

Three-Dimensional Molecular Theory of Solvation Coupled with Molecular Dynamics in Amber

Tyler Luchko,^{†,‡,§} Sergey Gusarov,[†] Daniel R. Roe,^{||} Carlos Simmerling,^{⊥,#}
David A. Case,^{▽,○} Jack Tuszynski,^{¶,◆} and Andriy Kovalenko*,^{†,‡}

National Institute for Nanotechnology, 11421 Saskatchewan Drive, Edmonton, Alberta T6G 2M9, Canada, Department of Physics, University of Alberta, Edmonton, Alberta T6G 2G7, Canada, Department of Mechanical Engineering, University of Alberta, Edmonton, Alberta T6G 2G8, Canada, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, Maryland 20899-8443, Department of Chemistry, Graduate Program in Biochemistry and Structural Biology, and Center for Structural Biology, Stony Brook University, Stony Brook, New York 11794-3400, Computational Science Center, Brookhaven National Laboratory, Upton, New York 11973, BioMaPS Institute and Department of Chemistry & Chemical Biology, Rutgers University, Piscataway, New Jersey 08854, and Department of Oncology, University of Alberta, Edmonton, Alberta T6G 2G8, Canada

Received August 18, 2009

Abstract: We present the three-dimensional molecular theory of solvation (also known as 3D-RISM) coupled with molecular dynamics (MD) simulation by contracting solvent degrees of freedom, accelerated by extrapolating solvent-induced forces and applying them in large multiple time steps (up to 20 fs) to enable simulation of large biomolecules. The method has been implemented in the Amber molecular modeling package and is illustrated here on alanine-dipeptide and protein-G.

1. Introduction

Molecular dynamics (MD) simulation with explicit solvent, in particular, available in the Amber molecular dynamics package,¹ yields accurate and detailed modeling of biomolecules (e.g., proteins and DNA) in solution, provided the processes to be described are within accessible time scales,

typically up to tens of nanoseconds. A major computational burden comes from the treatment of solvent molecules (usually water, sometimes cosolvent, and counterions/buffer or salt for electrolyte solutions), which typically constitute a large part of the system. Moreover, solvent enters pockets and inner cavities of the proteins through their conformational changes, which is a very slow process and nearly as difficult to model as protein folding.

Of no surprise, then, is the considerable interest in MD simulation with solvent degrees of freedom contracted by using implicit solvation approaches. In particular, of interest is the generalized Born (GB) model,² in which the solvent polarization effects are represented by a cavity in dielectric continuum (optionally, with Debye screening by the charge distribution of structureless ions in the form of the Yukawa screened potential), whereas the nonelectrostatic contributions are phenomenologically parametrized against the solvent-accessible area and excluded volume of the biomolecule. The cavity shape is formed by rolling a spherical probe, of a size to be parametrized for each solvent, over the surface of the

* Corresponding author e-mail: andriy.kovalenko@nrc-cnrc.gc.ca.
† National Institute for Nanotechnology.

‡ Department of Physics, University of Alberta.

§ Present affiliation: Department of Chemistry & Chemical Biology, and BioMaPS Institute, Rutgers University, Piscataway, NJ 08854.

|| National Institute of Standards and Technology.

▽ Stony Brook University.

○ Brookhaven National Laboratory.

△ BioMaPS Institute, Rutgers University.

◆ Department of Chemistry & Chemical Biology, Rutgers University.
◆ Department of Oncology, University of Alberta.

biomolecule. The polarization energy follows from the solution to the Poisson equation, which is computationally expensive, and is approximated in the GB model for fast calculation by algebraic expressions interpolating between the simple cases of two point charges in a spherical cavity. Conceptually transparent and computationally simple, the GB model has long been popular, including its implementations in the Amber molecular dynamics package.¹ However, it bears the fundamental drawbacks of implicit solvation methods: the energy contribution from solvation shell features such as hydrogen bonding can be parametrized but not represented in a transferable manner; the three-dimensional variations of the solvation structure, in particular, the second solvation shell, are lost; the volumetric properties of the solute are not well-defined; the nonelectrostatic solvation energy terms are empirically parametrized, and, therefore, effective interactions like hydrophobic interaction and hydrophobic attraction are not described from the first principles and thus are not transferable to new systems with complex compositions (e.g., with cosolvent and/or different buffer ions); and the entropic term is absent in continuum solvation, thus excluding from consideration the whole range of effects, such as the energy-entropy balance for the temperature control over supramolecular self-assembly in solution. To this end, the notion of a solvent-accessible surface, defined as that delineated by the center of the probe “rolled” over the surface, becomes meaningless for inner cavities of biomolecules hosting just a few solvent molecules.

An attractive alternative to continuum solvation is the three-dimensional molecular theory of solvation, also known as the 3D reference interaction site model (3D-RISM).^{3–10} Starting from an explicit solvent model, it operates with solvent distributions rather than individual molecules, but yields the solvation structure and thermodynamics from the first principles of statistical mechanics. It properly accounts for chemical specificities of both solute and solvent molecules, such as hydrogen bonding or other association and hydrophobic forces, by yielding the 3D site density distributions of solvent, similar to explicit solvent simulations. Moreover, it readily provides via analytical expressions all of the solvation thermodynamics, including the solvation free energy potential, its energetic and entropic decomposition, and partial molar volume and compressibility. The expression for the solvation free energy (and its derivatives) in terms of integrals of the correlation functions follows from a particular approximation for the so-called closure relation used to complete the integral equation for the direct and total correlation functions.¹¹ The 3D-RISM theory in the so-called hypernetted chain (HNC) closure approximation was sketched by Chandler and co-workers in their derivation of density functional theory for classical site distributions of molecular liquids.^{3,4} Beglov and Roux for the first time used the 3D-HNC closure to calculate the distribution of a monatomic Lennard-Jones (LJ) solvent in the neighborhood of solid substrates of arbitrary shape constructed from LJ centers¹² and introduced the 3D-RISM-HNC theory in the above way for polar molecules in liquid water.⁵ Kovalenko and Hirata derived the 3D-RISM integral equation from the six-dimensional molecular Ornstein–Zernike integral equation¹¹

for the solute–solvent correlation functions by averaging out the orientation degrees of freedom of solvent molecules while keeping the orientation of the solute macromolecule described at the three-dimensional level.^{6,7,10} They also developed an analytical treatment of the electrostatic long-range asymptotics of both the 3D site direct correlation functions (Coulomb tails) and the total correlation functions (screened Coulomb tails and constant shifts), including analytical corrections to the 3D site correlation functions for the periodicity of the supercell used in solving the 3D-RISM integral equation.^{8–10} This enabled 3D-RISM calculation of the solvation structure and thermodynamics of different ionic and polar macromolecules/supramolecules, for which distortion or loss of the long-range asymptotics for either of the correlation functions leads to huge errors in the 3D-RISM results for the solvation free energy (even for simple ions and ion pairs in water), while the analytical corrections/treatment of the asymptotics restores it to an accuracy of a small fraction of kcal/mol. Furthermore, Kovalenko and Hirata proposed the closure approximation (3D-KH closure) that couples the 3D-HNC treatment automatically applied to repulsive cores and other regions of density depletion due to repulsive interaction and steric constraints, and the 3D mean-spherical approximation (3D-MSA) applied to distribution peaks due to associative forces and other density enhancements, including long-range distribution tails for structural and phase transitions in fluids and mixtures.^{7,10} The 3D-KH approximation yields solutions to the 3D-RISM equations for polyionic macromolecules, solid–liquid interfaces, and fluid systems near structural and phase transitions, for which the 3D-HNC approximation is divergent and the 3D-MSA produces nonphysical areas of negative density distributions. (For the site–site OZ, or conventional RISM theory,¹¹ the corresponding radial 1D-KH version is available and capable of predicting phase and structural transitions in both simple and complex associating liquids and mixtures.¹⁰) The 3D-RISM-KH theory has been successful in analyzing a number of chemical and biological systems in solution,¹⁰ including structure of solid–liquid interfaces,⁷ structural transitions and thermodynamics of micromolecules in alcohol–water mixtures,^{13,14} structure and thermochemistry of various inorganic and (bio)organic molecules in different solvents,^{15,16} conformational equilibria, tautomerization energies, and activation barriers of chemical reactions in solution,¹⁶ solvation of carbon nanotubes,¹⁵ structure and thermodynamics of self-assembly, stability and conformational transitions of synthetic organic supramolecules (e.g., organic rosette nanotubes in different solvents)^{17–20} as well as peptides and proteins in aqueous solution,^{21–23} and molecular recognition and ligand–protein docking in solution.^{23,24} It constitutes a promising method to contract solvent degrees of freedom in MD simulation.

Miyata and Hirata²⁵ have introduced a coupling of 3D-RISM with MD in a multiple time step (MTS) algorithm, which can be formulated in terms of the RESPA^{26,27} method. It converges the 3D-RISM equations for the solvent correlations at the current snapshot of the solute conformation by using the accelerated iterative MDIIS solver, then performs several MD steps, and solves the 3D-RISM

equations over again. The MDIIS (modified direct inversion in the iterative subspace) procedure¹⁰ is a Krylov subspace type iterative solver for integral equations of liquid state theory, closely related to the DIIS approach of Pulay²⁸ for quantum chemistry equations and other similar algorithms, in particular, the GMRES solver.²⁹ The MTS approach was necessary to bring down the relatively large computational expenses of solving the 3D-RISM equations. Their implementation achieved stable simulation with the 3D-RISM equations solved at each fifth step of MD at most, which is not sufficient for realistic simulation of macromolecules and biomolecular structures of interest.

In this work, we couple the 3D-RISM solvation theory with MD in the Amber molecular dynamics package in an efficient way that includes a number of accelerating schemes. This includes several cutoffs for the interaction potentials and correlation functions, an iterative guess for the 3D-RISM solutions, and an MTS procedure with solvation forces at each MD step, which are extrapolated from the previous 3D-RISM evaluations. This coupled method makes modeling of biomolecular structures of practical interest, for example, proteins with water in inner pockets, feasible. As a preliminary illustration, we apply the method to alanine-dipeptide and protein-G in ambient water.

2. Theory and Implementation

2.1. Molecular Solvation. Solvation free energies, and their associated forces, are obtained for the solute from the 3D reference interaction site model (3D-RISM) for molecular solvation, coupled with the 3D version of the Kovalenko–Hirata (3D-KH) closure.¹⁰ 3D-RISM provides the solvent structure in the form of a 3D site distribution function, $g_{\gamma}^{\text{UV}}(\mathbf{r})$, for each solvent site, γ . With $g_{\gamma}(\mathbf{r}) \rightarrow 1$, the solvent density distribution $\rho_{\gamma}(\mathbf{r}) = \rho_{\gamma}g_{\gamma}(\mathbf{r})$ approaches the solvent bulk density ρ_{γ} . The 3D-RISM integral equation has the form:

$$h_{\gamma}^{\text{UV}}(\mathbf{r}) = \sum_{\alpha} \int d\mathbf{r}' c_{\alpha}^{\text{UV}}(\mathbf{r} - \mathbf{r}') \chi_{\alpha\gamma}^{\text{VV}}(r') \quad (1)$$

where superscripts “U” and “V” denote the solute and solvent species, respectively; $h(\mathbf{r}) = g(\mathbf{r}) - 1$ is the site–site total correlation function; $c_{\alpha}^{\text{UV}}(\mathbf{r})$ is the 3D direct correlation function for solvent site α having asymptotics of the interaction potential between the solute and solvent site: $c_{\alpha}^{\text{UV}}(\mathbf{r}) \propto -u_{\alpha}^{\text{UV}}(\mathbf{r})/(k_B T)$; and $\chi_{\alpha\gamma}^{\text{VV}}(r)$ is the site–site susceptibility of the solvent, given by

$$\chi_{\alpha\gamma}^{\text{VV}}(r) = \omega_{\alpha\gamma}^{\text{VV}}(r) + \rho_{\alpha} h_{\alpha\gamma}^{\text{VV}}(r) \quad (2)$$

Here, $\omega_{\alpha\gamma}^{\text{VV}}(r)$ is the intramolecular correlation function, representing the internal geometry of the solvent molecules, while $h_{\alpha\gamma}^{\text{VV}}(r)$ is the site–site radial total correlation function of the pure solvent calculated from the dielectrically consistent version of the 1D-RISM theory (DRISM).^{30,31} Equation 1 is complemented with the 3D-KH closure:

$$g_{\gamma}^{\text{UV}}(\mathbf{r}) = \begin{cases} \exp(d_{\gamma}^{\text{UV}}(\mathbf{r})) & \text{for } d_{\gamma}^{\text{UV}}(\mathbf{r}) \leq 0 \\ 1 + d_{\gamma}^{\text{UV}}(\mathbf{r}) & \text{for } d_{\gamma}^{\text{UV}}(\mathbf{r}) > 0 \end{cases} \quad (3)$$

where

$$d_{\gamma}^{\text{UV}}(\mathbf{r}) = -\frac{u_{\gamma}^{\text{UV}}(\mathbf{r})}{k_B T} + h_{\gamma}^{\text{UV}}(\mathbf{r}) - c_{\gamma}^{\text{UV}}(\mathbf{r})$$

and $u_{\gamma}^{\text{UV}}(\mathbf{r})$ is the 3D interaction potential of the solute acting on solvent site γ , given by the sum of the pairwise site–site potentials from all of the solute interaction sites i located at frozen positions \mathbf{R}_i :

$$u_{\gamma}^{\text{UV}}(\mathbf{r}) = \sum_i u_{i\gamma}^{\text{UV}}(|\mathbf{r} - \mathbf{R}_i|) \quad (4)$$

As with the 3D-HNC closure approximation, the 3D-RISM eq 1 with 3D-KH closure 3 possesses an exact differential of the free energy and thus has a closed analytical expression for the excess chemical potential of solvation:¹⁰

$$\Delta\mu = k_B T \sum_{\alpha} \rho_{\alpha} \int d\mathbf{r} \left\{ \frac{1}{2} (h_{\alpha}^{\text{UV}}(\mathbf{r}))^2 \Theta(-h_{\alpha}^{\text{UV}}(\mathbf{r})) - c_{\alpha}^{\text{UV}}(\mathbf{r}) - \frac{1}{2} h_{\alpha}^{\text{UV}}(\mathbf{r}) c_{\alpha}^{\text{UV}}(\mathbf{r}) \right\} \quad (5)$$

where $\Theta(x)$ is the Heaviside function, which results in $(h_{\alpha}(\mathbf{r}))^2$ being applied only in areas of site density depletion.

2.2. Analytical Solvent Forces for 3D-RISM. The solvation free energy $\Delta\mu$ is generally determined by the Kirkwood “charging” formula with thermodynamic integration over the parameter λ gradually “switching on” the solute–solvent interaction potential $\tilde{u}(r; \lambda)$ along some path from no interaction at $\lambda = 0$ to the full interaction potential $u(r)$ at $\lambda = 1$. In the case of the interaction site model, it has the form:

$$\Delta\mu = k_B T \sum_{\alpha} \rho_{\alpha} \int_0^1 d\lambda \int d\mathbf{r} g_{\alpha}^{\text{UV}}(\mathbf{r}; \lambda) \frac{\partial \tilde{u}_{\alpha}^{\text{UV}}(\mathbf{r}; \lambda)}{\partial \lambda} \quad (6)$$

The solvation free energy $\Delta\mu(\{\mathbf{R}_i\})$ dependent on protein conformation $\{\mathbf{R}_i\}$, determined by eq 6 and obtained as eq 5, is actually the potential of mean force. The expression for the mean solvent force acting on each atom i of the solute is defined as a derivative of the solvation free energy with respect to the atom coordinates \mathbf{R}_i . The mean solvent force can be obtained in the general form by differentiating the expression (6) modified in such a way that the thermodynamic integration is extended over the end point $\lambda = 1$ to the full interaction potential further changed by $du_{\alpha}^{\text{UV}}(\mathbf{r})$ due to infinitesimal shift $d\mathbf{R}_i$ of solute atom i :

$$\begin{aligned} \Delta\mu(\mathbf{R}_i + d\mathbf{R}_i) = & k_B T \sum_{\alpha} \rho_{\alpha} \left(\int_0^1 d\lambda \int d\mathbf{r} g_{\alpha}^{\text{UV}}(\mathbf{r}; \lambda) \frac{\partial \tilde{u}_{\alpha}^{\text{UV}}(\mathbf{r}; \lambda)}{\partial \lambda} + \right. \\ & \left. \int d\mathbf{r} g_{\alpha}^{\text{UV}}(\mathbf{r}; \lambda) \frac{\partial u_{\alpha}^{\text{UV}}(\mathbf{r})}{\partial \mathbf{R}_i} d\mathbf{R}_i \right) \end{aligned}$$

For the 3D site interaction potential (4), differentiation of this expression with respect to \mathbf{R}_i immediately gives the mean solvent force acting on solute site i as

$$\mathbf{f}^{\text{UV}}(\mathbf{R}_i) \equiv -\frac{\partial \Delta \mu}{\partial \mathbf{R}_i} = \sum_{\alpha} \rho_{\alpha} \int d\mathbf{r} g_{\alpha}^{\text{UV}}(\mathbf{r}) \frac{\partial u_{i\alpha}^{\text{UV}}(\mathbf{r} - \mathbf{R}_i)}{\partial \mathbf{R}_i} \quad (7)$$

where $u_{i\alpha}^{\text{UV}}(\mathbf{r} - \mathbf{R}_i)$ is the pairwise interaction potential between solute site i located at \mathbf{R}_i and solvent site γ at \mathbf{r} . It is obvious that the form (7) is valid for any closure approximation that yields the solvation free energy (at a frozen solute conformation $\{\mathbf{R}_i\}$) independent of a thermodynamic integration path, that is, that possesses an exact free energy differential. These are, in particular, the 3D-HNC and 3D-KH closures.¹⁰ The expression (7) has also been obtained, by directly differentiating a closure to the 3D-RISM equation, for the 3D-KH closure¹⁵ and for the 3D-HNC closure.^{15,25} The mean solvent force in the general form (7) still holds for any closure, subject to performing the thermodynamic integration along the path described above.

2.3. Computational Methods for Accelerating Dynamics. Modifications to the SANDER molecular dynamics module of Amber were minor. Other than calling the RISM3D subroutine, the only modifications were to add in calls for memory allocation and file input/output. A single 3D-RISM calculation is roughly 3 orders of magnitude slower than a single time step for a system solvated with the same solvent model at the same volume and density. This is not unexpected as 3D-RISM calculates the complete equilibrium distribution of solvent about the solute. To obtain meaningful sampling of solute conformations, it is necessary to reduce the computational expense of 3D-RISM calculations. To achieve this goal, three different optimization strategies were employed: (1) high-quality initial guesses to the direct correlation function were created from multiple previous solutions; (2) the pre- and postprocessing of the solute–solvent potentials, long-range asymptotics, and forces was accelerated using a cutoff scheme and minimal solvation box; and (3) direct calculation of the 3D-RISM solvation forces was avoided altogether by interpolating current force based off of atom positions from previous time steps.

2.3.1. Solution Propagation. Rapid convergence of an individual 3D-RISM calculation is facilitated by a high-quality initial guess. Given the nature of molecular dynamics simulations, it is possible to use solutions from previous time steps as the initial guess for current time step t_k . The simplest case is to use the solution from the previous time step. It is possible to improve on this by including numerically calculated derivatives:

$$c_{\alpha}^{\text{UV}}(\mathbf{r}; t_{k+1}) = c_{\alpha}^{\text{UV}}(\mathbf{r}; t_k) + (c_{\alpha}^{\text{UV}}(\mathbf{r}; t_k))' + (c_{\alpha}^{\text{UV}}(\mathbf{r}; t_k))'' + \dots \quad (8)$$

Derivatives may be calculated for each point on the grid using finite difference techniques. In this Article, we have used up to the fourth-order derivative to calculate an initial guess:

$$c_{\alpha}^{\text{UV},(k+1)} = c_{\alpha}^{\text{UV},(k)} \quad (9)$$

$$c_{\alpha}^{\text{UV},(k+1)} = 2c_{\alpha}^{\text{UV},(k)} - c_{\alpha}^{\text{UV},(k-1)} \quad (10)$$

$$c_{\alpha}^{\text{UV},(k+1)} = 3(c_{\alpha}^{\text{UV},(k)} - c_{\alpha}^{\text{UV},(k-1)}) + c_{\alpha}^{\text{UV},(k-2)} \quad (11)$$

$$c_{\alpha}^{\text{UV},(k+1)} = 4c_{\alpha}^{\text{UV},(k)} - 6c_{\alpha}^{\text{UV},(k-1)} + 4c_{\alpha}^{\text{UV},(k-2)} - c_{\alpha}^{\text{UV},(k-3)} \quad (12)$$

$$c_{\alpha}^{\text{UV},(k+1)} = 5c_{\alpha}^{\text{UV},(k)} - 10(c_{\alpha}^{\text{UV},(k-1)} - c_{\alpha}^{\text{UV},(k-2)}) - 5c_{\alpha}^{\text{UV},(k-3)} + c_{\alpha}^{\text{UV},(k-4)} \quad (13)$$

The order at which the propagation is terminated can be indicated by the number of previous solutions used, N^{cUV} .

2.3.2. Adaptive Solvation Box. The number of floating point entries that must be stored in memory for a 3D-RISM calculation is approximately where N_{FP} is the total number

$$N_{\text{FP}} = N_{\text{box}} \left[\underbrace{4}_{\text{asymptotics}} + N_{\text{solv}} \left\{ \underbrace{2N_{\text{MDIIS}}}_{c,\text{residual}} + \underbrace{2}_{g,h} \right\} \right] \quad (14)$$

of floating point entries, $N_{\text{box}} = N_x \times N_y \times N_z$ is the total number of grid points, N_{solv} is the number of solvent atom species, and N_{MDIIS} is the number of MDIIS vectors used to accelerate convergence. A full grid for g and h is required for each solvent species, and four grids are required to compute the long-range asymptotics. Memory, therefore, scales linearly with N_{box} , while computation time scales as $O(N_{\text{box}} \log(N_{\text{box}}))$ due to the requirements of calculating the 3D fast Fourier transform (3D-FFT).

For independent 3D-RISM calculations, solvation box dimensions can be selected to accommodate the particular shape of the solute. For MD, however, a solvation box of fixed size throughout the simulation must be cubic to accommodate rotations and large enough to handle changes in size and shape of the solute. Alternatively, the solvation box may be determined dynamically throughout the simulation. In this case, a linear grid spacing and minimal buffer distance between any atom of the solute and the edge of the solvent box is specified. The actual dimensions of the solvent box must satisfy the constraints of maintaining specified buffer distance and linear grid spacing. To calculate the required 3D-FFT and long-range asymptotics, each grid dimension must also be divisible by 2 and have factors of only 2, 3, or 5. Previous solutions may still be propagated by transferring the past solutions to the new grid. Past solutions are truncated or padded with zeroes as required by larger or small grid dimension.

2.3.3. Potential and Force Cutoffs. Both solute–solvent potential interactions and force calculations require interactions of every solute atom with every grid point for each solvent atom species. These calculations then scale as $O(N_{\text{box}} M^{\text{U}} M^{\text{V}})$, where M^{U} and M^{V} are the number of solute atoms and solvent species respectively. As each grid point must still be assigned a value, the use of cutoffs will not change how these calculations scale. However, computationally expensive distance-based potential calculations can be replaced with cheaper calculations outside the cutoff, reducing the computational cost by a constant factor.

As Lennard-Jones and Coulomb potentials have different long-range asymptotic behavior, the two potentials are treated

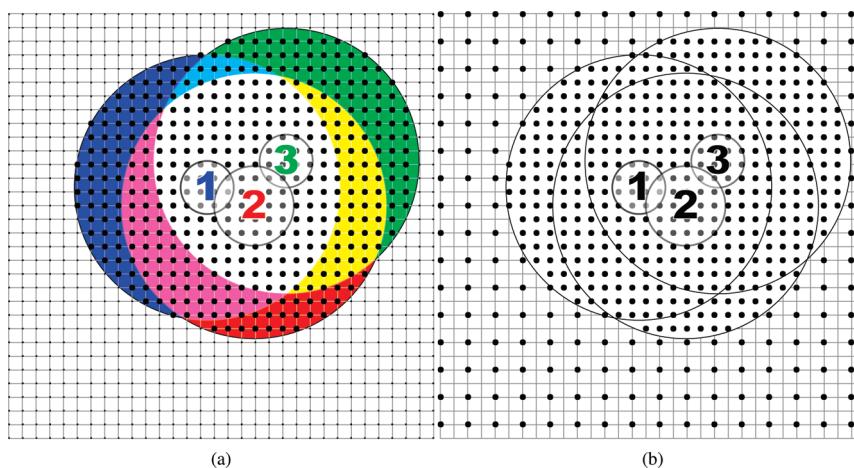


Figure 1. Cutoff schemes for grid-based (a) Lennard-Jones and (b) Coulomb potential and force calculations. Lennard-Jones calculations are performed for each solute atom only at grid sites within the cutoff distance of that atom. Grid sites within the cutoff distance of multiple solute sites take on the sum of these interactions. Coulomb interactions are calculated for every solute atom at grid sites in the union of all cutoff volumes. Grid sites outside the cutoff use explicit calculations or interpolation from surrounding values.

differently outside the cutoff radius. Lennard-Jones calculations use a hard cutoff for each solute atom. For both potential and force calculations, each solute atom only interacts with grid points within the cutoff distance, as is depicted in Figure 1a. In contrast, the long tail of the Coulomb interaction does not allow hard cutoffs to be used. Rather, within the union of the entire volume within the cutoff distance of all atoms, the entire interaction for all solute atoms is calculated at each grid point (see Figure 1b). Outside of this volume, where the interaction varies smoothly, only even grid points have the full potential calculated; that is, only one-eighth of the grid is visited. Values are then interpolated for grid points that have not been visited using a fast interpolation scheme.³²

Contributions to atomic forces from grid points outside the cutoff volume are calculated in an analogous treatment for both Lennard-Jones and Coulomb interactions. For Lennard-Jones forces on each solute atom, only the volume within the cutoff distance from that atom is included in the integration. Coulomb forces achieve the same low density sampling used in the potential calculation by doubling the integration step size outside of the cutoff volume, effectively visiting only one-eighth of the points in this region. However, for simplicity, the cutoff volume is taken as a rectangular prism rather than a sphere for Coulomb forces alone.

An alternate method for the electrostatic potential is Ewald summation,³³ which scales as $O(N_{\text{box}} \ln(N_{\text{box}})M^V)$. This scaling is generally better than the cutoff method with interpolation described as $\ln(N_{\text{box}}) < M^U$ for most systems. However, the scaling coefficients for the two methods are not equal, and the cutoff method significantly outperformed Ewald summation for systems in this study. Furthermore, Ewald summation necessarily provides a periodic potential, and a correction to this must be computed to maintain the assumption of infinite dilution,³⁴ adding to the overhead of the Ewald method. Of course, for a large enough solute, the Ewald method with periodic correction will be more efficient than the cutoff method.

2.3.4. Force Extrapolation. A variety of multiple time step (MTS) methods have been developed to limit the number of expensive force calculations required for MD. Specifically for 3D-RISM-HNC calculations, Miyata and Hirata²⁵ used RESPA MTS^{26,27} where slowly varying forces are only applied at an integer multiple of the base time step, effectively introducing large, periodic impulses to the dynamics. RESPA MTS has desirable properties, such as energy conservation; however, it is well-known that resonance artifacts limit the MTS step size to 5 fs for atomistic biomolecular simulations, after which the method becomes catastrophically unstable.^{35–37} An alternate approach, extrapolative MTS, applies a constant force over all intermediate time steps. There are no impulses in this method to cause resonance artifacts, but it does not conserve energy as the forces at intermediate time steps do not correspond to a conservative potential. LN MTS couples extrapolative MTS with Langevin dynamics to produce stable trajectories for MTS time steps up to tens or hundreds of femtoseconds, provided the forces being extrapolated are slow varying on these time scales.^{37–39} Unfortunately, the microscopic detail present in 3D-RISM calculations gives rise to forces that vary on too short a time scale to make use of LN MTS.

Inspired by LN MTS, we introduce force-coordinate extrapolation (FCE) MTS. Rather than applying a constant force, based on the last force calculations, we use previous atom configurations and forces to extrapolate what the forces should be at intermediate time steps. In this method, the forces on each of the M^U solute atoms for a current intermediate time step t_k given by the $3 \times M^U$ matrix of forces $\{\mathbf{F}\}^{(k)}$ are approximated as a linear combination of forces $\{\mathbf{F}\}^{(l)}$ at N previous time steps obtained in 3D-RISM calculations:

$$\{\mathbf{F}\}^{(k)} = \sum_{l=1}^N a_{kl} \{\mathbf{F}\}^{(l)}, l \in \text{3D-RISM steps} \quad (15)$$

The weight coefficients a_{kl} are obtained as the best representation of the arrangement of solute atoms at the current

time step k in terms of its projections onto the “basis” of N previous solute arrangements obtained from 3D-RISM, by minimizing the norm of the difference between the current $3 \times M^U$ matrix of coordinates $\{\mathbf{R}\}^{(k)}$ and the corresponding linear combination of the previous ones $\{\mathbf{R}\}^{(l)}$:

$$\text{minimize} \|\{\mathbf{R}\}^{(k)} - \sum_{l=1}^N a_l \{\mathbf{R}\}^{(l)}\|^2 \quad (16)$$

This is achieved by calculating the scalar products of the current coordinates matrix $\{\mathbf{R}\}^{(k)}$ and each basis coordinates matrix $\{\mathbf{R}\}^{(l)}$ and between all of the basis matrices:

$$P_{kl} = \sum_{i=1}^{M^U} (\mathbf{R}_i^{(k)} \cdot \mathbf{R}_i^{(l)}) \text{ and } S_{ll'} = \sum_{i=1}^{M^U} (\mathbf{R}_i^{(l)} \cdot \mathbf{R}_i^{(l')}) \quad (17)$$

where i is the solute atom index, and then solving the set of N linear equations for the weight coefficients a_{kl} :

$$\sum_{l=1}^N S_{ll'} a_{kl} = P_{kl} \quad (18)$$

Coefficients a_{kl} are then used in eq 15 to extrapolate forces at the current intermediate time step. Similarly, the known coordinates for the current time step can be approximated from previous time steps as

$$\{\mathbf{R}\}^{(k)} = \sum_{l=1}^N a_{kl} \{\mathbf{R}\}^{(l)} \quad (19)$$

These forces are approximate and do not correspond to a conservative potential; thus, MD simulations using these forces will not conserve energy. However, they provide a “smooth” transition between explicitly calculated forces. As in the LN MTS method, the resulting energy gains can be damped out with the use of Langevin dynamics to provide stable, constant temperature trajectories and enhance conformational sampling through increased efficiency.^{38,39}

With this method, one chooses a base time step, δt , and then calculates 3D-RISM at an integer number of base time steps, giving Δt between 3D-RISM calculations. Furthermore, RESPA MTS can also be applied to the intermediate, extrapolated forces, reducing the number of extrapolations required. As a concrete example, one can choose $\delta t = 2$ fs; after the specified number of previous coordinate sets with 3D-RISM forces has been calculated, extrapolated forces can be applied every 5 fs with new 3D-RISM solutions calculated every $\Delta t = 20$ fs.

Because the solvation forces on any particular solute atom typically correlate only with nearest neighbors, it is possible to use a cutoff for $\{\mathbf{R}\}$ and $\{\mathbf{F}\}$. Given the size of the systems in this Article, this was not used, although this capability is in our implementation.

2.3.5. Distributed Memory Parallelization. 3D-RISM calculations typically require large amounts of both computer time and memory. A distributed memory parallel implementation allows computation time to be decreased but also allows the aggregate memory of a distributed cluster to be utilized. The use of 3D-FFTs in calculating 3D-RISM solutions dictates that the memory model of the 3D-FFT

library must be adopted by 3D-RISM. As we use the FFTW 2.1.5 library,⁴⁰ memory decomposition is performed along the Z-axis for all 3D arrays (u^{UV} , g^{UV} , h^{UV} , c^{UV} , etc.). Communication between processes only occurs in the MDIIS, 3D-FFT routines and for the final summation of forces.

The force extrapolation method may also be parallelized. In anticipation of the use of cutoffs, coefficients for each solute atom in eq 19 are found independently. This is trivially distributed between processes.

2.4. Solvent Model. 1D- and 3D-RISM calculate the equilibrium distribution of an explicit solvent model. Two of the most popular models for water, SPC/E⁴¹ and TIP3P,⁴² do not include van der Waals terms for the hydrogens. The incomplete intramolecular correlation in RISM theory allows a catastrophic overlap between oxygen and hydrogen sites, preventing 1D-RISM from converging on a solution. The standard approach to this problem has been to apply a small Lennard-Jones potential to the hydrogen atoms:

$$U_{LJ} = 4\epsilon \left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^6 \right) = \epsilon \left(\left(\frac{\sigma^*}{r}\right)^{12} - 2\left(\frac{\sigma^*}{r}\right)^6 \right) \quad (20)$$

Common parameters used in the literature include those of Pettitt and Rossky, $\sigma = 0.4$ Å and $\epsilon = 0.046$ kcal/mol,⁴³ which we will refer to as PR-SPC/E and PR-TIP3P, and those often used by Hirata and co-workers, $\sigma = 1.0$ Å and $\epsilon = 0.05455$ kcal/mol.⁴⁴ As noted by Sato and Hirata,⁴⁵ van der Waals parameters are required to solve the RISM equations but also perturb the thermodynamics of the solution.

Alternative approaches to this problem do exist and involve corrective bridge functions^{46–48} or new formalisms that go beyond RISM theory to include orientational correlations and use proper diagrams.^{49–52} The major drawback of the corrective bridge function approach is that a new expression for the excess chemical potential must be derived, a nontrivial task. By including our correction in the potential, the standard closures and related thermodynamic expressions still hold. Including orientational correlations obviates the need for any “protective” Lennard-Jones potential and holds considerable promise. However, the computational complexity of these methods is even greater than that of RISM. Applying them to relatively simple systems presented here will require considerable further development of these methods.

To overcome shortcomings in previous Lennard-Jones parameters while maintaining an analytic expression for the excess chemical potential and mean solvation force, we introduce a general and transferable rule that can be applied to any model with embedded sites. Specifically, we choose

$$\frac{\sigma_e}{2} = \frac{\sigma_h}{2} - b_{he} \quad (21)$$

$$\epsilon_e = 0.1\epsilon_h \quad (22)$$

where σ_e is the radius of the embedded site, σ_h is the radius of the host site, and b_{he} is the bond length between the two. As the embedded radius is now coincident with the host radius along the bond vector, unphysical overlap between sites is prevented. The size of ϵ_e relative to ϵ_h balances deforming the

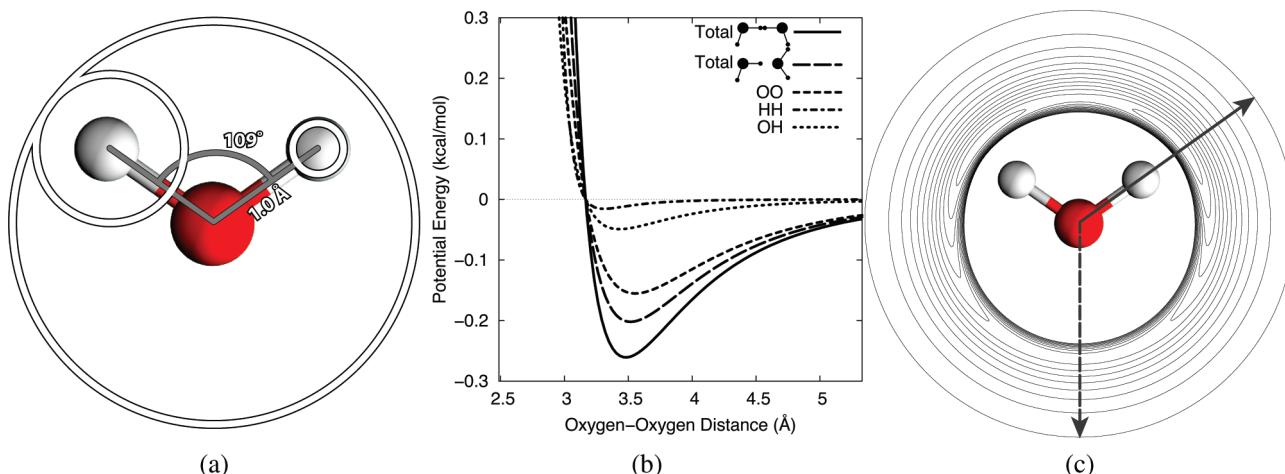


Figure 2. Modified water potential. (a) Schematic illustration of Lennard-Jones parameters for SPC/E water. Lennard-Jones radii, $\sigma/2$, are illustrated by white circles. The radius on the right-hand hydrogen corresponds to that of Pettitt and Rossky,⁴² while the left-hand hydrogen radius is from eqs 21 and 22. (b) Perturbation of water–water Lennard-Jones potential due to the hydrogen potential. The maximum perturbation (solid line) is for two waters with hydrogens aligned. The case of hydrogen bonding is given by the long-dashed line, while the original potential is given by the short dashed line. HH (dot-dashed line) and OH (dotted line) interactions are a result of the new parameters. (c) Angle-dependent water–water interaction. The second water is oriented such that a hydrogen is always pointing toward the central water. The solid and long dashed arrows correspond to the solid and long-dashed lines in (b). Contour lines are spaced 0.02 kcal/mol apart.

Table 1. Parameters for Standard and Modified SPC/E and TIP3P Water Models

model name	$\sigma_O, \text{\AA}$	$\epsilon_O, \text{kcal/mol}$	$\sigma_H, \text{\AA}$	$\epsilon_H, \text{kcal/mol}$	q_O, e	q_H, e	$r(OH), \text{\AA}$
SPC/E	3.1658	0.15530			-0.8476	0.4238	1.0000
cSPC/E	3.1658	0.15530	1.1658	0.01553	-0.8476	0.4238	1.0000
PR-SPC/E	3.1658	0.15530	0.4000	0.04600	-0.8476	0.4238	1.0000
TIP3P	3.1507	0.15200			-0.8340	0.4170	0.9572
cTIP3P	3.1507	0.15200	1.2363	0.01520	-0.8340	0.4170	0.9572
PR-TIP3P	3.1507	0.15200	0.4000	0.04600	-0.8340	0.4170	0.9572

potential of the host while proving a “stiff” enough potential to the embedded site to prevent overlaps. When applied to SPC/E and TIP3P, we refer to these models as coincident SPC/E (cSPC/E) and coincident TIP3P (cTIP3P). This is illustrated for SPC/E water in Figure 2a, and parameters for SPC/E and TIP3P water are given in Table 1.

Unlike the Pettitt and Rossky parameters, the large hydrogen site suggested here does slightly perturb the Lennard-Jones potential of the explicit model (Figure 2b and c). In particular, the well depth is increased in an orientationally dependent manner with hydrogen–hydrogen (solid line, Figure 2b) and hydrogen-bond (long dashed line) orientations becoming more favorable by 0.1 and 0.05 kcal/mol, respectively. Given the improvement in thermodynamics, this small perturbation is justified.

3. Computational Details

All simulations were carried out in a modified version of Amber 10¹ with the Langevin integrator⁵³ and SHAKE⁵⁴ on all bonds involving hydrogen. All 3D-RISM-KH, GB, and GBSA (GBNeck,⁵⁵ igb = 7, parameters in Amber 10) simulations for alanine-dipeptide and protein-G used free boundary conditions, no cutoff for long-range interactions, and a $\delta t = 2 \text{ fs}$ base time step. Explicit solvent calculations used periodic boundary conditions (PBC) with particle-mesh Ewald (PME) summation.⁵⁶

For all alanine-dipeptide simulations, the Amber03 force field⁵⁷ was used with neutral acetyl and *N*-methyl caps. Protein-G simulations used the Amber99SB force field⁵⁸ with an initial conformation from PDB ID: 1P7E.⁵⁹

3.1. Alanine-Dipeptide – Single Point. Grid resolution and residual tolerance effects on numerical artifacts and integration of forces, including net force, were characterized with single point SPC/E 3D-RISM-KH calculations on alanine-dipeptide. A fixed solvation box of $32 \text{ \AA} \times 32 \text{ \AA} \times 32 \text{ \AA}$ with grid spacings of 0.5, 0.25, 0.125, and 0.0625 Å was used to perform calculations with residual error tolerances of 10^{-2} , 10^{-3} , 10^{-4} , 10^{-5} , and 10^{-6} . Because equilibration does not have an impact on these calculations, the default structure for alanine-dipeptide from TLEAP was used. For technical reasons, we used the Numerical Recipes FFT⁶⁰ rather than FFTW for these calculations only.

3.2. Alanine-Dipeptide – Constant Energy. Constant energy simulations were performed on alanine-dipeptide using 3D-RISM-KH and GB solvation models with the standard leapfrog-Verlet integrator. Four 3D-RISM parameter spaces were explored with 8 ns MD simulations: (1) impulse MTS 3D-RISM for a fixed box size ($32 \text{ \AA} \times 32 \text{ \AA} \times 32 \text{ \AA}$), using three previous solutions, with variable grid spacing (0.5 Å, 0.25 Å) and residual tolerance (10^{-3} , 10^{-4} , 10^{-5}); (2) impulse MTS 3D-RISM for a fixed box size ($32 \text{ \AA} \times 32 \text{ \AA} \times 32 \text{ \AA}$), 0.5 Å grid spacing and variable tolerance (10^{-3} ,

10^{-4} , 10^{-5}), and zero to five previous solutions ($N^{cUV} = 0\ldots 5$); (3) dynamic solvation box impulse MTS 3D-RISM calculations with buffers and cutoffs of 4, 6, 8, 10, 12, 14, 16, and 18 Å, $N^{cUV} = 5$; and (4) force extrapolation impulse MTS 3D-RISM for a fixed box size ($32 \text{ \AA} \times 32 \text{ \AA} \times 32 \text{ \AA}$), 0.5 Å grid spacing, 10^{-5} tolerance, and full 3D-RISM solutions every $\Delta t = 2, 4, 6, 10$, and 20 fs.

3.3. Alanine-Dipeptide – Constant Temperature. Long sampling runs were carried out on alanine-dipeptide at constant temperature (300 K) with explicit (SPC/E and TIP3P), implicit (GBNeck), and cSPC/E 3D-RISM-KH solvents. The Langevin integrator⁵³ was used in all cases with $\gamma = 1 \text{ ps}^{-1}$ for explicit solvents, $\gamma = 5 \text{ ps}^{-1}$ for implicit solvents, and $\gamma = 5, 10$, and 20 ps^{-1} for 3D-RISM-KH. 3D-RISM-KH simulations were performed with and without extrapolated forces. Simulations without extrapolated forces had tolerances of 10^{-5} and 10^{-3} with $N^{cUV} = 3$, and one run with a tolerance of 10^{-3} and $N^{cUV} = 5$. Simulations with force extrapolation were performed with 1 and 2 fs time steps. $\delta t = 1 \text{ fs}$ time step runs were performed at $\gamma = 5, 10$, and 20 ps^{-1} , used 10 previous force/coordinate pairs, and $\Delta t = 10$ or 20 fs. $\delta t = 2 \text{ fs}$ time step runs were performed at $\gamma = 5, 10$, and 20 ps^{-1} , used 10 previous force/coordinate pairs, and performed full 3D-RISM calculations every $\Delta t = 4, 6, 8$, or 10 fs.

For all 3D-RISM simulations, a 14 Å cutoff was used for solvent–solute potential and force calculations. Explicit solvent simulations were carried out with both 8 and 14 Å cutoffs for direct nonbond calculations. There was a negligible difference in the results, and only the 14 Å results are presented here.

All simulations were at least 3 ns. Explicit solvent simulations were extended to 21 ns to obtain better sampling. Several other simulations were extended to test convergence of sampling quality. This included GBNeck, 3D-RISM-KH with a tolerance of 10^{-3} , $N^{cUV} = 5$, and $\Delta t = 0$, and 3D-RISM-KH with $\delta t = 1 \text{ fs}$, $\Delta t = 20 \text{ fs}$, and $\gamma = 20 \text{ ps}^{-1}$.

3.4. Sodium-Chloride. A Na^+Cl^- pair in an SPC/E solvent was simulated with 3D-RISM-KH-MD, and the distribution was compared to that expected from the potential of mean force (PMF). To prevent complete dissociation of the ion pair, a distance-based restraint was used:

$$U_{\text{rest}} = k(r - r_0)^2 \quad (23)$$

where $k = 1 \text{ kcal/mol}$ and $r_0 = 4 \text{ \AA}$. Simulations were carried out with both RESPA and FCE MTS. RESPA MTS simulations used $\Delta t = 5 \text{ ps}$ and $\gamma = 5 \text{ ps}^{-1}$. FCE MTS simulations used $\Delta t = 10 \text{ ps}$ and $\gamma = 5, 10$, or 20 ps^{-1} . An integration time step of $\delta t = 1 \text{ fs}$ was used in all cases for a total of 500 ps simulation time.

The PMF was calculated using single point calculations of a Na^+Cl^- pair with radial separations from 2 to 8 Å in 0.02 Å steps. The expected Boltzmann probability distribution is calculated as

$$P(r) \, dr = \frac{4\pi r^2 \exp(-\beta\omega(r)) \, dr}{\int_0^\infty 4\pi \exp(-\beta\omega(r))r^2 \, dr} \quad (24)$$

where $\omega(r)$ is the PMF as a function of r .

3.5. Protein-G. Explicit solvent (SPC/E and TIP3P), GBSA, and cSPC/E 3D-RISM-KH simulations were carried out on protein-G (PDB ID: 1P7E).⁵⁹ SPC/E and TIP3P simulations were both solvated with 16 895 water molecules and used a 8 Å cutoff for direct, nonbonded interactions. MBondi radii were applied for the GBSA (GBNeck) system. All systems were minimized for 1000 steps. Explicit solvent systems were heated to 300 K over 10 ps before production runs. Equilibrium NPT dynamics for the explicit solvent systems was run for 3 ns. GBSA and 3D-RISM-KH were each run for 600 ps. $\gamma = 1 \text{ ps}^{-1}$ was used for the explicit simulations, while $\gamma = 5 \text{ ps}^{-1}$ was used for GBSA. 3D-RISM-KH simulations used time steps of $\delta t = 1 \text{ fs}$ and $\Delta t = 10 \text{ fs}$. A 10 Å cutoff was used for solute–solvent calculations.

3.6. Deca-Alanine. MD, thermodynamic integration (TI), and implicit solvent free energy calculations for deca-alanine are described by Roe et al.⁶¹ As with the implicit solvent calculations, cTIP3P 3D-RISM-KH calculations were performed on each of 1000 frames for each conformation of the 5 ns TI calculation. To accelerate the convergence of 3D-RISM solutions for each frame, the structures in each individual frame were rotated such that the first principal axis was on the z -axis using PTRAJ. A $36 \text{ \AA} \times 36 \text{ \AA} \times 60 \text{ \AA}$ solvation box with a 0.5 Å grid spacing was used for all calculations.

4. Results and Discussion

4.1. Decoy Analysis. Comparison of 3D-RISM-KH MD simulations to explicit and implicit solvent calculations necessarily includes the quality of the pair potential used in the 3D-RISM-KH calculation. Thus, we begin by determining our ability to reproduce the SPC/E and TIP3P model with 1D- and 3D-RISM-KH.

As all thermodynamic properties of the solvent are ultimately calculated from the 1D radial distribution function (RDF), the RDFs of our cTIP3P model with PR-TIP3P, TIP3P MD, and experimental values⁶² are compared in Figure 3 (analogous SPC/E calculations show similar results). The cTIP3P parameters do not improve $g_{\text{OO}}(r)$ relative to PR-TIP3P (Figure 3a). Rather, we see that first peak has moved to a slightly larger radius, while the second peak, the so-called fingerprint of the tetrahedral hydrogen bonding of water,^{43,45} is qualitatively present in PR-TIP3P but completely lost for cTIP3P. $g_{\text{OH}}(r)$ and $g_{\text{HH}}(r)$, on the other hand, are noticeably improved. The first peak of $g_{\text{OH}}(r)$ (Figure 3b) is now at the correct separation (although the magnitude is slightly too low), while the second peak is relatively unchanged. For $g_{\text{HH}}(r)$ (Figure 3c), the first peak has moved to a slightly larger separation, but the magnitude, both absolute and relative to the second peak, is much improved.

The improved structure of liquid water seen in Figure 3 should also provide improved thermodynamics, as the ultimate goal of 3D-RISM (the accurate prediction of experimental solvation free energies) is achieved through accurately reproducing the results of the explicit pair potential used as input. For the purposes of such a comparison, it is

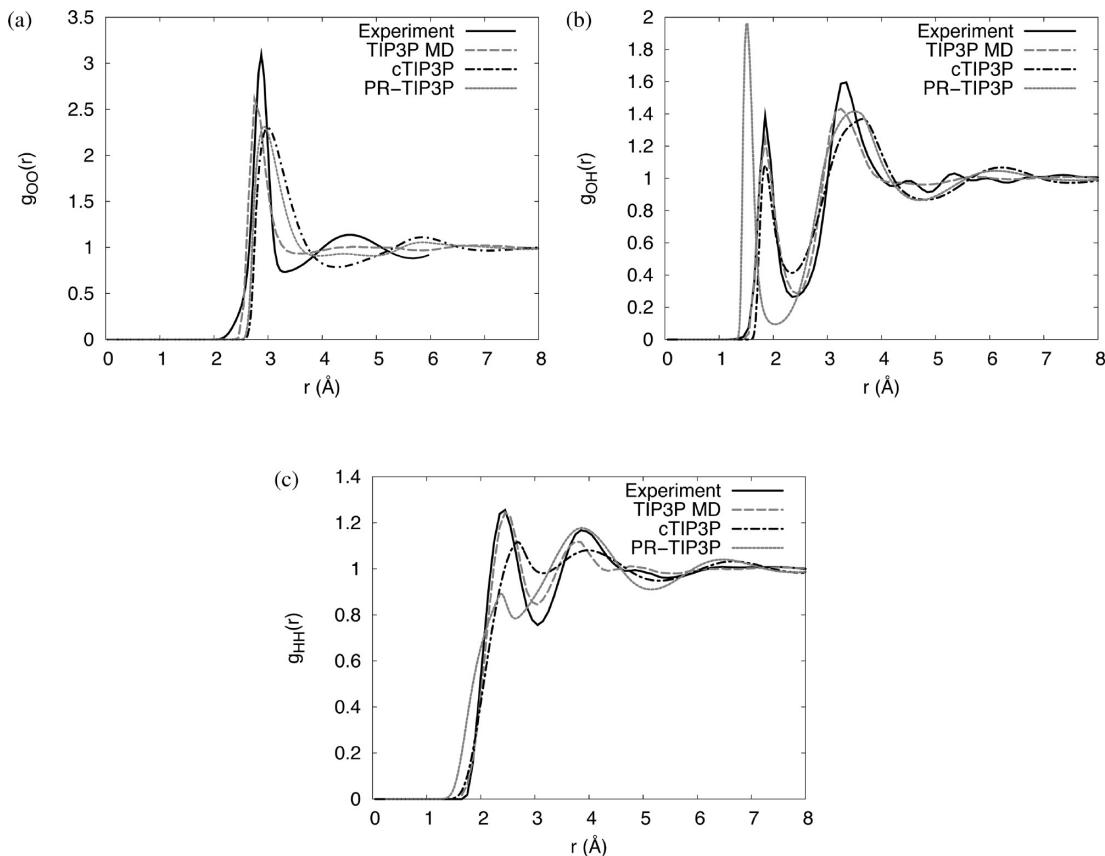


Figure 3. Water radial distribution functions from experiment, MD simulation, and 1D-RISM for (a) oxygen–oxygen, (b) oxygen–hydrogen, and (c) hydrogen–hydrogen.

useful to decompose the total solvation free energy into polar and nonpolar parts, following the standard definitions of the corresponding components in the literature:^{63–65}

$$\Delta G_{\text{sol}} = \Delta G_{\text{cav}} + \Delta G_{\text{vdW}} + \Delta G_{\text{pol}} \quad (25)$$

where G_{cav} , G_{vdW} , and G_{pol} are the free energies of cavity formation, van der Waals dispersion, and solvent polarization, respectively, and are all path-dependent quantities. 3D-RISM calculates G_{sol} directly, so to obtain each component for comparison with TI of explicit solvent it is necessary to follow the same path as used in the benchmark calculation. The free energy of solvent polarization with 3D-RISM-KH is then

$$\Delta G_{\text{pol}} = \Delta G_{\text{sol}} - \Delta G_{\text{sol}}^{\text{uncharged}} \quad (26)$$

where $G_{\text{sol}}^{\text{uncharged}}$ is the solvation free energy of the solute with all partial charges removed. Using this method, we can compare values for deca-alanine calculated by Roe et al.⁶¹ (Table 2). Absolute values of solvent polarization free energy are qualitatively correct for PR-TIP3P, with alpha > left > hairpin > PP2. Both absolute values for and relative difference between the different conformations are quantitatively poor. cTIP3P greatly improves on this with relative errors of 3% or less for each conformation and less than 1 kcal/mol rmsd in relative difference between conformations. Although this does not include nonpolar contributions, it does show a good agreement with the input model.

4.2. Net Force Drift Error. A necessary property of mean solvation forces, such as those calculated by 3D-RISM, is

the lack of a net force on the solute. As 3D-RISM is a grid-based method with an iterative solution, a zero net force is not guaranteed and is a function of the quality of the solution, in particular, the density of the grid and the residual tolerance of the solution. To quantify the net force error, we calculate the absolute force and root-mean-squared error (RMSE) in the force for a single point alanine-dipeptide 3D-RISM-KH solution (see Table 3).

The absolute force drift is the total force in each direction applied to the solute and should be zero for the mean solvation force. For convenience, we report the magnitude of this vector:

$$\left\| E_{\text{abs}(\mathbf{f})} \right\| = \left\| (\sum f_x, \sum f_y, \sum f_z) \right\| \quad (27)$$

Ideally, all components should be zero, although in numerical force calculations this is often not the case (for example, particle-mesh Ewald summation⁵⁶). In practice, artifacts associated with a nonzero net force can be minimized by subtracting the mass weighted average force from each atom:

$$\mathbf{f}'_i = \mathbf{f}_i - \frac{m_i}{M} \mathbf{f}_{\text{net}} \quad (28)$$

where m_i is the mass of the i th solute particle and M is the total mass of the solute. However, the error in the net force is also an indicator of inaccuracies in other components not as easily corrected, such as the net torque. Table 3a suggests that the residual tolerance used for the calculation should be no higher than 10^{-3} . Values lower than this have little impact unless the grid spacing is sufficiently small.

Table 2. Comparison of Explicit TIP3P ΔG_{pol} for Deca-Alanine with 3D-RISM-KH, Poisson Equation (PE), and Generalized Born (GB)^a

3D-RISM ^c							
	TIP3P ^b	cTIP3P	PR-TIP3P	PE ^b	GBHCT ^b	GBOBC ^b	GBNeck ^b
(a) ΔG_{pol}							
alpha	-44.08 ± 0.04	-44.91 ± 1.27	-55.79 ± 0.93	-47.97 ± 0.77	-51.69 ± 1.21	-49.38 ± 1.21	-43.26 ± 0.90
PP2	-76.39 ± 0.15	-76.82 ± 1.31	-93.60 ± 1.07	-78.05 ± 0.91	-77.35 ± 1.05	-78.07 ± 1.09	-77.59 ± 1.02
left	-51.30 ± 0.12	-51.60 ± 1.22	-61.81 ± 1.03	-54.85 ± 0.90	-55.05 ± 1.08	-52.67 ± 1.10	-48.19 ± 0.91
hairpin	-54.16 ± 0.25	-56.00 ± 1.17	-69.36 ± 1.31	-57.28 ± 1.13	-57.48 ± 1.45	-56.03 ± 1.47	-52.85 ± 1.29
(b) $\Delta\Delta G_{\text{pol}}$							
PP2-alpha	-32.31	-31.91	-37.81	-30.07	-25.67	-28.69	-34.33
PP2-left	-25.09	-25.22	-31.79	-23.19	-22.31	-25.40	-29.40
PP2-hairpin	-22.23	-20.82	-24.24	-20.77	-19.87	-22.03	-24.73
alpha-left	7.22	6.69	6.02	6.88	3.36	3.29	4.93
alpha-hairpin	10.08	11.09	13.57	9.31	5.80	6.66	9.60
left-hairpin	2.86	4.40	7.55	2.43	2.43	3.37	4.67
(c) $\Delta\Delta G_{\text{pol}}$ Root-Mean-Square Deviations							
overall	0.99	4.37	1.39	3.89	2.60	2.51	
PP2	0.85	5.14	1.89	4.37	2.10	3.11	
non-PP2	1.11	3.45	0.55	3.34	3.02	1.71	
hairpin	1.34	3.57	1.53	2.83	2.00	1.80	
nonhairpin	0.39	5.05	1.58	4.72	3.09	3.05	

^a Conformations are as in Roe et al.⁶⁰ (alpha, α -helix; PP2, polyproline II; left, left-hand helix; and hairpin, β -hairpin). Units are in kcal/mol, and errors are one standard deviation from the mean. ^b From Roe et al.⁶¹ ^c This work.

Table 3. (a) Net Force (kcal/mol/Å), (b) Root-Mean-Squared Error in the Force, and (c) Solvation Free Energy (kcal/mol) for Single Point 3D-RISM-KH Calculations of Alanine-Dipeptide

tolerance	grid spacing			
	0.5 Å	0.25 Å	0.125 Å	0.0625 Å
(a) Net Force				
10^{-2}	3.2	2.4	2.7	3.3
10^{-3}	1.6	0.35	0.093	0.30
10^{-4}	1.5	0.36	0.061	0.044
10^{-5}	1.5	0.37	0.041	0.0047
10^{-6}	1.5	0.37	0.042	0.0016
(b) Force rms Error				
10^{-2}	$7.1 \times 10^{+0}$	$6.3 \times 10^{+0}$	$6.3 \times 10^{+0}$	$8.3 \times 10^{+0}$
10^{-3}	3.4×10^{-1}	1.0×10^{-1}	1.2×10^{-1}	6.2×10^{-2}
10^{-4}	1.8×10^{-1}	7.5×10^{-3}	7.6×10^{-4}	8.7×10^{-4}
10^{-5}	1.8×10^{-1}	7.4×10^{-3}	5.1×10^{-5}	9.2×10^{-6}
10^{-6}	1.8×10^{-1}	7.6×10^{-3}	5.0×10^{-5}	
(c) Solvation Free Energy				
10^{-2}	7.5794	7.3873	7.4024	8.4253
10^{-3}	14.5614	14.4441	14.4574	14.3924
10^{-4}	14.6366	14.5097	14.5090	14.5092
10^{-5}	14.6382	14.5123	14.5121	14.5117
10^{-6}	14.6382	14.5125	14.5120	14.5116

Another method to quantify the numerical error in the forces is the RMSE.⁵⁶ For a set of “correct” forces, $\tilde{\mathbf{f}}$, we have

$$\text{rmse}_f = \sqrt{\frac{\sum(\mathbf{f} - \tilde{\mathbf{f}})^2}{N_{\text{sol}}} \quad (29)}$$

Because there is no analytic calculation of the forces available for comparison, we use the solution with the smallest grid spacing (0.0625 Å) and lowest tolerance (10^{-6}) as our benchmark. As with the net force calculations, the maximum tolerance permissible is dependent on the grid spacing used. While results do improve as finer grid spacings and smaller tolerances are used, results similar to other methods, for example, particle mesh Ewald,⁵⁶ are obtained for a residual tolerance of 10^{-4} and grid spacings of 0.5 or 0.25 Å.

This observation is also evident in the solvation free energies calculated. A minimum resolution of 0.5 Å provides agreement with high grid densities within 1%. Decreasing

the spacing to 0.25 Å improves this to four significant digits, but little is gained beyond this. In particular, a residual tolerance of 10^{-4} appears to be sufficient, although 10^{-3} can also be considered acceptable.

4.3. Energy Conservation. Numerical artifacts, such as those seen in the net force, typically have a large impact on energy conservation during simulation. Even after removal of the net force, all NVE simulations displayed small amplitude oscillations in the total energy about a linear decay. To quantify the linear decay, the equation

$$E_{\text{tot}} = a \cdot t + b \quad (30)$$

was fit to each data set with t representing the time in picoseconds and a corresponding to the rate of decay in kcal/mol/ps (Table 4). All calculations employed RESPA MTS, as the method is known to conserve energy for 3D-RISM time steps < 5 fs. A comparable calculation using GBNeck yields a decay rate of $-6.37 \pm 6 \times 10^{-3}$ kcal/mol/ps.

Table 4. Rate of Decay (kcal/mol/ps) of Constant Energy Simulations of Alanine-Dipeptide for (a) Variable Grid Spacing and Solution Tolerance, (b) Variable Solution Propagation and Solution Tolerance, (c) Variable Cutoff and Solvent Box Buffer, and (d) Variable Time Step for FCE RESPA MTS^a

		(a)	
		grid spacing	
tolerance		0.5 Å	0.25 Å
	10^{-4}	-0.4372(9)	-0.2207(6)
	10^{-5}	-0.0828(6)	-0.0824(6)
	10^{-6}	-0.0234(6)	-0.0122(5)

		(b)					
		N^{UV}					
tolerance	0	1	2	3	4	5	
1×10^{-3}	0.292(3)	22.62(6)	38.6(1)	9.89(1)	0.0686(4)	0.1127(9)	Energy Conservation
1×10^{-4}	0.0063(1)	0.651(1)	0.992(4)	0.0684(2)	-0.00321(6)	0.00282(9)	
1×10^{-5}	-0.00196(7)	0.01918(6)	0.01526(7)	0.00306(7)	-0.00590(6)	-0.00089(6)	
Average Number of 3D-RISM Iterations per Solution							
1×10^{-3}	47.5	18.3	21.4	28.9	29.4	35.3	
1×10^{-4}	73.2	27.5	30.5	28.0	32.6	35.8	
1×10^{-5}	95.7	47.1	42.8	40.4	40.1	44.1	

		(c)		(d)	
				δt	
cutoff and buffer	energy conservation	Δt		1 fs	2 fs
4 Å	0.623(4)	4 fs			-0.048(2)
6 Å	0.0706(4)	8 fs			0.132(1)
8 Å	-0.00218(9)	10 fs	0.139(4)		
10 Å	-0.00188(6)	12 fs			1.15(1)
12 Å	-0.00033(5)	15 fs	1.50(1)		
14 Å	-0.00198(6)	20 fs	2.26(4)	2.37(3)	
16 Å	-0.00112(6)	40 fs			75(3)
18 Å	-0.00139(7)				

^a Error in the least-squares fit for the last significant digit is given in parentheses.

The impact of grid density and residual tolerance on energy conservation is shown in Table 4a for practical grid densities. Despite differences in the net force and force RMSE produced by these two different spacings, there is negligible difference in the conservation of energy for the same residual tolerance. Considering Tables 3 and 4a suggests that the net force on the solute is primarily an artifact of the grid. The grid is part of the potential, and the tolerance determines the accuracy of the solution for this potential.

The issue is complicated by the fact that the 3D-RISM solution at each time step is not independent but is influenced by the previous solution(s) calculated and retained to seed the initial guess. Table 4b shows the effect on both energy conservation and the number of iterations required to converge on a solution for various truncations of eq 8 and residual tolerances of the 3D-RISM solution. Using zero previous solutions means that the solution at each time step is independent, $c^{UV} = 0$. A strong memory effect is observed when only one or two previous solutions are used. Increasing the number of solutions or decreasing the tolerance effectively erases this effect. Increasing the number of terms used from eq 8 increases the memory required and the number of iterations required to converge.

Two other time-saving methods introduced were cutoffs and a dynamic solvation box. In testing these methods, the cutoff was set equal to the buffer distance, effectively cutting

off the corners of the solvation box. Table 4c shows that once a minimal distance of 8 Å is used, energy conservation is not affected by these methods. It should be noted that the solvation free energy calculated will vary with buffer size.

In contrast to the methods already discussed, FCE RESPA MTS (Table 4d) is not expected to conserve energy. The ability of Langevin dynamics to compensate for this depends on the rate of energy gain. For example, if the time step between 3D-RISM solutions is limited to 20 fs, energy drifts comparable to 10^{-3} tolerance are obtained. Given that dynamics are necessarily perturbed by mean-field methods like 3D-RISM and by Langevin dynamics, some energy drift may be permissible as long as the temperature and sampling are not adversely effected. Figure 4 shows the average temperature for several solvent models and parameters. Note that the values for SCP/E and TIP3P include the solute and solvent. The combination of averaging over a larger system and longer simulation time results in smaller standard errors in the mean. Combined with a sufficiently large friction coefficient, γ , a number of different parameters for FCE RESPA MTS provide stable simulations at the target temperature.

The numerical quality of the 3D-RISM solution is controlled by two parameters: (a) the residual error tolerance in the 3D-RISM calculation and (b) the linear grid spacing of the grid that the solution is found on. To a large extent,

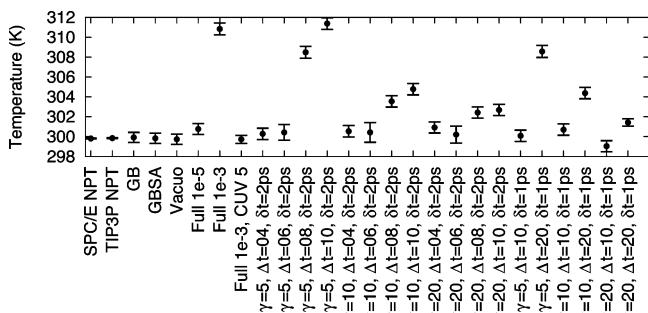


Figure 4. Average temperature for Langevin dynamics simulations of alanine-dipeptide. Error bars represent the standard error in the mean.

these two parameters independently control the conservation of energy and the net force error, respectively.

The 3D-RISM parameters used for MD/3D-RISM-KH depend on the objective of the simulation. If rigorous, constant energy simulations are desired, a residual tolerance of 10^{-5} or lower should be used with $N^{cUV} = 5$ and a buffer and cutoff of 8 Å or more. A larger buffer and cutoff, together with a finer grid spacing, provide better solvation accuracy. However, if the objective is efficient conformational sampling with solvation effects, FCE RESPA MTS can be introduced with $\Delta t = 20$ fs and a Langevin friction coefficient of $\gamma = 20$ ps⁻¹.

4.4. Sodium-Chloride. MD sampling of a Na^+Cl^- pair in solution with a weak restraint provides a simple test of the ability of FCE MTS to correctly sample a known distribution. The small size of the system (the smallest for which solvation effects will perturb the distribution) and the distance restraint near the largest potential barrier in the PMF (Figure 5a) ensure that the solvation forces play the largest possible role in the dynamics.

As expected for such a system, the FCE MTS method does cause heating that is effectively controlled by the Langevin damping coefficient. In particular, the distribution for $\gamma = 20 \text{ ps}^{-1}$ ((Figure 5b) is only slightly skewed from the expected distribution. Here, the distribution is shifted toward larger separations, although this is only clear by the small under sampling around global minimum.

4.5. Conformational Sampling. As 3D-RISM-KH uses an explicit solvent model as input, the conformational sampling should, ideally, be comparable to the underlying explicit solvent model used, in this case, SPC/E. Figure 6 shows free energy differences calculated from sampling distribution between SPC/E and TIP3P, GB, 3D-RISM-KH, and no model (vacuo). Figure 6a–c shows differences between other solvent models and SPC/E, providing context for comparisons with 3D-RISM-KH. Clearly, solvation effects are important, as demonstrated by Figure 6c. Even between very similar explicit models (Figure 6a), the impact can be observed with the TIP3P simulation sampling relatively more in regions of extended (-150° , 155°) and polyproline II conformations (-70° , 150°) than SPC/E. 3D-RISM-KH does see some minor deviations from the SPC/E model, with slightly more sampling of extended regions and slightly less α -helical (-58° , -47°) (Figure 6d–f). Overall, differences between 3D-RISM-KH with the cSPC/E water

model and SPC/E are similar to, if slightly less than, differences between TIP3P and SPC/E. Using FCE RESPAPMTS with 3D-RISM-KH also provides good results, although some softening of the potential barriers appears to occur (Figure 6f and g). This is evidenced by slightly increased sampling particularly between α -helical and polyproline II regions.

Both the quality of the sampling used for Figure 6 and the rate of convergence are shown in Figure 7. Following Lui et al.,⁶⁶ convergence of the Ramachandran sampling was calculated by dividing each trajectory into thirds and computing for each pair of trajectories, A and B:

$$(\chi^{AB})^2(t) = \frac{1}{mn} \sum_{i=1, j=1}^{m,n} (R_{i,j}^A(t) - R_{i,j}^B(t))^2 \quad (31)$$

where the Ramachandran plot at time t is discretized into an $m \times n$ grid. The average $\chi^2(t)$ of the three trajectory combinations for each solvent model is then shown in Figure 7. As mentioned in the methods section, some trajectories were extended to obtain better sampling (explicit SPC/E and TIP3P) or to confirm that convergence was not artificial or coincidental (3D-RISM-KH with 10^{-3} tolerance and $\Delta t = 20$ fs). As expected, the convergence rate of GBSA and 3D-RISM-KH calculations was faster than explicit solvent as friction from the solvent is removed. By this measure, 3D-RISM and GBSA sample 3–4 times more efficiently per simulation time than explicit solvent.

Electrostatic properties of the solute are strongly coupled to conformational sampling and influenced by the solvent. In particular, dielectric properties of the solvent can modify the dipole moment distribution of the solvent. The dipole moment distribution of various solvent models is shown in Figure 8 and tends to echo the results of the Ramachandran distributions. As Kwac et al.⁶⁷ have noted, peaks at 2.5, 4.5, and 7 D for alanine-dipeptide tend to correspond to extended, polyproline II, and α -helical conformations. As compared to SPC/E, all other solvent models show enhancement in extended regions and reductions in α -helical regions. Only TIP3P shows enhancement in polyproline II.

4.6. Speedup. As 3D-RISM computes the complete equilibrium solvent distribution for each solute structure it is applied to, its cost is relatively high per time step as compared to explicit solvent. To offset this, we have introduced a number of methods to reduce the number of computations required and distribute the work over multiple processors.

Serial optimizations for MD/3D-RISM-KH consist of multiple time step methods, solution propagation, cutoffs, and a dynamic solvation box. Two of these methods, MTS and solution propagation, have been previously introduced by Miyata and Hirata²⁵ but have been further extended here. By extending our solution propagation (eq 8) to higher derivatives, using additional previous time steps, computational efficiency has actually been slightly reduced from using only a single previous solution (Figure 9a: RESPA $N^{cUV} = 1$ and RESPA $N^{cUV} = 5$). However, as shown in Table 4b, this additional work greatly enhances energy conservation by eliminating memory effects. A moderate speedup is still

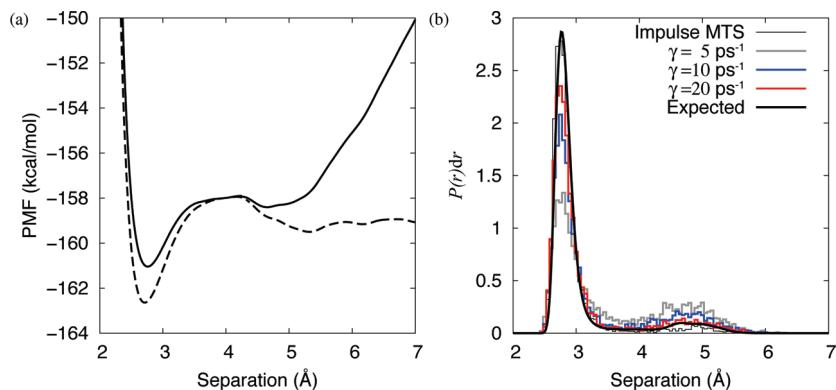


Figure 5. Na^+Cl^- pair in cSPC/E with a weak distance restraint. (a) PMF for the unrestrained pair (dash line) and restrained pair (solid line). (b) Site–site distance distribution for Na^+Cl^- with a weak harmonic restraint. The expected distribution from the potential of mean force is the thick black line; RESPA MTS is the thin black line; FCE MTS with Langevin damping coefficients of $\gamma = 5, 10$, and 20 ps^{-1} are colored gray, blue, and red, respectively.

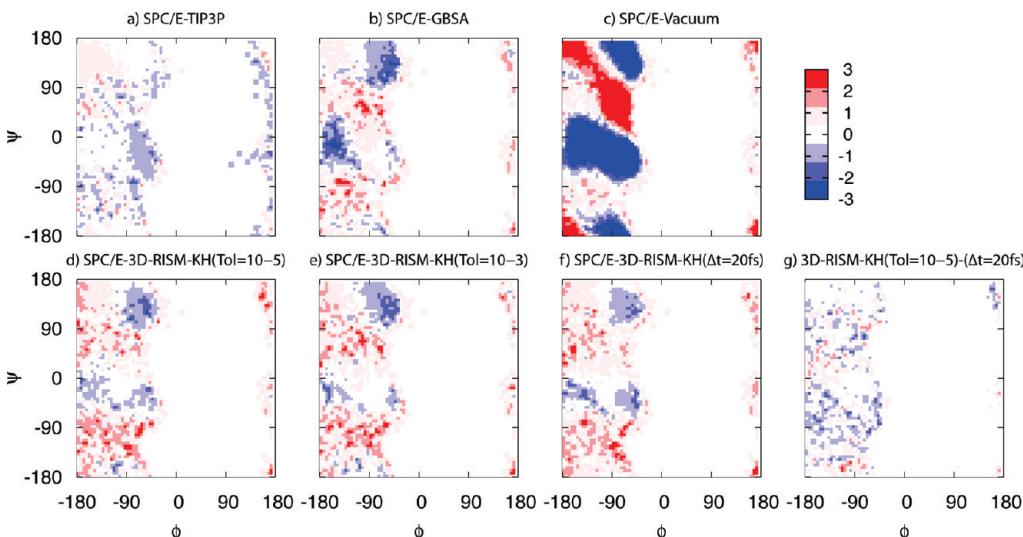


Figure 6. Ramachandran free energy differences of (a)–(f) of select solvation methods from explicit SPC/E water for alanine-dipeptide. (g) Difference of 3D-RISM-KH with a residual tolerance of 10^{-5} and 3D-RISM-KH with a FCE RESPA MTS time step of $\Delta t = 20 \text{ fs}$. Energy units are in kcal/mol.

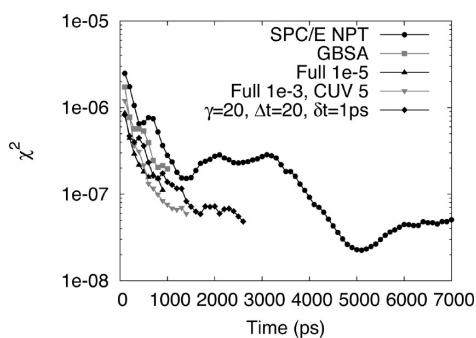


Figure 7. Convergence (χ^2) of Ramachandran plots over simulation time for select solvation methods.

achieved over no solution propagation (Figure 9a: SHAKE and SHAKE $N^{\text{cUV}} = 5$).

Additional computational savings can be achieved for grid-based solute–solvent potential and force calculations. While similar to the use of cutoffs for explicit simulations, cutoffs here can take advantage of the fixed grid spacing (no need for cutoff lists), and points outside of the cutoff can still be accounted for through simple interpolation. However, cutoff

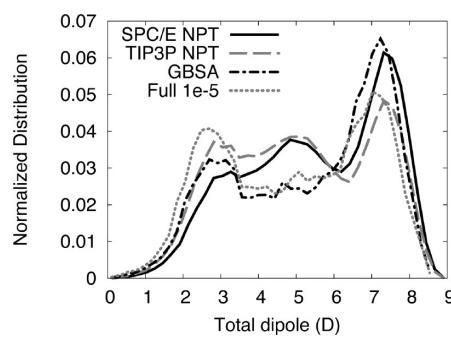


Figure 8. Dipole moment magnitude distributions of alanine-dipeptide for select solvation methods.

methods only offer computational reductions by a constant factor as all grid points must still be visited. The computational savings are due to the number of grid points requiring expensive calculations, involving all of the solute atoms, being considerably reduced. As the grid density and number of solute atoms increase, the cutoff optimizations become more valuable.

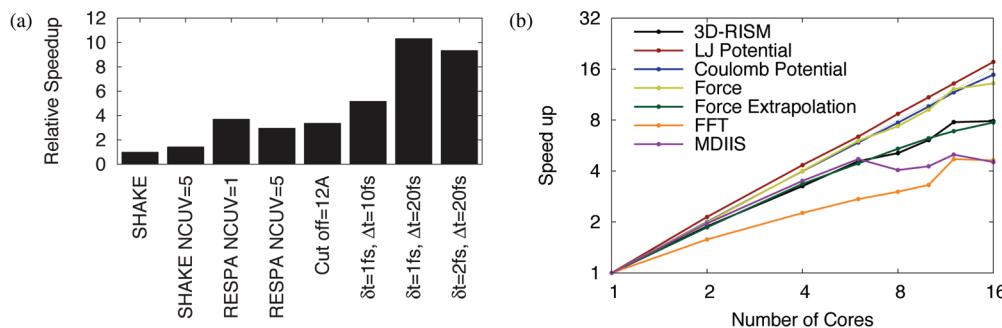


Figure 9. 3D-RISM execution speedup. (a) Serial calculations are shown with optimizations incrementally added. “SHAKE” refers to calculations where $\delta t = \Delta t$. “NCUV” indicates the number of previous solutions used for the initial guess. A cutoff of 12 Å was used for all other calculations. (b) The total parallel speedup is indicated by “3D-RISM”, while the relative speedups of critical subroutines are indicated by the colored lines.

A natural extension to cutoffs is the dynamic resizing of the solvation box. For globular solutes, this has little cost-saving effect and is mostly useful as a convenience; the user only needs to input the buffer distance from the solute and the grid spacing. As solutes become less spherical or undergo large conformational changes, the benefits of the adaptive box size grow by ensuring only the minimum number of grids points is used. Together, adaptive box sizes and cutoffs offer a small overall improvement for alanine-dipeptide (Figure 9a, RESPA $N^{cUV} = 5$ and cutoff = 12 Å).

The greatest computational savings can be achieved by avoiding 3D-RISM calculations altogether by using MTS methods. The nature of biomolecular systems does not allow RESPA MTS time steps to be larger than 5 fs as resonance artifacts are introduced.³⁷ It is possible to overcome this resonance barrier, however, by introducing a nonconservative force approximation at intermediate time steps and using Langevin dynamics to compensate. In the case of FCE RESPA MTS, 3D-RISM-KH solutions can be calculated once every 20 fs (Figure 4). Combined with the other cost-saving measures, a speedup over a basic implementation of 3D-RISM-KH of approximately 10 times is achieved (Figure 9a, SHAKE and $\Delta t = 10, 20$ fs). While it is true that increasing the friction coefficient has a negative impact on the accuracy of dynamics, the use of a mean-field method, 3D-RISM, means that the observed dynamics are not true dynamics in any case. Our goal is to increase sampling efficiency, and using a large friction coefficient is justified in this context.

While parallelization does not decrease the computational workload, it does decrease the wall time for calculations. Furthermore, the spatial decomposition, distributed memory model used here allows the calculation to be run on a network of computers and make use of the total aggregate memory available. Relative speedups as compared to single CPU are shown in Figure 9b. Parallel speedups used protein-G simulations with a total of 50 time steps. Of these, there were eight full 3D-RISM-KH calculations, and three were interpolated 3D-RISM-KH forces. Calculations were performed on a four CPU AMD Opteron machine with four cores per CPU. Grid-based potential and force calculations that were already accelerated with cutoffs and a dynamic solvation box show linear speedups with the number of cores. The force extrapolation method has increasing efficiencies comparable

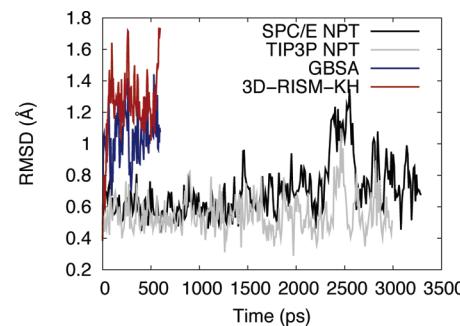


Figure 10. C_α RMSD of protein-G for explicit, implicit, and 3D-RISM-KH solvent models.

to the overall speedup of 3D-RISM. Overall parallel performance is heavily influenced by the scaling of the 3D-FFT and MDIIS routines, which also dominate the overall computation time. As we use FFTW 2.1.5 library for our 3D-FFT calculations, our speedup for the 3D-FFT part of the calculation is limited to scaling of the library.

4.7. Protein-G. Even with our decreased calculation costs, exhaustive conformational sampling of small proteins is still not accessible with 3D-RISM-KH-MD at this time. It is possible to compare different solvation models on the subnanosecond time scale for errors that may be introduced. In particular, differences in secondary and tertiary structure that are indicative of errors may be apparent in subnanosecond trajectories in 3D-RISM-KH due to the enhanced sampling that the method provides.

Figure 10 gives the root-mean-squared deviation (rmsd) of the C_α atoms from the crystal structure of protein-G as a function of simulation time. Both 3D-RISM-KH and GBSA quickly approach rmsd values of 1 Å or greater, with 3D-RISM-KH generally being higher. While these values are higher than those observed with either of the explicit models, they are comparable to previous works.^{68–70} Furthermore, the RMSD of longer explicit simulations continues to grow throughout, suggesting that the equilibrium value may be close to that of 3D-RISM-KH.

Radius of gyration (Figure 11) also shows quickly equilibrating, stable trajectories for 3D-RISM-KH and GBSA with similar values and distributions. Explicit solvent simulations show a smaller and steadily increasing radius of gyration. While it is not clear if the radius of gyration has equilibrated by the end

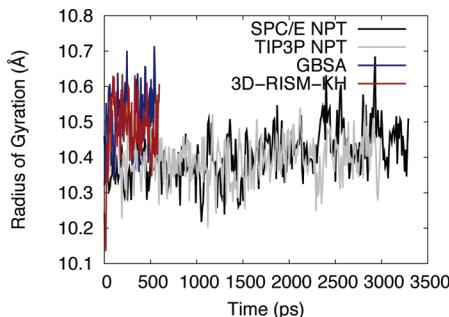


Figure 11. Radius of gyration of protein-G for explicit, implicit, and 3D-RISM-KH solvent models.

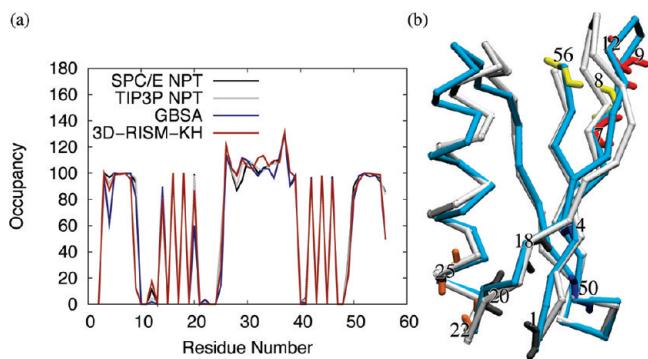


Figure 12. (a) Occupancies for internal backbone hydrogen bonding of protein-G for explicit, implicit, and 3D-RISM-KH solvent models. Occupancies >100% indicate bifurcated hydrogen bonds. (b) 3D trace of C_{α} atoms for NMR structure (PDB ID: 1P7E) in white and final 3D-RISM-KH structure in cyan. Backbone atoms are shown for residues with hydrogen bonding that differs from explicit solvent simulation. Images are made with VMD.^{71,72}

of the simulation (3 ns), it has approached values comparable to both 3D-RISM-KH and GBSA.

As well as providing stable dynamics, solvation methods should preserve both the secondary and the tertiary structures of the solute. Hydrogen-bond calculations were performed with PTRAJ using the default criteria: a distance cutoff of 3.5 Å and an angle cutoff of 120°. Secondary structure involves hydrogen bonding within the backbone of the protein. Figure 12a shows backbone NH groups occupied by hydrogen bonds from backbone CO groups over the entire trajectory. While all solvent models are generally in good agreement, six residues show differences in the occupancies between models (Figure 12b): LYS4, GLY9, LEU12, ALA20, THR25, and GLU56. We examine these case by case.

Hydrogen bonding between residues LYS4 and LYS50 is primarily an issue for GBSA. As this is at the end of a β -sheet, it may indicate some additional flexibility, even unzipping, of the sheet. If the hydrogen-bond cutoff criteria is extended to 4.0 Å from 3.5 Å, the occupancy exceeds 80%. Enhanced flexibility also appears to be the cause for reduced hydrogen bonding between THR25 (NH) and ASP22 (CO) for nonexplicit models with 14% and 28% occupancy for GBSA and 3D-RISM-KH as compared to 40% and 50% for SPC/E and TIP3P.

The loop consisting of residues 9–12 is a site of qualitative difference in structure (Figure 12b) and behavior (Figure 13)

of the 3D-RISM-KH simulation from other solvation models. While a stable hydrogen bond is seen for GBSA (81%), SPC/E (80%), and TIP3P (94%), 3D-RISM-KH shows an occupancy of only 12% for a GLY9 (NH) to LEU12 (CO). In contrast, 3D-RISM-KH also shows an occupancy of 17% for a LEU12 (NH) to GLY9 (CO), while the three other methods only show a 1–2% occupancy. This suggests that there is a oscillation between two weak hydrogen bonds. Indeed, in Figure 13b and c, the GLY9 (NH) to LEU12 (CO) hydrogen bond is disrupted by solvent. The overall effect is to bend this loop out from the protein core into the solvent (Figure 12b).

Residue ALA20 is another site where it would appear that GBSA has failed to capture the correct hydrogen bonding; however, the situation is somewhat more complex. The ALA20 (NH) site is 60% occupied in hydrogen bonding, but this bonding is strictly with THR18 (CO). For SPC/E and TIP3P, ALA20 (NH) has no hydrogen bonding with THR18, but 99% and 100%, respectively, with MET1. 3D-RISM-KH, however, has ALA20 (NH) binding to both THR18 and MET1, 54% and 33%, respectively. The correct behavior in this case is not clear. Clore and Gronenborn,⁷³ on the basis of nuclear magnetic resonance (NMR) data, proposed a three-site bifurcated hydrogen bond between ALA20 (NH), MET1 (CO), and a bound water molecule with residence time >1 ns. In an explicit solvent MD simulation, Sheinerman and Brooks⁶⁸ observed a long residence time water in this location, but, in this case, there was no direct hydrogen bond between ALA20 (NH) and MET1 (CO), and the water served as an intermediary between the two residues. No such long residence time water is observed in our explicit solvent simulations, although the residues are highly solvated (Figure 14a). For our 3D-RISM-KH simulation, however, the hydrogen bond is broken by the solvent (Figure 14c), reformed (Figure 14b), and broken again in the course of the 600 ps simulation.

As well, Clore and Gronenborn also proposed that a similar long residence time water would stabilize a hydrogen bond between TYR33 (NH) and ALA29 (CO). Neither the explicit simulations nor the 3D-RISM-KH simulation showed any water situated to do this, although the site was well hydrated. This is in agreement with the observations of Sheinerman and Brooks.

Both GBSA and 3D-RISM-KH have 50% occupancy for the hydrogen bond between GLU56 (NH) and ASN8 (CO) as compared to 75% and 79% for SPC/E and TIP3P. However, the reason for the low occupancy for 3D-RISM-KH is due to a larger systematic problem. As shown in Figure 15, a large, solvated cleft opens into the hydrophobic interior of the protein in the 3D-RISM-KH simulation. This allows the GLU56 (NH) and ASN8 (CO) pair to be solvated such that the hydrogen bond is disrupted. As pointed out by Kovalenko and Hirata,⁴⁸ this is likely due to the overestimation of solvent ordering around the hydrophobic side chains at the core of the protein. This is a shortcoming of the KH closure, although the same deficiency was originally identified in the HNC closure equation. As such, it is not a shortcoming of 3D-RISM and can be overcome with an improved closure, although such a development is not a trivial task.

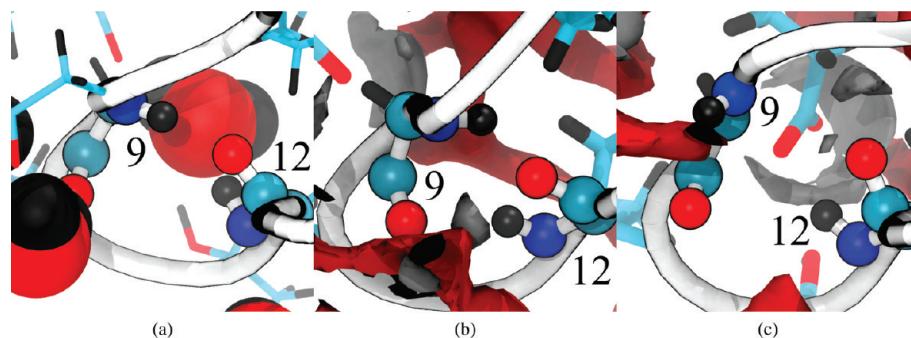


Figure 13. Backbone hydrogen bonding between residues 9 and 12 for representative structures of (a) explicit SPC/E, (b) 3D-RISM-KH with hydrogen bonding, and (c) 3D-RISM-KH without hydrogen bonding. Protein backbone drawn as a white tube, backbone atoms for residues 9 and 12 as spheres, and side chains as sticks. Carbons are cyan, oxygens red, hydrogens black, and nitrogens blue. Solvent density isosurfaces are shown at $g_O^{UV} = g_H^{UV} = 3$ for both oxygen (red) and hydrogen (gray). Images made with VMD.^{71,72}

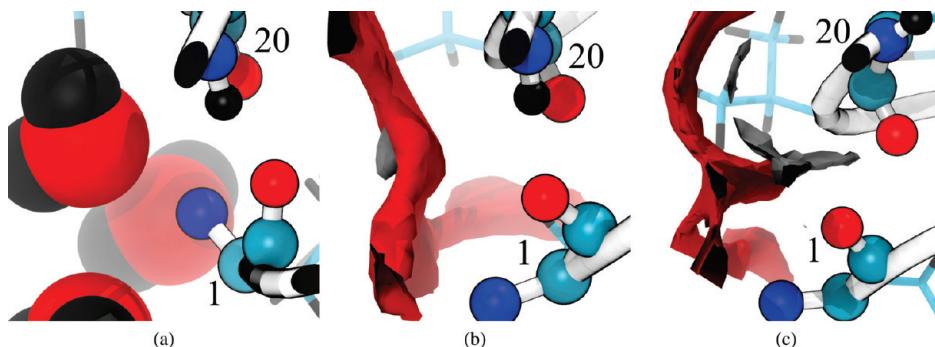


Figure 14. Backbone hydrogen bonding between residues 1 and 20 for representative structures of (a) explicit SPC/E, (b) 3D-RISM-KH with hydrogen bonding, and (c) 3D-RISM-KH without hydrogen bonding. Coloring as in Figure 13. Images made with VMD.^{71,72}

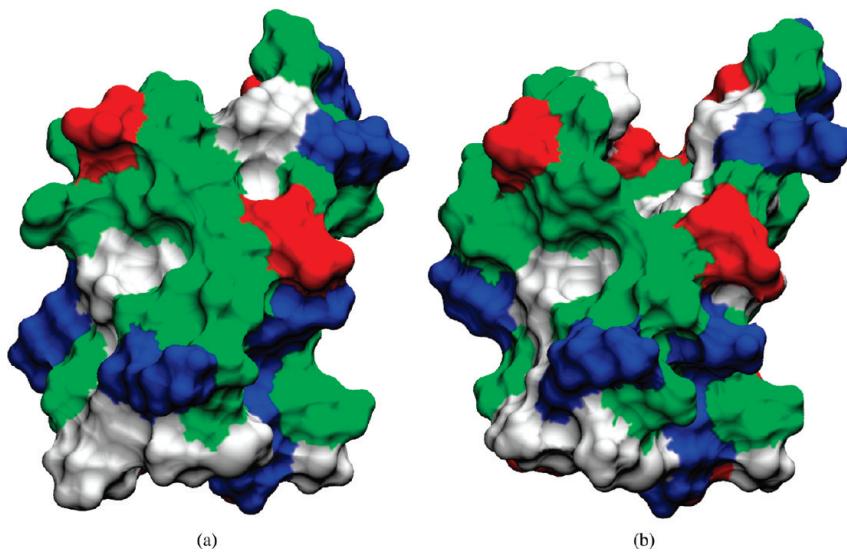


Figure 15. Solvent-accessible surface area for protein-G simulated with (a) explicit SPC/E and (b) 3D-RISM-KH. Surface is colored by residue type: acid (red), base (blue), polar (green), and nonpolar (white). Images made with VMD.^{71,72,74}

5. Conclusions

We have presented an efficient coupling of molecular dynamics simulation with the three-dimensional molecular theory of solvation (3D-RISM-KH), contracting the solvent degrees of freedom, and have implemented this multiscale method in the Amber molecular dynamics package.

The 3D-RISM-KH theory uses the first principles of statistical mechanics to provide a proper account of molecular specificity of both the solute biomolecule and the solvent. This includes such effects as hydrogen bonding both between solvent molecules and between the solute and solvent, hydrophobic hydration, and hydrophobic interaction. The 3D-

RISM-KH theory readily addresses electrolyte solutions and mixtures of liquids of given composition and thermodynamic conditions. As the solvation theory works in a full statistical-mechanical ensemble, the coupled method yields solvent distributions without statistical noise, and further gives access to slow processes like hydration of inner spaces and pockets of biomolecules.

The use of 3D-RISM, a mean-field method contracting the solvent degrees of freedom in a statistical-mechanical average, means that the solvent dynamics are lost and the observed trajectories in any case are not true dynamics of MD simulation with explicit solvent. They are driven largely by a solvent-mediated potential of mean force, that is, by the probability of finding the biomolecule in a particular conformation, sampled over an ensemble of solvation shell arrangements, which frequently require extremely long time to realize (e.g., opening of protein parts to let solvent molecules or ions in to the inner spaces or pockets, multiply repeated to reach proper statistics). However, such trajectories in a solvent potential of mean force preserve the thermodynamic properties such as conformational distribution of the biomolecule and efficiently sample the conformational space regions of interest in a number of molecular biology problems such as functioning of biomolecular structures (e.g., biological channels and chaperones), protein folding, aggregation, and ligand binding.

Arrangements of solution species in the solvation shells of the biomolecule, sampled by the 3D-RISM-KH theory, can include structural solvent and/or cosolvent molecules and other associating structures like salt bridges, buffer ions, and associated ligand molecules. In the latter case, ligand molecules (or their relatively small fragments) at a given concentration in solution are described as a component of solvent at the level of site–site RISM theory and then mapped onto the biomolecule surface by the 3D-RISM method identifying the most probable binding modes of ligand molecules.²⁴ Together with MD sampling of biomolecular conformations, this opens up a new computational method for fragment-based drug design, which provides a proper, statistical-mechanical account of solvation forces with self-consistent coupling of both nonpolar and polar components and which gives access to binding events accompanied by rearrangements of the biomolecule and solvent on a long-time scale.

The implementation includes several procedures to maximally speed up the calculation: (i) cutoff procedures for the Lennard-Jones and electrostatic potentials and the forces acting on the solute, (ii) cutoffs and approximations for the asymptotics of the 3D site correlation functions of solvent, (iii) an iterative guess for the solution to the 3D-RISM-KH equations by extrapolating the past solutions, and (iv) multiple time step (MTS) interpolation of solvation forces between the successive 3D-RISM-KH evaluations of the forces, which are then extrapolated forward at the MD steps until the next 3D-RISM evaluation.

As a preliminary validation, we have applied the method to alanine-dipeptide and protein-G in ambient water. Analysis of the accuracy of forces, energy, and temperature, including such known artifacts as net force drift, has been performed; factors affecting the accuracy have been quantified, and the

range of grid resolution and tolerance parameters ensuring reliable results has been outlined. The performance of the coupled method has been characterized and compared to MD with explicit and implicit solvent. This work is a preliminary but significant step toward the full-scale characterization and analysis of the new method and is a further improvement of its performance to address slow processes of large biomolecules in solution.

Acknowledgment. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the National Research Council (NRC) of Canada. All calculations were performed on the HPC cluster of the Center of Excellence in Integrated Nanotools (CEIN) at the University of Alberta. T.L. acknowledges financial support from the NSERC, NRC, and University of Alberta.

References

- Case, D.; Cheatham, T.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
- Still, W.; Tempczyk, A.; Hawley, R.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
- Chandler, D.; McCoy, J.; Singer, S. *J. Chem. Phys.* **1986**, *85*, 5971–5976.
- Chandler, D.; McCoy, J.; Singer, S. *J. Chem. Phys.* **1986**, *85*, 5977–5982.
- Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 7821–7826.
- Kovalenko, A.; Hirata, F. *Chem. Phys. Lett.* **1998**, *290*, 237–244.
- Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **1999**, *110*, 10095–10112.
- Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2000**, *112*, 10391–10402.
- Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2000**, *112*, 10403–10417.
- Kovalenko, A. In *Molecular Theory of Solvation*; Hirata, F., Ed.; Kluwer Academic Publishers: Norwell, MA, 2003; Vol. 24, Chapter 4, pp 169–276.
- Hansen, J.-P.; McDonald, I. *Theory of Simple Liquids*, 3rd ed.; Elsevier: Amsterdam, the Netherlands, 2006.
- Beglov, D.; Roux, B. *J. Chem. Phys.* **1995**, *103*, 360–364.
- Yoshida, K.; Yamaguchi, T.; Kovalenko, A.; Hirata, F. *J. Phys. Chem. B* **2002**, *106*, 5042–5049.
- Omelyan, I.; Kovalenko, A.; Hirata, F. *J. Theor. Comput. Chem.* **2003**, *2*, 193–203.
- Gusarov, S.; Ziegler, T.; Kovalenko, A. *J. Phys. Chem. A* **2006**, *110*, 6083–6090.
- Casanova, D.; Gusarov, S.; Kovalenko, A.; Ziegler, T. *J. Chem. Theory Comput.* **2007**, *3*, 458–476.
- Moralez, J.; Raez, J.; Yamazaki, T.; Motkuri, R.; Kovalenko, A.; Fenniri, H. *J. Am. Chem. Soc.* **2005**, *127*, 8307–8309.
- Johnson, R.; Yamazaki, T.; Kovalenko, A.; Fenniri, H. *J. Am. Chem. Soc.* **2007**, *129*, 5735–5743.
- Tikhomirov, G.; Yamazaki, T.; Kovalenko, A.; Fenniri, H. *Langmuir* **2008**, *24*, 4447–4450.

- (20) Yamazaki, T.; Fenniri, H.; Kovalenko, A. *ChemPhysChem* **2010**, *11*, 361–367.
- (21) Drabik, P.; Gusarov, S.; Kovalenko, A. *Biophys. J.* **2007**, *92*, 394–403.
- (22) Yamazaki, T.; Imai, T.; Hirata, F.; Kovalenko, A. *J. Phys. Chem. B* **2007**, *111*, 1206–1212.
- (23) Yoshida, N.; Imai, T.; Phongphanphanee, S.; Kovalenko, A.; Hirata, F. *J. Phys. Chem. B* **2009**, *113*, 873–886.
- (24) Imai, T.; Oda, K.; Kovalenko, A.; Hirata, F.; Kidera, A. *J. Am. Chem. Soc.* **2009**, *131*, 12430–12440.
- (25) Miyata, T.; Hirata, F. *J. Comput. Chem.* **2008**, *29*, 871–882.
- (26) Tuckerman, M.; Berne, B.; Martyna, G. *J. Chem. Phys.* **1991**, *94*, 6811–6815.
- (27) Tuckerman, M.; Berne, B.; Martyna, G. *J. Chem. Phys.* **1992**, *97*, 1990–2001.
- (28) Pulay, P. *Chem. Phys. Lett.* **1980**, *73*, 393–398.
- (29) Saad, Y.; Schultz, M. *J. Sci. Stat. Comput.* **1986**, *7*, 856–869.
- (30) Perkyns, J.; Pettitt, B. *J. Chem. Phys.* **1992**, *97*, 7656–7666.
- (31) Perkyns, J.; Pettitt, B. *Chem. Phys. Lett.* **1992**, *190*, 626–630.
- (32) Burkhardt, J. *BLEND: Transfinite Interpolation*; http://people.scs.fsu.edu/~burkardt/f_src/blend/blend.html, accessed on 6/12/2008.
- (33) Allen, M.; Tildesley, D. *Computer Simulation in Chemical Physics*; Oxford University Press: Oxford, UK, 1993; Chapter 4, pp 140–181.
- (34) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2000**, *112*, 10391–10402.
- (35) Barth, E.; Schlick, T. *J. Chem. Phys.* **1998**, *109*, 1633–1642.
- (36) Barash, D.; Yang, L.; Qian, X.; Schlick, T. *J. Comput. Chem.* **2003**, *24*, 77–88.
- (37) Schlick, T. *Molecular Modeling and Simulation: an Interdisciplinary Guide*, 1st ed.; Springer-Verlag New York, Inc.: Secaucus, NJ, 2002.
- (38) Batcho, P.; Case, D.; Schlick, T. *J. Chem. Phys.* **2001**, *115*, 4003–4018.
- (39) Qian, X.; Schlick, T. *J. Chem. Phys.* **2002**, *116*, 5971–5983.
- (40) Frigo, M. *SIGPLAN Not.* **1999**, *34*, 169–180.
- (41) Berendsen, H.; Grigera, J.; Straatsma, T. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
- (42) Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (43) Pettitt, B.; Rossky, P. *J. Chem. Phys.* **1982**, *77*, 1451–1457.
- (44) Hirata, F.; Levy, R. M. *Chem. Phys. Lett.* **1987**, *136*, 267–273.
- (45) Sato, H.; Hirata, F. *J. Chem. Phys.* **1999**, *111*, 8545–8555.
- (46) Cortis, C.; Rossky, P.; Friesner, R. *J. Chem. Phys.* **1997**, *107*, 6400–6414.
- (47) Du, Q.; Beglov, D.; Roux, B. *J. Phys. Chem. B* **2000**, *104*, 796–805.
- (48) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2000**, *113*, 2793–2805.
- (49) Dyer, K.; Perkyns, J.; Stell, G.; Pettitt, B. *J. Chem. Phys.* **2008**, *129*, 104512–104520.
- (50) Gendre, L.; Ramirez, R.; Borgis, D. *Chem. Phys. Lett.* **2009**, *474*, 366–370.
- (51) Sumi, T.; Sekino, H. *J. Chem. Phys.* **2006**, *125*, 034509–034509.
- (52) Urbic, T.; Vlachy, V.; Kalyuzhnyi, Y.; Dill, K. *J. Chem. Phys.* **2003**, *118*, 5516–5525.
- (53) Loncharich, R.; Brooks, B.; Pastor, R. *Biopolymers* **1992**, *32*, 523–535.
- (54) Ryckaert, J.; Ciccotti, G.; Berendsen, H. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (55) Mongan, J.; Simmerling, C.; McCammon, J.; Case, D.; Onufriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156–169.
- (56) Essmann, U.; Perera, L.; Berkowitz, M.; Darden, T.; Lee, H.; Pedersen, L. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (57) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999–2012.
- (58) Simmerling, C.; Strockbine, B.; Roitberg, A. *J. Am. Chem. Soc.* **2002**, *124*, 11258–11259.
- (59) Ulmer, T.; Ramirez, B.; Delaglio, F.; Bax, A. *J. Am. Chem. Soc.* **2003**, *125*, 9179–9191.
- (60) Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B. *Numerical Recipes in FORTRAN: the Art of Scientific Computing*, 2nd ed.; Cambridge University Press: New York, NY, 1992.
- (61) Roe, D.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C. *J. Phys. Chem. B* **2007**, *111*, 1846–1857.
- (62) Soper, A.; Phillips, M. *Chem. Phys.* **1986**, *107*, 47–60.
- (63) Still, W.; Tempczyk, A.; Hawley, R.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
- (64) Levy, R.; Zhang, L.; Gallicchio, E.; Felts, A. *J. Am. Chem. Soc.* **2003**, *125*, 9523–9530.
- (65) Gonçalves, P.; Stassen, H. *Pure Appl. Chem.* **2004**, *76*, 231–240.
- (66) Liu, P.; Kim, B.; Friesner, R.; Berne, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13749–13754.
- (67) Kwac, K.; Lee, K.-K.; Han, J. B.; Oh, K.-I.; Cho, M. *J. Chem. Phys.* **2008**, *128*, 105106.
- (68) Sheinerman, F.; C. B. III. *Proteins: Struct., Funct., Genet.* **1997**, *29*, 193–202.
- (69) Patel, S.; A. M. Jr.; C. B. III. *J. Comput. Chem.* **2004**, *25*, 1504–1514.
- (70) Calimet, N.; Schaefer, M.; Simonson, T. *Proteins: Struct., Funct., Genet.* **2001**, *45*, 144–158.
- (71) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.
- (72) Stone, J. M. Sc. thesis, Computer Science Department, University of Missouri-Rolla, 1998.
- (73) Clore, G.; Gronenborn, A. *J. Mol. Biol.* **1992**, *223*, 853–856.
- (74) Sanner, M.; Olsen, A.; Spehner, J.-C. *Fast and Robust Computation of Molecular Surfaces*. Proceedings of the 11th ACM Symposium on Computational Geometry, New York, 1995; pp C6–C7.