB charge-charge term given in Equation 2. As a further example, the GK interaction between two permanent dipoles $G_{GK}^{(1)}$ is given by

$$G_{GK}^{(1)} = \frac{1}{2}\left[\frac{1}{\varepsilon_h}\frac{2(\varepsilon_h - \varepsilon_s)}{2\varepsilon_s + \varepsilon_h}\right]\sum_{i,j} u_{i,\alpha} u_{j,\beta}\left[\frac{3 r_\alpha r_\beta (1 - f_{ij})}{f_{ij}^5} + \frac{\delta_{\alpha\beta}}{f_{ij}^3}\right]$$

Equation 4

where $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ are permanent dipole moment vectors, the α and β subscripts denote use of the Einstein summation convention, and $\delta_{\alpha\beta}$ is the Kronecker delta. Higher-order GK interaction tensors have been described previously[12] and can be generated to a desired order using a tensor recursion[63] for two Cartesian multipoles in the global frame or after rotation into their quasi-internal (QI) frame[64]. For the interaction between two multipoles truncated at quadrupole order,

use of the QI frame is ~30% faster for computing the pairwise GK energy, force and torque (despite the cost of rotating both multipoles from the global frame into the QI frame and the cost of rotating both the forces and torques back into the global frame). Reference GK tensor recursion code is available in the "multipole" package of Force Field X (https://ffx.biochem.uiowa.edu)[52].

For GB or GK approximations to concord with numerical solutions to the Poisson equation (PE), it has been demonstrated that effective radii should approach being *perfect*[65]. The reference perfect effective Born radius for an atom with a fixed partial charge (q) is defined based on its self-energy ΔG_{self} as determined using a numerical solution of the PE with all other atoms in the molecule uncharged

$$a_i = -\frac{1}{2}\left(\frac{1}{\varepsilon_h} - \frac{1}{\varepsilon_s}\right)\frac{q_i^2}{\Delta G_{self}}$$

Equation 5

Although *perfect* effective Born radii enforce that the electric potential at atomic centers match those from the numerical PE solutions, neither the electric field nor its gradient are guaranteed to be correct (*i.e.*, permanent dipole and quadrupole self-energy contributions computed using perfect effective Born radii generally deviate from their reference numerical PE values). As an alternative, the contribution of higher order atomic multipole moments to the self-energy can be included to calculate a perfect effective Kirkwood radius using the following equation[66]

$$\Delta G_{self} = \frac{1}{2}\left[\varepsilon_0 \frac{q_i^2}{a_i} + \varepsilon_1 \frac{u_{x,i}^2 + u_{y,i}^2 + u_{z,i}^2}{a_i^3} + \frac{2}{3}\varepsilon_2 \frac{\Theta_{xx,i}^2 + \Theta_{yy,i}^2 + \Theta_{zz,i}^2 + 2\left(\Theta_{xy,i}^2 + \Theta_{xz,i}^2 + \Theta_{yz,i}^2\right)}{a_i^5}\right]$$

Equation 6

where Θ is a traceless permanent quadrupole and the permittivity function ε_n for a multipole moment of order *n* is given by[67]

$$\varepsilon_n = \frac{1}{\varepsilon_h}\frac{(n+1)(\varepsilon_h - \varepsilon_s)}{(n+1)\varepsilon_s + n\varepsilon_h}$$

Equation 7

A *perfect* effective Kirkwood radius a_i can then be determined using Equation 6 and a simple numerical search. Note that the right-hand side of Equation 6 neglects polarization energy (*i.e.*, the interaction of an induced dipole at site *i* with the reaction field of its permanent dipole) and the computed ΔG_{self} is based on input of an AMOEBA permanent multipole with no induced dipole. In practice, *perfect* effective Born radii and *perfect* effective Kirkwood radii agree to within ~2% on average (see Table 1) and both represent the degree of burial of an atom within a molecule. Mean *perfect* effective Born radii and *perfect* effective Kirkwood radii for several biomolecules are shown in Table 1 and the full regression of radii for ubiquitin (PDB ID: 1UBQ) is shown in Supplementary Figure 1 as an example.

Table 1. Mean *perfect* Born radii and *perfect* Kirkwood radii for each tested molecule. On average, *perfect* Kirkwood radii are slightly smaller than *perfect* Born radii for AMOEBA permanent multipoles.

Molecule	Average *Perfect* Born Radius for All Atoms (Å)	Average *Perfect* Kirkwood Radius for all Atoms (Å)
1MIS	2.93	2.86
2JXQ	2.91	2.86
1F5G	2.93	2.88
2L8F	3.03	2.97
1ZIH	3.12	3.07
1SZY	2.94	2.88
2KOC	2.87	2.82
1D20	2.88	2.86
2HKB	2.91	2.88
1BPI	3.48	3.41
1L2Y	2.96	2.89
1UBQ	4.14	4.08
1UCS	4.00	3.92
1VII	3.12	3.05
1WM3	3.94	3.87
2OED	3.53	3.45
2PPN	4.16	4.09
7SKW	4.57	4.51
Average	3.36	3.30

The effective radius of an ion in solvent shrinks to its vdW radius ($a_i \approx \rho_i$) as it moves away from any other descreening atoms. For an atom in a molecule, the effective radius is larger than the vdW radius ($a_i > \rho_i$), particularly for a deeply buried atom in a biomolecule ($a_i \gg \rho_i$).

Element-Specific Overlap Scale Factors

For the calculation of effective radii, the GK implicit solvent model combines the analytic HCT pairwise descreening approximation[13] with the solvent field approximation (SFA) proposed by Grycuk[25]. For the calculation of effective radii, the GK implicit solvent model uses an analytic approach based on the HCT pairwise descreening approximation[13] in combination with insights from Grycuk[25]. As a part of this approximation, a unitless scale factor was set previously at 0.72 to account for the atomic overlaps that would otherwise lead to overestimated effective radii[32].

While a single scale factor worked well, it was straightforward to achieve a modest improvement in accuracy using element specific overlap scale factors. The current implementation of the GK pairwise descreening term with element specific scale factors is given in by

$$\boldsymbol{I}^{vdw}(r_{ij}) = \frac{\pi}{12}\left(\frac{3\left(r_{ij}^2 - \left(S_{HCT,j} * \rho_j\right)^2\right) + 6u^2 - 8ur_{ij}}{u^4 r_{ij}} - \frac{3\left(r_{ij}^2 - \left(S_{HCT,j} * \rho_j\right)^2\right) + 6l^2 - 8lr_{ij}}{l^4 r_{ij}}\right)$$

Equation 8

where ρ_j is the radius of atom j used for descreening, $S_{HCT,j}$ is the element-specific scaling factor for atom j, and u and l are the upper and lower integration bounds, determined based on the overlap of atoms i and j. A detailed description of how to determine the upper and lower integration bounds can be found in the original paper[13].

Pairwise Neck Interstitial Space Correction

Descreening for the AMOEBA small molecule implicit solvent was based on using a van der Waals definition of solute volume. This approximation is not appropriate for large biomolecules where interstitial spaces become increasingly important, which motivates the more physically realistic Lee-Richards molecular volume. Specifically, use of a van der Waals volume leads to underestimation of the effective radii for biomolecules by failing to account for descreening due to interstitial spaces that are too small to accommodate water molecules. Although these interstitial spaces are too small to accommodate an explicit water molecule, they are nonetheless "filled" by continuum water leading to artificially favorable electrostatic hydration. A

more accurate descreening integral requires a correction to account for these interstitial spaces and to properly exclude continuum water. The concept of "neck" regions between pairs of nearby atoms was first described by Mongan and co-workers for the $|\boldsymbol{r}|^{-4}$ integral[23] and later refined by Aguilar and co-workers for the $|\boldsymbol{r}|^{-6}$ integral[24]. The functional form of the latter is given by

$$\mathrm{I}^{\mathrm{neck}}(r_{ij}) = \frac{4\pi}{3} S_{neck,ij} * A_{ij}(r_{ij} - B_{ij})^4(\rho_i + \rho_j + 2\rho_w - r_{ij})^4$$

Equation 9

where, ρ_i and ρ_j are the vdW radii of atoms i and j, respectively, ρ_w is the radius of water (1.4 Å) and r_{ij} is the separation distance. The A_{ij} and B_{ij} constants were originally determined using benchmark values from a numerically exact method of calculating effective Born radii called NSR6, which has been described previously[24]. In this work, values of A_{ij} and B_{ij} were calculated for an expanded set of vdW radii using benchmark *perfect* effective radii values from APBS calculations. The procedure for determining A_{ij} and B_{ij} was otherwise analogous – for pairs of atoms, APBS was used to determine the value of the neck integral at various separation distances. The separation distance at which the value of the neck integral is at a maximum ($neck^{max}$) was recorded as r_{ij}^{max} for that pair of radii. The value of B_{ij} was then calculated as $B_{ij} = 2r_{ij}^{max} - (\rho_i + \rho_j + 2\rho_w)$ and A_{ij} was calculated such that $\mathrm{I}^{\mathrm{neck}}\left(r_{ij}^{max}\right) = neck^{max}$. A slight change to the determination of A_{ij} values in this work was to include the $\frac{4\pi}{3}$ constant explicitly in the neck integral equation instead of including it in the A_{ij} values – this was done to facilitate consistency in the components of the descreening integral. A full tabulation of these updated A_{ij} and B_{ij} constants is available in Supplementary Tables S1 and S2.

For a single pair of atoms, the neck between them perfectly describes the correction from the van der Waals volume to the Lee-Richards molecule volume. For more than two atoms, however, neck regions can overlap and lead to an overestimation of the interstitial volume. A scale factor (S_{neck}) is introduced to correct for these overlaps. The S_{neck} scale factor to correct for overcounting neck regions is analogous to the HCT scale factors for overlapping atoms during pairwise descreening. Additionally, neck contributions to molecular volume are only calculated for pairs of atoms that are close enough to exclude water between them. In other words, neck regions are calculated only between atoms whose separation distance (r_{ij}) satisfies the following criterion

$$r_{ij} \leq \rho_i + \rho_j + 2\rho_w$$

Equation 10

To calculate forces (*e.g.,* for optimization or molecular dynamics) the neck correction must be differentiable with respect to separation distance. This derivative is given by

$$\frac{\partial\, \mathrm{I}^{\mathrm{neck}}\left(r_{ij}^{max}\right)}{\partial\, r_{ij}} = \frac{16\pi}{3} * (S_{neck,ij} * A_{ij}\left(r_{ij} - B_{ij}\right)^3\left(\rho_i + \rho_j + 2\rho_w - r_{ij}\right)^4 - S_{neck,i} * A_{ij}\left(r_{ij} - B_{ij}\right)^4\left(\rho_i + \rho_j + 2\rho_w - r_{ij}\right)^3)$$

Equation 11

In previous implementations of the neck correction, a single S_{neck} scaling factor was used in all cases. In this work, we propose a modification to the neck scaling factor based on the number of heavy atoms bound to a particular atom of interest. If no heavy atoms are bound to atom *i*. then $S_{neck,i} = 1.0$. For all other cases, the scaling factor for atom i is calculated based on Equation 12:

$$S_{neck,i} = S_{neck} * \frac{5.0 - n_{heavy}}{4.0}$$

Equation 12

Where n_{heavy} is the number of heavy atoms bound to the atom of interest and S_{neck} is the maximum fit scale factor. A pictorial representation of this S_{neck} scheme is shown in Figure 2.

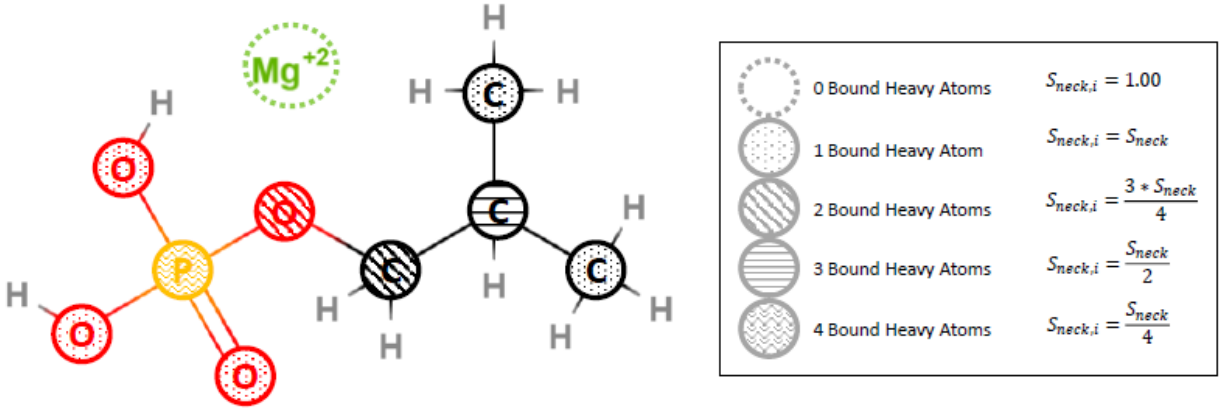

Figure 2 Diagram of bonding aware neck scaling scheme and associated equations used to calculate atomic $S_{neck,i}$ scale factors. The scaling factor is reduced proportionally to the number of heavy atoms bound to an atom of interest. If the atom of interest has no bound heavy atoms, the scale factor is set to 1.00; this is done to preserve accurate scaling for free ions in implicit solvent.

This modification to the treatment of interstitial space necks preserves accuracy for free ions (Figure 3) and results in atoms with fewer bound heavy atoms forming more necks than atoms with more bound heavy atoms. Finally, the following combining rule is used to weight the chemical environment of both atoms that form the neck

$$S_{neck,ij} = \left(S_{neck,i} + S_{neck,j}\right)/2$$

Equation 13

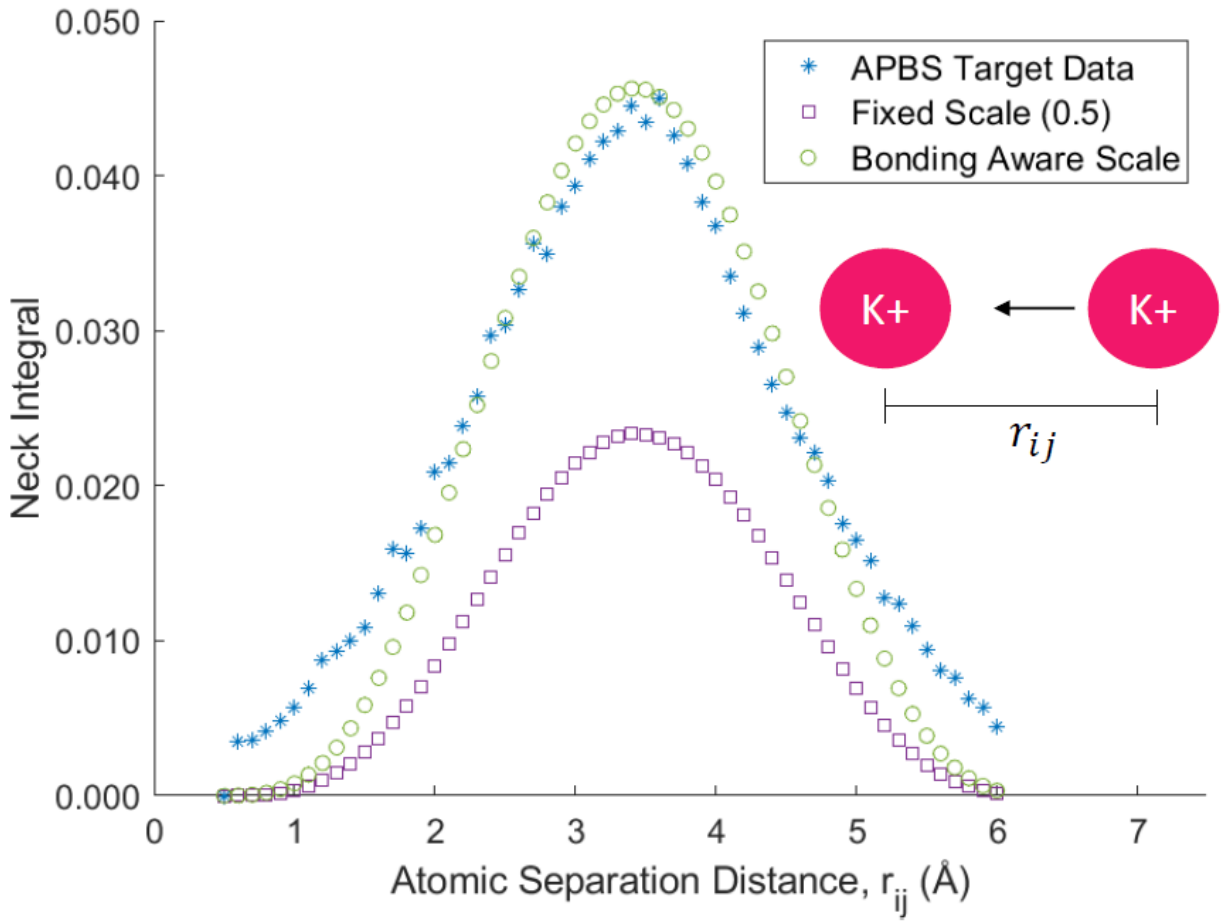

Figure 3 Demonstration of utility of the bonding aware neck scaling scheme for unbound ions in the AMOEBA GK implicit solvent. Using a single, fixed scaling factor leads to underestimation of the neck integral value at all separation distances for pairs of unbound ions while chemically aware scaling preserves accuracy for unbound ions.

Hyperbolic Tangent Interstitial Space Correction

The pairwise neck correction, described above, is helpful in accounting for short-range underestimation of molecular volume, but does not account for long-range underestimation. A version of the hyperbolic tangent (*tanh*) function is used to smoothly scale up effective Born radii, increasing the radii of deeply buried atoms more than the radii of atoms closer to the surface. This type of *tanh* rescaling function has been used previously as part of both $|\boldsymbol{r}|^{-4}$ CFA interstitial space corrections[23,65] and, more recently, $|\boldsymbol{r}|^{-6}$ SFA interstitial space corrections[24]. The *tanh*

correction used in the current implicit solvent model is a version of the previously used *tanh* function for $|\boldsymbol{r}|^{-6}$ interstitial space corrections. The maximum effective radius is capped at 30 Å and does not consider the electrostatic size of the solute[24]. The *tanh* rescaling function and associated component functions are presented below:

$$c_i = \frac{4\pi}{3}\left(\frac{1}{\rho_i^3} - \frac{1}{30^3}\right)$$

Equation 14

$$\Psi_i = \sum_{i \neq j} \mathrm{I}^{\mathrm{vdw}}(r_{ij}) + \sum_{i \neq j} \mathrm{I}^{\mathrm{neck}}(r_{ij})$$

Equation 15

$$\frac{1}{a_i} = \frac{3}{4\pi}\left(\frac{4\pi}{3}\rho_i^{-3} - c_i \tanh(\beta_0 \Psi_i \rho_i^3 - \beta_1(\Psi_i \rho_i^3)^2 + \beta_2(\Psi_i \rho_i^3)^3)\right)^{1/3}$$

Equation 16

Here a_i is the effective radius of atom i, β_0, β_1, and β_2 are tunable parameters, $\mathrm{I}^{\mathrm{vdw}}(r_{ij})$ is the $|\boldsymbol{r}|^{-6}$ integral over all van der Waals spheres in the solute (Equation 8) and $\mathrm{I}^{\mathrm{neck}}(r_{ij})$ are short-range pairwise neck contributions (Equation 9). The *tanh* correction to the effective radius and its derivative are given by

$$scale(\Psi_i) = c_i \tanh(\beta_0 \Psi_i \rho_i^3 - \beta_1(\Psi_i \rho_i^3)^2 + \beta_2(\Psi_i \rho_i^3)^3)$$

Equation 17

and

$$\frac{\partial\, scale(\Psi_i)}{\partial \Psi_{\boldsymbol{i}}} = c_i\left(\beta_0\rho_i^3 - 2\beta_1\Psi_i\rho_i^6 + 3\beta_2{\Psi_i}^2\rho_i^9\right)(1 - \tanh(\beta_0\Psi_i\rho_i^3 - \beta_1(\Psi_{\boldsymbol{i}}\rho_i^3)^2 + \beta_2(\Psi_i\rho_i^3)^3)^2)$$

Equation 18

Parameterization

Element-Specific Scale Factors

Element-specific scaling factors were determined using a limited memory BFGS optimizer and five different test molecules – two proteins, two RNA, and one DNA that were chosen from the set of biomolecules used to validate the small molecule implicit solvent models[32]. The optimizer target function is given by

$$E(\boldsymbol{P}) = W_{MUE}\sum_{i=1}^{n}\left(\Delta G_{i,self}^{PB} - \Delta G_{i,self}^{GK}\right)^2 + W_{MSE}\left(\sum_{i=1}^{n}\Delta G_{i,self}^{PB} - \sum_{i=1}^{n}\Delta G_{i,self}^{GK}\right)^2 + W_{Regularization}\sum_{i=1}^{N_{elements}}\left(S_{HCT}^{element} - 0.72\right)^2$$

Equation 19

where $W_{MUE} = 1.0$, $W_{MSE} = 10.0$, and $W_{Regularization} = 1.0E4$. Here $\Delta G_{i,self}$ is the self-energy for atom *i* calculated using either PB or GK for *n* atoms and $S_{HCT}^{element}$ is the element-specific scale factor for each element (C, N, O, P, S) where $N_{elements} = 5$. The HCT scale factors were optimized for each molecule individually and then averaged. Benchmark permanent AMOEBA electrostatic solvation energy values were calculated using APBS based on a van der Waals

definition of the solute volume. The decision to use permanent self-energy values was motivated by the expense of using APBS to compute the self-consistent reaction fields and by the relatively smaller contribution of self-polarization. Van der Waals radii were used for consistency between the PB and GK electrostatics models[32]. The benchmark self-energies, $\Delta G_{i,self}^{PB}$, used in the HCT scale factor optimizer were determined using monopoles without considering higher order multipole moments. These self-energies are consistent with *perfect* effective Born radii and promote transferability of the final $S_{HCT}^{element}$ to other force fields, including fixed charge varieties.

Starting from the initial small molecule scale factor, element-specific scaling factors were fit for C, N, O, P, and S. Due to the high degree of overlap with their bound heavy atom, the choice was made to exclude hydrogen atoms from contributing to descreening for the AMOEBA GK implicit solvent. For this reason, no scale factor was fit for hydrogen atoms (*i.e.*, the HCT scale factor for hydrogen atoms is 0). The final element-specific scaling factors ($S_{HCT}^{element}$) are shown in Table 2.

Table 2. Element-specific S_{HCT} scaling factors optimized for individual protein (1BPI and 1UCS), RNA (1MIS and 1ZIH), and DNA (1D20) molecules. Final scale factors are averages.

Element and Bondi[68] Radius (Å)	1BPI	1UCS	1MIS	1ZIH	1D20	Final
C (1.70)	0.7151	0.7294	0.6694	0.6975	0.6634	0.6950
N (1.55)	0.8348	0.7614	0.7659	0.7365	0.7377	0.7673
O (1.50)	0.7635	0.7785	0.8098	0.8048	0.8261	0.7965
P (1.80)	---	---	0.6173	0.6300	0.5878	0.6117
S (1.80)	0.7214	0.7194	---	---	---	0.7204

All $S_{HCT}^{element}$ scale factors were fit using a van der Waals description of solute volume, expanding directly on previous work with small molecules. Thus, only the *neck* and *tanh* corrections account for a molecular description of solute volume[18] for biomolecules (*i.e.,* the HCT scale factors do not implicitly account for interstitial spaces). The final scale factors agree with chemical intuition regarding heavy atom overlaps. Carbon atoms can form four bonds, often to other heavy atoms, necessitating a smaller scale factor than the base value of 0.72. Phosphorous atoms can form four bonds as well (*e.g.*, to oxygen atoms in the tested nucleic acids), and are larger than carbon atoms, which explains why phosphorous has the smallest HCT scale. Conversely, nitrogen and oxygen atoms are smaller than carbon and phosphorous, while also generally forming fewer bonds with other heavy atoms. This is consistent with fewer overlaps and explains the relative increase in those two scale factors during fitting. In the tested protein systems, sulfur atoms form either one or two bonds with heavy atoms. Due to sulfur atoms being larger than nitrogen or oxygen, but forming fewer bonds than carbon or phosphorous, an intermediate amount of overlap is expected and is consistent with the sulfur HCT scale remaining close to the base value. Final S_{HCT} scale factors were used to compute the electrostatic portion of solvation free energy differences for all 18 molecules used to validate the small molecule implicit solvent models[32]. When compared with APBS results, GK energy differences produced a slope of 1.0001 and an R-squared of 0.9971 (Figure 4).

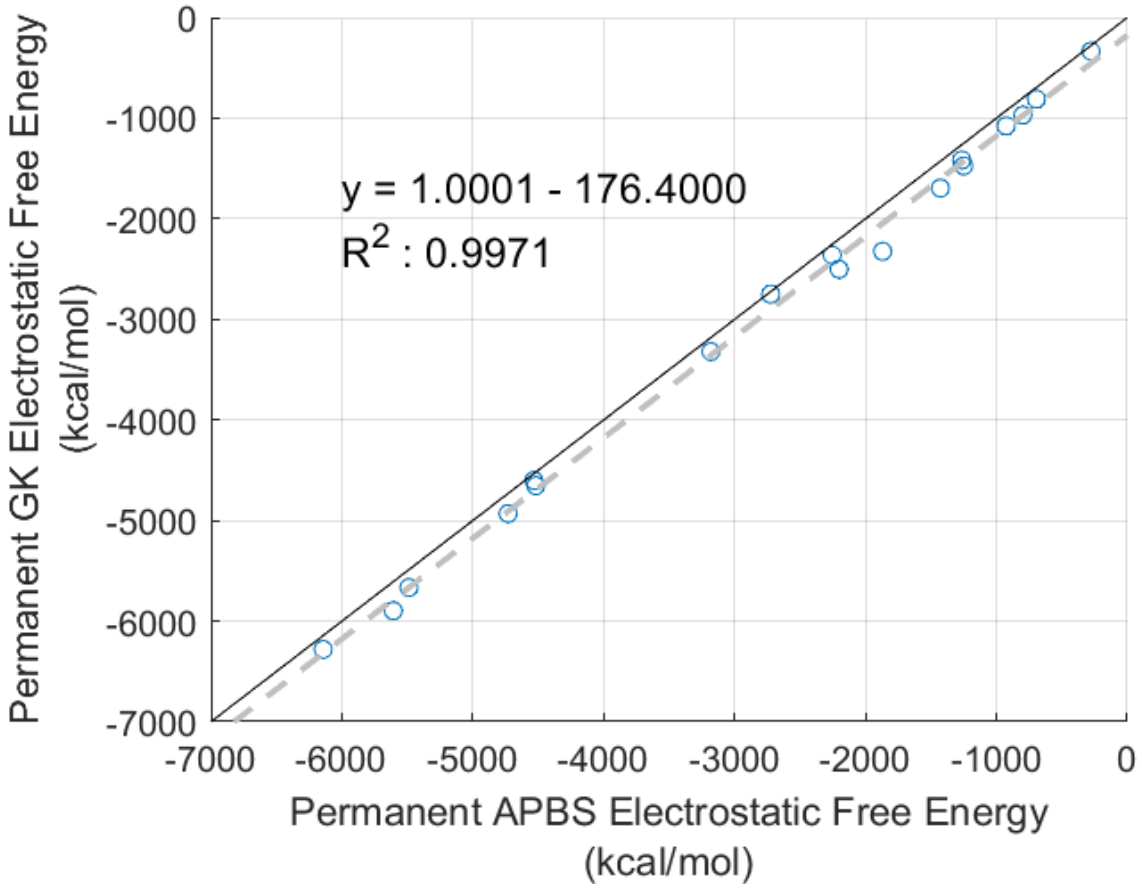

Figure 4 Comparison of permanent electrostatic energies for biomolecules calculated in APBS and GK. GK energies are calculated using van der Waals base radii to be consistent with APBS and element specific HCT scaling factors. The dashed gray line is the best fit regression line with slope = 1.0001 and $R^2 = 0.9971$. The solid black is x=y to guide the eye.

Neck and Tanh Interstitial Space Corrections

To facilitate the use of the GK implicit solvent model with mixed protein/nucleic acid simulations, a single set of *tanh* β parameters were fit for all tested biomolecules. During initial MD testing with the full biomolecule test set, proteins and nucleic acids showed distinct sensitivities to the magnitude of the neck scaling factor. The original goal of fitting a single neck scaling factor for use with both proteins and nucleic acids lead to a balancing of errors whereby folded protein electrostatic energies were too negative and folded nucleic acid electrostatic

energies were too positive (*i.e.*, compared to PB reference values). For this reason, neck scale factors for proteins and nucleic acids are expected to have different optimal values.

The $\{\beta_0, \beta_1, \beta_2\}$ parameters were initially fit simultaneously using a genetic algorithm. Each run of the genetic algorithm included 1000 generations of 500 individuals with the top 20% of individuals being carried over directly to the next generation and a mutation rate of 0.3. Permanent self-energies ($\Delta G_i^{Self,Perfect}$), permanent electrostatics energies ($\Delta G_i^{Elec,Perfect}$), and *perfect* effective Kirkwood radii ($\bar{R}_{perfect}$) were calculated using APBS and used as target data for optimization according to the following objective function

$$E(\boldsymbol{P}) = W_{MUE}\left(\sum_{i=1}^{n}\left|\Delta G_i^{Elec,GK} - \Delta G_i^{Elec,Perfect}\right|^2 + \sum_{i=1}^{n}\left|\Delta G_i^{Self,GK} - \Delta G_i^{Self,Perfect}\right|^2\right) + W_{MSE}\left(\sum_{i=1}^{n}\Delta G_i^{Elec,GK} - \sum_{i=1}^{n}\Delta G_i^{Elec,Perfect}\right)^2 + W_{Rad}\left(\sum_{i=1}^{n}\left(\bar{R}_{perfect} - \bar{R}_{GK}\right)^2\right)$$

Equation 20

where $W_{MUE} = 0.001$, $W_{MSE} = 1.0$ and $W_{Rad} = 1.0$. The parameters for each new (non-mutant) individual were selected randomly from uniform distributions across the following ranges: $\beta_0 \in \{0.5000, 1.5000\}$, $\beta_1 \in \{0.1000, 0.4000\}$, $\beta_2 \in \{0.0004, 0.2000\}$. The permanent energies and *perfect* effective Kirkwood radii used as benchmarks were from a set of nine proteins (1BPI, 1L2Y, 1UBQ, 1UCS, 1VII, 1WM3, 2OED, 2PPN, 7SKW) and nine nucleic acids (1MIS, 2JXQ, 1F5G, 2L8F, 1SZY, 1ZIH, 2KOC, 1D20, 2HKB). The fitting of *tanh* parameters is a multiple-minima problem[22,24,27] and for this reason a local optimization approach (*e.g.*, using an L-BFGS optimizer) will not explore the parameter space effectively. Instead, the genetic algorithm was used to control parameter ranges based on results from prior work[22,26,27] and trial and error.

The candidate parameter sets produced by the optimization runs were used to calculate GK permanent electrostatic energies for the biomolecule test set. Additionally, the GK effective radii were plotted against the *tanh* input (Ψ) to check the shape of the *tanh* function for given β parameter sets. Any *tanh* function that did not have a positive first derivative across the full functional range was eliminated. Output parameters from the genetic algorithm were slightly adjusted manually to improve the electrostatic energy regression for all tested biomolecules resulting in the following parameter set $\{\beta_0 = 0.9563, \beta_1 = 0.2578, \beta_2 = 0.0810\}$, which is plotted in Figure 5.

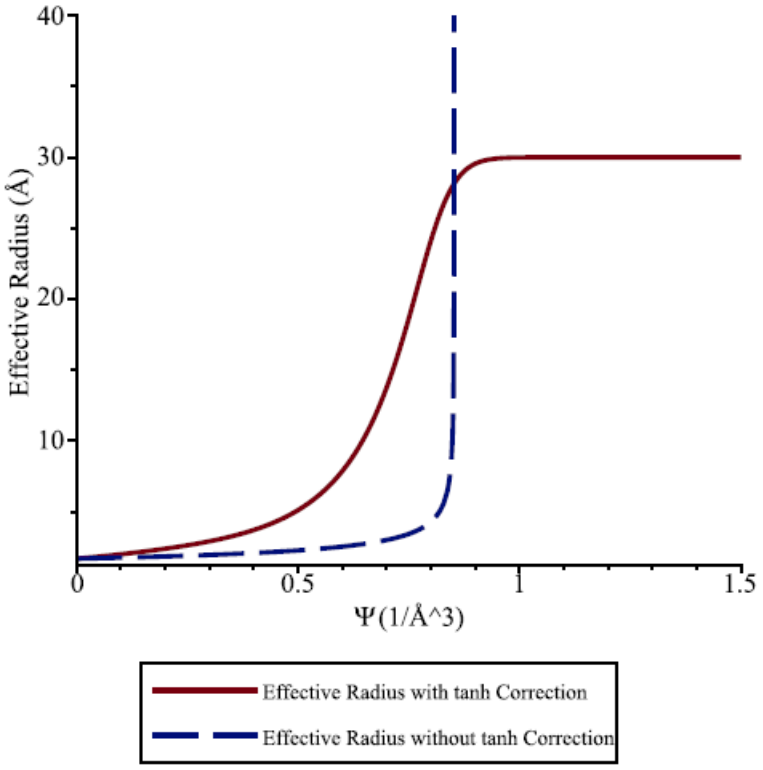

Figure 5 Effective Born radii for a molecule with a 1.7 Å base radius (carbon atom) without a *tanh* rescaling function (dashed line) and with the final fit tanh rescaling function using $\beta_0 = 0.9563$, $\beta_1 = 0.2578$, $\beta_2 = 0.0810$ (solid line) across a range of the scaled volume integral, Ψ

With the *tanh* β parameters fixed, final S_{neck} scaling factor for proteins was then determined using progressively finer S_{neck} scans. The bonding awareness scheme, described above, was used for all

scans. The final protein scale factor ($S_{neck,pr} = 0.1350$) helped improve permanent self-energies relative to only using a *tanh* correction (Figure 6). The final optimized parameters are given in Table 3.

Table 3 Final fit implicit solvent parameters. Neck scaling factor is applied in a bonding aware manner, described in Equation 9

Parameter	Value
β_0	0.9563
β_1	0.2578
β_2	0.0810
$S_{neck,protein}$	0.1350

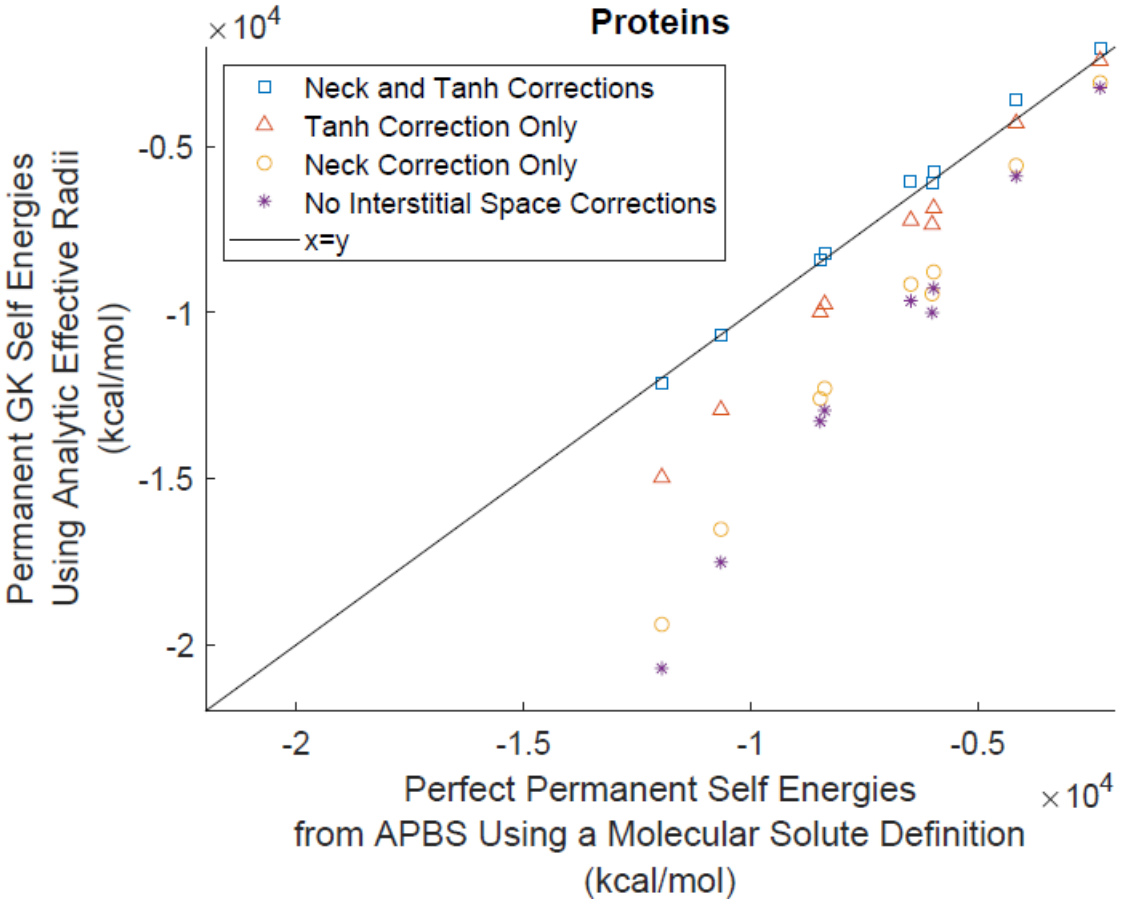

Figure 6 Comparison of permanent self energies with *perfect* effective multipole radii and GK fit radii for proteins. Energies calculated with only the tanh correction have a slope of 1.215 and R^2 of 0.996; energies calculated with the full correction (neck and tanh) have an improved slope of 1.072 and R^2 of 0.983; The solid black line is x=y to guide the eye.

Effective radii were calculated in GK using the protein parameter set and thioredoxin (2TRX), which was not used in fitting. These radii are plotted against *perfect* effective Kirkwood radii (calculated using APBS) in Figure 7.

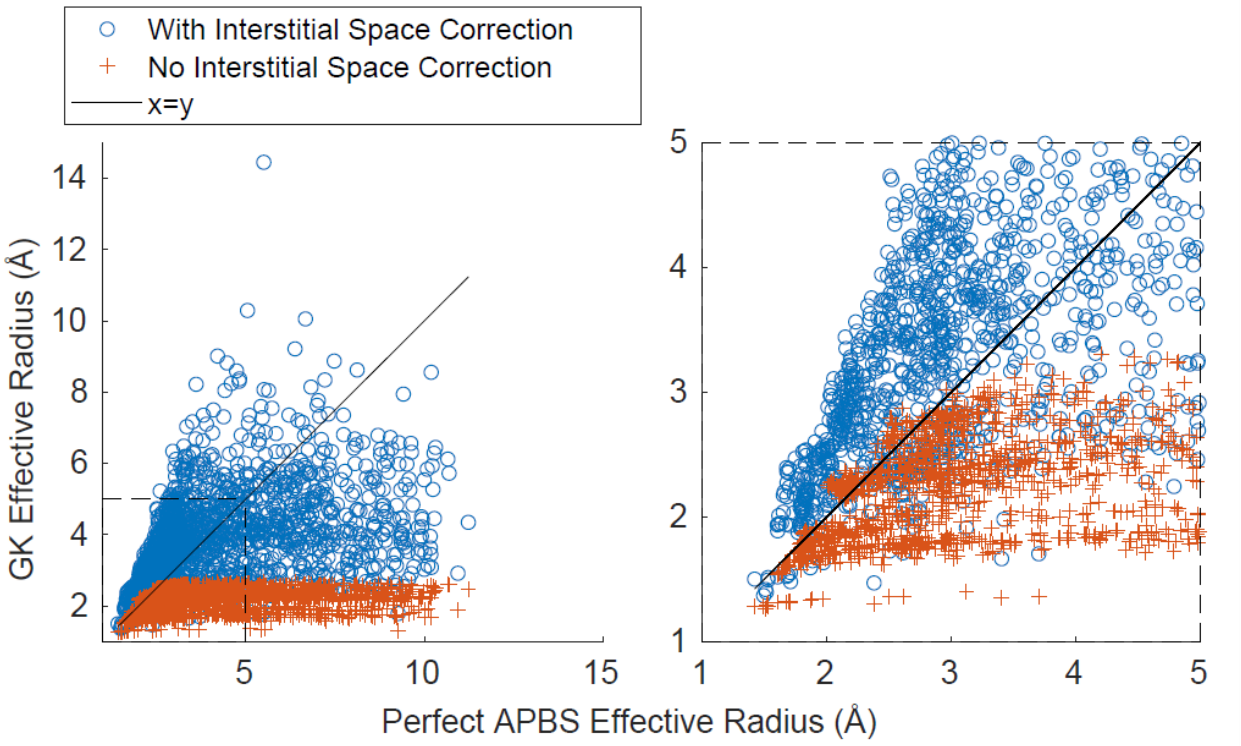

Figure 7 Comparison of 2TRX effective radii calculated in GK without interstitial space corrections (orange plusses) and with interstitial space corrections (blue circles) to *perfect* effective multipole radii calculated in APBS. Fit GK electrostatic radii are used for both series. The first plot (left) shows all effective radii for 2TRX (thioredoxin) while the second plot (right) shows the range of effective radii from 1-5 Å. The solid black line is x=y to guide the eye.

Total electrostatic hydration free energy differences calculated using APBS with a molecular surface are plotted against total electrostatic hydration free energy differences calculated using GK with the full interstitial space correction model for all test molecules in Figure 8.

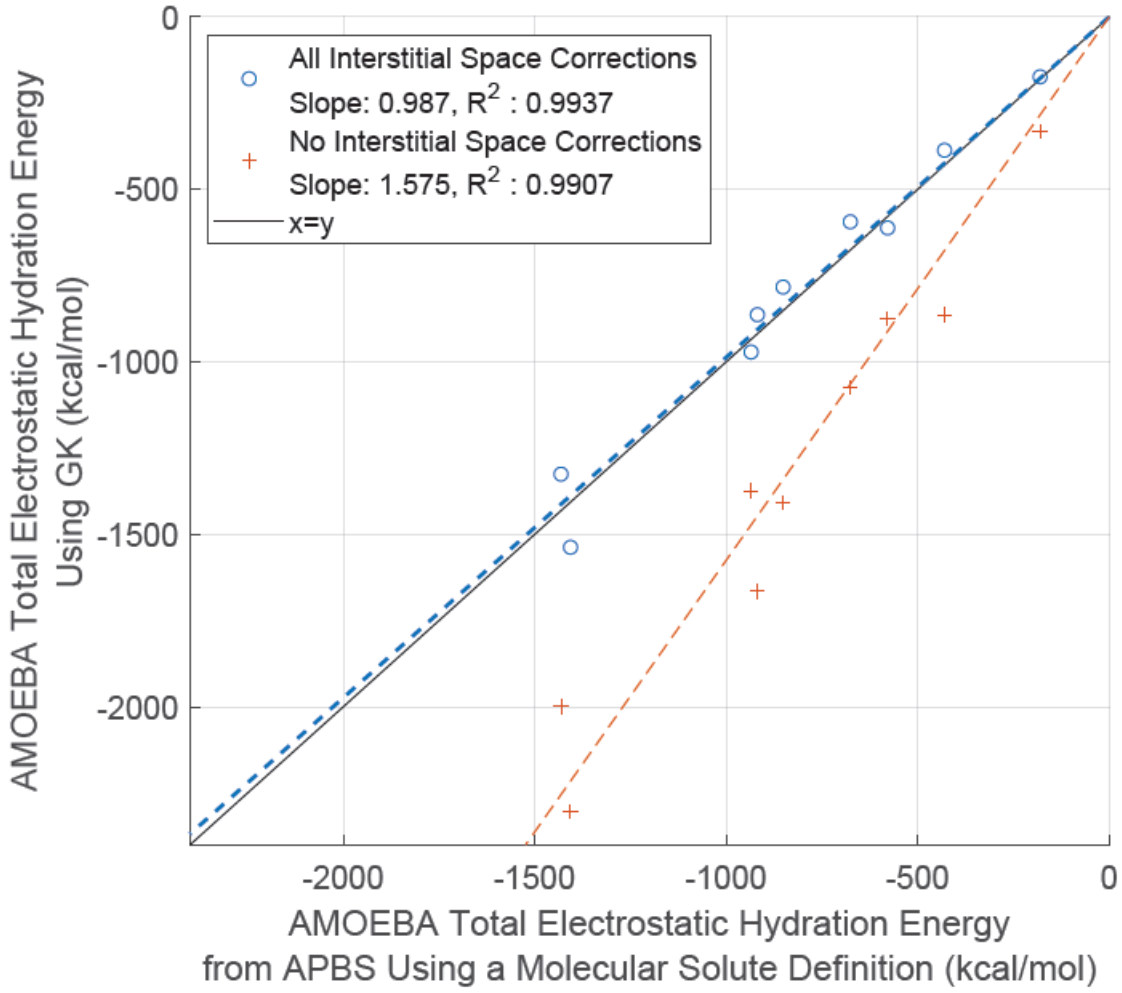

Figure 8 Total electrostatic energy for tested proteins calculated using APBS and GK. All energies were calculated using a full SCF and GK fit base radii. GK energies were calculated with (Slope: 0.987, R^2: 0.9937) and without (Slope: 1.575, R^2: 0.9907) interstitial space corrections.

Tuning Based on Molecular Dynamics Trajectories

Original base radii for the GK implicit solvent model were fit previously using small molecule solvation free energy differences[32]. During initial molecular dynamics tests on biomolecules, the simulations exhibited overcounting in the pairwise descreening integrals. This

was traced to the $1/r^6$ descreening integral being the largest for small separation distances (*i.e.*, the HCT overlap scale factors are appropriate on average, but can be too large for some overlaps at very short range). To alleviate this, a small descreening offset of 0.3 Å was added to push the beginning of the descreening integral away from the atomic center. It was also observed that repeated backbone atoms tended to favor intramolecular interactions (such as hydrogen bonding) over interactions with the GK continuum. This is in part due to the lack of hydrogen bonding within the fitting test set of small molecule solvation free energy differences. Slight alterations to selected atomic base radii for proteins helped alleviate the incorrect preference of certain groups to form intramolecular hydrogen bonds with backbone groups in place of interacting with implicit water. Radii for protein carbonyl carbon and oxygen atoms, asparagine and glutamine amide nitrogen atoms, and lysine and arginine HN atoms were modified slightly – a tabulation of the updated GK base radii is available in Supplementary Table S4.

A known limitation of the current implicit solvent model involves the use of GaussVol to determine the surface area term used in cavitation free energy calculations. GaussVol is comparable to the more general Connolly method[69,70] when calculating a van der Waals surface area, but underestimates the molecular surface area compared to a Connolly molecular surface area. However, the latter is not yet available on GPUs. For this reason, underestimation of molecular surface area due to use of GaussVol van der Waals surfaces leads to an underestimation of the cavitation free energy term, which reduces the energetic penalty of unfolding. To help correct for the underestimation, the atomic radii used for GaussVol were scaled up by 15%. This modest increase preserved simulation efficiency, while even small additional increases to 20 or 25% reduced simulation speed by almost a factor of 2 (due to nonlinear increase in GaussVol atomic overlaps as a function of atomic radii). Future work to extend the GaussVol approach to

efficiently handle molecular surface areas will benefit both fixed charge and polarizable implicit solvents.

Results and Discussion

Parameterization of the current biomolecular implicit solvent model was designed to enable simulations of both proteins and nucleic acids. Results for proteins are presented here, while those for nucleic acids will be described in a later contribution. MD simulations were performed for the set of nine proteins used to validate the small molecule implicit solvent model[32] using the GK implicit solvent with interstitial space corrections. The 7SKW structure of lysozyme was used in this work in place of the 6LYT structure used previously due to the improved resolution of 7SKW across the same lysozyme sequence. Final implicit solvent parameters (Table 3) and updated GK base radii (Supplementary Table S4) were used for all simulations. Each molecule was simulated continuously for at least 500 ns. Explicit neutralizing chloride ions (nine total ions) were added to the lysozyme simulation (7SKW) and restrained using flat-bottom potentials. These restraints help maintain the neutralizing ion cloud around the solute and keep ions from diffusing away towards entropically favored states. The restraints enforced a maximum separation distance of 45.0 Å from the center of mass of the lysozyme protein, while no minimum distance penalty was used. Explicit ions were energy-minimized to an RMS gradient of 1.0 kcal/(mol Å) and then equilibrated for 1 ns at 100K, 1 ns at 200K, and 1 ns at 300 K with the position of the biomolecule fixed before systems began the simulation protocol for all test molecules.

All test systems were first energy-minimized to an RMS gradient of 1.0 kcal/(mol Å) then equilibrated for 1 ns at 100K, 1 ns at 200K, and 1 ns at 300K. During equilibration, protein C-alpha atoms were fixed to promote relaxation of side chains before allowing the full biomolecule

to move. Production runs were 500 ns at 298.15K. A 2 fsec Langevin multiple time step integrator was used for all simulations, along with mass repartitioning from heavy atoms to bound hydrogen atoms.

Output MD trajectories were compared to base structures for all test systems. Where experimental structures consisted of NMR ensembles, the first structure in the ensemble was selected as the base structure. RMSDs for proteins are reported in Table 4. All RMSDs reported are backbone heavy atom RMSDs, which consider heavy atoms in the peptide backbone, and were calculated in FFX using the Superpose utility script. Average backbone heavy atom RMSDs for proteins were 2.74Å for all residues and 2.14Å for non-terminal residues (Table 4). Average RMSDs after 30ns of simulation are also available in Table 4 in order to more directly compare to previous protein simulations in explicit AMOEBA water[29]. The average backbone RMSD for explicit AMOEBA water simulations reported in the original paper was 1.33 Å[29], while the average backbone RMSD for implicit AMOEBA water simulations was 2.09 Å (Table 4). Part of this difference is likely due to the reduced viscosity of continuum solvent, which results in faster kinetics. Snapshots along each trajectory were clustered based on non-terminal backbone heavy atom RMSD into ten clusters and representative structures from the largest clusters are presented superposed with the base experimental structures in Figure 9. Representative structures were the minimum RMSD structure from the second half of the trajectory (250 ns and beyond) in the largest cluster. RMSD trajectories across the full 500ns simulation time are presented in Figure 10.

Table 4 Average Backbone (BB) RMSD values across 500 ns MD trajectories for protein molecules. BBRMSDs include heavy atoms in the protein backbone – values are presented in Angstroms (Å).

PDB ID	# Residues	Formal Charge	Average BBRMSD with All Residues (Å)		Average Non-Terminal BBRMSD (Å)	
			30 ns	500 ns	30 ns	500 ns
1BPI	58	6.0	1.73	1.69	1.37	1.57
1L2Y	20	1.0	1.99	2.86	1.24	2.10
1UBQ	76	1.0	2.98	3.97	1.48	1.94
1UCS	64	0.0	1.22	1.76	0.83	1.12
1VII	76	2.0	2.18	2.76	2.01	2.56
1WM3	88	1.0	1.51	3.23	1.48	3.08
2OED	56	-2.0	1.80	2.03	1.78	2.02
2PPN	107	4.0	2.89	3.47	1.65*	2.05*
7SKW	129	9.0	2.53	2.86	2.44	2.73
Average			2.09	2.74	1.59	2.13

*Also excludes flexible loop (residues 82-96)

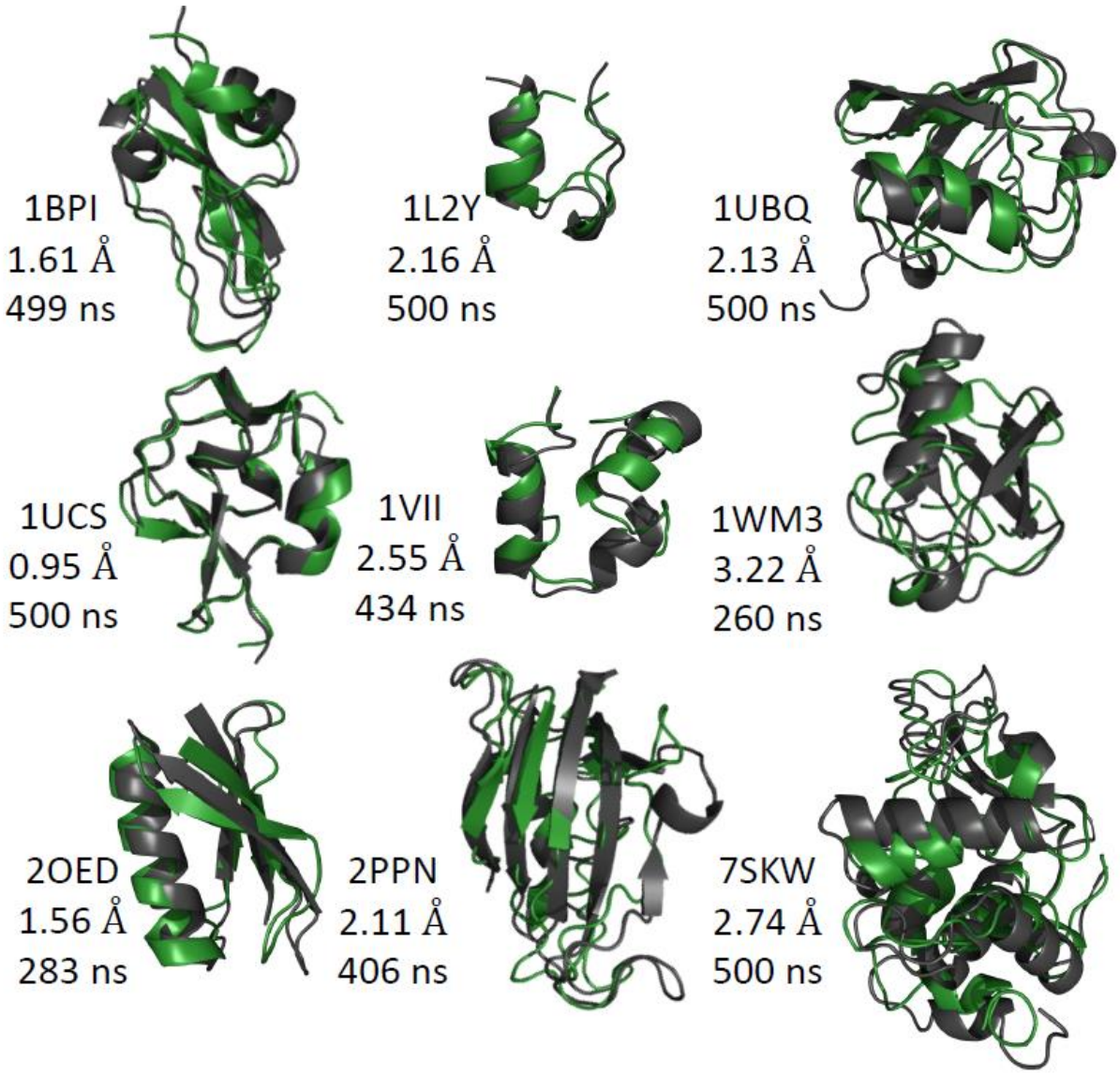

Figure 9 Superposition of the deposited X-ray crystallography or NMR structure (gray) with the lowest-RMSD structure from the largest cluster (green). The time step in the 500 ns trajectory that the representative snapshot was taken from as well as the RMSD to the deposited structure are displayed beside the structures. All representative snapshots were taken from the second half of the trajectories.

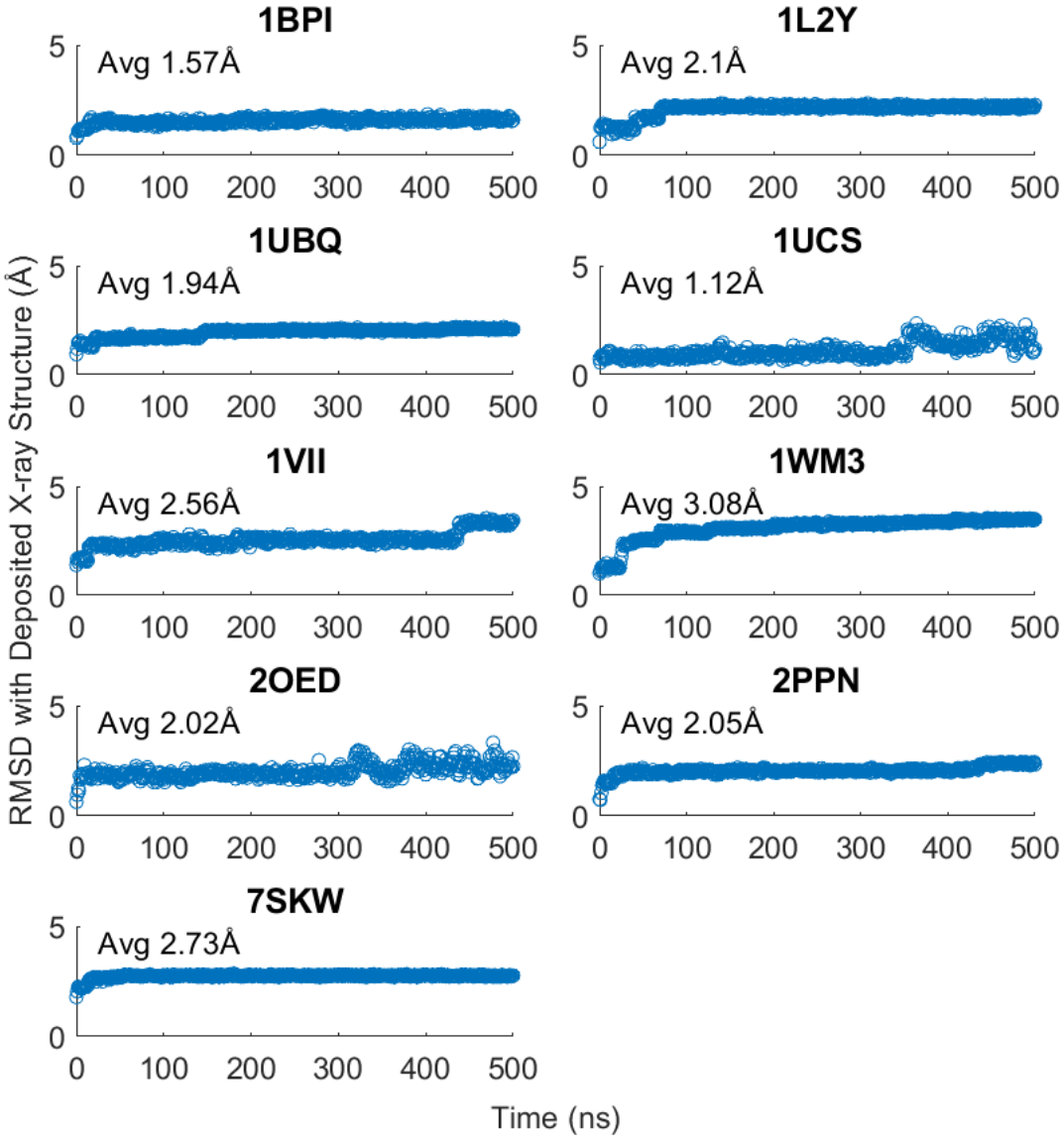

Figure 10 Protein non-terminal backbone (BB) RMSDs across 500ns production trajectories. All trajectory structures are compared to the base X-ray crystallography or NMR structure. Backbone RMSDs include heavy atoms in the peptide backbone and are all reported in Angstroms (Å).

Testing of the GBNeck2 implicit solvent model[26] included simulations of the trp-cage protein, which was also simulated here (1L2Y). A histogram of RMSD probability across two trajectories at 300K was presented, analogous to the 1L2Y histogram (top row, center) in Figure 11. GBNeck2 trajectories, reported in the original paper in supplementary Figure S9, are for 160 ns trajectories at 300K with enhanced sampling from replica exchange molecular dynamics

(REMD). The use of REMD simulations facilitates comparison to our 500 ns trajectories that do not include enhanced sampling. For both models, trp-cage RMSDs at or below 2.0Å are the most probable, though the distributions are have different features. RMSD histograms for all proteins tested in this work are shown in Figure 11. The percentage of snapshots across the 500ns trajectory that fall into each of the RMSD bins used to create the RMSD histograms is tabulated in supplementary Table S4.

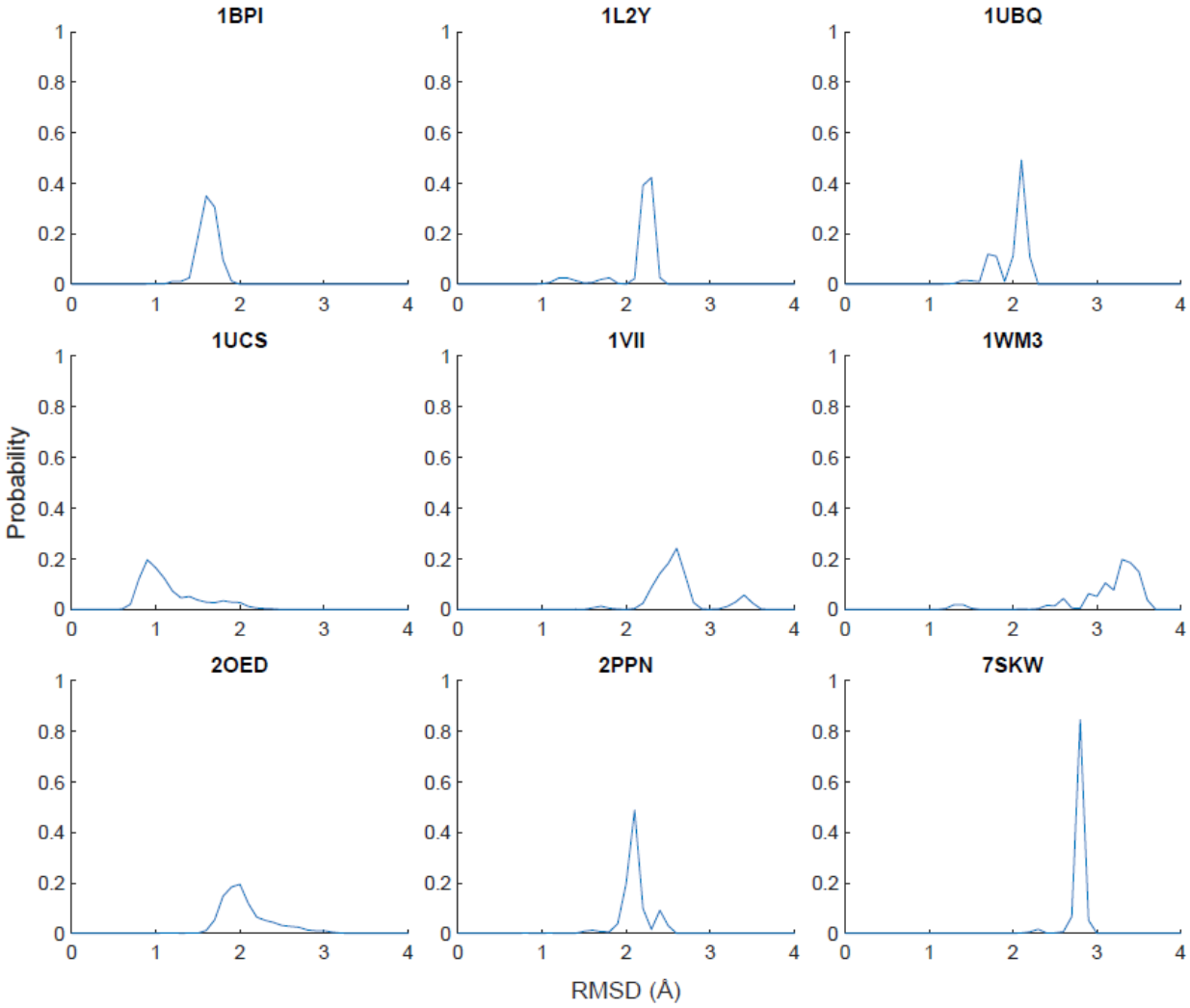

Figure 11 Histograms of RMSD values across 500 ns of simulation in GK implicit solvent for all tested proteins.

Average dipole moment magnitudes across production MD trajectories were calculated for all proteins in both vacuum and GK implicit solvation conditions. Dipole moment magnitudes in GK implicit solvent were calculated with and without interstitial space corrections. Average magnitudes in GK implicit solvent were ~30-35% larger than those in vacuum with the addition of interstitial space corrections slightly reducing average dipole moment magnitudes (Figure 12). The change between vacuum and condensed phase dipole moment magnitudes can only be captured using polarizable force fields, such as the AMOEBA and Drude models.

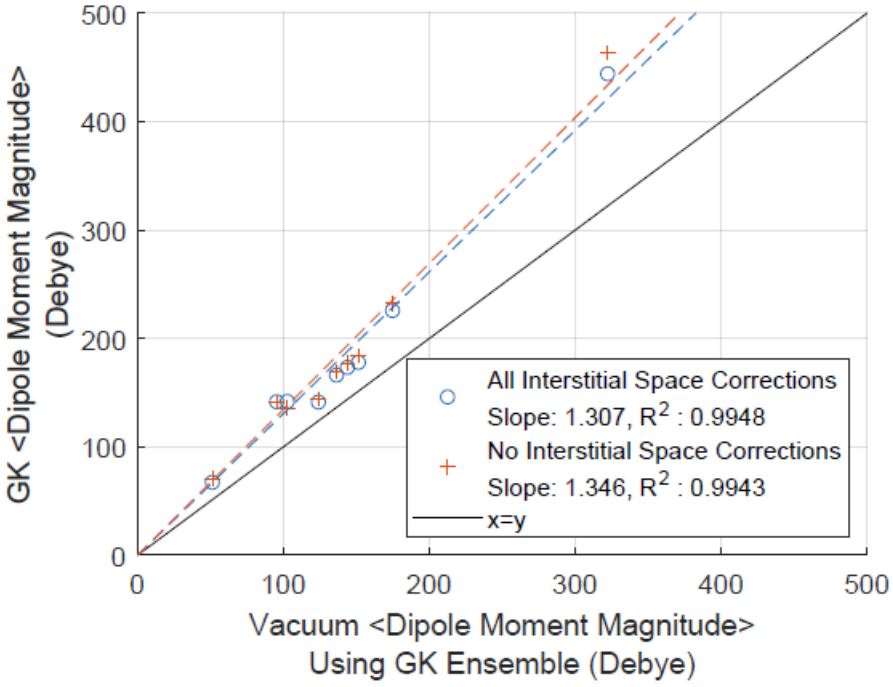

Figure 12 Comparison of average dipole moment magnitudes across production MD trajectories in vacuum and in GK implicit solvent. Implicit solvent dipole moment magnitudes are reported with (blue circles) and without (orange pluses) interstitial space corrections. Dashed lines are the best fit regression lines for GK dipole moment magnitudes, solid black line is x=y to guide the eye

Conclusions

In this work, updates to the AMOEBA GK implicit solvent to correct for interstitial spaces have been described. The updated implicit solvent model was tested with over 4 μs of MD

simulation using a varied set of proteins. Tested molecules were generally stable across 500 ns trajectories, with an average (non-terminal) backbone RMSD of 2.14 Å. The current fitting of GK radii was based on small molecule solvation free energy differences, which do not feature repetitive elements such as those in a protein or nucleic acid backbone. For this reason, additional tuning of base radii to account for repeated groups was performed. Tensor recursion formulations for GK using both Cartesian and QI frames have been made available to ease implementation in software packages such FFX[52], OpenMM[53] and Tinker[54,55]. This recursive scheme will help to facilitate $n \cdot \log(n)$ implicit solvent implementations based multipolar methods[71], including for fixed charge GB models[72].

Future work may benefit from force matching[73,74] data from biomolecular explicit solvent simulations to augment traditional fitting based on small molecule solvation free energy differences. One method of force matching parameterization was described for 16 GROMOS atom types by Kleinjung[75] and similar procedures could be used with AMOEBA atom types as an alternative method of implicit solvent model fitting. The parameterization of interstitial space correction terms was designed to ensure that the implicit solvent model can be used to simulate proteins and nucleic acids simultaneously. Fitting of nucleic acid specific parameters, including the nucleic acid neck scale factor and tuning of previously fit electrostatic radii for nucleic acid atom types to account for repetitive backbone chemistries will be addressed in future work.

The implicit solvent model for proteins is currently being used in the development of new protein optimization and design methods within FFX, including a family of side-chain optimization methods. These algorithms use a many-body energy expansion to determine optimal side chain conformations and titration states (*e.g.,* for LYS, HIS, ASP, and GLU residues) from a set of low-energy conformations known as rotamers. Typically, the polypeptide backbone remains

fixed, while the side chains are moved through their rotamers. Use of a continuum solvent is essential to eliminate steric clashes with explicit solvent molecules as the energy of each rotamer (or pair of rotamers) is computed. This approach can be used in conjunction with new techniques for experimental structure determination (*e.g.*, CryoEM, time resolved X-ray crystallography), which are rapidly increasing the number of biomolecular structures in the Protein Data Bank. However, experimental resolution is rarely high enough to assign titratable amino acid protons. Additionally, manual placement of side chains during model building is time-consuming and can result in energetically nonoptimal structures. Using global sidechain optimization methods built on rotamer libraries[76], it is possible to optimize both side-chain conformations and their titration states during model building and refinement.

This GK implicit solvent for proteins can also facilitate the development of constant pH molecular dynamics (CpHMD) algorithms for the AMOEBA force field. Implicit solvents have already been shown to work well with CpHMD methods using fixed charge force fields[49,77-79]. For example, GB-CpHMD[80] in the AMBER simulation package has been used to predict pK_a shifts using the Amber ff14sb. The advantage of using an implicit solvent for CpHMD simulations is two-fold. First, the number of atoms being simulated is reduced. A second advantage is that solvent relaxes instantaneously to changes in ionization state, which avoids the relatively slow kinetics associated with water reorientation. This is especially apparent for enhanced sampling methods such as pH replica exchange, where different ionization states are immediately accommodated by continuum water during exchanges (*i.e.*, promoting efficient pH replica exchange rates).

An implicit solvent was recently used in conjunction with a deep learning approach to calculate the absolute binding free energy difference of a host-guest system via the DeepBAR method presented by Ding and Zhang[81]. With the implicit solvent model described here, the

DeepBAR approach could now be applied to the series of host-guest systems modeled successfully by the AMOEBA polarizable force field in the context of the SAMPL challenges [82,83]. Additional applications that stand to benefit from implicit solvation, such as the simulation of protein/nucleic acid complexes[27] or intrinsically disordered proteins[84,85], may be explored in future work. While the current model is parameterized for use with the AMOEBA force field, it will also be adapted for force fields with similar electrostatics models (*e.g.,* AMOEBA+[86] and Hippo[87]) as they are developed. This will ensure transferability and continued use with more advanced force fields. Overall, the stability of the current model for a broad array of proteins with the addition of only a few new parameters shows promise for the expanded use of GK implicit solvent for biomolecular simulations.

Funding

Author RAC was supported by the NSF (National Science Foundation) Graduate Research Fellowship under Grant No. 000390183. Authors JWP and PR were supported by NIH grants R01GM114237 and R01GM106137. Author MJS was supported by NIH grant R01DC012049 and NSF grant CHE-1751688.

Author Declarations

J. Ponder and P. Ren are cofounders of Qubit Pharmaceuticals.

Data Availability

Molecular dynamics trajectories generated during this work are available upon request and can be regenerated using Force Field X (https://github.com/SchniedersLab/forcefieldx, https://ffx.biochem.uiowa.edu/)

References

1 Snow, C. D., Sorin, E. J., Rhee, Y. M. & Pande, V. S. How well can simulation predict protein folding kinetics and thermodynamics? *Annu. Rev. Biophys. Biomol. Struct.* **34**, 43-69 (2005).

2 Dill, K. A. & MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science* **338**, 1042-1046 (2012). https://doi.org:10.1126/science.1219021

3 Gallicchio, E., Lapelosa, M. & Levy, R. M. Binding Energy Distribution Analysis Method (BEDAM) for Estimation of Protein−Ligand Binding Affinities. *J. Chem. Theory Comput.* **6**, 2961-2977 (2010). https://doi.org:10.1021/ct1002913

4 Gallicchio, E. & Levy, R. M. in *Adv. Protein Chem. Struct. Biol.* Vol. 85 (ed Christo Christov) 27-80 (Academic Press, 2011).

5 Michael, E. & Simonson, T. How much can physics do for protein design? *Curr. Opin. Struct. Biol.* **72**, 46-54 (2022). https://doi.org:https://doi.org/10.1016/j.sbi.2021.07.011

6 Schnieders, M. J., Baker, N. A., Ren, P. Y. & Ponder, J. W. Polarizable atomic multipole solutes in a Poisson-Boltzmann continuum. *J. Chem. Phys.* **126**, 124114 (2007). https://doi.org:10.1063/1.2714528

7 Cooper, C. D., Bardhan, J. P. & Barba, L. A. A biomolecular electrostatics solver using Python, GPUs and boundary elements that can handle solvent-filled cavities and Stern layers. *Computer Physics Communications* **185**, 720-729 (2014). https://doi.org:10.1016/j.cpc.2013.10.028

8 Cooper, C. D. A boundary-integral approach for the poisson-boltzmann equation with polarizable force fields. *J. Comput. Chem.* **40**, 1680-1692 (2019). https://doi.org:10.1002/jcc.25820

9 Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127-6129 (1990).

10 Qiu, D., Shenkin, P. S., Hollinger, F. P. & Still, W. C. The GB/SA continuum model for solvation: a fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* **101**, 3005-3014 (1997).

11 Onufriev, A. V. & Case, D. A. in *Annual Review of Biophysics, Vol 48* Vol. 48 *Annual Review of Biophysics* (ed K. A. Dill) 275-296 (Annual Reviews, 2019).

12 Schnieders, M. J. & Ponder, J. W. Polarizable atomic multipole solutes in a generalized Kirkwood continuum. *J. Chem. Theory Comput.* **3**, 2083-2097 (2007).

13 Hawkins, G. D., Cramer, C. J. & Truhlar, D. G. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* **246**, 122-129 (1995).

14 Im, W. P., Lee, M. S. & Brooks, C. L. Generalized Born model with a simple smoothing function. *J. Comput. Chem.* **24**, 1691-1702 (2003).

15 Arthur, E. J. & Brooks, C. L. Parallelization and improvements of the Generalized Born model with a Simple sWitching function for modern graphics processors. *J. Comput. Chem.* **37**, 927-939 (2016). https://doi.org:10.1002/jcc.24280

16 MacKerell, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586-3616 (1998). https://doi.org:10.1021/jp973084f

17 Best, R. B. *et al.* Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *J. Chem. Theory Comput.* **8**, 3257-3273 (2012). https://doi.org:10.1021/ct300400x

18 Richards, F. M. Areas, volumes, packing and protein structure. *Annual review of biophysics and bioengineering* **6**, 151-176 (1977). https://doi.org:10.1146/annurev.bb.06.060177.001055

19 Lee, M. S., Salsbury, F. R. & Brooks, C. L. Novel Generalized Born methods. *J. Chem. Phys.* **116**, 10606-10614 (2002).

20 Lee, M. S., Feig, M., Salsbury, F. R. & Brooks, C. L. New analytic approximation to the standard molecular volume definition and its application to generalized born calculations. *J. Comput. Chem.* **24**, 1348-1356 (2003).

21 Hornak, V. *et al.* Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712-725 (2006). https://doi.org:10.1002/prot.21123

22 Onufriev, A., Bashford, D. & Case, D. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* **55**, 383-394 (2004). https://doi.org:10.1002/prot.20033

23 Mongan, J., Simmerling, C., McCammon, J. A., Case, D. A. & Onufriev, A. Generalized Born model with a simple, robust molecular volume correction. *J. Chem. Theory Comput.* **3**, 156-169 (2007).

24 Aguilar, B., Shadrach, R. & Onufriev, A. V. Reducing the secondary structure bias in the generalized Born model via R6 effective radii. *J. Chem. Theory Comput.* **6**, 3613-3630 (2010). https://doi.org:10.1021/ct100392h

25 Grycuk, T. Deficiency of the Coulomb-field approximation in the Generalized Born model: An improved formula for Born radii evaluation. *J. Chem. Phys.* **119**, 4817-4826 (2003).

26 Nguyen, H., Roe, D. R. & Simmerling, C. Improved generalized Born solvent model parameters for protein simulations. *J. Chem. Theory Comput.* **9**, 2020-2034 (2013). https://doi.org:10.1021/ct3010485

27 Nguyen, H., Perez, A., Bermeo, S. & Simmerling, C. Refinement of generalized Born implicit solvation parameters for nucleic acids and their complexes with proteins. *J. Chem. Theory Comput.* **11**, 3714-3728 (2015). https://doi.org:10.1021/acs.jctc.5b00271

28 Maple, J. R. *et al.* A polarizable force field and continuum solvation methodology for modeling of protein-ligand interactions. *J. Chem. Theory Comput.* **1**, 694-715 (2005).

29 Shi, Y. *et al.* Polarizable atomic multipole-based AMOEBA force field for proteins. *J. Chem. Theory Comput.* **9**, 4046-4063 (2013). https://doi.org:10.1021/ct4003702

30 Zhang, C. *et al.* AMOEBA polarizable atomic multipole force field for nucleic acids. *J. Chem. Theory Comput.* **14**, 2084-2108 (2018). https://doi.org:10.1021/acs.jctc.7b01169

31 Lipparini, F. *et al.* Polarizable Molecular Dynamics in a Polarizable Continuum Solvent. *J. Chem. Theory Comput.* **11**, 623-634 (2015). https://doi.org:10.1021/ct500998q

32 Corrigan, R. A. *et al.* Implicit Solvents for the Polarizable Atomic Multipole AMOEBA Force Field. *J. Chem. Theory Comput.* **17**, 2323-2341 (2021). https://doi.org:10.1021/acs.jctc.0c01286

33 Miertus, S., Scrocco, E. & Tomasi, J. Electrostatic interaction of a solute with a continuum - a direct utilization of abinitio molecular potentials for the prevision of solvent effects. *Chemical Physics* **55**, 117-129 (1981). https://doi.org:10.1016/0301-0104(81)85090-2

34 Cramer, C. J. & Truhlar, D. G. An SCF solvation model for the hydrophobic effect and absolute free-energies of aqueous solvation. *Science* **256**, 213-217 (1992). https://doi.org:10.1126/science.256.5054.213

35 Klamt, A. & Schuurmann, G. COSMO - a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *Journal of the Chemical Society-Perkin Transactions 2*, 799-805 (1993). https://doi.org:10.1039/p29930000799

36 Klamt, A. Conductor-like screening model for real solvents - a new approach to the quantitative calculation of solvation phenomena. *J. Phys. Chem.* **99**, 2224-2235 (1995).

37 Cances, E., Mennucci, B. & Tomasi, J. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *J. Chem. Phys.* **107**, 3032-3041 (1997).

38 Cramer, C. J. & Truhlar, D. G. Implicit solvation models: Equilibria, structure, spectra, and dynamics. *Chem. Rev.* **99**, 2161-2200 (1999).

39 Ponder, J. W. & Case, D. A. in *Adv. Protein Chem.* Vol. 66 27-85 (Academic Press, 2003).

40 Tomasi, J. Thirty years of continuum solvation chemistry: a review, and prospects for the near future. *Theor. Chem. Acc.* **112**, 184-203 (2004).

41 Tomasi, J., Mennucci, B. & Cammi, R. Quantum mechanical continuum solvation models. *Chem. Rev.* **105**, 2999-3093 (2005). https://doi.org:10.1021/cr9904009

42 Kelly, C. P., Cramer, C. J. & Truhlar, D. G. SM6: A density functional theory continuum solvation model for calculating aqueous solvation free energies of neutrals, ions, and solute-water clusters. *J. Chem. Theory Comput.* **1**, 1133-1152 (2005).

43 Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **113**, 6378-6396 (2009). https://doi.org:10.1021/jp810292n

44 Lemkul, J. A., Huang, J., Roux, B. & MacKerell, A. D. An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem. Rev.* **116**, 4983-5013 (2016). https://doi.org:10.1021/acs.chemrev.5b00505

45 Lemkul, J. A. & MacKerell, A. D. Polarizable force field for RNA based on the classical drude oscillator. *J. Comput. Chem.* **39**, 2624-2646 (2018). https://doi.org:10.1002/jcc.25709

46 Poier, P. P. & Jensen, F. Including implicit solvation in the bond capacity polarization model. *J. Chem. Phys.* **151**, 6 (2019). https://doi.org:10.1063/1.5120873

47 Poier, P. P. & Jensen, F. Polarizable charges in a generalized Born reaction potential. *J. Chem. Phys.* **153**, 10 (2020). https://doi.org:10.1063/5.0012022

48 Aleksandrov, A., Lin, F.-Y., Roux, B. & MacKerell Jr., A. D. Combining the polarizable Drude force field with a continuum electrostatic Poisson–Boltzmann implicit solvation model. **39**, 1707-1719 (2018). https://doi.org:doi:10.1002/jcc.25345

49 Aleksandrov, A., Roux, B. & MacKerell, A. D. pKa Calculations with the Polarizable Drude Force Field and Poisson–Boltzmann Solvation Model. *J. Chem. Theory Comput.* **16**, 4655-4668 (2020). https://doi.org:10.1021/acs.jctc.0c00111

50 Zhang, B., Kilburg, D., Eastman, P., Pande, V. S. & Gallicchio, E. Efficient gaussian density formulation of volume and surface areas of macromolecules on graphical processing units. *J. Comput. Chem.* **38**, 740-752 (2017). https://doi.org:10.1002/jcc.24745

51 Weeks, J. D., Chandler, D. & Andersen, H. C. Role of repulsive forces in determining the equilibrium structure of simple liquids. *The Journal of Chemical Physics* **54**, 5237-5247 (1971). https://doi.org:10.1063/1.1674820

52 Schnieders, M. J. *Force Field X, Version 1.0*, <https://ffx.biochem.uiowa.edu> (2021).

53 Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, 17 (2017). https://doi.org:10.1371/journal.pcbi.1005659

54 Rackers, J. A. *et al.* Tinker 8: Software Tools for Molecular Design. *J. Chem. Theory Comput.* **14**, 5273-5289 (2018). https://doi.org:10.1021/acs.jctc.8b00529

55 Lagardère, L. *et al.* Tinker-HP: a massively parallel molecular dynamics package for multiscale simulations of large complex systems with advanced point dipole polarizable force fields. *Chemical Science* **9**, 956-972 (2018). https://doi.org:10.1039/C7SC04531J

56 Roux, B. & Simonson, T. Implicit solvent models. *Biophys. Chem.* **78**, 1-20 (1999).

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

58 Gallicchio, E., Zhang, L. Y. & Levy, R. M. The SGB/NP hydration free energy model based on the surface Generalized Born solvent reaction field and novel nonpolar hydration free energy estimators. *J. Comput. Chem.* **23**, 517-529 (2002).

59 Gallicchio, E., Kubo, M. M. & Levy, R. M. Enthalpy-entropy and cavity decomposition of alkane hydration free energies: Numerical results and implications for theories of hydrophobic solvation. *J. Phys. Chem. B* **104**, 6271-6285 (2000).

60 Gallicchio, E., Paris, K. & Levy, R. M. The AGBNP2 Implicit Solvation Model. *J. Chem. Theory Comput.* **5**, 2544-2564 (2009). https://doi.org:10.1021/ct900234u

61 Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 10037-10041 (2001).

62 Jurrus, E. *et al.* Improvements to the APBS biomolecular solvation software suite. *Protein Sci.* **27**, 112-128 (2018). https://doi.org:10.1002/pro.3280

63 Challacombe, M., Schwegler, E. & Almlof, J. Recurrence relations for calculation of the Cartesian multipole tensor. *Chem. Phys. Lett.* **241**, 67-72 (1995).

64 Simmonett, A. C., Pickard, F. C., Schaefer, H. F. & Brooks, B. R. An efficient algorithm for multipole energies and derivatives based on spherical harmonics and extensions to particle mesh Ewald. *The Journal of Chemical Physics* **140**, 184101 (2014). https://doi.org:10.1063/1.4873920

65 Onufriev, A., Case, D. A. & Bashford, D. Effective Born radii in the Generalized Born approximation: The importance of being perfect. *J. Comput. Chem.* **23**, 1297-1304 (2002).

66 Kirkwood, J. G. Theory of solutions of molecules containing widely separated charges with special application to zwitterions. *J. Chem. Phys.* **2**, 351-361 (1934).

67 Böttcher, C. J. F. *Dielectrics in Static Fields*. 1 edn, Vol. 1 (Elsevier Pub. Co., 1952).

68 Bondi, A. Van der Waals volumes + radii *J. Phys. Chem.* **68**, 441-+ (1964). https://doi.org:10.1021/j100785a001

69 Connolly, M. Analytical molecular surface calculation. *Journal of Applied Crystallography* **16**, 548-558 (1983). https://doi.org:doi:10.1107/S0021889883010985

70 Connolly, M. L. Computation of molecular volume. *J. Am. Chem. Soc.* **107**, 1118-1124 (1985). https://doi.org:10.1021/ja00291a006

71 Barnes, J. & Hut, P. A hierarchical O(N log N) force-calculation algorithm. *Nature* **324**, 446-449 (1986).

72 Bajaj, C. & Zhao, W. Fast Molecular Solvation Energetics and Forces Computation. *SIAM Journal on Scientific Computing* **31**, 4524-4552 (2010).

73 Kleinjung, J. & Fraternali, F. Design and application of implicit solvent models in biomolecular simulations. *Current Opinion in Structural Biology* **25**, 126-134 (2014). https://doi.org:10.1016/j.sbi.2014.04.003

74 Bottaro, S., Lindorff-Larsen, K. & Best, R. B. Variational optimization of an all-atom implicit solvent force field to match explicit solvent simulation data. *J. Chem. Theory Comput.* **9**, 5641-5652 (2013). https://doi.org:10.1021/ct400730n

75 Kleinjung, J., Scott, W. R. P., Allison, J. R., van Gunsteren, W. F. & Fraternali, F. Implicit solvation parameters derived from explicit water forces in large-scale molecular dynamics simulations. *J. Chem. Theory Comput.* **8**, 2391-2403 (2012). https://doi.org:10.1021/ct200390j

76 LuCore, Stephen D. *et al*. Dead-end elimination with a polarizable force field repacks PCNA structures. *Biophys. J.* **109**, 816-826 (2015). https://doi.org:http://dx.doi.org/10.1016/j.bpj.2015.06.062

77 Khandogin, J. & Brooks, C. L. Constant pH molecular dynamics with proton tautomerism. *Biophys. J.* **89**, 141-157 (2005). https://doi.org:10.1529/biophysj.105.061341

78 Lee, M. S., Salsbury, F. R. & Brooks, C. L. Constant-pH molecular dynamics using continuous titration coordinates. *Proteins* **56**, 738-752 (2004). https://doi.org:10.1002/prot.20128

79 Baptista, A. M., Martel, P. J. & Petersen, S. B. Simulation of protein conformational freedom as a function of pH: Constant-pH molecular dynamics using implicit titration. *Proteins-Structure Function and Genetics* **27**, 523-544 (1997). https://doi.org:10.1002/(sici)1097-0134(199704)27:4<523::aid-prot6>3.3.co;2-9

80 Harris, R. C. & Pettitt, B. M. Effects of geometry and chemistry on hydrophobic solvation. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 14681-14686 (2014). https://doi.org:10.1073/pnas.1406080111

81 Ding, X. Q. & Zhang, B. DeepBAR: A fast and exact method for binding free energy computation. *J. Phys. Chem. Lett.* **12**, 2509-2515 (2021). https://doi.org:10.1021/acs.jpclett.1c00189

82 Chung, M. K. J., Miller, R. J., Novak, B., Wang, Z. & Ponder, J. W. Accurate host-guest binding free energies using the AMOEBA polarizable force field. *J. Chem Inf. Model.*, 14 https://doi.org:10.1021/acs.jcim.3c00155

83 Amezcua, M., Setiadi, J., Ge, Y. & Mobley, D. L. An overview of the SAMPL8 host–guest binding challenge. *J. Comput. Aided Mol. Des.* **36**, 707-734 (2022). https://doi.org:10.1007/s10822-022-00462-5

84 Click, T. H., Ganguly, D. & Chen, J. H. Intrinsically disordered proteins in a physics-based world. *Int. J. Mol. Sci.* **11**, 5293-5309 (2010). https://doi.org:10.3390/ijms11125292

85 Shea, J. E., Best, R. B. & Mittal, J. Physics-based computational and theoretical approaches to intrinsically disordered proteins. *Current Opinion in Structural Biology* **67**, 219-225 (2021). https://doi.org:10.1016/j.sbi.2020.12.012

86 Liu, C. W., Piquemal, J. P. & Ren, P. Y. AMOEBA plus classical potential for modeling molecular interactions. *J. Chem. Theory Comput.* **15**, 4122-4139 (2019). https://doi.org:10.1021/acs.jctc.9b00261

87 Rackers, J. A., Silva, R. R., Wang, Z. & Ponder, J. W. Polarizable water potential derived from a model electron density. *J. Chem. Theory Comput.* **17**, 7056-7084 (2021). https://doi.org:10.1021/acs.jctc.1c00628