IMPROVING STATISTICAL MECHANICAL SOLVATION MODELS FOR BIOMOLECULAR APPLICATIONS

BY JESSE A. JOHNSON

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Computational Biology and Molecular Biophysics

Written under the direction of

Dr. David A. Case

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2016

ABSTRACT OF THE DISSERTATION

Improving Statistical Mechanical Solvation Models For Biomolecular Applications

by Jesse A. Johnson

Dissertation Director: Dr. David A. Case

0.1 Introduction

Chapter 1 contains a basic introduction to solvation models. Special attention is given to the Ornstein-Zernike and RISM statistical mechanical solvation models used throughout this work.

0.2 Correction of 3D-RISM Solvation Thermodynamics for Small Molecules

Implicit solvent models offer a fast way to estimate the effects of solvation on solute without the complications of explicit simulations. One common test of model accuracy is to compute the transfer energy from gas to liquid for a variety of small molecules, since many of these values have been experimentally measured. Studies of the temperature dependence of these values can provide additional insights into the performance of implicit solvent models. In this work the temperature derivatives of solvation energies for the 3D-RISM integral equation approach are computed. Results for 1123 small drug-like molecules (both neutral and charged) in water are compared to results from molecular dynamics simulations and experiment. The uncorrected results are rather poor, but it

is known that errors are strongly correlated with the partial molar volumes of the solutes. Several linear solvation energy corrections are examined and extended to deal with solvation enthalpies and entropies. A new temperature-dependent linear correction is introduced.

0.3 Crystal Structure Refinement with Periodic 3D-RISM

X-ray scattering measurements from macromolecular crystals can provide valuable information about the solvent environment around biomolecules, but conventional refinement techniques use only very simplified solvation models. In this work solvent distributions for six protein structures are computed using molecular dynamics or integral equation (3D-RISM) solvation models. Bragg intensities for both models are in better agreement with experiment at all resolution ranges than those computed using the default "flat" solvent model in the refmac5 refinement program, with the greatest improvement in the 1.5 to 2.5 Å range. Results from MD simulation are generally closer to experiment than those from 3D-RISM, but the differences are small and should be balanced against the much larger computational resources required for MD simulations. The 3D-RISM solvent distributions can be derived in seconds (for unit cells with 50 Å sides), and could be updated regularly during the course of crystallographic refinement.

Acknowledgements

Chapter 2 is a collaborative effort with Tyler Luchko and David Case and has been recently submitted for publication in the Journal of Physics: Condensed Matter. Chapter 3 is a collaborative effort with David Case and Timothy Giese and is in preparation for publication.

While many have helped me reach this point in my studies, I will single out two people for their major roles in my graduate school career.

I thank Professor David Case for his patience, encouragement, and guidance. "Make hay while the sun is shining." Advice I strive to live by.

I thank Professor Tyler Luchko for helping orient me both into my field of research and academic life in general.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Solvation

Solvation is the process of molecules (known as solute) becoming surrounded by a molecular fluid (known as solvent) through their mutual interactions. The effect of solvation on chemical processes is of great practical importance in science, medicine, and industry. Nearly all known biological processes occur in a salt water solution, where the liquid solvent often plays a critical role in altering biochemical reactions. In addition, many biomolecules require a stable solvent environment to maintain their form and function. Several major examples of solvation affecting biology include:

1. Protein conformational transitions can be more or less favorable depending on the solvent environment [86, 18].

2. Protein and nucleic acid binding of virtually all varieties can be aided or hindered depending on their solvent environment. This includes ligand binding, complex formation, and DNA recognition and binding [69].

Drug solubility, efficacy, and side effects can critically depend on the surrounding solvent. Thus modern drug design must take solvation into consideration. In industry, solvation is commonly used to control reaction rates and reduce wear on manufacturing equipment. Most industrialized methods of chemical synthesis and purification rely on regulated solvation environments to economically produce large product volumes [44]. Thus, in principle, a successful solvation model could reduce costs and increase success rates of many activities: designing drugs to interact with their targets, investigating the mechanisms of biochemical interactions which may reveal new drug targets, and engineering industrial solvents and the materials they interface with, to name a few.

Due to the importance of solvation interactions in a variety of fields, much effort has been expended attempting to simulate solvation so its effects can be more accurately predicted. Simulating the motion of molecules in a fluid and interactions with their environment poses both theoretical and computational challenges. For mildly dilute solutions, the ratio of solvent atoms to solute atoms is rather large, being on the order of 10^4:1 for water:albumin in the average human blood sample [1]. Even in more concentrated cellular environments there remain far more water molecules than solute. Without the use of clever methods and algorithms, even modern supercomputers could not handle systems with such large numbers of mutually interacting moving objects. Attempting to theoretically comprehend the quantitative and qualitative behavior of extremely large systems requires sophisticated mathematical models whose complexity can easily obscure their applicability and explanatory power. Thus modern solvation models trend towards simplicity and computational efficiency, though this may be done at the cost of losing generality and overlooking subtle solvation behavior which only occurs in complex or computationally expensive models.

1.2 Solvation Models

Like all physical processes, solvation is ultimately governed by physics. At the heart of most solvation models are solvent-solute interactions which can be modeled using a number of approaches varying in complexity, computational cost, and physical realism. Since thermodynamics and statistical mechanics are the tools of choice when dealing with very large numbers of interacting particles, most modern solvation models have as their goal the calculation of thermal quantities which predict the statistically averaged behavior of solvation dynamics. Thus solvation interactions form the starting point of solvation models and serve as a common tongue to bind them together. Similarly, thermodynamics and mechanical dynamics are the the typical output of these models. Thus a natural starting point for discussing solvation models is discussing the various known solvation interactions.

1.2.1 Solvation Interactions

The standard solvation model involves a total solute-solvent interaction energy which can be divided into physically distinct additive component energies.

$$E_{\text{solv}} = E_{\text{coul}} + E_{\text{pol}} + E_{\text{disp}} + E_{\text{exc}} + E_{\text{CT}} \tag{1.1}$$

The total interaction energy E_{solv} is the **total solvation interaction energy** and can be used when deriving most thermodynamic quantities of interest related to solvation. The Coulombic solvation energy E_{coul} results from the electrostatic interaction between charged particles of the solute and solvent. It can be attractive or repulsive (i.e., negative or positive respectively) depending on the signs of interacting charged particles. All magnetic and electrodynamic interactions are typically ignored since they are assumed to have negligible effect on solvation. The polarization energy E_{pol} arises from the solute and solvent causing mutually induced electric dipole moments in one another (often called a induced dipole-induced dipole interaction). The dispersion energy E_{disp} is due to charged particle motion leading to instantaneous dipole moments of the solute or solvent which in turn creates an induced dipole moment in the other (often called an instantaneous dipole-induced dipole interaction). Both polarization and dispersion interactions are always attractive since the resulting induced dipoles increase separation of oppositely charged particles between the solute and solvent. The exchange energy E_{exc} results from the Pauli exclusion force between solute and solvent particles and is always positive (i.e., repulsive) since Pauli exclusion prevents fermions (e.g., electrons, protons, and neutronst) from occupying the same quantum state. The polarization, dispersion, and exchange interactions are collectively known as the **Van der Waals interactions**. Finally, the charge transfer energy E_{CT} is the energy from charges being exchanged between solute and solvent. While charge transfer is important in many solvation reactions where solvent chemically interact with the solute, it has been largely ignored in published solvation models, in part due to it requiring quantum methods which carry high computational cost and added complexity.

The total solvation interaction is used by solvation models to calculate solvation properties such as solvation thermodynamics. How exactly this is done depends on the model, including the specific thermodynamic ensemble used by the model. Common thermodynamic quantities of interest that solvation models can compute include the solvation excess chemical potential, pressure of solvation, partial molar volume of the solvent, and more. In addition, some models can provide a spatial distribution of the solvent. In biophysics these solvation quantities can be applied in calculating the effect of the solvent environment on reaction rate constants, ligand binding affinity, and solubility. In addition, most models allow solvation forces to be calculated, which can be used in molecular dynamics calculations, such as protein folding and ligand docking simulations.

1.2.2 Solvation Models

All solvation models discussed in this work provide a method of calculating one or more of the solvation energies in equation (1.1). These models differ in

- physical assumptions
- accuracy
- computational cost

No existing practical solvation model is universally applicable as they must make physical assumptions in order to balance accuracy and computational cost. As new and improved computing technologies become available, this situation may change and a unified solvation model may become feasible. As of this writing there are a large number of competing models. Which model a researcher chooses depends on all three above factors with respect to their application, though computational cost is often the dominant factor and thus has been a major focus of recent solvation models.

Solvation models can be roughly divided into three categories (roughly in order of greatest to least general computational cost): quantum, classical, and quasi- or non-physical.

Quantum solvation models are the most computationally expensive, often prohibitively so for practical application, but they are necessary when solute-solvent and solvent-solvent

covalent bonding or electron transfer must be considered. The most popular quantum solvation model is the polarizable continuum model (PCM) which makes quantum modeling tractable by treating the solvent as a continuum of some kind (dielectric, conductive, etc.) rather than individual molecules. However, the same continuum approximation which makes the PCM computationally attractive also reduces its ability to accurately model covalent and electron transfer reactions, partly defeating one of the primary uses of quantum models. Consequently PCM models are primarily useful when electrostatic forces dominate solvation. For a general review of quantum solvation models, see [84].

Unlike quantum models, classical solvation models ignore electron transfer entirely and focus an electrostatic effects and wave-free approximations for the van der Waals interactions. Classical solvation models can be divided into explicit solvent models, where every solvent molecule is individually modeled, and implicit solvent models, such as where the solvent is represented as a continuum (similar to the PCM). By far the most common explicit solvent model is the all-atom molecular dynamics simulation which places solvent molecules in the simulation box with the solute and performs a full Newtonian force calculation for all atoms. This is very computationally expensive due to the sheer number of solvent atoms involved and the fact that they typically must interact with one another as well as with the solute in order to obtain realistic solvent distributions. Nonetheless, explicit solvent models have shown excellent agreement with experiment in cases where electrostatic effects dominate solvation and have the added benefit of allowing a consistent model to be used for both solvation and solute-solute interactions while retaining computational feasibility.

The most popular classical implicit solvent model is the Poisson-Boltzmann equation (PBE) and its linearized form, Debye-Hückle theory. When electrostatic effects dominate, this family of models generally agree with explicit solvent models while requiring a tiny fraction of the computational cost. Whereas most explicit and quantum solvent simulations take hours to complete, PBE simulations take minutes or even seconds. However, implicit solvent models tend to have greater error when compared to experiment than explicit solvent models, though whether the size of this error is too great depends on the application. Nevertheless, PBE models tend to be the solvent model of choice in practical

applications due to their reasonable accuracy and very low computational cost.

Most of the models mentioned so far have experimentally determined parameters. The few models which have little or no experimental parameters are called 'ab initio' models, though even among these so-called parameter free models there usually are a few experimental parameters. Virtually no practically useful solvation model relies entirely on fundamental physical constants. In contrast to ab inito models, there are quasi- or non-physical solvation models. Examples include the Solvent Accessible Surface Area (SASA) model and the Generalized Born (GB) model. These models tend to use primitive geometric approximations which, when well parameterized, can potentially give extremely fast results that approach or exceed the accuracy of implicit solvation models. However, since these models rely heavily on parameterization, it can be difficult to accurately estimate their error when applied to a new molecule not contained in the parameterization set. Thus these models tend to be primarily used in specialized cases for which they were parameterized, such as solvation of small molecules or solvation of DNA-like helices.

There are limitations shared by all current practical solvation models. Generally, hydrophobic effects have proven difficult to capture using computationally accessible models. Viscosity of fluids is typically poorly modeled or ignored. Perhaps most important to practical applications, especially in biophysics, hydrogen bonding often is not included in models or inaccurately reproduced. For reactions involving ionization, they typically cannot be modeled without resorting to the more sophisticated (and computationally expensive) quantum solvation models.

This work focuses on a currently underutilized family of classical solvation models based on integral equations from statistical mechanics. They are in many ways similar to implicit solvent models, treating the solvent as a continuum whose density distribution must minimize an energy function, but unlike most implicit solvent models they also consider the molecular structure of the solvent as well as solvent-solvent correlations. These statistical mechanical models allow fast computation of accurate anisotropic thermally averaged solvent distributions, even in 3D for the case of the 3D-RISM theory, something not presently possible using traditional implicit solvent models.

1.3 Statistical mechanics of fluids

A classical statistical mechanical model of fluids has proven useful for modeling solvation. A brief general overview of the theory will be presented here. For a more thorough presentation see [66] or [57], and for a relatively complete reference see [29].

For a given fluid, consider a function which counts the average number of particles at a given position relative to a reference particle in the fluid. This is called the **radial distribution function** (**RDF**) $g(\mathbf{r})$, and can be used to obtain the particle number density ρ at a position $\mathbf{r}$ given some reference unperturbed (e.g., bulk) particle density ρ_0, $\rho(\mathbf{r}) = g(\mathbf{r})\rho_0$. For a homogeneous isotropic system of evenly distributed particles, the RDF is dependent only on radial distance from the reference particle, $g(r)$, whereas in inhomogeneous systems where local density varies based on angle with respect to the reference particle, the RDF may be a fully 3D function of relative position.

The RDF can be used to calculate the partition function, which bridges statistical mechanics to thermodynamics. With the partition function, the total thermodynamic system energy can be calculated, which in turn allows for the Helmholtz energy to be calculated. With these energies and the partition function, most desired thermodynamic quantities can be computed, including the entropy, pressure, and heat capacity, among others. The RDF approach to fluids extends naturally to solvation, where typically the reference particles are solute molecules (or their atoms) and the RDFs represent the local density of solvent molecules (or their atoms) about the solute.

Unfortunately it is often difficult or impossible to measure fluid RDFs experimentally. Thus theoretical methods are frequently employed to calculate the RDF of a system using known parameters. Doing so requires knowledge about properties of the constituent particles of the system, including their particle-particle interactions. These properties may be theoretically or experimentally determined, though whether this is practically possible depends on the system.

One approach to calculating the RDF is by dividing it into two component correlation functions. This is traditionally done by defining a new correlation function which represents the excess of the RDF over bulk, $h(r) \equiv g(r) - 1$, named the **total correlation**

function (**TCF**). Then $h(r)$ is divided into a sum of component correlations functions

$$h(r_{12}) = c(r_{12}) + t(r_{12}) = c(r_{12}) + \rho \int d\mathbf{r}_3 \, c(r_{13}) \, h(r_{23}) \tag{1.2}$$

where r_{12} is the distance between particles 1 and 2 and ρ is the bulk particle density. Equation (1.2) is know as the **Ornstein-Zernike (OZ) equation**. The OZ equation expresses the TCF $h(r_{12})$ between particles 1 and 2 as the sum of their pair-wise **direct** and **indirect correlation functions** $c(r_{12})$ and $t(r_{12})$ (**DCF** and **ICF** respectively). The DCF represents the direct probabilistic correlation between two particles without any intermediate particles, while the ICF is the sum of all correlations between two particles mediated by a number of intermediate correlated particles. To calculate the ICF for each possible mediating particle 3, the direct correlation of particle 1 and mediator 3 must be multiplied by the *total* correlation between particle 2 and 3, hence the form of the second term on the right side of equation (1.2). The hope of the OZ equation is that by separating the direct and indirect interactions, one of the two interactions can be modeled in some way and this model can be used with the OZ equation to obtain solutions for the TCF.

Note that the inclusion of the TCF in calculating the ICF makes the OZ equation recursive, meaning an infinite set of mediating particles must be included when calculating the TCF. Further, there are two unknowns ($h(r)$, $c(r)$) and only one equation, so the equation is undetermined. Still more, it is unclear how best to calculate the convolution integral introduced to the OZ equation by the ICF. Thus the ICF is both what makes the OZ equation interesting and challenging.

The problem of underdetermination is commonly resolved by artificially introducing a second equation known as the **closure relation**, which effectively serves as a model of the ICF:

$$t(r_{12}) = h(r_{12}) - c(r_{12}) = \exp\left[-\beta u(r_{12}) - h(r_{12}) + B(r_{12})\right] \tag{1.3}$$

Here $u(r_{12})$ is the interaction potential energy between particles 1 and 2 and $\beta \equiv$

$1/k_BT$. The form the bridge function $B(r_{12})$ takes is left to the imagination of the researcher; without its definition, the equation is underdetermined. Roughly, the closure relation states that the ICF exponentially decreases with the interaction and the TCF, while the bridge function offers the means by which the ICF may exponentially increase (or perhaps another cause of its exponential decrease depending on the sign of the bridge function). There are some arguments attempting to justify the introduction of the closure relation and the particular form it is given, including arguments for particular choices of bridge functions, but all these arguments bely an underlying truth: the closure relation is introduced artificially in an attempt to save a promising theory. Surprisingly, despite the artificiality of the closure relation, it allows the OZ equation to produce physically plausible results. Unsurprisingly, the quasi-physical nature of the closure relation leads to difficulty diagnosing and treating sources of error in the resulting OZ equation solutions. Thus the closure relation is a double-edged sword that must be treated with care in order to obtain useful results from the OZ theory.

This does not resolve the infinitely recursive convolution integral. Most OZ solution methods resolve this by solving the OZ equation in frequency space:

$$\tilde{h}(k) = \tilde{c}(k) + \rho\tilde{h}(k)\,\tilde{c}(k) \tag{1.4}$$

By the convolution theorem the convolution integral is converted to multiplication in frequency space, resolving the infinite recursion. Matrix algebra can then be used to solve the equation for the TCF.

The standard algorithm for solving the OZ equation uses iterative convergence:

1. Guess the values for one of the correlation functions (typically the DCF).
2. Calculate one of the other correlation functions using the closure relation.
3. Calculate the remaining correlation function using the OZ equation.
4. Check for consistency between correlation functions holds within an error tolerance ξ (e.g., $h(r_{12}) - c(r_{12}) - t(r_{12}) \leq \xi$).

5. If the error greater than the tolerance, repeat the process using a newly computed correlation function guess using the calculated values of other correlation functions.

Many practical computational problems must be solved to make this approach tenable, but in principle it allows solving the OZ equation to obtain the values of the associated correlation functions, which can in turn be used to compute solvent distributions and solvation thermodynamics.

1.4 RISM

The most popular form of the OZ equation is the **molecular Ornstein-Zernike (MOZ)** theory, which treats molecules as particles by assigning them molecular correlation functions. Unfortunately MOZ is computationally expensive to solve, in large part due to its six degrees of freedom (relative position and orientation of interacting molecules). One way to reduce the cost is by averaging the orientational contribution to the correlation functions so that only a radial component remains, then assuming that the molecular correlation functions are a linear combination of atomic correlation functions. Doing so gives rise to the **reference interaction site model** (**RISM**), which is the primary theoretical model used in this work. The RISM retains many of the theoretical benefits of the OZ, including producing spatial particle density distributions, while being significantly cheaper to compute. The RISM can make use of many of the same closure relations as the OZ equation, including the previously mentioned KH, HNC, and PSE-n closures. In addition, similar algorithms can be employed to obtain its solution.

1.4.1 1D-RISM

The derivation of the RISM equation begins with the MOZ equation, which is very similar to the OZ equation, but includes a dependence on relative molecular orientation

$$h\left(\mathbf{r}_{AB}, \mathbf{\Omega}_A, \mathbf{\Omega}_B\right) = c\left(\mathbf{r}_{AB}, \mathbf{\Omega}_A, \mathbf{\Omega}_B\right) + \frac{\rho}{8\pi^2} \int \int_{-\infty}^{\infty} d\mathbf{r}_C d\mathbf{\Omega}_C \; c\left(\mathbf{r}_{AC}, \mathbf{\Omega}_A, \mathbf{\Omega}_C\right) h\left(\mathbf{r}_{CB}, \mathbf{\Omega}_C, \mathbf{\Omega}_B\right)$$

where A, B, and C are molecules.

The RISM can be obtained by two operations.

First, the molecule-molecule TCF of MOZ is averaged over its orientations at a fixed intermolecular distance, resulting in a site-site TCF:

$$h_{12}(\mathbf{r}) = \frac{1}{\Omega^2} \int d\Omega_A d\Omega_B \delta\left(\mathbf{r}_1^A\right) \delta\left(\mathbf{r}_2^B - \mathbf{r}\right) h_{AB}\left(r_{AB}, \Omega_A, \Omega_B\right)$$

where A and B are molecules and Ω is the number of angles (2π per angular degree of freedom in the case of continuous angular integrals). The deltas effectively place site 1 at the origin and site 2 at r, such that the contributions of all orientations and intermolecular distances which maintain the site-site distance are averaged.

Second, the fundamental approximation of RISM is that the molecule-molecule DCF is the sum of its respective atomic site-site DCFs

$$c_{AB}(r_{AB}) = \sum_{1 \in A, 2 \in B} c_{12}(r_{12}) \tag{1.5}$$

Combining the site-site approximation and the orientational averaging with the OZ equation, the site-site Ornstein Zernike (SSOZ) equation, also called the one-dimensional reference interaction site model (1D-RISM) equation, is obtained [12]:

$$h_{12}(r) = \sum_{3,4} \omega_{13}(r) * c_{34}(r) * \omega_{42}(r) + \rho \sum_{3,4} \omega_{13}(r) * c_{34}(r) * h_{42}(r) \tag{1.6}$$

where L_{ab} is the bond length between sites a and b, δ_{ab} is unity if site a and b are the same species and nil otherwise, and ω is the intramolecular correlation matrix

$$\omega_{12}(r) = \delta_{12}\delta(r) + \frac{(1 - \delta_{12})}{4\pi L_{12}^2} \delta_{12}(r - L_{12})$$

This matrix is equal to unity for correlation of particles with their own species at zero distance, $1/4\pi L_{12}^2$ when the species are different and separated by a distance L_{ab}, and nil otherwise. It represents the rigid molecular structure of the solvent in matrix form so that the inter-atomic distances have influence on the resulting TCF when solving the RISM equation. A more complete discussion of this term is given in [30].

As with the OZ equation, in practice the 1D-RISM is solved in k-space since it simplifies calculation of the convolutions:

$$\tilde{h}_{12}(k) = \sum_{3,4} \tilde{\omega}_{13}(k)\,\tilde{c}_{34}(k)\,\tilde{\omega}_{42}(k) + \rho \sum_{3,4} \tilde{\omega}_{13}(k)\,\tilde{c}_{34}(k)\,\tilde{h}_{42}(k) \tag{1.7}$$

In this work the primary importance of the 1D-RISM is that its solution can be used to calculate the solvent-solvent susceptibility function of a solvent system at a fixed temperature and solvent density:

$$\chi^{vv}_{\alpha\gamma}(r) = \omega^{vv}_{\alpha\gamma}(r) + \rho^{v}_{\alpha} h^{vv}_{\alpha\gamma}(r) \tag{1.8}$$

where $\omega^{vv}_{\alpha\gamma}(r)$ is the solvent intramolecular coordination function (a matrix which models the solvent molecular geometry by being zero everywhere except where r is the distance between two sites in the same solvent molecule), and ρ^{v}_{α} is the solvent site bulk density.

The utility of $\chi^{vv}_{\alpha\gamma}(r)$ is revealed in the next section.

1.4.2 3D-RISM

If instead of averaging the orientations of both molecules in the OZ equation, the position and orientaiton of the central molecule is fixed and only the orientational freedom of the test molecule is averaged:

$$h_{12}(\mathbf{r}) = \frac{1}{\Omega} \int d\Omega_B dr_{AB} \delta\left(\mathbf{r}_1^A\right) \delta\left(\mathbf{r}_2^B - \mathbf{r}\right) h_{AB}(r_{AB}, 0, \Omega_B)$$

while the RISM approximation is assumed the same as for the 1D-RISM (see equation (1.5)), then the 3D-RISM equation is obtained [6, 46, 48, 45]:

$$h^{uv}_{\gamma}(\mathbf{r}) = \sum_{\alpha} \int d\mathbf{r}'\, c^{uv}_{\alpha}\left(\mathbf{r} - \mathbf{r}'\right) \chi^{vv}_{\alpha\gamma}\left(r'\right) \tag{1.9}$$

The superscript uv indicates an interaction between a solute molecule u and solvent molecule v, the subscripts α and γ indicate a given solvent site within the solvent molecule v, $\mathbf{r}$ and $\mathbf{r}'$ are Cartesian position vectors, $h^{uv}_{\gamma}(\mathbf{r})$ is the total correlation correlation function (TCF) (related to the radial distribution function (RDF) by $g^{uv}_{\alpha}(\mathbf{r}) = h^{uv}_{\alpha}(\mathbf{r}) + 1$), $c^{uv}_{\alpha}(\mathbf{r})$ is the direct correlation function (DCF) (which is asymptotically proportional to the potential, $c^{uv}_{\alpha}(\mathbf{r}) \propto -u^{uv}_{\alpha}(\mathbf{r}) / (k_B T)$), and $\chi^{vv}_{\alpha\gamma}$ is the solvent-solvent susceptibility function

obtained from equation (1.8) using the 1D-RISM. Note that $\chi^{vv}_{\alpha\gamma}$ implicitly contains influence from the intramolecular correlation matrix and hence introduces the influence of the atomic structure of the solvent into the 3D-RISM equation. It is this structural influence which primarily separates the 3D-RISM solvation model from other solvation models which almost universally ignore solvent structure both when calculating thermodynamics and solvent distributions.

The 3D-RISM is defined on a 3D grid due to the fixed orientation of the central molecule, hence the "3D" in the 3D-RISM. By Fourier transform the 3D-RISM can be expressed in k-space which allows the convolution integral to be more efficiently computed as simple multiplication:

$$\hat{h}_\alpha(\mathbf{k}) = \sum_\gamma \hat{c}_\gamma(\mathbf{k})\,\hat{\chi}^{\mathrm{VV}}_{\gamma\alpha}(k) \tag{1.10}$$

The algorithm for solving the 3D-RISM equation is almost identical to the one previously outlined for solving the OZ equation. Just as with the OZ equation, a closure equation is needed in order to solve the 3D-RISM for the DCF and TCF. Solvation thermodynamics and forces can then be directly computed using the TCF and DCF obtained from solving the 3D-RISM equation for a particular closure.

1.5 Closures

The OZ equation requires an expression for the closure equation in order to calculate a solution and provide an expression for the excess chemical potential and associated thermodynamic variables. The most popular closure equations are the hypernetted chain (HNC)[62], Kovalenko-Hirata (KH)[46] and the partial series expansion of order-n[41]. Since the results for the HNC and KH equations can be obtained from PSE-n when $n = \infty$ and $n = 1$ respectively, only the PSE-n will be considered here. Temperature derivatives of all closures are in table 2.2.

The PSE-n expression for the closure is given as

$$\mathbf{g}(\mathbf{r}) = \begin{cases} \exp(\mathbf{t}^*(\mathbf{r})) & \mathbf{t}^*(\mathbf{r}) < 0 \\ \sum_{i=0}^{n} \frac{\mathbf{t}^*(\mathbf{r})^i}{i!} & \mathbf{t}^*(\mathbf{r}) \geq 0 \end{cases} \tag{1.11}$$

$$\mathbf{t}^*(\mathbf{r}) \equiv -\beta \mathbf{u}(\mathbf{r}) + \mathbf{h}(\mathbf{r}) - \mathbf{c}(\mathbf{r})$$

The PSE-n closure assumes that the general form of the closure relation with the bridge function set to zero is accurate when $\mathbf{t}^*(\mathbf{r}) < 0$ (and hence $\mathbf{t}(\mathbf{r}) < 1$), but when $\mathbf{t}^*(\mathbf{r}) \geq 0$ (i.e., $\mathbf{t}(\mathbf{r}) \geq 1$) a partial series expansion of the exponential is more numerically stable. Surprisingly this relatively simple closure and its aforementioned relatives have become the dominant closure model due to their simplicity and success in applications. As will be shown in Chapter chapter 2, there are known physical deficiencies inherent to these closures which manifest themselves in a number of ways, yet so far no other published closure relation has managed to achieve such broad utility.

Given a specific closure equation, thermodynamic quantities can be computed for the 3D-RISM. For example, the corresponding excess chemical potential is

$$\Delta\mu^{\text{PSE-}n} = kT \sum_{\gamma} \rho_\gamma \int \frac{h_\gamma^2(\mathbf{r})}{2} - c_\gamma - \frac{h_\gamma(\mathbf{r}) c_\gamma(\mathbf{r})}{2} - \frac{t_\gamma^*(\mathbf{r})^{n+1}}{(n+1)!} \Theta(h_\gamma(\mathbf{r})) \mathrm{d}\mathbf{r} \tag{1.12}$$

where Θ is the Heaviside step function and γ is a solvent site.

The remainder of this work will attempt apply the 3D-RISM and its closures to practical problems involving solvation. Chapter 2 focuses on calculating accurate solvation energies for small drug-like molecules, while Chapter 3 introduces a new version of 3D-RISM extended to periodic solute and applies it to crystal structure refinement.

Chapter 2

Correction of 3D-RISM Solvation Thermodynamics for Small Molecules

Accurate solvation free energies and entropies are critical to correctly predicting and understanding the outcome of most clinically relevant biochemical processes, including drug binding affinity and reaction rates. During drug development, experimental measurement of solvation thermodynamics is cost prohibitive due to the large number of candidate molecules that must be synthesized and tested. This has led to interest in calculating solvation thermodynamics using computer simulations. Unfortunately the most accurate simulation methods which use explicit atomic models of the solvent, such as molecular dynamics (MD) and ab initio quantum mechanical methods (QM), are computationally expensive and often take weeks or months to complete a single calculation [88, 73, 42, 59].

In response, faster, less accurate simulation methods have been developed which model the solvent as an implicit continuum, such as the generalized Born (GB) [83] and Poisson-Boltzmann (PB) [32, 5] methods. While PB in particular has had some success in predicting experimental and MD solvation free energies, it is unable to predict the location of the solvent molecules about the solute, and generally yields rather poor estimates for temperature derivatives [74, 35]. The spatial distribution of solvent molecules is often critical to understanding how solvation affects a particular reaction and can help improve the design of drug candidates. Statistical mechanical methods from liquid-state theory [29], such as density functional theory (DFT) [21, 93] or integral equation theories like molecular Ornstein-Zernike (MOZ) [8, 37, 38] and the reference interaction site model (RISM) [12, 31, 67], fill the gap between explicit and implicit solvent simulations.

These methods typically make use of an atomic model of the solvent without explicitly modeling its motion in solution, allowing the methods to predict accurate solvent distributions, similar to explicit solvent models, while retaining the relatively low computational cost of implicit solvent models.

One promising integral equation method is the 3D-RISM [47, 7, 71], which extends the RISM to calculate three-dimensional solvent distributions about a solute, at a fraction of the computational expense of explicit solvent simulations. A known limitation of the 3D-RISM is its poor agreement with experimental solvation energies for small neutral molecules. Linear corrections for HNC-like closures have been proposed which increase the accuracy of the 3D-RISM solvation energies to be comparable with those of explicit solvent MD. Two such linear corrections are the Universal Correction (UC) [64, 70] and the Ng bridge correction (NgB) [85], both having correction terms related to the partial molar volume of the solvent. These linear corrections have found application both to DFT theories and the 3D-RISM. Though useful, these corrections have not yet been satisfactorily explained on physical grounds and require experimental parameterization to obtain best results. Recent work by Sergiievsky *et al.* [80, 60] has lead to a parameter free correction of similar quality to UC and NgB for which several physical explanations have been proposed [80, 15, 52, 79]. Gaining physical insight into why these corrections are needed may provide a deeper understanding of integral equation methods and point towards an analytic means of increasing their accuracy beyond what is possible with *ad hoc* corrections.

In this work we introduce 3D-RISM as a practical method to calculate solvation enthalpies and entropies. Used in combination with the aforementioned corrections, good quantitative agreement with experiment is achieved. Previously, the only way to obtain such a decomposition was using exceptionally taxing MD simulations [68, 26, 33], but now the 3D-RISM can be used to calculate accurate solvation energies and entropies for small molecules in a fraction of the time. Further, when compared to experiment, the decomposition indicates that the linear solvation energy corrections are mostly correcting the entropic term. Possible implications of this on the physical basis for the linear corrections and its relation to HNC-like closures will be discussed. These results provide insights into

the physical realism of the 3D-RISM and suggest a path a to further improvements in the method.

2.1 Theory

2.1.1 Energy and Entropy

Decomposition of excess chemical potentials into energy and entropic contributions using temperature derivatives comes from Yu, Roux and Karplus [91, 92] and has been previously applied in a few applications [14, 89, 90].

In the canonical ensemble, the excess chemical potential, $\Delta\mu$, due to a solvent site α is composed of the excess partial molar entropy, $\Delta s_{T,V}$, and partial molar total system energy, $\Delta\epsilon_{T,V}$[92]

$$\Delta\mu_\alpha = \Delta\epsilon_{\alpha,T,V} - T\Delta s_{\alpha,T,V}. \tag{2.1}$$

To simplify the notation, we will assume a canonical ensemble and omit the T, V subscript from this point on. The entropy can be expressed as the temperature derivative of the the excess chemical potential,

$$T\Delta s_\alpha = -T\left(\frac{\partial\Delta\mu_\alpha}{\partial T}\right)_\rho = -\delta_T\Delta\mu_\alpha \tag{2.2}$$

where

$$\delta_T \equiv T\left(\frac{\partial}{\partial T}\right)_\rho. \tag{2.3}$$

Inserting equation (2.2) into equation (2.1) we have

$$\Delta\epsilon_\alpha = \Delta\mu_\alpha - \delta_T\Delta\mu_\alpha \tag{2.4}$$

and

$$-T\Delta s_\alpha = \Delta\mu_\alpha - \Delta\epsilon_\alpha.$$

The 3D-RISM can be used to calculate $\Delta\mu_\alpha$ (*e.g.*, using equation (1.12) for the PSE-n closure). The temperature derivative is then obtained by applying equation (2.3) to the 3D-RISM equation, equation (1.10), giving

$$\delta_T\hat{\mathbf{h}} = \{\delta_T\hat{\mathbf{c}}\}\,\hat{\boldsymbol{\chi}}^{\mathrm{VV}} + \hat{\mathbf{c}}\delta_T\hat{\boldsymbol{\chi}}^{\mathrm{VV}} \tag{2.5}$$

where $\delta_T \hat{\boldsymbol{\chi}}^{\mathrm{VV}} = \boldsymbol{\rho} \delta_T \hat{\mathbf{h}}^{\mathrm{VV}}$ is obtained from the 1D-RISM.

Applying equation (2.3) and the PSE-n expression of the excess chemical potential (equation (1.12)) gives

$$\begin{aligned}
\Delta\epsilon^{\text{PSE-}n} &= \Delta\mu^{\text{PSE-}n} - \delta_T \Delta\mu^{\text{PSE-}n} \\
&= -kT \sum_\gamma \rho_\gamma \int h_\gamma(\mathbf{r})\, \delta_T h_\gamma(\mathbf{r}) - \delta_T c_\gamma(\mathbf{r}) \\
&\qquad - \frac{1}{2}\left[\{\delta_T h_\gamma(\mathbf{r})\}\, c_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r})\, \delta_T c_\gamma(\mathbf{r})\right] \\
&\qquad - \frac{t_\gamma^{*\,n}(\mathbf{r})}{n!}\left[\beta u(\mathbf{r}) + \delta_T h_\gamma(\mathbf{r}) - \delta_T c_\gamma(\mathbf{r})\right] \Theta\left(h_\gamma(\mathbf{r})\right) d\mathbf{r}.
\end{aligned} \tag{2.6}$$

The temperature derivative of the closure, necessary to solve the temperature derivative integral equation, is

$$\begin{aligned}
\delta_T \mathbf{h}(\mathbf{r}) = \delta_T \mathbf{g}(\mathbf{r}) &= \begin{cases} \mathbf{g}(\mathbf{r})\, \delta_T t^*(\mathbf{r}) & \mathbf{t}^*(\mathbf{r}) < 0 \\ \sum_{i=0}^{n-1} \dfrac{\mathbf{t}^*(\mathbf{r})^i}{i!} \delta_T t^*(\mathbf{r}) & \mathbf{t}^*(\mathbf{r}) \geq 0 \end{cases} \\
\delta_T \mathbf{t}^*(\mathbf{r}) &= \beta \mathbf{u}(\mathbf{r}) + \delta_T \mathbf{h}(\mathbf{r}) - \delta_T \mathbf{c}(\mathbf{r}).
\end{aligned} \tag{2.7}$$

2.1.2 Solvation Energy Corrections

While the excess chemical potential given by equation (1.12) is consistent with the PSE-n closure, in practice, solvation free energies calculated from this expression are too high, which has been linked to the non-polar component[16, 24, 40]. In response to this, a number corrections have been proposed that use a modified form of the excess chemical potential while leaving the predicted solvent distributions unchanged.

For brevity, we will focus on the Universal Correction (UC) [65] and initial state correction (ISc) [80, 60]. Details of the Gaussian fluctuations correction (GF) [13, 36], and the Ng Bridge Correction (NgB) [85] are presented in 4.1. Temperature derivatives of these corrected free energy expressions yield an expression for the solvation energy, much like equation (2.6). Expressions for all these corrections can be found in table 2.1 and table 2.2.

Closure	Closure Relation
KH	$\mathbf{g}(\mathbf{r}) = \begin{cases} \exp(\mathbf{t}^*(\mathbf{r})) & \mathbf{t}^*(\mathbf{r}) < 0 \\ \mathbf{1} + \mathbf{t}^*(\mathbf{r}) & \mathbf{t}^*(\mathbf{r}) \geq 0 \end{cases}$
HNC	$g(\mathbf{r}) = \exp(\mathbf{t}^*(\mathbf{r}))$
PSE-n	$\mathbf{g}(\mathbf{r}) = \begin{cases} \exp(\mathbf{t}^*(\mathbf{r})) & \mathbf{t}^*(\mathbf{r}) < 0 \\ \sum_{i=0}^{n} \frac{\mathbf{t}^*(\mathbf{r})^i}{i!} & \mathbf{t}^*(\mathbf{r}) \geq 0 \end{cases}$
	$\mathbf{t}^*(\mathbf{r}) = -\beta\mathbf{u}(\mathbf{r}) + \mathbf{h}(\mathbf{r}) - \mathbf{c}(\mathbf{r})$
Closure	**Excess Chemical Potential**
KH	$\Delta\mu^{\mathrm{KH}} = kT\sum_\gamma \rho_\gamma \int \frac{h_\gamma^2(\mathbf{r})}{2}\Theta(-h_\gamma(\mathbf{r})) - c_\gamma(\mathbf{r}) - \frac{h_\gamma(\mathbf{r})c_\gamma(\mathbf{r})}{2}\mathrm{d}\mathbf{r}$
HNC	$\Delta\mu^{\mathrm{HNC}} = kT\sum_\gamma \rho_\gamma \int \frac{h_\gamma^2(\mathbf{r})}{2} - c_\gamma(\mathbf{r}) - \frac{h_\gamma(\mathbf{r})c_\gamma(\mathbf{r})}{2}\mathrm{d}\mathbf{r}$
PSE-n	$\Delta\mu^{\text{PSE-}n} = kT\sum_\gamma \rho_\gamma \int \frac{h_\gamma^2(\mathbf{r})}{2} - c_\gamma(\mathbf{r}) - \frac{h_\gamma(\mathbf{r})c_\gamma(\mathbf{r})}{2} - \frac{t_\gamma^*(\mathbf{r})^{n+1}}{(n+1)!}\Theta(h_\gamma(\mathbf{r}))\mathrm{d}\mathbf{r}$
Correction	**Excess Chemical Potential**
GF	$\Delta\mu^{\mathrm{GF}} = kT\sum_\gamma \rho_\gamma \int -c_\gamma(\mathbf{r}) - \frac{h_\gamma(\mathbf{r})c_\gamma(\mathbf{r})}{2}\mathrm{d}\mathbf{r}$
UCT	$\Delta\mu^{\mathrm{UC}} = \Delta\mu^{\mathrm{RISM}} + av + b$
NgBT	$\Delta\mu^{\mathrm{NgB}} = \Delta\mu^{\mathrm{RISM}} + \frac{kT\rho_{\mathrm{O}}}{2}(1-\gamma)\int c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\,d\mathbf{r}$
ISc	$\Delta\mu^{\mathrm{ISc}} = \Delta\mu^{\mathrm{RISM}} - \frac{1}{2}kTv\left(\frac{1}{\chi_T kT} + \rho_{\mathrm{Tot}}\right)$
ISc*	$\Delta\mu^{\mathrm{ISc}^*} = \Delta\mu^{\mathrm{RISM}} - \frac{1}{2}kTv\left(\frac{1}{\chi_T kT} - \rho_{\mathrm{Tot}}\right)$

Table 2.1: Closure expressions and excess chemical potential equations for various common closures and corrections.

Closure	Temperature Derivative
KH	$\delta_T \mathbf{h}(\mathbf{r}) = \begin{cases} \mathbf{g}(\mathbf{r})\,\delta_T \mathbf{t}^*(\mathbf{r}) & \text{for } \mathbf{t}^* < 0 \\ \delta_T \mathbf{t}^*(\mathbf{r}) & \text{for } \mathbf{t}^* \geq 0 \end{cases}$
HNC	$\delta_T \mathbf{h}(\mathbf{r}) = \mathbf{g}(\mathbf{r})\,\delta_T t^*(\mathbf{r})$
PSE-n	$\delta_T \mathbf{h}(\mathbf{r}) = \begin{cases} \mathbf{g}(\mathbf{r})\,\delta_T t^*(\mathbf{r}) & \mathbf{t}^*(\mathbf{r}) < 0 \\ \sum_{i=0}^{n-1} \frac{\mathbf{t}^*(\mathbf{r})^i}{i!} \delta_T t^*(\mathbf{r}) & \mathbf{t}^*(\mathbf{r}) \geq 0 \end{cases}$
	$\delta_T \mathbf{t}^*(\mathbf{r}) = \beta \mathbf{u}(\mathbf{r}) + \delta_T \mathbf{h}(\mathbf{r}) - \delta_T \mathbf{c}(\mathbf{r})\,.$

Closure	Excess Solvation Energy
KH	$\Delta\epsilon^{\mathrm{KH}} = -kT \sum_\gamma \rho_\gamma \int h_\gamma(\mathbf{r})\,\delta_T h_\gamma(\mathbf{r})\,\Theta(-h_\gamma(\mathbf{r})) - \delta_T c_\gamma(\mathbf{r}) - \frac{1}{2}\left[\{\delta_T h_\gamma(\mathbf{r})\}\, c_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r})\,\delta_T c_\gamma(\mathbf{r})\right] d\mathbf{r}$
HNC	$\Delta\epsilon^{\mathrm{HNC}} = -kT \sum_\gamma \rho_\gamma \int h_\gamma(\mathbf{r})\,\delta_T h_\gamma(\mathbf{r}) - \delta_T c_\gamma(\mathbf{r}) - \frac{1}{2}\left[\{\delta_T h_\gamma(\mathbf{r})\}\, c_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r})\,\delta_T c_\gamma(\mathbf{r})\right] d\mathbf{r}$
PSE-n	$\Delta\epsilon^{\text{PSE-}n} = -kT \sum_\gamma \rho_\gamma \int h_\gamma(\mathbf{r})\,\delta_T h_\gamma(\mathbf{r}) - \delta_T c_\gamma(\mathbf{r}) - \frac{1}{2}\left[\{\delta_T h_\gamma(\mathbf{r})\}\, c_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r})\,\delta_T c_\gamma(\mathbf{r})\right] - \frac{t_\gamma^{*n}(\mathbf{r})}{n!}\left[\beta u(\mathbf{r}) + \delta_T h_\gamma(\mathbf{r}) - \delta_T c_\gamma(\mathbf{r})\right]\Theta(h_\gamma(\mathbf{r}))\, d\mathbf{r}$

Correction	Excess Solvation Energy
GF	$\Delta\epsilon^{\mathrm{GF}} = -kT \sum_\gamma \rho_\gamma \int -\delta_T c_\gamma(\mathbf{r}) - \frac{1}{2}\left[\{\delta_T h_\gamma(\mathbf{r})\}\, c_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r})\,\delta_T c_\gamma(\mathbf{r})\right] d\mathbf{r}$
UCT	$\Delta\epsilon^{\mathrm{UCT}} = \Delta\epsilon^{\mathrm{RISM}} + a(v - \delta_T v) - Ta_1 v + b_0$
NgBT	$\Delta\epsilon^{\mathrm{NgB}} = \Delta\epsilon^{\mathrm{RISM}} - \frac{kT\rho_{\mathrm{O}}}{2}\left\{(1-\gamma)\int \delta_T c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\, d\mathbf{r} - \gamma_1 T \int c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\, d\mathbf{r}\right\}$
ISc	$\Delta\epsilon^{\mathrm{ISc}} = \Delta\epsilon^{\mathrm{RISM}} + \frac{1}{2}kT\delta_T v\left(\frac{1}{\chi_T kT} + \rho_{\mathrm{Tot}}\right) - \frac{1}{2}kTv\left(\frac{1}{\chi_T kT}\right)^2 (\chi_T kT + \delta_T \chi_T kT)$
ISc*	$\Delta\epsilon^{\mathrm{ISc}^*} = \Delta\epsilon^{\mathrm{RISM}} + \frac{1}{2}kT\delta_T v\left(\frac{1}{\chi_T kT} - \rho_{\mathrm{Tot}}\right) - \frac{1}{2}kTv\left(\frac{1}{\chi_T kT}\right)^2 (\chi_T kT + \delta_T \chi_T kT)\,.$

Table 2.2: Closure expression temperature derivatives and excess solvation energy equations for various common closures and corrections.

2.1.2.1 Universal Correction

The Universal Correction (UC) is a simple empirical correction to the RISM excess chemical potential [65],

$$\Delta\mu^{\mathrm{UC}} = \Delta\mu^{\mathrm{RISM}} + av + b \tag{2.8}$$

where a and b are parameterized from experimental data and v, the partial molar volume (PMV), is calculated from equation (2.9). In the original presentation, the Gaussian fluctuation approximation (GF) [13, 36] was used for $\Delta\mu^{\mathrm{RISM}}$, but subsequent studies have used the closure specific excess chemical potential with improved results [85, 34]. In what follows either the closure specific or Gaussian fluctuation approximation may be used (see 4.1.1 for details of the GF temperature derivative). The parameterization must be repeated for any change in solvent composition, temperature or density. As will be revealed in the results seciton, the expression shows considerable improvement for small, non-polar molecules in pure water.

[14] provide the following convenient expression for the PMV

$$v = k_{\mathrm{B}}T\chi_T \left(1 - \sum_{\gamma}\rho_\gamma \int c_\gamma\left(\mathbf{r}\right) d\mathbf{r}\right) \tag{2.9}$$

where χ_T is the isothermal compressibility for the bulk solvent, calculated as[20, 57, 28]

$$\chi_T = \frac{\beta}{\rho_{\mathrm{Tot}} - \sum_\alpha \sum_\gamma \rho_\alpha \rho_\gamma \hat{c}_{\alpha\gamma}\left(0\right)}. \tag{2.10}$$

For uncorrected excess chemical potential, the non-polar component, $\Delta\mu^{\mathrm{RISM}}_{\mathrm{NP}}$, can be obtained by setting all partial charges to zero. The polar component is then

$$\Delta\mu^{\mathrm{RISM}}_{\mathrm{Pol}} = \Delta\mu^{\mathrm{RISM}} - \Delta\mu^{\mathrm{RISM}}_{\mathrm{NP}}.$$

For UC, the non-polar component is

$$\Delta\mu^{\mathrm{UC}}_{\mathrm{NP}} = \Delta\mu^{\mathrm{RISM}}_{\mathrm{NP}} + av_{\mathrm{NP}} + b$$

where v_{NP} is the partial molar volume of the chargeless solute. The polar component is computed as

$$\begin{aligned}\Delta\mu^{\mathrm{UC}}_{\mathrm{Pol}} &= \Delta\mu^{\mathrm{UC}} - \Delta\mu^{\mathrm{UC}}_{\mathrm{NP}} \\ &= \Delta\mu^{\mathrm{RISM}}_{\mathrm{Pol}} + av_{\mathrm{Pol}}.\end{aligned}$$

Since solvent polarization component of the partial molar volume, v_{Pol}, is relatively small, the polar component of the excess chemical potential is only slightly changed.

In the original formulation, a and b are constants with no temperature dependence. However, as we show in Results, including a linear temperature dependence for these coefficients,

$$\begin{aligned} a &= a_0 + a_1 T, \\ b &= b_0 + b_1 T, \end{aligned} \tag{2.11}$$

provides significantly improved results compared to experiment. Applying equation (4.5) to equation (2.8) we have

$$\delta_T \Delta\mu^{\text{UC}} = \delta_T \Delta\mu^{\text{RISM}} + a_1 T v + a \delta_T v + b_1 T \tag{2.12}$$

The temperature derivative of the PMV is

$$\delta_T v = v + kT \left\{\delta_T \chi_T\right\} \left(1 - \sum_\gamma \rho_\gamma \int c_\gamma \left(\mathbf{r}\right) d\mathbf{r}\right) - kT\chi_T \left(\sum_\gamma \rho_\gamma \int \left\{\delta_T c_\gamma\right\} d\mathbf{r}\right) \tag{2.13}$$

where $\delta_T \chi_T$ is also pre-calculated with DRISM,

$$\begin{aligned} \delta_T \chi_T \quad &= -\chi_T + \beta \left[\rho_{\text{Tot}} - \textstyle\sum_\alpha \sum_\gamma \rho_\alpha \rho_\gamma \hat{c}_{\alpha\gamma}\left(0\right)\right]^{-2} \left[\textstyle\sum_\alpha \sum_\gamma \rho_\alpha \rho_\gamma \int \left\{\delta_T c_{\alpha\gamma}\right\} d\mathbf{r}\right] \\ &= -\chi_T + \tfrac{\chi_T^2}{\beta} \left[\textstyle\sum_\alpha \sum_\gamma \rho_\alpha \rho_\gamma \int \left\{\delta_T c_{\alpha\gamma}\right\} d\mathbf{r}\right]. \end{aligned} \tag{2.14}$$

After some algebra, we have

$$\Delta\epsilon^{\text{UC}} = \Delta\epsilon^{\text{RISM}} + a\left(v - \delta_T v\right) - a_1 T v + b_0.$$

Including this form of temperature dependence does not change the fitting procedure to determine a and b needed for solvation free energies. Only two new parameters need to be fit, a_1 and b_0, and can determined by fitting against empirical enthalpies or entropies at a single temperature.

The Ng Bridge Correction (NgB) [85] is similar in spirit to UC but contains an explicit temperature dependence and only one free parameter. Details of NgB can be found in 4.1.2.

2.1.2.2 Initial State Correction

The so-called initial state correction (ISc) [80],

$$\Delta\mu^{\mathrm{ISc}} = \Delta\mu^{\mathrm{RISM}} - \left[\rho_{\mathrm{Tot}}kT - \frac{1}{2}\rho_{\mathrm{Tot}}^2 kT\hat{c}\left(k=0\right)\right]v, \tag{2.15}$$

and ISc*,

$$\Delta\mu^{\mathrm{ISc}^*} = \Delta\mu^{\mathrm{RISM}} + \frac{1}{2}\rho_{\mathrm{Tot}}^2 kT\hat{c}\left(k=0\right)v, \tag{2.16}$$

are similar to UC and NgB corrections, but differ in that they are analytic, parameter free corrections. There has been considerable discussion as to the physical meaning of these corrections and which is appropriate for 3D-RISM [60, 15, 79]. The factor in square brackets in equation (2.15) is the pressure of the solvent derived from the molecular density functional theory (MDFT) free energy expression [80],

$$P^{\mathrm{MDFT}} = \rho_{\mathrm{Tot}}kT - \frac{1}{2}\rho_{\mathrm{Tot}}^2 kT\hat{c}\left(k=0\right). \tag{2.17}$$

It is well known that the HNC family of closures overestimate solvent pressures by several orders of magnitude [28, 39] and the ISc correction can be seen as compensating for the additional work required to insert the solute into the liquid. With this in mind, [79] have derived

$$P^{\mathrm{3DRISM}} = \frac{N_{\mathrm{site}}+1}{2}\rho_{\mathrm{Tot}}kT - \frac{1}{2}\rho_{\mathrm{Tot}}^2 kT\hat{c}\left(k=0\right) \tag{2.18}$$

specifically for 3D-RISM using a density functional approach. N_{site} is the number of solvent sites and, for water, $N_{\mathrm{site}} = 3$, giving $P^{\mathrm{3DRISM}} = P^{\mathrm{MDFT}} + \rho_{\mathrm{Tot}}kT$. [79] argue that $P^{\mathrm{3DRISM}}v$ should be subtracted from $\Delta\mu^{\mathrm{RISM}}$ and the ideal part added back in, in which case, ΔG^{ISc} is recovered.

While it is tempting to interpret this as a pressure correction, P^{3DRISM} is not the true pressure predicted by RISM. The ideal gas contribution incorrectly depends on the number of internal degrees of freedom (N_{site}) of the solvent molecules. Furthermore, MDFT cannot be used to derive properties for 3D-RISM since there is no known way to obtain molecular $\rho\left(\mathbf{r}\right)$ from site-site distributions [28]. We also observe that both equation (2.17) and equation (2.18) are quite different from the RISM expression for pressure derived from

the free energy route [81]

$$P^{\mathrm{RISM}} = kT\rho + 2\pi kT \sum_{\alpha}\sum_{\gamma}\rho_{\alpha}\rho_{\gamma}\int\left[\frac{h_{\alpha\gamma}^{2}(r)}{2} - c_{\alpha\gamma}(r) - \frac{\left(t_{\alpha\gamma}^{*}(r)\right)^{n+1}}{(n+1)!}\Theta\left(t_{\alpha\gamma}^{*}(r)\right)\right]r^{2}\,dr$$
$$- \frac{kT}{(2\pi)^{2}}\int\left\{\ln\left[\det\left[\mathbf{1} - \rho\hat{\boldsymbol{\omega}}(k)\,\hat{\mathbf{c}}(k)\right]\right] + \mathrm{Tr}\left[\rho\hat{\boldsymbol{\omega}}(k)\,\hat{\mathbf{c}}(k)\left[\mathbf{1} - \rho\hat{\boldsymbol{\omega}}(k)\,\hat{\mathbf{c}}(k)\right]^{-1}\right]\right\}k^{2}\,dk.$$

The two expression share only two terms in common and, for water under ambient conditions, $P^{\mathrm{RISM}} \approx 0.6P^{\mathrm{MDFT}}$. The physical interpretation of equation (2.15) is still unclear but the correction has been demonstrated to greatly improve 3D-RISM hydration free energies for ambient conditions.

For the purposes of practical calculation, it is convenient to substitute in the expression, $\rho\hat{c}(k=0) = 1 - \frac{1}{\rho kT\chi_{T}}$, where χ_{T} is the isothermal compressibility of the solvent, giving

$$\Delta\mu^{\mathrm{ISc}} = \Delta\mu^{\mathrm{RISM}} - \frac{1}{2}kTv\left(\frac{1}{\chi_{T}kT} + \rho_{\mathrm{Tot}}\right) \tag{2.19}$$

and

$$\Delta\mu^{\mathrm{ISc}^{*}} = \Delta\mu^{\mathrm{RISM}} - \frac{1}{2}kTv\left(\frac{1}{\chi_{T}kT} - \rho_{\mathrm{Tot}}\right). \tag{2.20}$$

As with UC, the polar/non-polar decomposition is straightforward:

$$\Delta\mu_{\mathrm{Pol}}^{\mathrm{ISc}} = \Delta\mu_{\mathrm{Pol}}^{\mathrm{RISM}} - \tfrac{1}{2}kTv_{\mathrm{Pol}}\left(\tfrac{1}{\chi_{T}kT} + \rho_{\mathrm{Tot}}\right) \quad \text{and} \quad \Delta\mu_{\mathrm{NP}}^{\mathrm{ISc}} = \Delta\mu_{\mathrm{NP}}^{\mathrm{RISM}} - \tfrac{1}{2}kTv_{\mathrm{NP}}\left(\tfrac{1}{\chi_{T}kT} + \rho_{\mathrm{Tot}}\right).$$

The solvation energy is then

$$\Delta\epsilon^{\mathrm{ISc}} = \Delta\epsilon^{\mathrm{RISM}} + \frac{1}{2}kT\delta_{T}v\left(\frac{1}{\chi_{T}kT} + \rho_{\mathrm{Tot}}\right) - \frac{1}{2}kTv\left(\frac{1}{\chi_{T}kT}\right)^{2}\left(\delta_{T}\chi_{T}kT + \chi_{T}kT\right).$$

Similarly, for ISc* we have

$$\Delta\epsilon^{\mathrm{ISc}^{*}} = \Delta\epsilon^{\mathrm{RISM}} + \frac{1}{2}kT\delta_{T}v\left(\frac{1}{\chi_{T}kT} - \rho_{\mathrm{Tot}}\right) - \frac{1}{2}kTv\left(\frac{1}{\chi_{T}kT}\right)^{2}\left(\delta_{T}\chi_{T}kT + \chi_{T}kT\right).$$

2.2 Methods

2.2.1 Data Sets and Hydration Energy Data

Several sets of small molecule structures and their experimental hydration energies were obtained from previous publications for use in this work. These sets are labeled after the last name of one of their publication authors: Abagyan [10], Mobley [61], Rizzo [72], and

Palmer [65]. Only the Rizzo set contains ionic molecules. Additionally, a solute set of 9 alkali halide ions was created using parameters from [40]. These sets were combined into a single small molecule database, including duplicate molecules whose structures differ due to use of different relaxation techniques among published sets.

In total the small molecule database contains 1123 molecules, consisting of 1075 neutral molecules, 39 monovalent ionic molecules, and 9 monovalent monoatomic ions, all with associated experimental Gibbs energies of hydration. To allow decomposing the enthalpic and entropic contributions to the Gibbs energies, experimental enthalpic and entropic energies of hydration were collected from the literature. Due to the relative sparsity of experimental entropic and enthalpic energies of hydration, only 74 molecules have their full experimental energy decomposition data. Datasets are again labeled using the last name of the first author: Abraham [4] and Cabani [11] (59 neutral molecules), Fawcett [22] (7 monovalent ionic molecules), and Marcus [55, 56] (8 monovalent monoatomic molecules). All experimental values are reported as being measured in standard thermodynamic conditions with temperatures between 298 and 298.15 K.

2.2.2 Solute Preparation

Antechamber was used to assign partial charges to all molecular atoms using the AM1-BCC semi-empirical model and the Amber GAFF [87] force field parameters, except for alkali halides which used Joung and Cheatham TIP3P parameters [40]. No structural alterations were made to the published molecules. A small set of molecules whose 3D-RISM calculations failed to converge were not used for correction model fitting.

2.2.3 Hydration Free Energy Calculations

All RISM calculations were performed using a modified version of AmberTools 15. Modifications will be released as part of AmberTools 16.

Parameter	cSPC/E H Value	cSPC/E O Value
mass (u)	1.008	16
charge (e^+)	0.4238	-0.8476
Lennard-Jones ε (J/C)	0.01553	0.1553
Lennard-Jones $r_{\mathrm{min}}/2$ (Å)	0.654237952	1.7767
H-O bond length (Å)	1	
H-O-H bond angle (degrees)	109.47°	

Table 2.3: Parameters of water models used in 1D-RISM calculations.

2.2.3.1 1D-RISM Calculations

One 1D-RISM calculation was performed for each desired closure (KH, HNC, and PSE-3) for a total of three 1D-RISM calculations. Each calculation used the cSPC/E water model (see Table table 2.3) at 298 K on a simulation grid of 16,384 grid points separated by a grid spacing of 0.025 Å. Calculations were performed with a solvent dielectric of 78.497, water density of 55.345 M, and a target residual tolerance for the MDIIS solver set to 1E-12.

2.2.3.2 3D-RISM Calculations

The 3D-RISM calculations were performed for the KH, HNC, and PSE-3 closures. The 3D-RISM equation and its closure relation were solved on a 3D grid with infinite dilution boundary conditions. The simulation box was a cube with 30 Å side length and 0.3 Å grid spacing. Interaction potentials were given an infinite cutoff to avoid cutoff approximation error. The modified direct inversion of iterative subspaces (MDIIS) solver was used to increase the rate of convergence of the integral equation solution[49]. In order to overcome convergence problems with PSE-3 and HNC closure 'bootstrapping' was used. Here, a solution was obtained using a lower order closure and this solution was used as a starting point for solving the target closure. For PSE-3 the lower order closure was PSE-2, while for HNC both PSE-2 and PSE-3 were used as lower order closures.

2.2.4 Parameter Fitting

Correction model parameters were fitted by bootstrap ordinary least squares (OLS) linear regression using the Python statsmodels module (version 0.6.1) [78]. Parameter fitting

used only empirical data for neutral molecules from the Mobley, Abagyan, Rizzo and Palmer datasets. As with bootstrap analysis described below, the original data was re-sampled with replacement to obtain a new data set. An independent OLS fitting was performed on each resampled data set. The final values and confidence interval for each parameter was taken as the mean and standard error over all best fit parameters for each set or resampled data.

2.2.5 Model Testing

Testing of corrected and uncorrected expressions, with and without fit parameters, was done independently from parameter fitting and employed 1,000 rounds of bootstrap analysis and k-fold cross validation.

2.2.5.1 k-Fold Cross-Validation

The expected goodness of fit of each model, independent of the data used to train it, was estimated using k-fold cross-validation. To perform k-fold cross-validation, the sample is randomly divided into k equally sized subsamples. Each subsample is used once as test data for the model produced using the other $k - 1$ subsamples as training data. The statistics of the resulting k models and their test regressions form distributions which can be used to calculate the root mean squared error of the R^2 regression factor. The error in the R^2 factor provides a statistical estimate for the effect the particular small molecular solute sample has on goodness of fit of the correction regression. For this work $k = 10$ was chosen and the average taken over 1,000 10-fold cross-validations.

2.2.5.2 Bootstrap Analysis

Bootstrap regression analysis was used to obtain the confidence intervals for the fitted model parameters. In bootstrap analysis, a random sample of size N is obtained from the original sample, allowing the same sample member to be selected more than once (*i.e.*, sampling with replacement). Regression is performed using the resampled data. This procedure is repeated many times. The statistics of the resulting regression models form

Correction	a	a_0	a_1	b	b_0	b_1
UC(T)$^{\mathrm{KH}}$	$-0.1498(8)$	$0.009(7)$	$-0.00053(2)$	$-0.1(1)$	$-3.2(9)$	$0.010(3)$
UC(T)$^{\mathrm{PSE3}}$	$-0.1185(7)$	$0.032(7)$	$-0.00051(2)$	$-0.3(1)$	$-3.2(9)$	$0.010(3)$
UC(T)$^{\mathrm{HNC}}$	$-0.1186(7)$	$0.033(7)$	$-0.00051(2)$	$-0.2(1)$	$-3.3(9)$	$0.010(3)$
Correction	γ		γ_0		γ_1	
NgB(T)$^{\mathrm{KH}}$	$0.333(1)$		$0.38(1)$		$-0.00015(4)$	
NgB(T)$^{\mathrm{PSE3}}$	$0.366(1)$		$0.31(1)$		$0.00019(4)$	
NgB(T)$^{\mathrm{HNC}}$	$0.364(1)$		$0.31(1)$		$0.00020(4)$	

Table 2.4: Fit parameters for UC(T) and NgB(T) corrections. Standard error in the last digit is given in parentheses.

distributions which can be used to calculate the desired confidence intervals for the fitting parameters. In this work N was chosen to be equal to the original sample size and 1,000 resamples were taken.

2.3 Results

In total, eight different corrections were tested with three different closures. Of these, three corrections performed particularly well against all closures: UCT, NgB and ISC. As will be discussed, UCGF(T) and ISC* only performed well for the KH closure. Despite adding temperature dependence to the fit coeffcient, NgBT did no better than NgB. This can be seen in parameterization (see table 2.4) where γ_0 is the dominant contribution to γ. In the case of UCT, the addition of temperature dependence to the coefficient is necessary and a_1 is the dominant contribution to a at room temperature. In this, UCT, NgB and ISC all give a linear temperature dependence to the PMV correction.

None of the three closures used performed significantly better or worse overall. PSE-3 generally has the good agreement with experiment and molecular dynamics and typically has the same convergence properties as KH when the closure bootstrapping protocol described in Methods is used. For this reason, our discussion focuses on PSE-3 but complete results for all closures and corrections can be found in Correction of 3D-RISM solvation thermodynamics for small drug-like molecules.

	ΔG				
	Slope	y-intercept	R^2	RMSE	MUE
PSE-3	1.20(8)	20.1(4)	0.305(1)	20.277(7)	19.537(8)
UC^{PSE3}	0.96(2)	−0.23(6)	0.8407(4)	1.261(2)	0.917(1)
NgB^{PSE3}	1.12(2)	0.20(6)	0.8509(4)	1.448(3)	1.037(1)
ISc^{PSE3}	0.96(2)	−0.75(6)	0.8322(4)	1.432(2)	1.053(1)
MD	0.99(2)	0.64(5)	0.8865(3)	1.249(1)	1.025(1)

Table 2.5: Bootstrap statistical comparison between predicted and empirical hydration free energies for neutral molecules (Mobley, Abagyan, Rizzo and Palmer datasets). As described in Methods, values are the mean of all resampled data. RMSE: root-mean-squared-error. MUE: mean unsigned error. Standard error in the last digit is given in parentheses.

2.3.1 Hydration Free Energies

As expected, uncorrected 3D-RISM provides poor predictions of the SFE of small neutral molecules for all closures (for PSE-3, see figure 2.1 and table 2.5 and for all closures, see table 4.2). R^2 ranged from 0.218(1) to 0.305(1) for the three closures and the y-intercept, MUE and RMSE were approximately 20 kcal/mol or higher.

UC^{PSE-3}, NgB^{PSE-3} and ISC^{PSE-3} all compare favorably to experiment and are statistically quite close to MD. Results for UC^{PSE-3} and ISC^{PSE-3} are extremely close for the PSE-3 closure and have identical slopes. The errors are, overall all, lower for UC^{PSE-3}, which is to be expected since UC^{PSE-3} has been fit to the data. However, while ISC performs well for PSE-3 and HNC, it is significantly worse with the KH closure, giving errors more than 2 kcal/mol (see table 4.2). ISC*, on the other hand, performs best for the KH closure, significantly outperforming ISC, but shows significant errors for PSE-3 and HNC. UCGF, the original correction proposed by [65], gives results similar to UC for KH but shows large systematic errors for higher order closures with slopes of 1.30.

There is no clear best overall correction and closure combination, though UC has the lowest errors across all closures. Corrections which should be avoided include UCGF for all closures (even for KH, UC is better), as well as ISC*. Using ISC for KH is also not recommended due to the relatively large errors.

UC^{PSE-3}, NgB^{PSE-3} and ISC^{PSE-3} all capture the slope of ionic solutes approximately as well as they do for neutral solutes and have improve R^2 values. In absolute terms, the

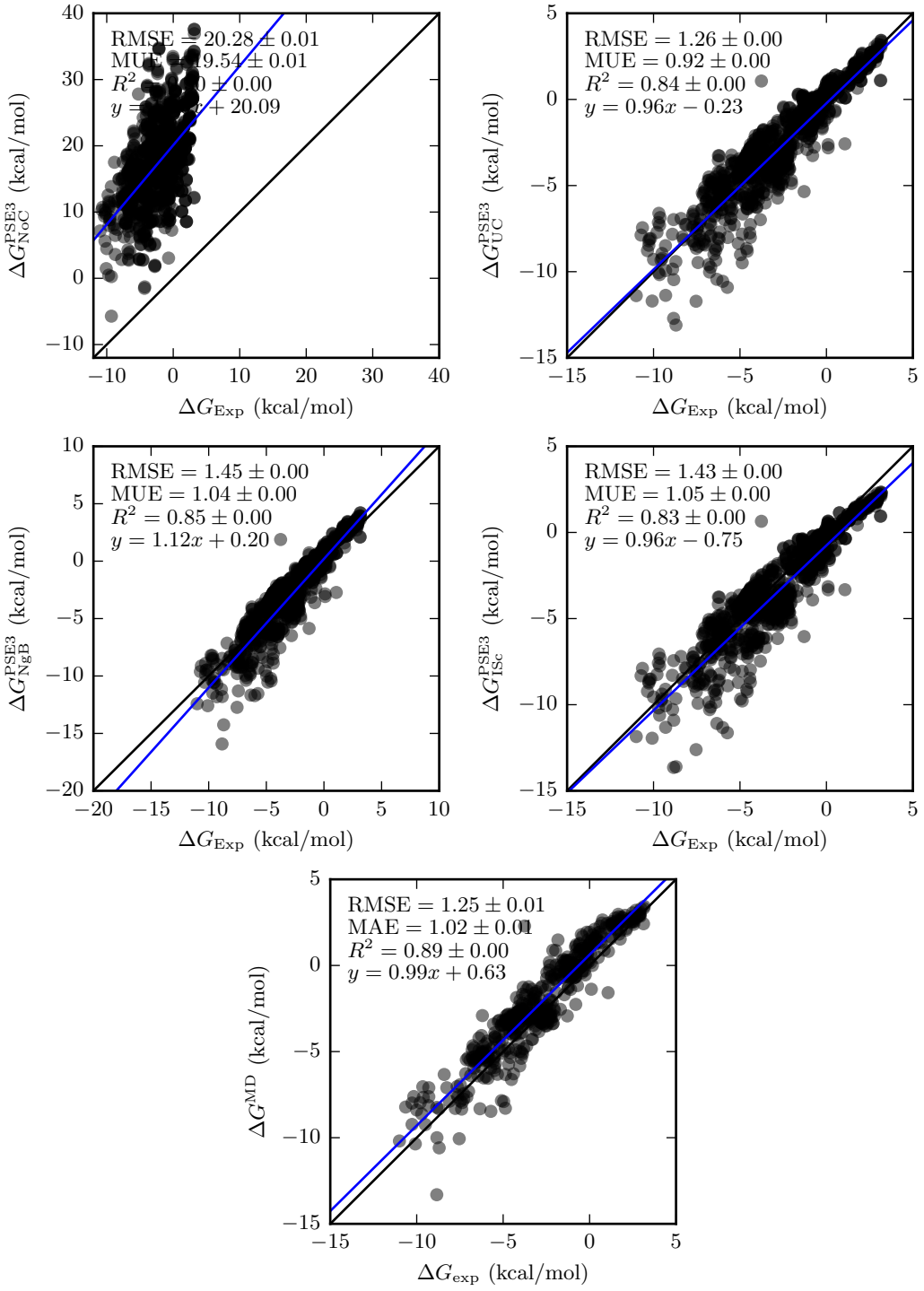

Figure 2.1: Hydration free energies of neutral molecules (semi-transparent circles) from 3D-RISM-PSE-3 and MD vs. experiment (Mobley, Abagyan, Rizzo and Palmer datasets).

	ΔG				
	Slope	y-intercept	R^2	RMSE	MUE
PSE-3	1.14(6)	17(5)	0.8952(8)	10.62(3)	9.26(2)
UC^{PSE3}	0.93(5)	−6(4)	0.8915(9)	6.53(2)	4.87(2)
NgB^{PSE3}	1.05(6)	1(4)	0.868(1)	8.24(3)	6.22(2)
ISc^{PSE3}	0.92(5)	−7(4)	0.8903(9)	6.57(2)	4.90(2)

Table 2.6: Bootstrap statistical comparison between predicted and empirical hydration free energies for ions (Rizzo dataset).

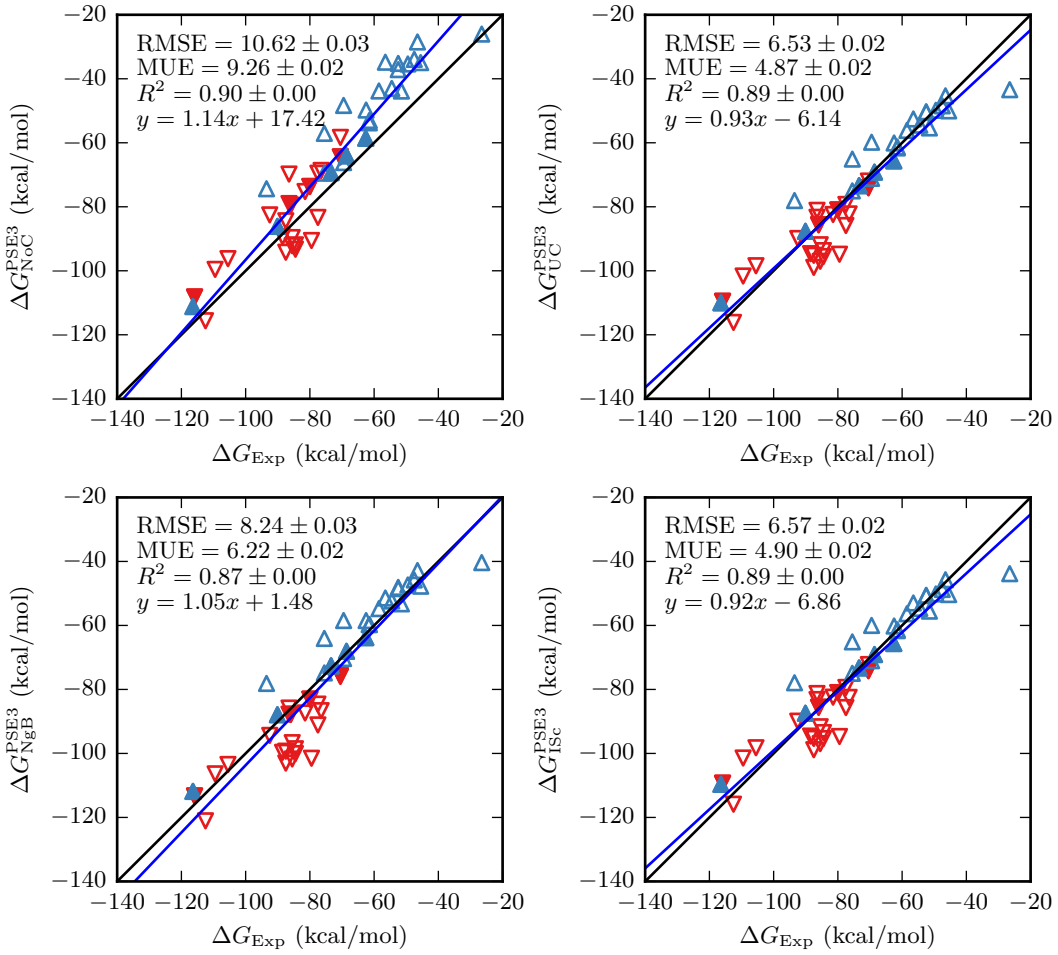

Figure 2.2: Hydration free energies of ions from 3D-RISM-PSE-3 vs. experiment (Rizzo dataset). Positive ions are blue triangles pointing up and negative ions are red triangles pointing down. Filled symbols are alkali-halide ions.

	ΔG				
	Slope	y-intercept	R^2	RMSE	MUE
PSE-3	1.00(9)	18.9(3)	0.232(1)	19.695(8)	18.861(8)
UC^{PSE3}	0.97(1)	−0.87(3)	0.9322(3)	1.146(2)	0.885(1)
NgB^{PSE3}	1.13(1)	−0.54(3)	0.9465(2)	1.231(2)	0.885(1)
ISc^{PSE3}	0.97(1)	−1.37(3)	0.9394(3)	1.519(2)	1.319(1)

Table 2.7: Bootstrap statistical comparison between predicted and molecular dynamics hydration free energies for neutral molecules (Mobley dataset). As described in Methods, R^2 bootstrap is the mean of all resampled data and R^2 k-fold is the mean over all training sub-samples. RMSE: root-mean-squared-error. MUE: mean unsigned error.

RMSE and MUE are all significantly worse (see figure 2.2 and table 4.3) but the relative errors are similar to those of the neutral compounds. The y-intercept also appears much worse but the statistical error is also much higher due to the smaller data set. Even so, the relative error for ions is somewhat smaller than that for neutral molecules.

It is not immediately clear how much of this error is due to 3D-RISM and how much should be attributed to the force field or errors in the experimental data. The relatively simple case of monovalent ions highlights the problem (filled triangles in figure 2.2). Empirical values are available from Refs. [77, 2, 53] and have a RMS difference of roughly 3 kcal/mol. In this work Joung-Cheatham parameters were used [39], which are fit to data from Ref. [77], but for comparison values from Ref. [2, 53] were also used. Other ions in the data set have not received the same attention as the alkali-halide ions, contributing to the large absolute errors. Due to these uncertainties, the ion HFEs were not used in fitting UC, UCGF or NgB corrections.

2.3.1.1 Comparison Against Molecular Dynamics

Since 3D-RISM and MD calculations share the same force field, the 3D-RISM may be expected to reproduce MD results better than experiment. Comparing table 4.4 and table 4.2, this is the case with R^2 values improve for all corrections. Of all the corrections considered only $ISC^{*PSE\text{-}3}$ and ISC^{*HNC} have $R^2 < 0.9$.

Despite the improved correlations, errors with respect to MD are, in some cases, increased. Notably, UC and ISC have increased RMSE, MUE and a larger absolute y-intercept, despite having improved correlation coefficients and slightly improved slopes.

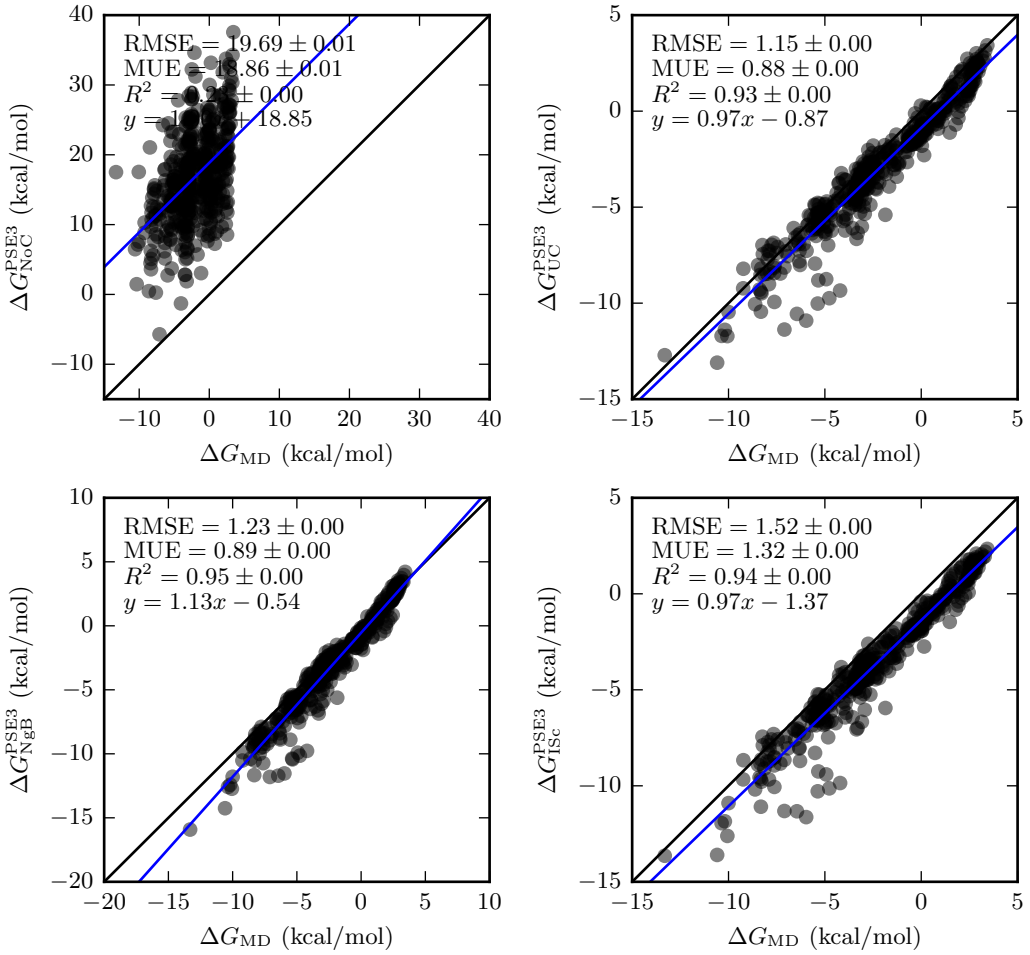

Figure 2.3: Hydration free energies of neutral molecules from 3D-RISM-PSE-3 vs. MD (Mobley dataset). Coloring as in figure 2.1.

ΔG_{Pol}					
	Slope	y-intercept	R^2	RMSE	MUE
PSE-3	1.16(1)	−0.16(4)	0.9530(3)	1.160(2)	0.831(1)
$\text{UC}_{\text{Pol}}^{\text{PSE3}}$	1.07(1)	−0.20(3)	0.9469(3)	0.841(2)	0.5377(9)
$\text{NgB}_{\text{Pol}}^{\text{PSE3}}$	1.16(1)	−0.16(4)	0.9531(2)	1.162(2)	0.833(1)
$\text{ISc}_{\text{Pol}}^{\text{PSE3}}$	1.06(1)	−0.20(4)	0.9459(3)	0.837(2)	0.5330(9)

Table 2.8: Bootstrap statistical comparison between predicted and molecular dynamics polar hydration free energies for neutral molecules (Mobley dataset).

For both corrections, the y-intercept was negative relative to the experimental data while MD results over estimated experimental data by $0.64(5)$ kcal/mol. This then contributes to the RMSE and MUE values. UC was fit against experimental data, so this result is not surprising.

MD data also allow comparison of polar and non-polar contributions to the free energy. Previous work has suggested that the polar component calculated by 3D-RISM is in good agreement with MD[54, 85, 39] while the non-polar contribution is the primary source of error[24, 85]. Indeed, uncorrected 3D-RISM data is much better for just the polar component (see table 2.8 and figure 2.4) and is generally poor for the non-polar component (see table 2.9 and figure 2.5).

All of the corrections perform as well as or better than uncorrected 3D-RISM for the polar SFE with the exception of $\text{UCGF}^{\text{PSE-3/HNC}}$ and NgB (see table 4.5). Even with these included, $R^2 > 0.92$ for all corrections and closures. That NgB does not show improvement is due to the nature of the correction, which can only effect the non-polar contribution (see equation (4.4)). The improvement in the other corrections is due to the electrostriction effect, which accounts for the polar component of the PMV. In particular, $\text{UC}^{\text{PSE-3}}$ and $\text{ISC}^{\text{PSE-3}}$ show improvement over uncorrected 3D-RISM with MUE ≈ 0.54 kcal/mol and RMSE ≈ 0.84 kcal/mol.

As previously noted, the non-polar contribution to solvation free energy predicted by 3D-RISM is extremely poor (see table 2.9 and figure 2.5). All corrections substantially improve the prediction of ΔG_{NP} though the extent to which they do varies considerably. NgB performs the best across all closures with slopes and y-intercepts within error of 1.0 and 0.0. NgB also has the lowest RMSE and MUE and the highest R^2. $\text{UC}^{\text{PSE-3}}$ and

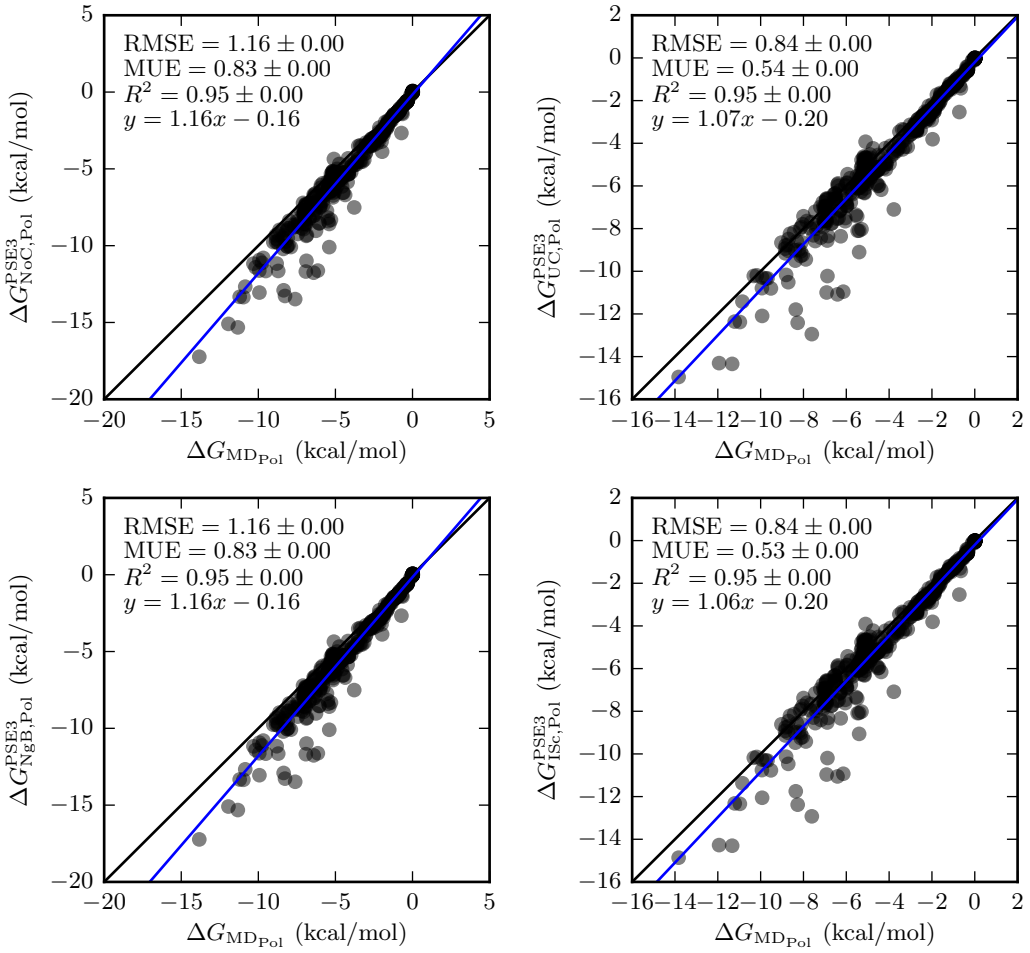

Figure 2.4: Hydration free energies of solvent polarization for neutral molecules from 3D-RISM-PSE-3 vs. MD (Mobley dataset). Coloring as in table 2.5.

	ΔG_{NP}				
	Slope	y-intercept	R^2	RMSE	MUE
PSE-3	2.5(4)	16.8(7)	0.0967(8)	20.471(8)	19.687(8)
$\mathrm{UC}^{\mathrm{PSE3}}_{\mathrm{NP}}$	0.64(3)	0.36(5)	0.542(1)	0.5856(6)	0.4614(5)
$\mathrm{NgB}^{\mathrm{PSE3}}_{\mathrm{NP}}$	0.99(2)	0.02(4)	0.7937(8)	0.3576(6)	0.2456(4)
$\mathrm{ISc}^{\mathrm{PSE3}}_{\mathrm{NP}}$	0.56(2)	−0.01(3)	0.7284(9)	0.9318(6)	0.8528(5)

Table 2.9: Bootstrap statistical comparison between predicted and molecular dynamics non-polar hydration free energies for neutral molecules (Mobley dataset).

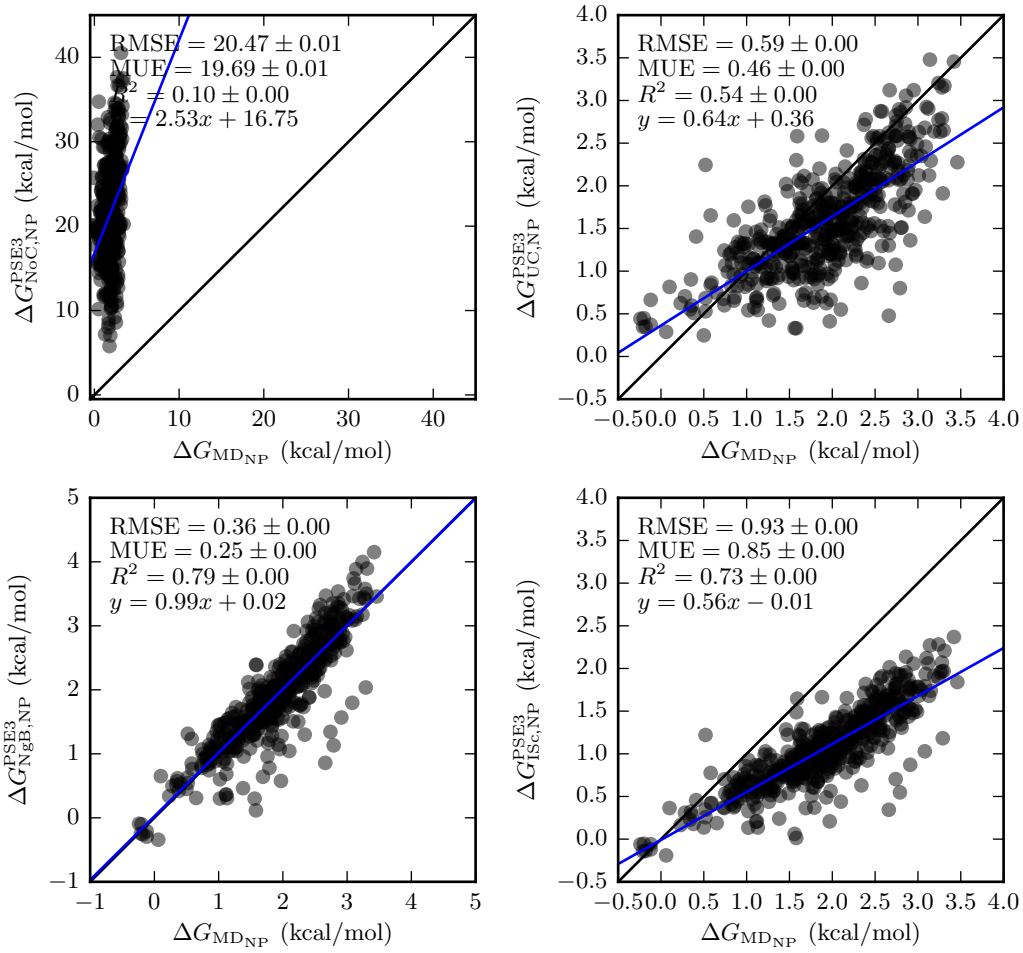

Figure 2.5: Non-polar hydration free energies of neutral molecules from 3D-RISM-PSE-3 vs. MD (Mobley dataset). Coloring as in table 2.5.

$\Delta H / \Delta \epsilon$

	Slope	y-intercept	R^2	RMSE	MUE
PSE-3	1.19(7)	1.2(8)	0.798(2)	2.79(1)	2.045(8)
UCT^{PSE3}	0.89(6)	−1.4(6)	0.806(1)	1.849(5)	1.504(4)
NgB^{PSE3}	1.19(7)	1.2(7)	0.802(1)	2.77(1)	2.030(8)
ISc^{PSE3}	0.98(6)	0.3(7)	0.799(1)	2.117(6)	1.667(6)

Table 2.10: Bootstrap statistical comparison between predicted ΔH (all UC and NgB corrections) or $\Delta\epsilon$ (uncorrected 3D-RISM and parameter free corrections) and ΔH from experiment for neutral molecules (Abraham and Cabani datasets).

$ISC^{PSE\text{-}3}$ do improve greatly over uncorrected 3D-RISM but do not have the predictive power of $NgB^{PSE\text{-}3}$. $UC^{PSE\text{-}3}$ has both poor slope and R^2 but the limited range of the data means that the RMSE and MUE are still reasonably good. In contrast, $ISC^{PSE\text{-}3}$ has an R^2 only slightly smaller than $NgB^{PSE\text{-}3}$ but the slope, RMSE and MUE are all worse than $UC^{PSE\text{-}3}$. Of course, UC and ISC have the same dependence on PMV but differ in how the coefficients are obtained. If UC was fit against MD data instead of experiment, we would expect UC results to be at least as good as ISC in this comparison.

2.3.2 Hydration Energies and Entropies

2.3.2.1 Solvation Energies/Enthalpies

Care must be taken when comparing 3D-RISM data against solvation enthalpy and entropy from experiment. As discussed in section §2.1 and [92], the temperature derivative data presented here are properly the solvation energy and entropy at constant volume. This difference will be small and is estimated by [92] to be on the order of 1 kcal/mol. Since the the coefficients for UCT, UCGFT, and NgBT corrections were fit against enthalpy data, we can claim that these are models that predict enthalpy and the associated constant pressure entropy.

Uncorrected 3D-RISM performs reasonably well for neutral molecules with all three closures (see table 2.10, table 4.7 and figure 2.6) and much better than for SFE. For all three closures $R^2 \approx 0.80$, RMSE < 2.8 kcal/mol and MUE < 2.4 kcal/mol, which is good considering that we are comparing enthalpies and energies and expect the error to be on the

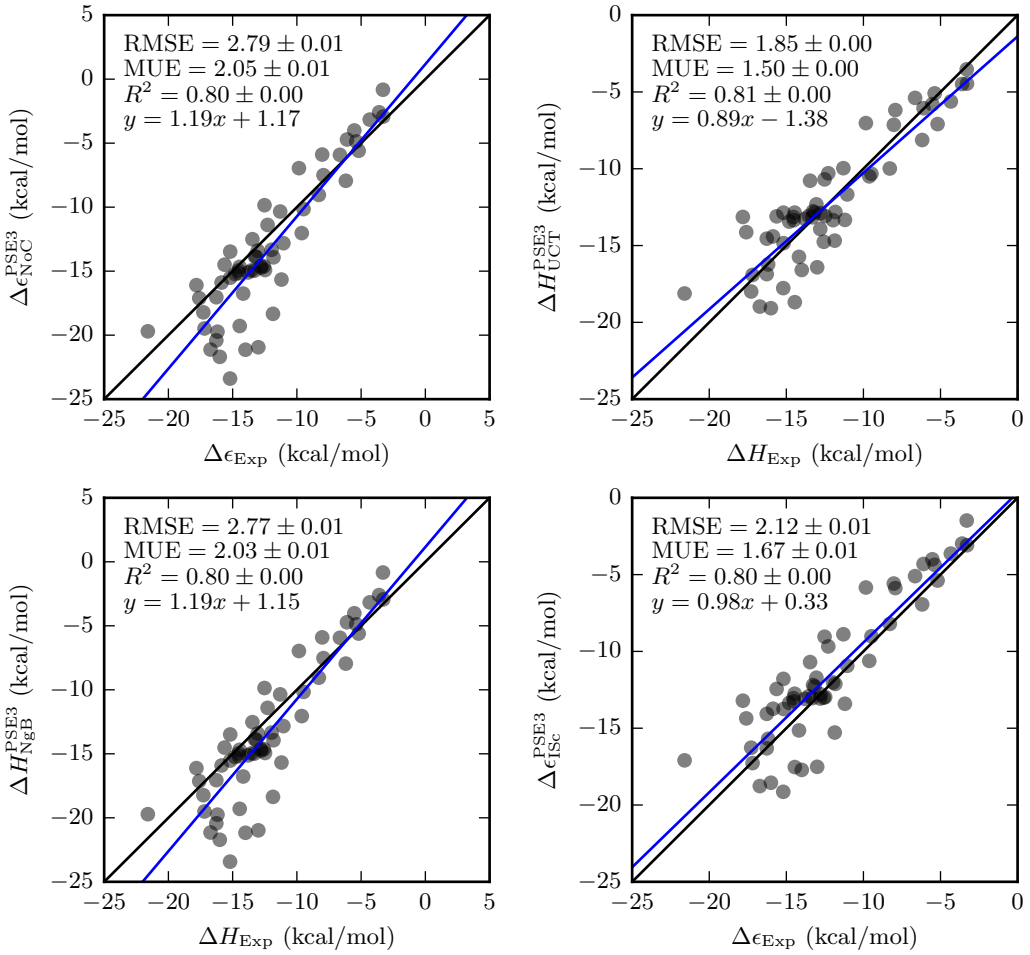

Figure 2.6: Hydration energies/enthalpies of neutral molecules from 3D-RISM-PSE-3 vs. experiment (Abraham and Cabani datasets). Coloring as in table 2.5.

$\Delta H / \Delta \epsilon$

	Slope	y-intercept	R^2	RMSE	MUE
PSE-3	0.88(9)	−12(8)	0.853(2)	8.16(6)	6.35(4)
UCT^{PSE3}	0.83(9)	−17(8)	0.867(2)	7.96(6)	6.12(4)
NgB^{PSE3}	0.88(9)	−12(9)	0.850(3)	8.15(6)	6.34(4)
ISc^{PSE3}	0.80(8)	−17(8)	0.867(2)	7.89(5)	6.04(4)

Table 2.11: Bootstrap statistical comparison between predicted ΔH (all UC and NgB corrections) or $\Delta\epsilon$ (uncorrected 3D-RISM and parameter free corrections) and ΔH from experiment for ions (Fawcett and Marcus datasets).

order of 1 kcal/mol at a minimum. NgB and NgBT provide nearly identical results with each other and with uncorrected 3D-RISM. Combined with the parameterization of γ_0 and γ_1 (see table 2.4) it is clear that NgB has the correct temperature dependence. UC, on the other hand, has RMSE $>$ 19 kcal/mol for all closures and requires temperature dependence to be added to both coefficients (see table 4.7). With this temperature dependence added, UCT performs quite well and has the lowest RMSE and MUE of any correction. R^2 and y-intercept from UC are both quite close to values from no correction and NgB. ISC also performs well with slope and y-intercept within error of 1.0 and 0.0, an R^2 identical to no correction, NgB and UCT, and RMSE and MUE close to those of UCT. Overall, the quality of 3D-RISM's treatment of enthalpies is similar to its treatment of the polar SFE. In both cases only small correction are needed and are made.

When we consider the energy/enthalpy of ionic solutes, the absolute errors are much larger, just as they were for the SFE, but the relative errors are similar to those of neutral compounds and correlation is improved. Again, there is little difference between uncorrected 3D-RISM and the corrections and NgB is nearly identical to uncorrected 3D-RISM is all metrics. Similarly, ISC and UC are slightly better than NgB in terms of RMSE and MUE. The values of the RMSE, MUE and y-intercept are all larger than observed for the SFE of ionic solutes. The errors in the y-intercept are particularly large but, again, so are the statistical errors as the sample size is becoming much smaller.

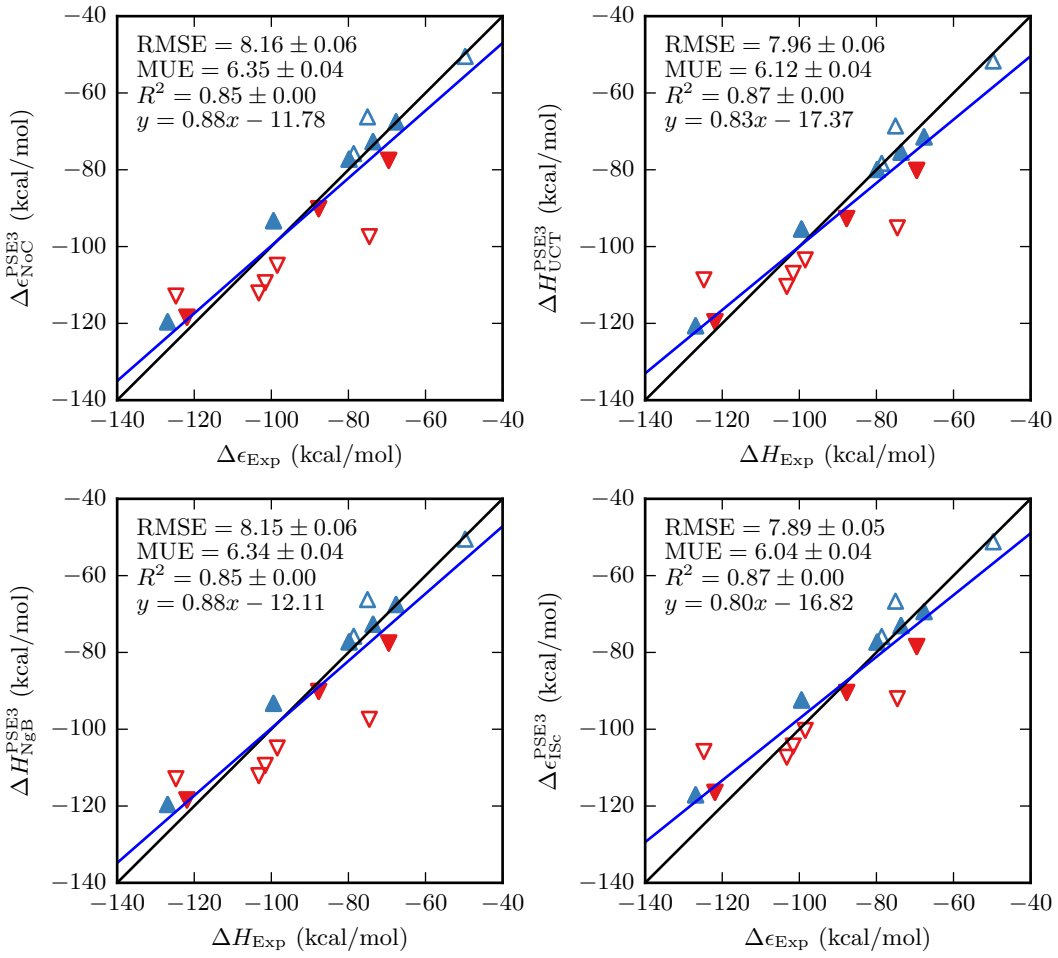

Figure 2.7: Hydration energies/enthalpies of ions from 3D-RISM-PSE-3 vs. experiment. Coloring as in figure 2.2.

$T\Delta S$

	Slope	y-intercept	R^2	RMSE	MUE
PSE-3	2.6(5)	−4(4)	0.439(3)	18.52(3)	17.13(3)
UCT^{PSE3}	0.56(8)	−3.7(7)	0.517(3)	1.497(5)	1.157(4)
NgB^{PSE3}	1.0(1)	−1(1)	0.495(3)	2.346(8)	1.807(6)
ISc^{PSE3}	0.7(1)	−1.3(9)	0.507(3)	1.930(6)	1.552(5)

Table 2.12: Bootstrap statistical comparison between predicted $T\Delta S_P$ (all UC and NgB corrections) or $T\Delta S_V$ (uncorrected 3D-RISM and parameter free corrections) and $T\Delta S_P$ from experiment for neutral molecules (Abraham and Cabani datasets).

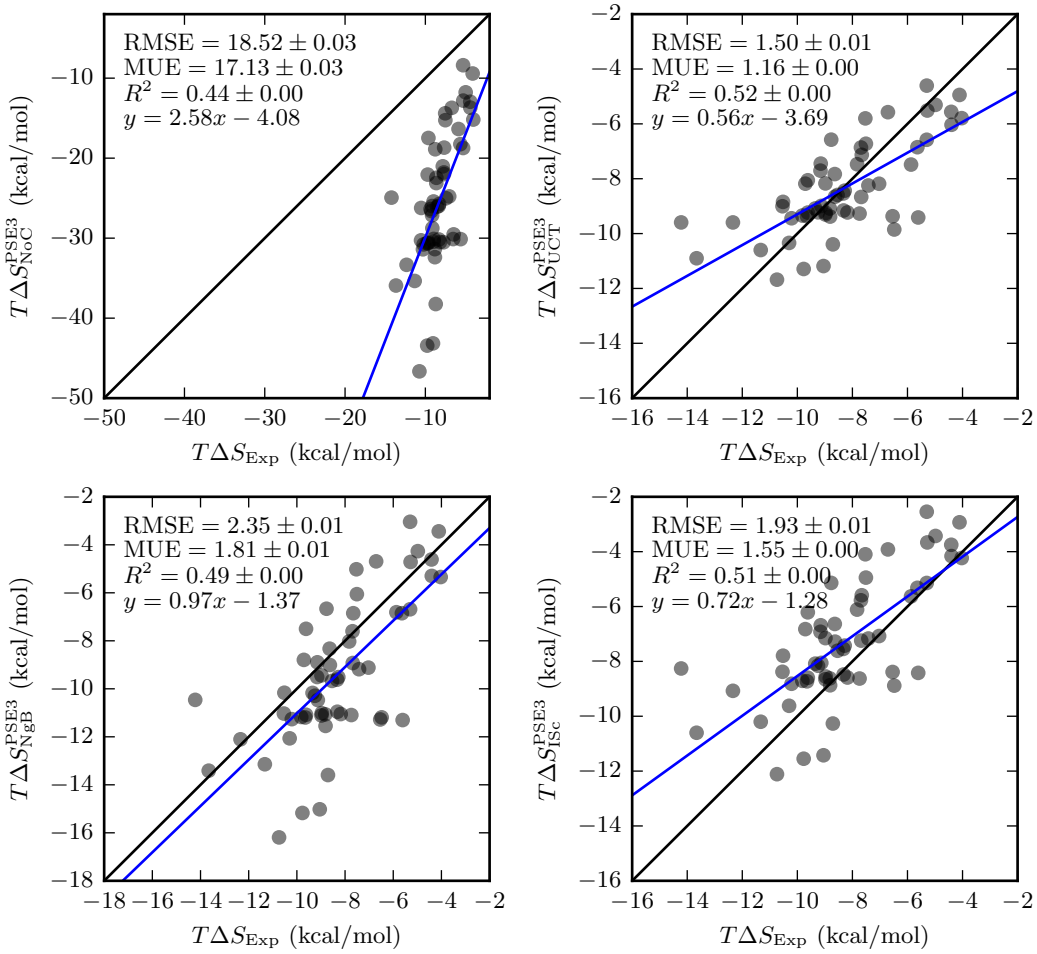

Figure 2.8: Hydration entropies of neutral molecules from 3D-RISM-PSE-3 vs. experiment (Abraham and Cabani datasets). Coloring as in table 2.5.

$T\Delta S$

	Slope	y-intercept	R^2	RMSE	MUE
PSE-3	1.0(5)	−4(3)	0.316(6)	6.19(3)	5.16(3)
UCT^{PSE3}	0.7(1)	−3.9(10)	0.570(5)	2.21(1)	1.77(1)
NgB^{PSE3}	0.8(1)	−0.3(9)	0.689(4)	1.91(1)	1.49(1)
ISc^{PSE3}	0.6(1)	−1.7(9)	0.616(5)	2.004(7)	1.758(8)

Table 2.13: Bootstrap statistical comparison between predicted $T\Delta S_P$ (all UC and NgB corrections) or $T\Delta S_V$ (uncorrected 3D-RISM and parameter free corrections) and $T\Delta S_P$ from experiment for ions (Fawcett and Marcus datasets).

2.3.2.2 Solvation Entropies

Entropies, like the non-polar SFE contributions, are poorly handled by 3D-RISM and are a large source of error (see table 4.9 and figure 2.8). Of the corrections, only ISC* has an R^2 significantly larger than 0.5. Of the three most successful corrections, NgB has the best slope and intercept but, given the low R^2, this may not be meaningful. ISC and UCT both have lower errors than NgB. While the errors for UCT are only slightly larger in magnitude than those for the SFE of small neutral molecules (see figure 2.1) the low R^2 means that comparing relative solvation entropies is not useful.

The solvation entropies of ionic solutes is quite similar to that of neutral molecules. The magnitudes of the values and errors as well as the quality of the corrections are qualitatively the same between the two data sets. The major differences for the corrections are that the R^2 values are slightly higher for ionic solutes and that NgB has the lowest errors and UCT the highest instead of the other way around.

2.4 Conclusions

We have presented a new implementation of temperature derivatives in 3D-RISM, capable of efficiently calculating solvation energies and entropies of charged and neutral small molecules. Accuracy comparable to that of explicit solvent, all-atom molecular dynamics simulations is achieved through the use of different correction methods. While a number of corrections have been proposed in the literature, we found that only UC(T) and NgB are applicable to all closures, while ISC works with PSE-n closures for $n \geq 3$. UC with the

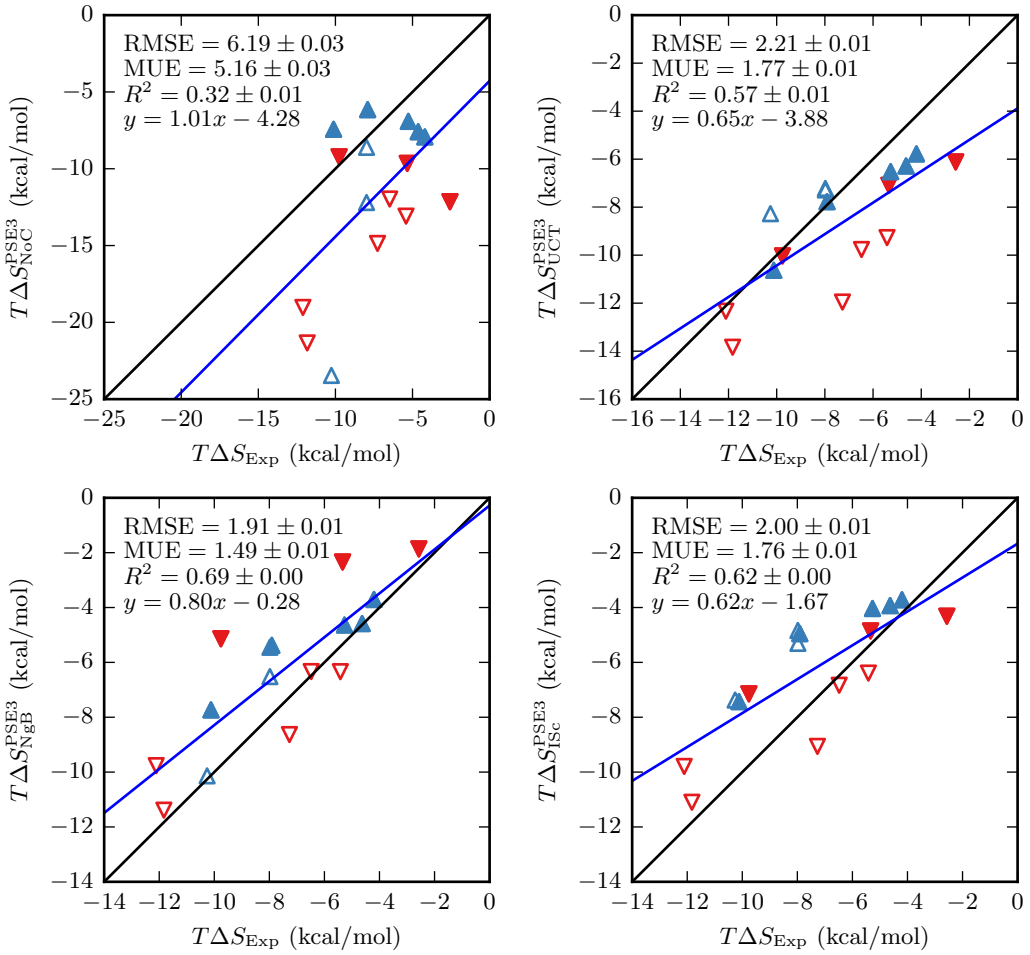

Figure 2.9: Hydration entropies of ions from 3D-RISM-PSE-3 vs. experiment. Coloring as in figure 2.2 (Fawcett and Marcus datasets).

Gaussian fluctuation free energy functional only works with the KH closure while ISC* performs poorly regardless of closure.

The physical basis of these corrections is to mitigate the effects of the excessively high pressures predicted by the HNC-like closures used here. Due to the over-estimation of pressure, additional mechanical work is included in the 3D-RISM excess chemical potential calculation and must be subtracted off. The PMV accounts for the change in volume required to accommodate the solute. Since the leading contribution to the PMV is the van der Waals size of the solute, UCT and ISC corrections primarily improve the non-polar and entropic components of the SFE. However, the PMV also depends on the charge state of the solute through the electrostriction effect. UCT and ISC also improve the polar and energetic components of the SFE as a result while NgB does not. To compensate for the extra mechanical work, several expressions for pressure are available, including the bulk pressure from the compressibility and energy routes, and the contact pressure. However, only the contact pressure on the solute, as used in ISC, successfully mitigates the excess mechanical work. The fact that these pressure expressions give different quantitative results is symptomatic of the larger inconsistencies in the HNC closure.

While these corrections are of practical use in calculating solvation free energies and their decomposition, they do not address the underlying deficiencies in the closures. Further improvements in 3D-RISM and related methods will require new closure approximations that avoid these inconsistencies.

Chapter 3

Crystal Structure Refinement with Periodic 3D-RISM

3.1 Periodic Interactions

The 3D-RISM equation as formulated by Kovalenko and Hirata does not assume a particular boundary condition in its derivation [48]. Boundary conditions arise solely from the interaction potential $u_{\alpha}^{uv}(\mathbf{r})$ contained in the closure equation. If the potential has periodic boundary conditions, such as in a crystal lattice, then the resulting density correlation functions obtained from the 3D-RISM will be periodic. Similarly, a potential which assumes the infinite dilution case (i.e., solute being infinitely far apart) will lead to the potential asymptotically approaching zero far from the solute. For such potentials the correlation functions will reflect infinite dilution boundary conditions as long as the box is sufficiently larger than the solute so that the potential smoothly approaches zero near the box boundaries. Thus to produce a periodic 3D-RISM, a periodic potential function is required.

In non-periodic systems, it is common to use the Coulomb inverse-square law to model electrostatic interactions. From these laws the electrostatic interaction on a point charge q_i exerted by a set of point charges Q (where $q_i \in Q$) is

$$u_i(\mathbf{r}_i) = q_i V(\mathbf{r}_i) = q_i \frac{1}{4\pi\epsilon_r} \sum_{j \in Q, j \neq i} \frac{q_j}{r_{ij}} \tag{3.1}$$

$$\mathbf{F}_i(\mathbf{r}_i) = q_i \mathbf{E}(\mathbf{r}_i) = q_i \frac{1}{4\pi\epsilon_r} \sum_{j \in Q, j \neq i} \frac{q_j}{\left|r_{ij}\right|^2} \hat{\mathbf{r}_{ij}} \tag{3.2}$$

where $r_{ij} = \left|\mathbf{r}_i - \mathbf{r}_j\right|$.

Additionally, the Lennard-Jones 12-6 equation is a popular model for the so-called Van der Waals interactions, including the attractive long-range induced dipole-induced dipole

(r^{-6}) and short-range Pauli exclusion between electron orbitals (r^{-12}). Unlike the Colombic potential, the Lennard-Jones equation requires two atom-atom interaction parameters whose values depend on the types of atoms interacting. These parameters are typically fitted using quantum chemistry calculations. The Lennard-Jones potential energy and force experienced between interacting atoms i and j is

$$u_{ij}\left(r_{ij}\right) = \varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{3.3}$$

$$\mathbf{F}_{ij}\left(\mathbf{r}_i,\, \mathbf{r}_j\right) = \varepsilon_{ij}\frac{12}{r_{ij}^2}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]\mathbf{r}_{ij} \tag{3.4}$$

In combination, the Coulomb and Lennard-Jones interactions are the most popular model for interparticle interactions in all of computational chemistry due to their balance of computational efficiency and accuracy.

The most direct means of calculating electrostatic potential energy of a periodic system is to sum the Coulomb interactions exerted by the charges in all unit cells on the charged particles in some reference unit cell. Excluding trivial toy cases, the solution of this approach is indeterminant due to the non-convergent infinite series of unit cell contributions. A popular method known as the Ewald sum avoids these convergence issues by solving the long range periodic contribution to the electostatic potential using a Fourier series of periodic Guassian charges. This method only features point charges in the central unit cell. This approach conditionally guarantees that the series will converge in a finite number of terms.

Alternative periodic potentials to the Ewald sum include particle-mesh methods, such as the particle-particle particle-mesh (PPPM) method, and fast multipole methods. The choice of potential is primarily a matter of computational efficiency and ease of implementation. The computational efficiency of these potentials, as well as the range of atom counts for which they are most efficient, is given in Table 3.1. In terms of implementation difficulty, the Ewald sum is by far the easiest while PPPM tends to be the most complex, with fast multipole in between.

What follows is a derivation of the Ewald sum solvation potential and force as it is

Table 3.1: Computational efficiency of popular periodic potentials [23]. Here N is the number of charged particles in the system.

periodic potential	asymptotic computational complexity	efficiency cutoff range of charged particle count
Ewald sum	$\mathcal{O}\left(N^{3/2}\right)$	$2 - 10^2$
particle-particle particle-mesh	$\mathcal{O}\left(N \log N\right)$	$10^3 - 10^5$
fast multipole	$\mathcal{O}\left(N\right)$	10^6 and up

applied to the RISM, followed by a derivation of the Particle Mesh Ewald (PME) method to expedite the calculation of the long-range Ewald sum term. The PME method is then employed to develop the periodic 3D-RISM, a variation of the 3D-RISM which can handle periodic solute. Finally, the periodic 3D-RISM is applied to X-ray crystal structure refinement and its results compared to those of traditional solvent models used in refinement.

3.2 Ewald Sum

3.2.1 Solvation Potential Energy

The Ewald sum is a means of solving the Poisson equation for a periodic charge distribution. It achieves this by treating the charges (e.g., partially charged atoms) within the reference periodic unit cell as a set of point charges, while charges in other unit cells are approximated as a set of infinitely periodic Gaussian charges, one Gaussian charge per point charge. The Poisson equation can be solved independently for the short and long-range charge distributions. Adding the solutions together produces the Ewald sum. To handle the infinite periodicity of the Gaussian charges while solving the Poisson equation, the Fourier transform is employed.

The Poission equation can be written as

$$-\nabla^2 \phi\left(\mathbf{r}\right) = 4\pi\rho\left(\mathbf{r}\right) \tag{3.5}$$

where $\phi\left(\mathbf{r}\right)$ is the electric potential at a Cartesian point $\mathbf{r}$ produced by an electric charge density distribution $\rho\left(\mathbf{r}\right)$, and $\nabla^2 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)$ is the Laplace operator or Laplacian in three-dimensional Cartesian coordinates.

The Ewald sum solves the Poisson equation for an infinitely periodic charge distribution by splitting the charge distribution into short-range charges, which are the set of point charges in some reference unit cell, and long-range charges, which are the set of periodic Gaussian charges in all other unit cells,

$$\rho\left(\mathbf{r}\right) = \rho_{\mathrm{SR}}\left(\mathbf{r}\right) + \rho_{\mathrm{LR}}\left(\mathbf{r}\right)$$

$$\rho_{\mathrm{SR}}\left(\mathbf{r}\right) = \sum_{i}^{N} q_i \delta\left(r - r_i\right) \tag{3.6}$$

$$\rho_{\mathrm{LR}}\left(\mathbf{r}\right) = \sum_{j=1}^{N} \sum_{\mathbf{n}} q_j \left(\alpha/\pi\right)^{\frac{3}{2}} \exp\left[-\alpha \left|\mathbf{r} - \left(\mathbf{r}_j + \mathbf{n}L\right)\right|^2\right] \tag{3.7}$$

where ρ_{SR} and ρ_{LR} are the short and long-range charge distributions respectively, q_i is the charge on point charge i located at position r_i, N is the total number of point charges in the reference cell, $\delta\left(r\right)$ is the Dirac delta function, L is the unit cell lattice vector (in the form of three ordered scalar values), $\mathbf{n}$ is the unit cell lattice iteration vector (iterating from zero to infinity in all directions), and α is the Gaussian charge 'smear' coefficient. As can be seen from equation (3.7), the Gaussian charge distribution is infinitely periodic in all directions and, to simplify the math, it includes the reference unit cell. The Gaussians originating from the reference cell will later be subtracted so that it does not contribute to the final Ewald sum.

Due to the distributivity of the Laplacian, this leads to separate solutions for the short-range and long-range potentials,

$$\phi\left(\mathbf{r}\right) = \phi_{\mathrm{SR}}\left(\mathbf{r}\right) + \phi_{\mathrm{LR}}\left(\mathbf{r}\right)$$

The potential resulting from the infinitely peroidic Gaussian charges will be solved first. To handle the infinite periodicity, the Poisson equation will be solved in frequency space (k-space) and later translated back to Cartesian space (r-space).

$$-\nabla^2\phi\left(\mathbf{r}\right) = -\nabla^2\left(\frac{1}{V_{\mathrm{cell}}}\sum_{\mathbf{k}}\tilde{\phi}\left(\mathbf{k}\right)e^{i\mathbf{r}\cdot\mathbf{k}}\right)$$

$$= \frac{1}{V_{\text{cell}}} \sum_{\mathbf{k}} k^2 \tilde{\phi}(\mathbf{k}) e^{i\mathbf{r}\cdot\mathbf{k}} \tag{3.8}$$

$$\rho(\mathbf{r}) = \frac{1}{V_{\text{cell}}} \sum_{\mathbf{k}} \tilde{\rho}(\mathbf{k}) e^{i\mathbf{r}\cdot\mathbf{k}} \tag{3.9}$$

Given the definition of the complex Fourier series and its inverse

$$f(\mathbf{r}) = \frac{1}{V_{\text{cell}}} \sum_{\mathbf{k}=-\infty}^{\infty} \tilde{f}(k) e^{i\mathbf{k}\cdot\mathbf{r}} \tag{3.10}$$

$$\tilde{f}(\mathbf{k}) = \int_V d\mathbf{r}\, f(\mathbf{r}) e^{-i\mathbf{k}\cdot\mathbf{r}} \tag{3.11}$$

where $k = |\mathbf{k}| = 2\pi/\lambda$ is the spatial angular frequency (i.e., wave number) and $\mathbf{k}$ is the associated wave vector, then the Poisson equation in k-space is

$$-\nabla^2 \phi(\mathbf{r}) = 4\pi\rho(\mathbf{r})$$

$$\frac{1}{V} \sum_{\mathbf{k}} k^2 \tilde{\phi}(\mathbf{k}) e^{i\mathbf{r}\cdot\mathbf{k}} = \frac{4\pi}{V} \sum_{\mathbf{k}} \tilde{\rho}(\mathbf{k}) e^{i\mathbf{r}\cdot\mathbf{k}}$$

$$\tilde{\phi}(k) = \frac{4\pi}{k^2} \tilde{\rho}(k) \tag{3.12}$$

The Fourier transform of the infinitely periodic Gaussian charge density is

$$\tilde{\rho}_{\text{LR}}(\mathbf{k}) = \int_V d\mathbf{r}\, e^{-i\mathbf{k}\cdot\mathbf{r}} \rho_{\text{LR}}(\mathbf{r})$$

$$= \int_V d\mathbf{r}\, e^{-i\mathbf{k}\cdot\mathbf{r}} \sum_{j=1}^{N} \sum_{\mathbf{n}} q_j (\alpha/\pi)^{\frac{3}{2}} \exp\left[-\alpha \left|\mathbf{r} - (\mathbf{r}_j + \mathbf{n}L)\right|^2\right]$$

$$= \int_{\text{all space}} d\mathbf{r}\, e^{-i\mathbf{k}\cdot\mathbf{r}} \sum_{j=1}^{N} q_j (\alpha/\pi)^{\frac{3}{2}} \exp\left[-\alpha \left|\mathbf{r} - \mathbf{r}_j\right|^2\right]$$

$$= \sum_{j=1}^{N} q_j e^{-i\mathbf{k}\cdot\mathbf{r}_j} e^{-k^2/4\alpha}]$$

Substituting this into the k-space Poisson equation allows for the solution of the long-range portion of the Ewald sum potential in k-space.

$$\tilde{\phi}_{\mathrm{LR}}(k) = \frac{4\pi}{k^2}\tilde{\rho}_{\mathrm{LR}}(k)$$

$$\tilde{\phi}_{\mathrm{LR}}(\mathbf{k}) = \frac{4\pi}{k^2}\sum_{j=1}^{N} q_j e^{-i\mathbf{k}\cdot\mathbf{r}_j} e^{-k^2/4\alpha} \tag{3.13}$$

The k-space potential can then be Fourier transformed back into r-space.

$$\phi_{\mathrm{LR}}(\mathbf{r}) = \frac{1}{V_{\mathrm{cell}}}\sum_{k\neq 0}\tilde{\phi}_{\mathrm{LR}}(\mathbf{k})\, e^{i\mathbf{k}\cdot\mathbf{r}}$$

$$= \frac{1}{V_{\mathrm{cell}}}\sum_{k\neq 0}\sum_{j=1}^{N}\frac{4\pi q_j}{k^2} e^{i\mathbf{k}\cdot\left(\mathbf{r}-\mathbf{r}_j\right)} e^{-k^2/4\alpha} \tag{3.14}$$

Before solving the electric potential produced by the point carges, the contribution to the long-range potential due to the Gaussian charge distribution in the reference unit cell can be removed so that the reference cell only contains point charges. This can be done by deriving the potential produced by a single Gaussian charge in the reference cell,

$$\rho_{\mathrm{Gauss}}(r) = q_i\left(\alpha/\pi\right)^{\frac{3}{2}} e^{-\alpha r^2} \tag{3.15}$$

Substituting this into r-space Poisson equation and solving it in spherical coordinates provides the resulting electric potential of the single Gaussian charge.

$$-\frac{1}{r}\frac{\partial^2 r\phi_{\mathrm{Gauss}}(r)}{\partial r^2} = 4\pi\rho_{\mathrm{Gauss}}(r)$$

$$-\frac{\partial^2 r\phi_{\mathrm{Gauss}}(r)}{\partial r^2} = 4\pi r\rho_{\mathrm{Gauss}}(r)$$

$$-\frac{\partial r\phi_{\mathrm{Gauss}}(r)}{\partial r} = \int_{\infty}^{r} dr\; 4\pi r\rho_{\mathrm{Gauss}}(r)$$

$$= -2\pi q_i\left(\alpha/\pi\right)^{\frac{3}{2}}\int_{r}^{\infty} dr^2\; e^{-\alpha r^2}$$

$$= -2q_i \left(\alpha/\pi\right)^{\frac{1}{2}} e^{-\alpha r^2}$$

$$r\phi_{\text{Gauss}}\left(r\right) = 2q_i \left(\alpha/\pi\right)^{\frac{1}{2}} \int_0^r dr\ e^{-\alpha r^2}$$

$$= q_i \text{erf}\left(\sqrt{\alpha} r\right)$$

$$\phi_{\text{Gauss}}\left(r\right) = \frac{q_i}{r} \text{erf}\left(\sqrt{\alpha} r\right)$$

where $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ is the error function.

The short range portion of the Ewald sum is thus the potential contributed by the set of point charges in the reference cell subtracted by their corresponding Gaussian charges to remove the undesired extra Gaussian charge added by the long-range potential.

$$\phi_{\text{SR}}\left(r\right) = \sum_i^N \frac{q_i}{r} - \frac{q_i}{r} \text{erf}\left(\sqrt{\alpha} r\right)$$

$$= \sum_i^N \frac{q_i}{r} \text{erfc}\left(\sqrt{\alpha} r\right)$$

where $\text{erfc}\left(x\right) \equiv 1 - \text{erf}\left(x\right)$ is the complementry error function.

The long range portion of the Ewald sum contains a self-interaction term where the Gaussian charge interacts with itself

$$\phi_{\text{self}}\left(r \to 0\right) = 2q_i \left(\alpha/\pi\right)^{\frac{1}{2}}$$

This erroneous interaction term can be subtracted in the final Ewald sum.

Combining the above results, the full Ewald sum is, for a smear parameter $\eta \equiv 1/\sqrt{\alpha}$ (the form used in the Amber implementation of RISM),

$$\phi_{\text{Ewald}}\left(\mathbf{r}\right) = \phi_{\text{LR}}\left(\mathbf{r}\right) + \phi_{\text{SR}}\left(r\right) - \phi_{\text{self}}\left(r\right) \tag{3.16}$$

$$\phi_{\mathrm{LR}}(\mathbf{r}) = \frac{1}{V_{\mathrm{cell}}} \sum_{k \neq 0} \sum_{j=1}^{N} \frac{4\pi q_j}{k^2} e^{i\mathbf{k}\cdot\left(\mathbf{r}-\mathbf{r}_j\right)} e^{-k^2\eta^2/4}$$

$$\phi_{\mathrm{SR}}(r) = \sum_{i}^{N} \frac{q_i}{r} \mathrm{erfc}\left(\frac{r}{\eta}\right)$$

$$\phi_{\mathrm{self}}(r \to 0) = 2q_i \left(1/\pi\eta^2\right)^{\frac{1}{2}}$$

Notice that in the long range term there is a singularity at $k = 0$ due to the k^2 term in the denominator. The traditional method of avoiding this issue is to employ so called tin-foil (i.e., conductive) 'boundary' conditions by assuming that $\tilde{\phi}_{\mathrm{LR}}(k = 0) = 0$. This causes a negligible change in results for electically neutral systems since they have a quickly decaying long-range term, but for non-neutral systems this approximation introduces a potentially large error term. However, this term is irrelevant so long as the system is net neutral ($\sum_i^N q_i = 0$) since in that case $\tilde{U}(k = 0) = \sum_i^N q_i \tilde{\phi}_{\mathrm{LR}}(0) = 0$ regardless of the value of $\tilde{\phi}_{\mathrm{LR}}(0)$.

However, for the 3D-RISM the calculation is performed with a possibly non-netural solute acting on solvent sites whose density distribution is initially assumed to be equal to bulk solvent. For neutral solvent this is not an issue, but for ionic solvent this leads to a potentially charged unit cell prior to the solution of the 3D-RISM equation. Thus the $\tilde{\phi}_{\mathrm{LR}}(0)$ term is not truly zero for the periodic 3D-RISM with ionic solute, but nevertheless it is ignored because assuming $\tilde{\phi}_{\mathrm{LR}}(0) = 0$ forces the 3D-RISM to converge the solvent density distribution towards charge neutrality on its own. Unfortunately, when using this procedure the 3D-RISM fails to produce solvent distributions which fully neutralize the solute, resulting in non-neutral unit cells. If the $\tilde{\phi}_{\mathrm{LR}}(0)$ term is added to the potential prior to solving the 3D-RISM, the final system is neutralized, but the resulting atom count disagrees significantly with molecular dynamics simulations. At the moment no method has been devised which leads to both a fully neutralizing solvent distribution and an accurate solvent atom count, so this is an open and unresolved problem with the approach to the periodic 3D-RISM outlined in this work.

The minimum image convention must be applied when calculating short range interactions in periodic systems, including the short range term of the Ewald sum and the Lennard-Jones potential energy. The convention simply states that when calculating the electrostatic contribution of a periodic charged particle at a given point, the closest periodic instance of the charge should be used as the source, and a spherical cutoff should be employed to prevent interactions with particles farther than half the shortest perpendicular width of the unit cell. See [58] for a more detailed discussion of the minimum image convention and its necessity for periodic interactions.

In the 3D-RISM only the solute is periodic and only solute-solvent interactions occur. Thus there are no self-interactions and ϕ_{self} can be ignored. Therefore the Ewald sum electric potential energy used by the 3D-RISM when dealing with periodic solute is given by

$$u_{\text{Ewald}}^{\text{UV}} = u_{\text{LR}}^{\text{UV}} + u_{\text{SR}}^{\text{UV}} \tag{3.17}$$

$$u_{\text{LR}}^{\text{UV}} = \frac{1}{V_{\text{cell}}} \sum_{\mathbf{k}\in\mathbb{K}^3,\mathbf{k}\neq 0} \sum_{u\in U} \sum_{v\in V} q_u q_v \frac{4\pi}{k^2} \exp\left(i\mathbf{k}\cdot\mathbf{r}_{uv} - \frac{k^2\eta^2}{4}\right)$$

$$u_{\text{SR}}^{\text{UV}} = \sum_{u\in U} \sum_{v\in V} \frac{q_i}{|\mathbf{r}_{uv}|} \text{erfc}\left(|\mathbf{r}_{uv}|/\eta\right)$$

where u and v are atoms from solute and solvent molecules U and V respectively, $\mathbf{r}_{uv} = r_v - r_u$ is the distance vector of a solvent atom v from a solute atom u, η is the solvent Gaussian charge smearing parameter, and V_{cell} is the volume of the unit cell.

3.3 Particle Mesh Ewald

The Particle Mesh Ewald (PME) method applies the Fast Fourier Transform (FFT) to the long range component of the Ewald sum. Doing so decreases the asymptotic computational time complexity of the long range term from $O\left(N^2\right)$ to $O\left(N\log N\right)$ where N is the number of solute atoms. The PME method has some overhead, so for small N solute it is possible PME may be equally fast or, in rare cases, slower than the original Ewald sum. The PME method does not change the short-range portion of the Ewald sum. Since the

short-range term has an asymptotic complexity of $O(N)$ and is relatively quick to compute compared to the long-range term, fewer efforts have been made to optimize it.

3.3.1 Solvation Potential

The PME method is essentially a way of rewriting the long range portion of the Ewald sum to make it amenable to computation on a grid (i.e., mesh) using the FFT. To achieve this, a series of somewhat intuitive mathematical theorems must be used to produce a not-so-intuitive rewriting of the long range term. The bulk of the mathematical reasoning behind the SPME derivation that follows is from [9], though a few corrections in the derivation have been made and the notation has been altered.

Beginning with the long range portion of the Ewald sum potential in k-space (eq. 3.13):

$$\tilde{\phi}_{\mathrm{LR}}(\mathbf{k}) = \frac{4\pi}{k^2} e^{-k^2/4\alpha} \sum_{j}^{N} q_j e^{-i\mathbf{k}\cdot\mathbf{r}_j}$$

Recall that the sum is over the k-space point charges, while the exponential outside the sum is the Fourier coefficient of the Gaussian smear used to make the point charges both diffuse and continuous, which helps speed up the convergence of the Fourier series. Fourier transforms of Dirac delta functionals, such as point-charge distribution functions, are notoroiously slow to converge since it attempts to map continuous sinusoidal waves to a discontinous step function - a worse case scenario for convergence. Wavelet transforms could potentially be a better choice if one wishes to avoid the Gaussian smear approximation at the heart of the Ewald sum method.

Writing the long-range Ewald sum potential energy of a collection of N interacting point charges as

$$u_{\mathrm{LR}} = \frac{1}{2} \sum_{i,\, j\neq i}^{N} q_i q_j \theta\left(\left|\mathbf{r}_i - \mathbf{r}_j\right|\right) + \frac{1}{2} \sum_{\mathbf{n}\in\mathbb{Z}^3, \mathbf{n}\neq 0} \sum_{i,\, j}^{N} q_i q_j \theta\left(\left|\mathbf{r}_i - \mathbf{r}_j + L\mathbf{n}\right|\right)$$

$$= \frac{1}{2} \sum_{\mathbf{n}\in\mathbb{Z}^3} \sum_{i,\, j}^{N} q_i q_j \theta\left(\left|\mathbf{r}_i - \mathbf{r}_j + L\mathbf{n}\right|\right) - \frac{1}{2}\theta(r\to 0) \sum_{i}^{N} q_i^2 \tag{3.18}$$

where $\theta(\mathbf{r}) \equiv \phi(\mathbf{r})/q$ for source charge q, $\theta(r\to 0) \equiv \lim_{r\to 0}\theta(r)$, and L is the unit cell lattice vector matrix. The first term of the first expression is the contribution from

the charges in the reference unit cell, while the second term is the contribution from their infinitely periodic images. The first term of the second expression is the contribution from all pairs of atoms, while the second term removes self-interactions which appear in the first term. The reason for writing the Ewald sum potential energy in this way is that the first term of the second expression can be rewritten using the relation

$$\sum_{\mathbf{n}\in\mathbb{Z}^3}\sum_{i,j}^{N} q_i q_j \theta\left(\left|\mathbf{r}_i-\mathbf{r}_j+L\mathbf{n}\right|\right)=\frac{1}{V}\sum_{\mathbf{m}\neq 0}\hat{\theta}\left(m\right)\left|\hat{\rho}\left(\mathbf{m}\right)\right|^2+\frac{1}{V}\hat{\theta}\left(m\to 0\right)\left(\sum_{i}^{N} q_i\right)^2 \tag{3.19}$$

where $\rho\left(\mathbf{r}\right)=\sum_{i=1}^{N} q_i\delta\left(\mathbf{r}-\mathbf{r}_i\right)$ is the point charge distribution. This relation essentially states that the potential energy contribution from the N interacting point charges is the same whether it is calculated in r-space or k-space, but the particular form of the k-space expression on the right can be computed very efficiently, as will be shown. Proving this relation is the bulk of the mathematical content of the PME method and will be performed in several steps. To begin, the following relation must be proved:

$$\sum_{\mathbf{n}\in\mathbb{Z}^3}\sum_{i,j}^{N} q_i q_j \theta\left(\left|\mathbf{r}_i-\mathbf{r}_j+L\mathbf{n}\right|\right)=\int \theta\left(r\right)\left(\rho\star\eta\star D\right)\left(\mathbf{r}\right)\,d\mathbf{r}$$

where $\star$ is the convolution operator (i.e., $f\left(\mathbf{x}\right)\star g\left(\mathbf{x}\right)\equiv\int f\left(\mathbf{s}\right)g\left(\mathbf{x}-\mathbf{s}\right)\,d\mathbf{s}$), $\eta\left(\mathbf{r}\right)\equiv\rho\left(-\mathbf{r}\right)$, and $D\equiv\sum_{\mathbf{n}\in\mathbb{Z}^3}\delta\left(\mathbf{r}-L\mathbf{n}\right)$.

Proof: First,

$$\rho\star\eta\left(\mathbf{r}\right)=\int\rho\left(\mathbf{s}\right)\eta\left(\mathbf{r}-\mathbf{s}\right)\,d\mathbf{s}$$

$$=\int\rho\left(\mathbf{s}\right)\rho\left(\mathbf{s}-\mathbf{r}\right)\,d\mathbf{s}$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{N} q_i q_j\int\delta\left(\mathbf{s}-\mathbf{r}_i\right)\delta\left(\mathbf{r}-\left(\mathbf{s}-\mathbf{r}_j\right)\right)\,d\mathbf{s}$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{N} q_i q_j\delta\left(\mathbf{r}-\left(\mathbf{r}_i-\mathbf{r}_j\right)\right)$$

By the above expression and the associative property of the the convolution operation,

$$\rho \star \eta \star D\left(\mathbf{r}\right) = \int \rho \star \eta\left(\mathbf{s}\right) D\left(\mathbf{s}-\mathbf{r}\right) d\mathbf{s}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} q_i q_j \int \delta\left(\mathbf{s}-\left(\mathbf{r}_i-\mathbf{r}_j\right)\right) D\left(\mathbf{s}-\mathbf{r}\right)\ d\mathbf{s}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} q_i q_j D\left(\mathbf{r}-\left(\mathbf{r}_i-\mathbf{r}_j\right)\right)$$

$$= \sum_{\mathbf{n}\in\mathbb{Z}^3}\sum_{i=1}^{N}\sum_{j=1}^{N} q_i q_j \delta\left(\mathbf{r}-\left(\mathbf{r}_i-\mathbf{r}_j+L\mathbf{n}\right)\right)$$

At last,

$$\int \theta\left(r\right)\left(\rho \star \eta \star D\right)\left(\mathbf{r}\right)\ d\mathbf{r} = \int \theta\left(r\right) \sum_{\mathbf{n}\in\mathbb{Z}^3}\sum_{i=1}^{N}\sum_{j=1}^{N} q_i q_j \delta\left(\mathbf{r}-\left(\mathbf{r}_i-\mathbf{r}_j+L\mathbf{n}\right)\right)\ d\mathbf{r}$$

$$= \sum_{\mathbf{n}\in\mathbb{Z}^3}\sum_{i=1}^{N}\sum_{j=1}^{N} q_i q_j \theta\left(\left|\mathbf{r}_i-\mathbf{r}_j+L\mathbf{n}\right|\right)$$

Using this newly proven relation, equation (3.18) can be rewritten as

$$u_{\mathrm{LR}} = \frac{1}{2}\sum_{\mathbf{n}\in\mathbb{Z}^3}\sum_{i,\,j}^{N} q_i q_j \theta\left(\left|\mathbf{r}_i-\mathbf{r}_j+L\mathbf{n}\right|\right) - \frac{1}{2}\theta\left(r\to 0\right)\sum_{i}^{N} q_i^2$$

$$= \frac{1}{2}\int \theta\left(r\right)\left(\rho \star \eta \star D\right)\left(\mathbf{r}\right)\ d\mathbf{r} - \frac{1}{2}\theta\left(r\to 0\right)\sum_{i}^{N} q_i^2$$

By the Plancherel theorem and the fact that the $\rho \star \eta \star D$ is real-valued,

$$\int \theta\left(r\right)\left(\rho \star \eta \star D\right)\left(\mathbf{r}\right)\ d\mathbf{r} = \int \hat{\theta}\left(k\right)\widehat{\left(\rho \star \eta \star D\right)}\left(\mathbf{k}\right)\ d\mathbf{k}$$

Using this relation with the current potential energy expression

$$u_{\mathrm{LR}} = \frac{1}{2}\int \theta\left(r\right)\left(\rho \star \eta \star D\right)\left(\mathbf{r}\right)\ d\mathbf{r} - \frac{1}{2}\theta\left(r\to 0\right)\sum_{i}^{N} q_i^2$$

$$= \frac{1}{2}\int \hat{\theta}\left(k\right)\widehat{\left(\rho \star \eta \star D\right)}\left(\mathbf{k}\right)\ d\mathbf{k} - \frac{1}{2}\theta\left(r\to 0\right)\sum_{i}^{N} q_i^2$$

Poisson's summation formula is given by

$$D(r) = \sum_{n\in\mathbb{Z}^3} \delta(\mathbf{r} - L\mathbf{n}) \rightarrow \hat{D}(\mathbf{k}) = \frac{1}{V} \sum_{\mathbf{d}\in\mathbb{Z}^3} \delta\left(\mathbf{k} - L^{-T}\mathbf{d}\right)$$

The Poisson's summation formula combined with the Fourier transform convolution identity ($a \hat{\star} b = \hat{a}\hat{b}$),

$$\widehat{(\rho \star \eta \star D)}(\mathbf{k}) = \hat{\rho}(\mathbf{k})\, \hat{\eta}(\mathbf{k})\, \hat{D}(\mathbf{k})$$

$$= \hat{\rho}(\mathbf{k})\, \hat{\rho}(-\mathbf{k}) \frac{1}{V} \sum_{\mathbf{d}\in\mathbb{Z}^3} \delta\left(\mathbf{k} - L^{-T}\mathbf{d}\right)$$

$$= |\hat{\rho}(\mathbf{k})|^2 \frac{1}{V} \sum_{\mathbf{d}\in\mathbb{Z}^3} \delta\left(\mathbf{k} - L^{-T}\mathbf{d}\right)$$

This expression is close to the desired form. To finish,

$$\int \hat{\theta}(k)\, \widehat{(\rho \star \eta \star D)}(\mathbf{k})\, d\mathbf{k} = \int \hat{\theta}(k)\, |\hat{\rho}(\mathbf{k})|^2 \frac{1}{V} \sum_{\mathbf{d}\in\mathbb{Z}^3} \delta\left(\mathbf{k} - L^{-T}\mathbf{d}\right) d\mathbf{k}$$

$$= \frac{1}{V} \sum_{\mathbf{d}\in\mathbb{Z}^3} \hat{\theta}\left(\left|L^{-T}\mathbf{d}\right|\right) \left|\hat{\rho}\left(L^{-T}\mathbf{d}\right)\right|^2$$

$$= \frac{1}{V} \sum_{\mathbf{m}} \hat{\theta}(m)\, |\hat{\rho}(\mathbf{m})|^2$$

$$= \frac{1}{V} \sum_{\mathbf{m}\neq 0} \hat{\theta}(m)\, |\hat{\rho}(\mathbf{m})|^2 + \frac{1}{V} \lim_{m\to 0} \hat{\theta}(m)\, |\hat{\rho}(0)|^2$$

$$= \frac{1}{V} \sum_{\mathbf{m}\neq 0} \hat{\theta}(m)\, |\hat{\rho}(\mathbf{m})|^2 + \frac{1}{V} \lim_{m\to 0} \hat{\theta}(m) \left(\sum_i^N q_i\right)^2$$

where $\mathbf{m} \equiv L^{-T}\mathbf{d}$. At last, it has been proven that

$$\int \hat{\theta}(k)\, \widehat{(\rho \star \eta \star D)}(\mathbf{k})\, d\mathbf{k} = \frac{1}{V} \sum_{\mathbf{m}\neq 0} \hat{\theta}(m)\, |\hat{\rho}(\mathbf{m})|^2 + \frac{1}{V} \hat{\theta}(m \to 0) \left(\sum_i^N q_i\right)^2$$

where $\hat{\theta}(m \to 0) \equiv \lim_{m\to 0} \hat{\theta}(m)$.

Thus equation (3.19) has been proven and so the long-range potential energy of the Ewald sum can be written as

$$u_{\mathrm{LR}} = \frac{1}{2}\sum_{i,j}^{N} q_i q_j \theta\left(\left|\mathbf{r}_i - \mathbf{r}_j\right|\right) + \frac{1}{2}\sum_{\mathbf{n}\in\mathbb{Z}^3}\sum_{i,j}^{N} q_i q_j \theta\left(\left|\mathbf{r}_i - \mathbf{r}_j + L\mathbf{n}\right|\right)$$

$$= \frac{1}{2V}\sum_{\mathbf{m}\neq 0} \hat{\theta}\left(m\right)\left|\hat{\rho}\left(m\right)\right|^2 + \frac{1}{2V}\hat{\theta}\left(m\to 0\right)\left(\sum_{i}^{N} q_i\right)^2 - \frac{1}{2}\theta\left(r\to 0\right)\sum_{i}^{N} q_i^2 \tag{3.20}$$

The first term in the second expression is the contribution from all periodic images in k-space, the second term is the 'background charge' contribution term for $m \to 0$ where the net source charge is smeared evenly over the whole unit cell volume, and the third term is the removal of self-interactions as previously described.

As is, this method is no faster to compute than the standard Ewald sum since the first term requires calculating the total electrostatic potential and performing a standard Fourier transform. However, with a few additional modifications the first term can be much more quickly computed by using the fast fourier transform (FFT). Since the FFT operates on a regularly spaced grid, the charge distribution must first be interpolated onto the grid before the FFT can be performed. The chosen method for smearing the charge is the main differentiator between the various flavors of PME. A particularly popular approach called smooth PME (SPME) uses a cardinal b-spline to interpolate the source charge to the grid. The b-spline interpolation has a roughly Gaussian character at high polynomial orders, thus approaching a Gaussian charge distribution. It also has the desirable trait that integration of its weights over the region of interpolation equals 1, thus retaining the value of the original point charge that was interpolated (if not its exact position in the case of interpolated charges overlapping the same grid points). Further, b-splines are fast to compute and add negligible overhead to the Ewald sum calculation.

The b-spline interpolation of point charges to a grid using a Gaussian distribution function can be done by rewriting the Gaussian charge distribution in terms of cardinal b-splines

$$\hat{\rho}(\mathbf{m}) = \sum_{j}^{N} q_j \exp\left(2\pi i \mathbf{m} \cdot \mathbf{r}_j\right) \approx \sum_{j}^{N} q_j \sum_{\mathbf{l} \in \mathbb{Z}^3} W(\mathbf{u} - \mathbf{l}) \exp\left(2\pi i \mathbf{d}^T K^{-1} \mathbf{l}\right) \equiv \hat{Q}(k)$$

where $W(\mathbf{x})$ are the cardinal b-spline weights and $\hat{Q}(k)$ is the b-spline interpolated charge grid. While the unmodified Gaussian could be directly evaluated at grid points, discretization error could be significant unless the grid was very fine. Interpolating using the cardinal b-spline avoids this issue by ensuring that the full charge is spread over the grid points so that integrating over all the interpolated grid points recovers the value of the source point charge.

The modified potential $\hat{\theta}(m)$ can be calculated in k-space as

$$\hat{G}(k) = \frac{\hat{\theta}(k)}{|B(k)|^2} = \frac{4\pi}{k^2} \frac{e^{-k^2/4\alpha}}{|B(k)|^2}$$

where $B(k) = b_x(k)\, b_y(k)\, b_z(k)$ is the 3D b-spline Fourier coefficient. The division by the b-spline Fourier coefficients is performed to eliminate the contribution of the b-spline function to the convolution such that the b-spline is used purely for interpolation and is not part of the physical model.

Thus the product of the Gaussian charge and its modified potential can be approximated by the b-spline interpolated charge grid $\hat{Q}(k)$ and the modified potential $\hat{G}(k)$:

$$\hat{\rho}(m)\, \hat{\theta}(m) \approx \hat{Q}(k)\, \hat{G}(k)$$

where $m = \left|L^{-T}\mathbf{d}\right|$ as before. This is the fundamental approximation of the SPME method. The error introduced by the approximation is dependent on the order of the cardinal b-spline used for interpolation as well as the fineness of the simulation grid. Substituting this expression into equation (3.20), the full SPME long range potential is obtained.

$$u_{\mathrm{LR}} = \frac{1}{2V} \sum_{j \in V} q_j \sum_{\mathbf{m} \neq 0} \hat{G}(\mathbf{m})\, \hat{Q}(\mathbf{m}) + \frac{1}{2V} \hat{\theta}(m \to 0) \left(\sum_{i \in U} q_i\right)^2 - \frac{1}{2} \theta(r \to 0) \sum_{i}^{N} q_i^2 \quad (3.21)$$

From this the long range potential of the solute acting on the solvent can be calculated:

$$u_{\mathrm{LR}}^{U\to V} = \frac{1}{V}\sum_{j\in V} q_j \sum_{\mathbf{m}\neq 0} \hat{G}^{U\to V}(\mathbf{m})\,\hat{Q}_U(\mathbf{m}) + \frac{1}{V}\hat{\theta}(m\to 0)\left(\sum_{i\in U} q_i\right)^2 \tag{3.22}$$

$$\hat{G}^{U\to V}(k) = \frac{4\pi}{k^2}\frac{e^{-k^2/4\alpha}}{B(k)}$$

Note that the self-interaction term is dropped since no solute-solute terms are considered. For the same reason, there is no longer double counting of interactions from the sums and hence the $1/2$ coefficients are removed. Finally, since the solvent charges remain as point charges, the b-spline Fourier coefficient in the denominator of $\hat{G}^{U\to V}$ is not squared.

Simliarly, the long range potential of the solvent acting on the solute is given by:

$$''u_{\mathrm{LR}}^{\mathrm{V}\to\mathrm{U}} = \frac{1}{V}\sum_{i\in U} q_i \sum_{\mathbf{m}\neq 0} \hat{G}^{\mathrm{V}\to\mathrm{U}}(\mathbf{m})\,\hat{Q}_V(\mathbf{m}) + \frac{1}{V}\hat{\theta}(m\to 0)\left(\sum_{i\in U} q_i\right)^2 \tag{3.23}$$

$$\hat{G}^{V\to U}(k) = \hat{G}^{U\to V}(k)$$

To summarize, the PME algorithm for calculating the long range term of the Ewald sum is as follows:

1. Interpolate point source charges to a Cartesian grid. The SPME uses a b-spline to perform interpolation.
2. Convert the source charge grid from real space to frequency space using a FFT.
3. Convolute the source charge grid with the electrostatic interaction Green function. In frequency space the convolution is a simple multiplication.
4. Convert the convoluted kernel from frequency space to real space using a FFT.
5. Calculate the electrostatic potential grid, dividing by the unit cell volume for plane wave normalization and applying a uniform background charge correction.

3.3.2 Solvation Force

The solvation force of the solvent acting on the solute can be directly derived from equation (3.23) and the standard relation between electrical potential and force:

$$\mathbf{F}_{\text{LR PME}}(\mathbf{r}) = -q\nabla\phi_{\text{LR PME}}(\mathbf{r}) = -\nabla U_{\text{LR PME}}(\mathbf{r})$$

$$\nabla U_{\text{LR}}^{V\to U} = \nabla\left[\frac{1}{V}\sum_{i\in U} q_i\beta(\mathbf{r}_i)\left[\sum_{\mathbf{m}\neq 0}\hat{G}^{V\to U}(\mathbf{m})\,\hat{Q}_V(\mathbf{m})\right] + \frac{1}{V}\hat{\theta}(m\to 0)\left(\sum_{j\in V} q_j\right)^2\right]$$

$$= \frac{1}{V}\sum_{i\in U} q_i\nabla\beta(\mathbf{r}_i)\left[\sum_{\mathbf{m}\neq 0}\hat{G}^{V\to U}(\mathbf{m})\,\hat{Q}_V(\mathbf{m})\right] + \nabla\frac{1}{V}\hat{\theta}(m\to 0)\left(\sum_{j\in V} q_j\right)^{2\to 0}$$

$$= \frac{1}{V}\sum_{i\in U} q_i\nabla\beta(\mathbf{r}_i)\left[\sum_{\mathbf{m}\neq 0}\hat{G}^{V\to U}(\mathbf{m})\,\hat{Q}_V(\mathbf{m})\right]$$

where $\beta(\mathbf{r}_i)[f(\mathbf{r})]$ is the b-spline interpolation of $f(\mathbf{r})$ onto grid point i at Cartesian position $\mathbf{r}_i$. Hence $\nabla\beta(\mathbf{r}_i)[f(\mathbf{r})]$ is the Cartesian gradient of the b-spline interpolation. By combining the b-spline weights with their Cartesian gradient, interpolation can be accomplished simultaneously with computing the gradient required to obtain the force in Cartesian coordinates. The PME force gradient can be calculated several other ways, each with their own benefits and downsides. The implementation of periodic 3D-RISM in Amber uses the b-spline analytic gradient primarily because of its speed of computation and reasonable accuracy.

Since solving the 3D-RISM equation produces a solvent density grid $g(r)$ (the radial distribution function), the solvent charge does not need to be interpolated to the grid as it is already on a grid. To obtain the solvent site charge grid, the solvent site density grid must merely be multiplied by its site charge. The charge grid can then be used to calculate a modified form of the solvent electric potential based on equation (3.23),

$$\phi'_{\text{LR PME}}(\mathbf{r}) = Q_U(r) \star G'(r) = \frac{1}{V_{\text{cell}}}\text{IFFT}\left[\tilde{Q}_U(k)\,\tilde{G}'(k)\right]$$

$$\tilde{G}^{'}(k) = \frac{4\pi}{k^2}\frac{1}{B(k)}$$

where $\phi^{'}_{\text{LR PME}}(\mathbf{r})$ is the modified PME potential which lacks the Gaussian coefficient, $G^{'}$ is the modified Green function with the Gaussian charge removed. The Gaussian charge has been removed from the Green function since the 3D-RISM solves for a discretized version of a continuous solvent density distribution, and hence point charges are not present to produce Gaussians.

An aside on removing the Gaussian from the Green function: it is unclear to the author if removing the Gaussian is the best choice, but it seems to work in practice. The solute point charges that are interpolated to a grid during the solute-to-solvent PME (prior to solving the 3D-RISM) are only treated as Gaussians *after* interpolation, so their resulting grid does not represent the true Gaussian source charges either. Further, the purpose of adding the periodic Gaussian is to increase the convergence rate of the Fourier transform by introducing a smoothly varying periodic function. Without the Gaussian the FFT is acting on a step function, which is a worst case scenario for Fourier transforms and may lead to failed convergence. Nontheless, leaving the Gaussian out does not seem to produce poor results when calculating the solvent-to-solute force, possibly because the continuous solvent density grid produced by the 3D-RISM is sufficiently smooth and periodic-like (i.e., low average discontinuities across boundaries) so as to make 'smoothing' by including the Gaussian Fourier coefficients unnecessary. A consequence of not including the Gaussian distribution is that the Gaussian smearing coefficient is not used for solvent-to-solute force calculations.

Using the modified electric potential, the solvent-to-solute atom force can be calculated by simultaneously interpolating the potential to the solute atom and calculating the force via the b-spline gradient method:

$$\mathbf{F}^{u}_{\text{LR PME}}(\mathbf{r}_u) = -q_u \nabla \phi^{'}_{\text{LR PME}}(\mathbf{r}_u)$$

$$\approx q_u \phi^{'}_{\text{LR PME}}(\mathbf{r}_u)\, \mathbf{L} \circ \nabla \mathbf{W}(\mathbf{r}_u)$$

$$\nabla \mathbf{W}(\mathbf{r}) = \partial w_x(r_x)\, w_y(r_y)\, w_z(r_z)\, \mathbf{i} + w_x(r_x)\, \partial w_y(r_y)\, w_z(r_z)\, \mathbf{j} + w_x(r_x)\, w_y(r_y)\, \partial w_z(r_z)\, \mathbf{k}$$

$$\mathbf{L} = L_x \mathbf{i} + L_y \mathbf{j} + L_z \mathbf{k}$$

where $\nabla \mathbf{W}\left(\mathbf{r}^{'}\right)$ is the b-spline interpolation gradient vector at a grid point $\mathbf{r}^{'}$ within order N grid points around the point of interpolation $\mathbf{r}$ (i.e. solute position), $\partial w_x(r_x) \equiv dw_x(r_x)/dx$ is the b-spline weight derivative along axis x (calculated using a simple central difference along that axis about the grid point), $w_i(r)$ is the 1D b-spline weight at grid point r, $\circ$ is the element-wise vector multiplication operator, $\mathbf{L}$ is a vector of the unit cell side lengths. The multiplication by $\mathbf{L}$ is done since $\mathbf{W}$ must be calculated in reciprocal space to handle triclinic unit cells and thus multipyling by $\mathbf{L}$ needed to convert the weights back to real space.

Thus the modified SPME force method used in Amber for periodic 3D-RISM calculations is:

1. Convert the solvent site density grid to a charge grid by multiplying each grid point by its site charge.

2. Convert the solvent charge grid from real space to frequency space using a FFT.

3. Convolute the solvent charge grid with the electrostatic interaction Green function. In frequency space the convolution is a simple multiplication.

4. Convert the convolution kernel from frequency space to real space using a FFT.

5. Multiply the resulting grid by the appropriate constants to obtain the solvent electrostatic potential field grid.

6. Obtain the solvation force on each solute atom by interpolating the solvent potential field grid onto the solute atom positions using an analytic gradient b-spline.

As previously mentioned, the Green function in step 3 has been modified by removing the Gaussian Fourier coefficient.

3.4 Periodic 3D-RISM Theory

To perform a 3D-RISM calculation for a periodic solute, two primary changes must be made to standard 3D-RISM theory:

- a periodic electrostatic interaction must be used, both for electric potential and force calculations
- the 3D-RISM solver algorithm no longer needs to contend with a $k = 0$ term

The first modification has already been covered. The periodic electrostatic interaction of choice is the Ewald sum. When the PME method is employed to calculate the long range term in the Ewald sum, the calculation can be performed with reasonable timings comparable to that of the aperiodic potentials. Long range asymptotic calculations which are critical for infinite dilution 3D-RISM calculations are already included in the long range term of the Ewald sum and thus do not need to be separately calculated.

The second modification is relatively minor and uses the following simplified 3D-RISM solver algorithm:

1. Make an initial guess for $c\left(r\right)$ (only affects rate of convergence, not the solution).
2. FFT: $c\left(r\right)$ to $\hat{c}\left(k\right)$.
3. Solve the 3D-RISM equation in k-space for $\hat{h}\left(k\right)$.
4. IFFT: $\hat{h}\left(k\right)$ to $h\left(r\right)$.
5. Solve the closure relation for $g\left(r\right)$ using the given $h\left(r\right)$ and $c\left(r\right)$.
6. Calculate the DCF residual $\Delta c\left(r\right)$, which is equal to the RDF residual: $\Delta c\left(r\right) = g\left(r\right) - 1 - h\left(r\right)$.
7. If $\Delta c\left(r\right)$ is less than the user specified error tolerance, cease iterating. Otherwise calculate a new $c^{'}\left(r\right) = c\left(r\right) + \Delta c\left(r\right)$ and go to Step 2.

This procedure is nearly identical to the standard 3D-RISM solver algorithm except long range asymptotics at $k = 0$ are no longer removed or restored between Fourier transforms.

Unlike infinite dilution 3D-RISM, periodic 3D-RISM uses tinfoil boundary conditions (i.e., $\hat{\phi}(k=0)=0$) and hence there is no explosive $k=0$ term to contend with.

These combined modifications to the standard 3D-RISM produce the periodic 3D-RISM, which can treat the solvation of solute periodic in all three dimensions (e.g., crystals).

3.5 Periodic 3D-RISM Implementation

There are a few practical matters which must be considered when implementing and using the periodic 3D-RISM to perform simulations.

Many crystal structures have triclinic unit cells (i.e., unit cells with all non-90 degree interior angles). Nothing in the 3D-RISM theory assumes a specific box geometry, so triclinic unit cell support is an implementation detail. The most notable points in the implementation where tricilinic cells must be considered is when calculating the Cartesian coordinates of grid points and when calculating k-space wave vectors. All distances should typically be calculated in Cartesian space, including when applying the minimum image convention. Fortunately the FFT requires no modification to treat triclinic unit cells as it makes no assumption of grid geometry, but the b-spline interpolation of charges to the grid must be performed in reciprocal space to ensure the correct distances are used in the interpolation.

For crystal solute the simulation box dimensions are typically fixed to match the unit cell dimensions. The grid spacing can be user defined, so the user still has control over simulation resolution. The elimination of freedom in box dimensions is a notable simplification over the infinite dilution 3D-RISM, which requires careful choice of box size to balance capturing short-range interactions with computational efficiency.

The implementation created for this work will be released in April, 2016 as part of AmberTool 16, an open source collection of molecular simulation software. The implementation was based upon an existing non-periodic RISM code that was primarily developed by Tyler Luchko, David Case, and Andriy Kovalenko [54]. The code is written in Fortran 90. Distributed parallelization using the MPI communications protocol was

added for the purpose of handling macromolecular solute that require too much physical memory or computation time for a single computer system to simulate. Parallelization is accomplished by dividing the simulation box along the z-axis according to the number of process nodes so that each process node performs a simluation on its assigned grid slab (a process commonly referred to as slab decomposition). Since the FFTW library was used to perform Fourier transforms, the MPI functions in that library were used to facilitate distributed Fourier transforms of the entire simulation grid, avoiding the practical complications and significant performance issues of performing the transform on a single compute node.

3.6 Crystal structure refinement

One practical application of the periodic 3D-RISM is X-ray crystal structure refinement. Traditionally structure refinement is performed as an iterative nonlinear error minimization process between experimental and theoretical structure data using a chemical interaction model:

$$\Delta E = E_{\text{calc}} - w_{\text{obs}} E_{\text{obs}} \tag{3.24}$$

where E_{calc} and E_{obs} are the calculated and observed structure data respectively, and w_{obs} are the observed data weights being optimized to minimize error ΔE. Typically E are the 3D electron density map or structure factor amplitudes. Calculating E_{calc} is usually done by performing a energy minimization molecular dynamics simulation with periodic boundary conditions and a chosen solvation model. Traditionally the solvation model is either an explicit solvent, an implicit solvent model such as Generalized Borne or Poisson-Boltzmann, or a 'flat' solvent model where the solvent is uniformly distributed in the unit cell similar to bulk solvent. As an alternative, the periodic 3D-RISM can be used as a solvation model. This allows potentially shorter computation times than explicit solvent models, while possibly providing higher accuracy than implicit and 'flat' solvation models. Further, since a solution of the 3D-RISM produces 3D solvent density distribution grids, electron density maps and associated structure factors can be conveniently

calculated with little or no additional theoretical approximations.

The 3D-RISM electron density grid is obtained from the radial distribution function grid. To do this for water, the oxygen RDF grid is interpreted as a water RDF. A new electron density grid is created from this by summing the charge contribution of each water RDF grid point, 'smearing' them over a user defined number of neighboring grid points using an experimental 1D solvent electron RDF to guide the interpolation.

Structure factors can then be calculated from the electron density map using a simple Fourier transform,

$$F(r) = \int \sigma_e^{\alpha}(r)\, e^{2\pi i \mathbf{k}\cdot\mathbf{r}} dV$$

where V is the simulation box volume, and $\sigma_e^{\alpha}(r)$ is the electron density for solvent species α. Structure factors are proportional to intensity of the reflected beam by $I_{hkl} \propto |F(hkl)|^2$ and thus can be calculated directly from experimental X-ray scattering intensity data.

The R-factor is a common measure of agreement between a crystallographic model and X-ray diffraction data and is a popular measure of structure refiniment quality. It is calculated by

$$R = \frac{\sum |F_{\text{obs}}| - |F_{\text{calc}}|}{\sum |F_{\text{obs}}|}$$

where F_{obs} and F_{calc} are the observed and calculated structure factors respectively.

To demonstrate the practical utility of the periodic 3D-RISM in crystal structure refinement, R-factors were calculated for four solvated crystal structures obtained from the PDB with associated experimental structure factors. For purpose of comparison, three different solvent models were used to calculate refined F_{calc}: explicit solvent, the periodic 3D-RISM solvent, and a "flat" solvent featured in the refmac5 refinement software application.

3.6.1 Methods

3.6.1.1 General

The six macromolecule crystal structures being refined have Protein Data Bank (PDB) IDs 1AHO (scorpion toxin protein), 1BZR (whale myoglobin), 2IGD (protein G IGG-binding domain II), 2LZT (lysozyme), 4LZT (hen egg white lysozyme), and 4YUL / 3K0N (Cyclophilin A - CypA).

For both 3D-RISM and explicit solvent, MD energy minimization calculations were performed using the sander application based on a developmental version of AmberTools derived from the AmberTools 15 release. The source code will be made available in the AmberTools 16 release. AmberTools is a free and open source molecular simulation toolset. The cSPC/E water model was used for both explicit and 3D-RISM solvents. The cSPC/E parameters are given in table 2.3. The long-range electrostatic interaction was calculated using the smooth particle mesh Ewald (SPME) method.[17, 19]

Refinement calculations were performed using the Refmac 5 software application. The refinement procedure requires two input structures: the solvent density distribution (as calcluated by the solvent model) and the energy minimized solute structure. During refinement, the solvent density is held constant (except for an overall scaling factor, which is refined), while the atomic positions and B-factors of the solute are modified to achieve best agreement with the observed diffraction intensities. The final R-factor is obtained after 40 refinement cycles.

3.6.1.2 Periodic 3D-RISM solvent

All 3D-RISM solvent model calculations were performed with the Kovalenko-Hirata (KH) closure on a uniform 0.35 Å spaced grid. A 10^{-6} correlation function convergence error tolerance was enforced at each MD time step. The periodic 3D-RISM implementation directly produced solvent distributions which were used in Refmac refinement calculations.

3.6.1.3 Explicit solvent

For explicit solvent model calculations, each solute was immersed in a pre-equilibriated cubic water box with a buffer distance of 20 Å. Counter-ions were added to neutralize the protein systems. Non-bonded interaction cutoff was set at 9.0 Å. Equations of motion were integrated by employing the leap-frog Verlet algorithm with a 2 fs time step. Covalent bond lengths involving hydrogen atoms were constrained using SHAKE.[75] The system was first minimized with 2000 steps of steepest descent, followed by 3000 steps of conjugate gradient method to remove bad contacts. The system was then equilibrated at 298.15 K and 1 atm with the solute kept fixed (with the restraint of 10.0 kcal/(mol.$Å^2$)) for 25ns. Temperature was regulated by using Langevin thermostat with a collision frequency of 2.0 ps^{-1} while pressure was maintained using Berendsen barostat. All simulations were performed using the GPU accelerated pmemd code (`pmemd.cuda`).[25, 76, 50] Only the last 20ns of trajectories were kept and their solvent distributions averaged to produce solvent distributions suitable for use in Refmac refinement calculations.

3.6.2 Results

The R-factor and R-free after 40 cycles of Refmac refinement for the six single conformer protein structures is shown in table 3.2. There is an average drop in R of 0.019 between flat and explicit MD, and an average drop of 0.011 between flat and 3D-RISM. Further, R-factor values for the 3D-RISM solvent model are typically close to in between those of the flat and explicit solvent models, with the notable exceptions of lysozyme 4LZT and 2LZT where 3D-RISM nearly matches the explicit R-factor. Thus in terms of R-factors, the 3D-RISM seems to have accuracy between the explicit and flat solvent models, with the curious exception of lysozymes.

The R-factor and R-free for 1AHO in both single and multiple conformations is shown in table 3.3. As in table 3.2, the R-factors of 3D-RISM are roughly in between flat and explicit solvent models, indicating that this trend is not merely a byproduct of only considering a single protein conformation.

Plots of the R-factors at multiple refinement resolutions for five of the proteins is

Protein	*scorpion toxin*	*GB3*	*myoglobin*	*lysozyme*	*lysozyme*	*cyclophilin*
PDB ID/resol.	1AHO/0.96	2IGD/1.10	1BZR/1.15	4LZT/0.95	2LZT/1.97	4YUL/1.42
flat (Refmac)	.209/.214	.220/.233	.200/.208	.196/.205	.167/.216	.201/.224
3D-RISM	.197/.211	.213/.224	.194/.206	.190/.197	.154/.201	.185/.202
explicit MD	.189/.198	.191/.209	.186/.192	.191/.202	.153/.214	.172/.185

Table 3.2: Solvent models with a single protein configuration; each block shows R/Rfree after 40 cycles of refmac refinement.

Solvent Model	R	R-free
single protein conformation		
flat (default)	.209	.214
RISM	.197	.211
explicit	.189	.198
multiple protein conformation		
flat (default)	.178	.190
RISM	.158	.174
explicit	.144	.167

Table 3.3: R-factor and R-free values for single and multiple protein conformations of 1AHO refined in water using the flat, periodic 3D-RISM, and explicit solvation models.

shown in figure 3.2. In general these plots hold no surprises. At coarse resolutions of around 5 Å and up, 3D-RISM tends to perform slightly worse than the three solvent models, but at reasonable refinement resolutions of about 1 to 4 Å 3D-RISM is roughly in the middle of the flat and explicit solvent model R-factors, matching the trend observed in table 3.2 and table 3.3.

Timings for single snapshot 3D-RISM solvation calculations are shown in table 3.4. Even for the relatively large grid of 140 x 120 x 90 Å for 1AHO, the calculation time is only 7.74 minutes. When a parallel execution is performed on two MPI processes the total calculation times are roughly halved, as expected. Thus the single-process computation times are short, and the times can be further linearly curtailed via parallelization.

3.7 Conclusions

The 3D-RISM was extended to handle periodic solute (e.g., crystals). This involved replacing the Coulombic potential with the Ewald sum to handle the periodic interactions, using the Smooth Particle Mesh Ewald (SPME) potential to speed up calculation of the

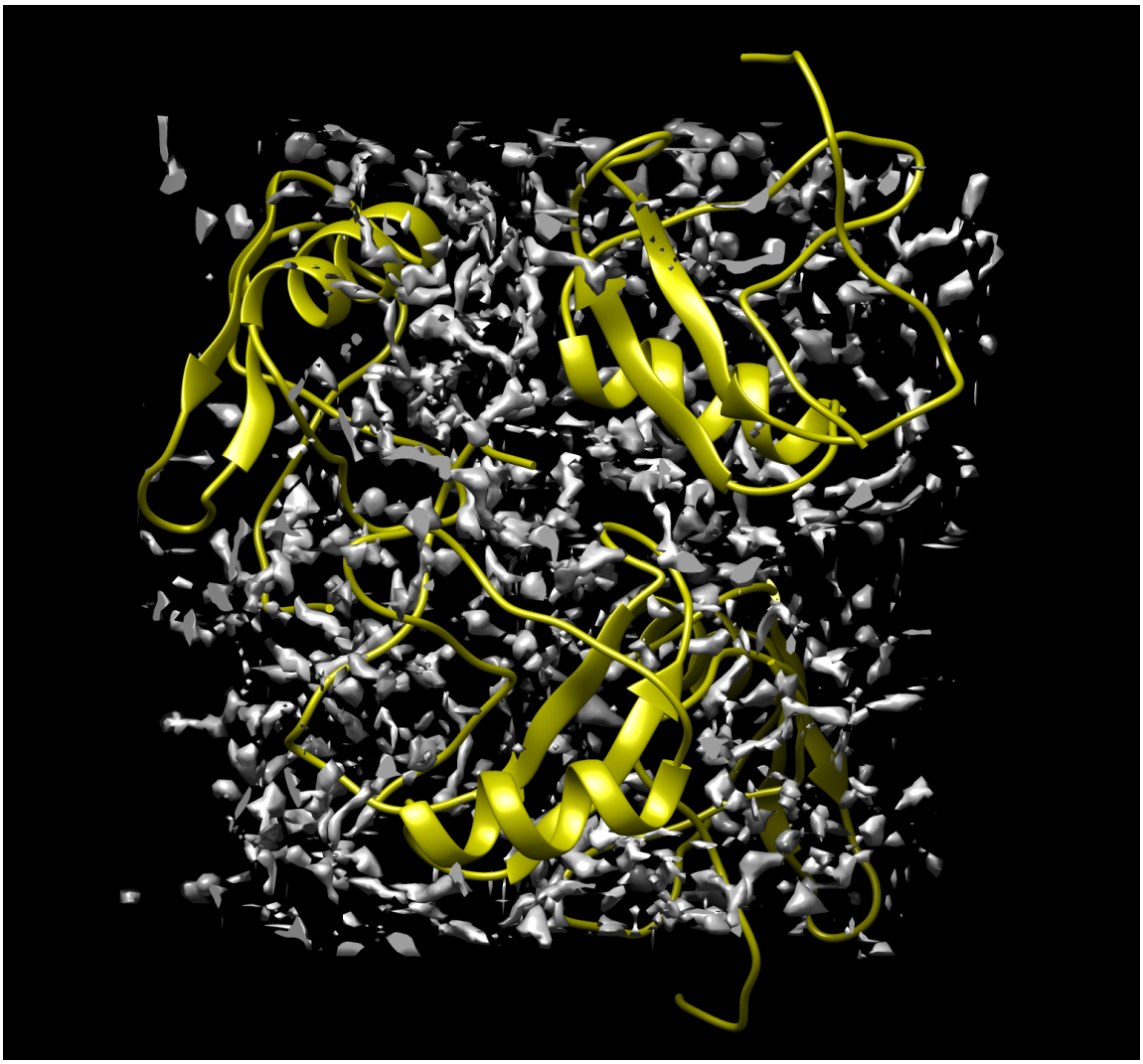

Figure 3.1: Water density distribution about a scorpion toxin protein (PDB ID 1AHO).

Protein	atom count	grid dimensions (Å)	threads	time (s)
1AHO	3848	140 x 120 x 90	1	464.4
			2	247.5
2IGD	3708	72 x 84 x 90	1	166.1
			2	86.6
4LZT	1984	56 x 64 x 70	1	58.42
			2	29.72

Table 3.4: Timings of single snapshot 3D-RISM solvation calculations using varying numbers of execution threads. CPU: Intel Core i7-5700HQ @ 2.7 GHz.

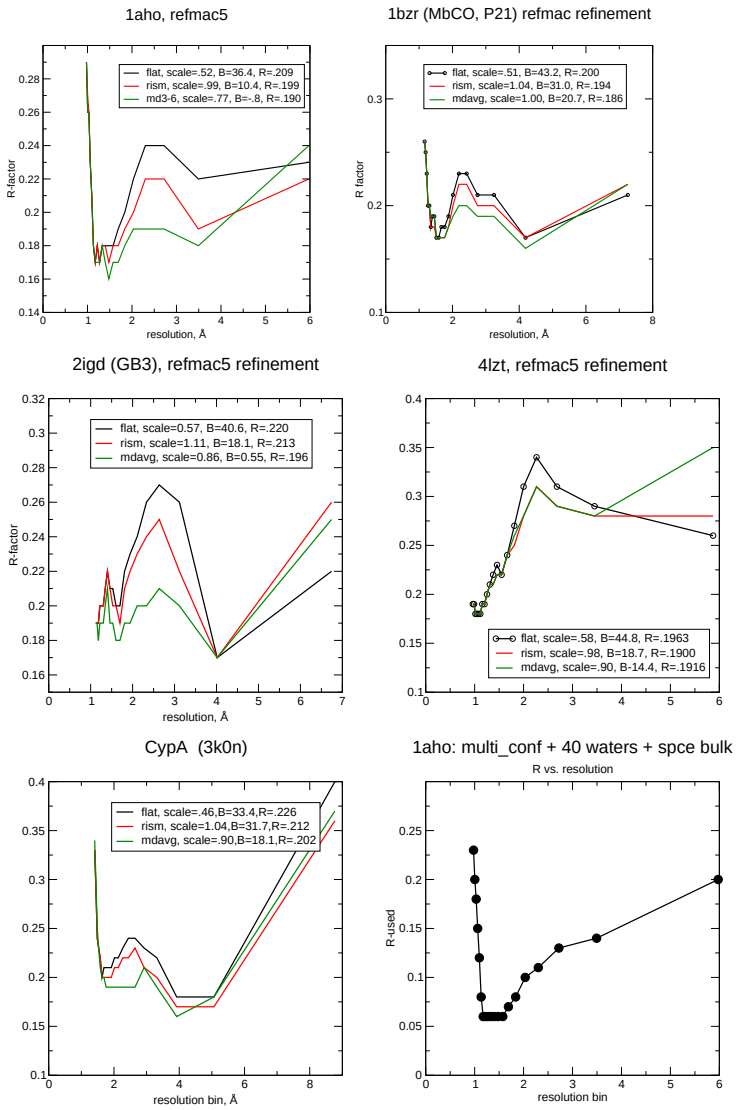

Figure 3.2: R-factor values for various solute structures refined in water using the explicit, periodic 3D-RISM, and flat solvation models.

long range term in the Ewald sum, and introducing the minimum image convention to the short range term of the Ewald sum and Lennard-Jones potential. In addition, a few minor details of the standard 3D-RISM solver were simplified, particularly there is no longer need to be concerned with divergence at $k = 0$ due to the use of tinfoil boundary conditions in the calculation of the SPME. This new periodic 3D-RISM does not correctly balance ionic solvent counts in order to neutralize the solute, an issue which will need to be addressed in future work.

The periodic 3D-RISM was applied to crystallographic structure refinement of six protein structures. It was found that the average R-factor using the periodic 3D-RISM solvation model is an improvement over the flat solvent model and sits roughly between the R-factors of the explicit solvent and flat solvent. Considering that the 3D-RISM simulations only take seconds to minutes to complete, whereas the equivalent MD simulations require hours or days, the periodic 3D-RISM solvent model poses an interesting compromise of accuracy and computational efficiency that may be useful to crystallographers seeking rapid structure refinement protocols where solvent distributions may be updated frequently.

Chapter 4

Appendix

4.1 Correction of 3D-RISM solvation thermodynamics for small drug-like molecules

4.1.1 Gaussian Fluctuation Approximation

The Gaussian fluctuation approximation (GF) [13, 36] has been shown to produce better results than HNC in many [36, 51] but not all[24, 34] cases. It has the form

$$\Delta\mu^{\mathrm{GF}} = kT\sum_{\gamma}\rho_{\gamma}\int -c_{\gamma}\left(\mathbf{r}\right) - \frac{h_{\gamma}\left(\mathbf{r}\right)c_{\gamma}\left(\mathbf{r}\right)}{2}\mathrm{d}\mathbf{r}. \tag{4.1}$$

Applying equation (4.5) we have

$$\delta_T\Delta\mu^{\mathrm{GF}} = \Delta\mu^{\mathrm{GF}} + kT\sum_{\gamma}\rho_{\gamma}\int -\delta_T c_{\gamma}\left(\mathbf{r}\right) - \frac{1}{2}\left[\left\{\delta_T h_{\gamma}\left(\mathbf{r}\right)\right\}c_{\gamma}\left(\mathbf{r}\right) + h_{\gamma}\left(\mathbf{r}\right)\delta_T c_{\gamma}\left(\mathbf{r}\right)\right]d\mathbf{r}. \tag{4.2}$$

This gives

$$\begin{aligned}\Delta\epsilon^{\mathrm{GF}} &= \Delta\mu^{\mathrm{GF}} - \delta_T\Delta\mu^{\mathrm{GF}} \\ &= -kT\sum_{\gamma}\rho_{\gamma}\int -\delta_T c_{\gamma}\left(\mathbf{r}\right) \\ &\qquad - \frac{1}{2}\left[\left\{\delta_T h_{\gamma}\left(\mathbf{r}\right)\right\}c_{\gamma}\left(\mathbf{r}\right) + h_{\gamma}\left(\mathbf{r}\right)\delta_T c_{\gamma}\left(\mathbf{r}\right)\right]d\mathbf{r}.\end{aligned} \tag{4.3}$$

4.1.2 Ng Bridge Correction

Closely related to the UC, [85] proposed the following correction

$$\Delta\mu^{\mathrm{NgB}} = \Delta\mu^{\mathrm{RISM}} + \frac{kT\rho_{\mathrm{O}}}{2}\left(1-\gamma\right)\int c_{\mathrm{O}}^{\mathrm{np}}\left(\mathbf{r}\right)d\mathbf{r} \tag{4.4}$$

where $c_{\mathrm{O}}^{\mathrm{np}}$ is the non-polar CDF of oxygen – calculated with the solute charges turned off – ρ_{O} is the bulk number density of oxygen and γ is an adjustable parameter. While the KH closure was originally used, here we extend the correction to use any PSE-n closure, which includes KH when $n = 1$.

Since NgB uses only non-polar terms to correction $\Delta\mu^{\mathrm{RISM}}$, the polar/non-polar decomposition is simply

$$\begin{aligned}\Delta\mu_{\mathrm{Pol}}^{\mathrm{NgB}} &= \Delta\mu_{\mathrm{Pol}}^{\mathrm{RISM}} \\ \Delta\mu_{\mathrm{NP}}^{\mathrm{NgB}} &= \Delta\mu_{\mathrm{NP}}^{\mathrm{RISM}} + \frac{kT\rho_{\mathrm{O}}}{2}(1-\gamma)\int c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\,d\mathbf{r}.\end{aligned}$$

From this we see that the correction does not change the polar component of the solvation free energy.

As with UC, we apply a linear temperature dependence to the fit parameter,

$$\gamma = \gamma_0 + \gamma_1 T.$$

We will denote the linear temperature dependence as NgBT. Taking the temperature derivative we have

$$\delta_T\Delta\mu^{\mathrm{NgBT}} = \delta_T\Delta\mu^{\mathrm{RISM}} + \frac{kT\rho_{\mathrm{O}}}{2}\left\{(1-\gamma)\left[\int c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\,d\mathbf{r} + \int \delta_T c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\,d\mathbf{r}\right] - \gamma_1 T\int c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\,d\mathbf{r}\right\}$$

We will denote the linear temperature dependence as NgBT. Taking the temperature derivative we have

$$\delta_T\Delta\mu^{\mathrm{NgBT}} = \delta_T\Delta\mu^{\mathrm{RISM}} + \frac{kT\rho_{\mathrm{O}}}{2}\left\{(1-\gamma)\left[\int c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\,d\mathbf{r} + \int \delta_T c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\,d\mathbf{r}\right] - \gamma_1 T\int c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\,d\mathbf{r}\right\}$$

Using for equation (2.6) we have

$$\begin{aligned}\Delta\epsilon^{\mathrm{NgBT}} &= \Delta\mu^{\mathrm{NgB}} - \delta_T\Delta\mu^{\mathrm{NgBT}} \\ &= \Delta\mu^{\mathrm{RISM}} + \frac{kT\rho_{\mathrm{O}}}{2}(1-\gamma)\int c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\,d\mathbf{r} - \delta_T\Delta\mu^{\mathrm{RISM}} \\ &\quad - \frac{kT\rho_{\mathrm{O}}}{2}\left\{(1-\gamma)\left[\int c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\,d\mathbf{r} + \int \delta_T c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\,d\mathbf{r}\right] - \gamma_1 T\int c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\,d\mathbf{r}\right\} \\ &= \Delta\epsilon^{\mathrm{RISM}} - \frac{kT\rho_{\mathrm{O}}}{2}\left\{(1-\gamma)\int \delta_T c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\,d\mathbf{r} - \gamma_1 T\int c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\,d\mathbf{r}\right\}.\end{aligned}$$

Without the temperature dependence, we have

$$\Delta\epsilon^{\mathrm{NgB}} = \Delta\epsilon^{\mathrm{RISM}} - \frac{kT\rho_{\mathrm{O}}}{2}\left\{(1-\gamma)\int \delta_T c_{\mathrm{O}}^{\mathrm{np}}(\mathbf{r})\,d\mathbf{r}\,(\mathbf{r})\,d\mathbf{r}\right\}.$$

As with UC, the temperature dependence of γ does not change the fitting procedure to obtain accurate solvation free energies and only one new parameter needs to be fit against empirical enthalpies and entropies. For testing purposes, we will denote the original correction NgB and NgBT when γ_1 is included.

4.1.3 Temperature derivatives

4.1.3.1 1D-RISM

The analytic expression for the temperature derivative of the RISM equation has been derived by [92] and gives the expression

$$\delta_T \hat{\mathbf{h}}^{\mathrm{VV}} = \hat{\boldsymbol{\omega}} \delta_T \hat{\mathbf{c}}^{\mathrm{VV}} \hat{\boldsymbol{\omega}} + \hat{\boldsymbol{\omega}} \delta_T \hat{\mathbf{c}}^{\mathrm{VV}} \boldsymbol{\rho} \hat{\mathbf{h}}^{\mathrm{VV}} + \hat{\boldsymbol{\omega}} \hat{\mathbf{c}}^{\mathrm{VV}} \boldsymbol{\rho} \delta_T \hat{\mathbf{h}}^{\mathrm{VV}}$$

in reciprocal space, where we have used the functional isochoric temperature derivative

$$\delta_T \equiv T \left(\frac{\partial}{\partial T} \right)_{\rho}. \tag{4.5}$$

Note that this definition provides the useful relations $\delta_T kT = kT$ and $\delta_T \beta = -\beta$.

For DRISM, we apply δ_T to equation (1.7) to get

$$\begin{aligned}
\delta_T \hat{\mathbf{h}}'^{\mathrm{VV}} = \quad & \left\{ \delta_T \hat{\boldsymbol{\omega}}' \right\} \hat{\mathbf{c}}^{\mathrm{VV}} \hat{\boldsymbol{\omega}}' + \hat{\boldsymbol{\omega}}' \left\{ \delta_T \hat{\mathbf{c}}^{\mathrm{VV}} \right\} \hat{\boldsymbol{\omega}}' + \hat{\boldsymbol{\omega}}' \hat{\mathbf{c}}^{\mathrm{VV}} \delta_T \hat{\boldsymbol{\omega}}' \\
& + \left\{ \delta_T \hat{\boldsymbol{\omega}}' \right\} \hat{\mathbf{c}} \rho \hat{\mathbf{h}}'^{\mathrm{VV}} + \hat{\boldsymbol{\omega}}' \left\{ \delta_T \hat{\mathbf{c}}^{\mathrm{VV}} \right\} \rho \hat{\mathbf{h}}'^{\mathrm{VV}} + \hat{\boldsymbol{\omega}}' \hat{\mathbf{c}}^{\mathrm{VV}} \rho \delta_T \hat{\mathbf{h}}'^{\mathrm{VV}} \\
= & \left[\left\{ \delta_T \hat{\boldsymbol{\omega}}' \right\} \hat{\mathbf{c}}^{\mathrm{VV}} + \hat{\boldsymbol{\omega}}' \delta_T \hat{\mathbf{c}}^{\mathrm{VV}} \right] \left[\hat{\boldsymbol{\omega}}' + \rho \hat{\mathbf{h}}'^{\mathrm{VV}} \right] + \hat{\boldsymbol{\omega}}' \hat{\mathbf{c}}^{\mathrm{VV}} \left[\delta_T \hat{\boldsymbol{\omega}}' + \rho \delta_T \hat{\mathbf{h}}'^{\mathrm{VV}} \right].
\end{aligned}$$

Now we use

$$\delta_T \hat{\mathbf{h}}'^{\mathrm{VV}} = \delta_T \hat{\mathbf{h}}^{\mathrm{VV}} - \delta_T \hat{\mathbf{D}}$$

and

$$\delta_T \hat{\boldsymbol{\omega}}' = \delta_T \hat{\boldsymbol{\omega}} + \rho \delta_T \hat{\mathbf{D}} = \rho \delta_T \hat{\mathbf{D}}$$

to get

$$\begin{aligned}
\delta_T \hat{\mathbf{h}}^{\mathrm{VV}} - \delta_T \hat{\mathbf{D}} &= \left[\left\{ \delta_T \hat{\boldsymbol{\omega}}' \right\} \hat{\mathbf{c}}^{\mathrm{VV}} + \hat{\boldsymbol{\omega}}' \delta_T \hat{\mathbf{c}}^{\mathrm{VV}} \right] \left[\hat{\boldsymbol{\omega}} + \cancel{\rho \hat{\mathbf{D}}} + \rho \hat{\mathbf{h}}^{\mathrm{VV}} - \cancel{\rho \hat{\mathbf{D}}} \right] + \hat{\boldsymbol{\omega}}' \hat{\mathbf{c}}^{\mathrm{VV}} \left[\cancel{\rho \delta_T \hat{\mathbf{D}}} + \rho \delta_T \hat{\mathbf{h}}^{\mathrm{VV}} - \cancel{\rho \delta_T \hat{\mathbf{D}}} \right] \\
&= \left[\left\{ \delta_T \hat{\boldsymbol{\omega}}' \right\} \hat{\mathbf{c}}^{\mathrm{VV}} + \hat{\boldsymbol{\omega}}' \delta_T \hat{\mathbf{c}}^{\mathrm{VV}} \right] \hat{\boldsymbol{\chi}} + \hat{\boldsymbol{\omega}}' \hat{\mathbf{c}}^{\mathrm{VV}} \rho \delta_T \hat{\mathbf{h}}^{\mathrm{VV}} \\
\delta_T \hat{\mathbf{h}}^{\mathrm{VV}} &= \left[\mathbf{1} - \hat{\boldsymbol{\omega}}' \hat{\mathbf{c}}^{\mathrm{VV}} \rho \right]^{-1} \left[\left[\left\{ \delta_T \hat{\boldsymbol{\omega}}' \right\} \hat{\mathbf{c}}^{\mathrm{VV}} + \hat{\boldsymbol{\omega}}' \delta_T \hat{\mathbf{c}}^{\mathrm{VV}} \right] \hat{\boldsymbol{\chi}} + \delta_T \hat{\mathbf{D}} \right].
\end{aligned}$$

4.1.3.2 3D-RISM

For 3D-RISM, we apply the temperature derivative to the full 3D-RISM equation, equation (1.10), giving

$$\begin{aligned}\delta_T\hat{\mathbf{h}} &= \{\delta_T\hat{\mathbf{c}}\}\left(\hat{\boldsymbol{\omega}} + \boldsymbol{\rho}\hat{\mathbf{h}}^{\mathrm{VV}}\right) + \hat{\mathbf{c}}\boldsymbol{\rho}\delta_T\hat{\mathbf{h}}^{\mathrm{VV}} \\ &= \{\delta_T\hat{\mathbf{c}}\}\,\hat{\boldsymbol{\chi}}^{\mathrm{VV}} + \hat{\mathbf{c}}\delta_T\hat{\boldsymbol{\chi}}^{\mathrm{VV}}\end{aligned} \tag{4.6}$$

where

$$\delta_T\hat{\boldsymbol{\chi}}^{\mathrm{VV}} = \boldsymbol{\rho}\delta_T\hat{\mathbf{h}}^{\mathrm{VV}}$$

is obtained from 1D-RISM.

4.1.4 Long-Range Asymptotics

In order to compute solutions to 1.9 and 4.6, it is necessary to account for the long-range behavior of electrostatic interactions, which cannot be Fourier transformed due to divergence at small k. The use of long-range asymptotics[82, 63, 3] has been described for 1D-RISM and 3D-RISM [43, 45, 27].

As we are only concerned with pure water, we need only consider the long-range behavior of $c_\alpha(\mathbf{r})$, which is approximated as

$$c_\alpha^{(\mathrm{as})}(\mathbf{r}) = -\beta q_\alpha \sum_i^{N^{\mathrm{U}}} \frac{Q_i}{|\mathbf{r}-\mathbf{R}_i|}\mathrm{erf}\left(\frac{|\mathbf{r}-\mathbf{R}_i|}{\eta}\right), \tag{4.7}$$

where q_α is the charge of solvent site α, Q_i and $\mathbf{R}_i$ are the partial charge and position of site i of N^{U} solute sites, η is the charge smearing coefficient and erf is the error function. $c_\alpha^{(\mathrm{as})}(\mathbf{r})$ is subtracted from $c_\alpha(\mathbf{r})$ before performing a forward Fourier transform and then

$$\hat{c}_\alpha^{(\mathrm{as})}(\mathbf{k}) = -4\pi\beta q_\alpha \sum_i^{N^{\mathrm{U}}} Q_i \frac{e^{-\left(\frac{k\eta}{2}\right)^2 - i\mathbf{k}\cdot\mathbf{R_i}}}{k^2} \tag{4.8}$$

is added back in reciprocal space.

equation (4.6) also requires the temperature derivatives of equation (4.7) and 4.8:

$$\delta_T c_\alpha^{(\mathrm{as})}(\mathbf{r}) = \beta q_\alpha \sum_i^{N^{\mathrm{U}}} \frac{Q_i}{|\mathbf{r}-\mathbf{R}_i|}\mathrm{erf}\left(\frac{|\mathbf{r}-\mathbf{R}_i|}{\eta}\right)$$

and

$$\delta_T \hat{c}_\alpha^{(\mathrm{as})}(\mathbf{k}) = 4\pi\beta q_\alpha \sum_i^{N^{\mathrm{U}}} Q_i \frac{e^{-\left(\frac{k\eta}{2}\right)^2 - i\mathbf{k}\cdot\mathbf{R_i}}}{k^2}. \tag{4.9}$$

4.1.5 Bootstrap Analysis

Correction	a	a_0	a_1	b	b_0	b_1
UC^{KH}	$-0.1498(8)$	$0.009(7)$	$-0.00053(2)$	$-0.1(1)$	$-3.2(9)$	$0.010(3)$
UC^{PSE3}	$-0.1185(7)$	$0.032(7)$	$-0.00051(2)$	$-0.3(1)$	$-3.2(9)$	$0.010(3)$
UC^{HNC}	$-0.1186(7)$	$0.033(7)$	$-0.00051(2)$	$-0.2(1)$	$-3.3(9)$	$0.010(3)$
$UCGF^{KH}$	$-0.1044(8)$	$0.018(7)$	$-0.00041(2)$	$0.6(1)$	$-3.1(9)$	$0.012(3)$
$UCGF^{PSE3}$	$-0.041(1)$	$0.064(7)$	$-0.00036(2)$	$1.1(2)$	$-3(1)$	$0.013(3)$
$UCGF^{HNC}$	$-0.038(1)$	$0.069(8)$	$-0.00036(3)$	$1.1(2)$	$-3(1)$	$0.013(4)$

Correction	γ	γ_0	γ_1
NgB^{KH}	$0.333(1)$	$0.38(1)$	$-0.00015(4)$
NgB^{PSE3}	$0.366(1)$	$0.31(1)$	$0.00019(4)$
NgB^{HNC}	$0.364(1)$	$0.31(1)$	$0.00020(4)$

Table 4.1: Fit parameters for UC and NgB corrections. Standard error in the last digit is given in parentheses.

	ΔG				
	Slope	y-intercept	R^2	RMSE	MUE
KH	1.2(1)	24.6(5)	0.218(1)	24.964(10)	24.030(10)
PSE-3	1.20(8)	20.1(4)	0.305(1)	20.277(7)	19.537(8)
HNC	1.21(9)	20.0(4)	0.301(1)	20.149(8)	19.415(8)
UC^{KH}	0.98(2)	−0.20(5)	0.8413(4)	1.272(2)	0.918(1)
UC^{PSE3}	0.96(2)	−0.23(6)	0.8407(4)	1.261(2)	0.917(1)
UC^{HNC}	0.92(2)	−0.28(5)	0.8625(4)	1.098(1)	0.837(1)
$UCGF^{KH}$	1.04(2)	−0.04(5)	0.8442(4)	1.354(2)	0.965(1)
$UCGF^{PSE3}$	1.30(3)	0.62(7)	0.8432(4)	1.917(4)	1.334(2)
$UCGF^{HNC}$	1.30(2)	0.67(6)	0.8612(5)	1.745(3)	1.296(2)
NgB^{KH}	1.10(2)	0.12(6)	0.8507(4)	1.409(2)	1.018(1)
NgB^{PSE3}	1.12(2)	0.20(6)	0.8509(4)	1.448(3)	1.037(1)
NgB^{HNC}	1.08(2)	0.14(5)	0.8777(4)	1.191(2)	0.905(1)
ISc^{KH}	0.96(2)	−2.17(6)	0.8045(5)	2.494(2)	2.091(2)
ISc^{PSE3}	0.96(2)	−0.75(6)	0.8322(4)	1.432(2)	1.053(1)
ISc^{HNC}	0.92(2)	−0.66(5)	0.8584(4)	1.188(1)	0.919(1)
ISc^{*KH}	0.99(2)	1.08(6)	0.8380(4)	1.705(1)	1.453(1)
ISc^{*PSE3}	1.00(2)	2.59(8)	0.8162(5)	2.951(2)	2.689(2)
ISc^{*HNC}	0.96(2)	2.67(7)	0.8232(5)	3.057(2)	2.811(2)
MD	0.99(2)	0.64(5)	0.8865(3)	1.249(1)	1.025(1)

Table 4.2: Bootstrap statistical comparison between predicted and empirical hydration free energies for neutral molecules (Mobley, Abagyan, Rizzo and Palmer datasets). As described in Methods, values are the mean of all resampled data. RMSE: root-mean-squared-error. MUE: mean unsigned error. Standard error in the last digit is given in parentheses.

	ΔG				
	Slope	y-intercept	R^2	RMSE	MUE
KH	1.14(5)	20(4)	0.8938(7)	12.76(3)	11.07(3)
PSE-3	1.14(6)	17(5)	0.8952(8)	10.62(3)	9.26(2)
HNC	1.16(7)	18(5)	0.9099(7)	10.16(3)	8.86(3)
UC^{KH}	0.91(5)	−7(3)	0.8918(9)	6.58(2)	4.94(2)
UC^{PSE3}	0.93(5)	−6(4)	0.8915(9)	6.53(2)	4.87(2)
UC^{HNC}	0.95(6)	−5(4)	0.9150(7)	6.07(2)	4.51(2)
$UCGF^{PSE3}$	1.10(6)	2(5)	0.8850(9)	9.90(4)	6.97(3)
$UCGF^{KH}$	0.93(5)	−6(4)	0.8959(8)	6.59(2)	4.89(2)
$UCGF^{HNC}$	1.12(7)	2(5)	0.9024(7)	10.54(4)	7.62(4)
NgB^{KH}	1.02(6)	−0(4)	0.8751(10)	7.80(3)	5.97(2)
NgB^{PSE3}	1.05(6)	1(4)	0.868(1)	8.24(3)	6.22(2)
NgB^{HNC}	1.07(7)	3(5)	0.9013(7)	7.71(3)	5.81(3)
ISc^{KH}	0.90(5)	−9(3)	0.8900(9)	6.73(2)	5.07(2)
ISc^{PSE3}	0.92(5)	−7(4)	0.8903(9)	6.57(2)	4.90(2)
ISc^{HNC}	0.94(6)	−6(5)	0.9146(7)	6.07(3)	4.52(2)
ISc^{*KH}	0.93(5)	−6(4)	0.8939(9)	6.51(2)	4.91(2)
ISc^{*PSE3}	0.96(5)	−3(4)	0.8931(9)	6.51(2)	5.03(2)
ISc^{*HNC}	0.98(6)	−2(5)	0.9162(7)	5.97(2)	4.75(2)
MD	0.9(1)	−8.8(1)	0.952(2)	2.933(8)	2.861(9)

Table 4.3: Bootstrap statistical comparison between predicted and empirical hydration free energies for ions (Rizzo dataset). Only the six Joung-Cheatham monovalent ions[39] are included for MD.

	ΔG				
	Slope	y-intercept	R^2	RMSE	MUE
KH	0.9(1)	23.2(4)	0.1474(1)	24.40(1)	23.36(1)
PSE-3	1.00(9)	18.9(3)	0.232(1)	19.695(8)	18.861(8)
HNC	1.03(9)	18.7(3)	0.225(1)	19.483(8)	18.648(8)
UC^{KH}	0.98(1)	−0.85(3)	0.9329(3)	1.151(2)	0.886(1)
UC^{PSE3}	0.97(1)	−0.87(3)	0.9322(3)	1.146(2)	0.885(1)
UC^{HNC}	0.95(1)	−0.93(3)	0.9301(3)	1.145(2)	0.909(1)
$UCGF^{KH}$	1.05(1)	−0.74(4)	0.9347(2)	1.214(2)	0.912(1)
$UCGF^{PSE3}$	1.31(2)	−0.25(4)	0.9333(3)	1.715(3)	1.171(2)
$UCGF^{HNC}$	1.32(2)	−0.28(5)	0.9089(6)	1.785(4)	1.221(2)
NgB^{KH}	1.10(1)	−0.62(3)	0.9423(2)	1.233(2)	0.913(1)
NgB^{PSE3}	1.13(1)	−0.54(3)	0.9465(2)	1.231(2)	0.885(1)
NgB^{HNC}	1.11(1)	−0.63(3)	0.9451(2)	1.187(2)	0.883(1)
ISc^{KH}	0.99(1)	−2.76(3)	0.9339(3)	2.849(1)	2.728(1)
ISc^{PSE3}	0.97(1)	−1.37(3)	0.9394(3)	1.519(2)	1.319(1)
ISc^{HNC}	0.95(1)	−1.29(3)	0.9387(3)	1.400(2)	1.215(1)
ISc^{*KH}	0.98(2)	0.39(5)	0.9065(3)	1.078(1)	0.8556(1)
ISc^{*PSE3}	0.97(2)	1.86(7)	0.8578(4)	2.286(2)	2.023(2)
ISc^{*HNC}	0.96(2)	1.93(7)	0.8369(4)	2.359(2)	2.091(2)

Table 4.4: Bootstrap statistical comparison between predicted and molecular dynamics hydration free energies for neutral molecules (Mobley dataset). As described in Methods, R^2 bootstrap is the mean of all resampled data and R^2 k-fold is the mean over all training sub-samples. RMSE: root-mean-squared-error. MUE: mean unsigned error.

	ΔG_{Pol}				
	Slope	y-intercept	R^2	RMSE	MUE
KH	1.14(1)	−0.19(3)	0.9516(2)	1.068(2)	0.750(1)
PSE-3	1.16(1)	−0.16(4)	0.9530(3)	1.160(2)	0.831(1)
HNC	1.14(1)	−0.23(3)	0.9535(3)	1.052(2)	0.7741(10)
$\mathrm{UC}_{\mathrm{Pol}}^{\mathrm{KH}}$	1.06(1)	−0.21(3)	0.9472(3)	0.832(2)	0.5344(9)
$\mathrm{UC}_{\mathrm{Pol}}^{\mathrm{PSE3}}$	1.07(1)	−0.20(3)	0.9469(3)	0.841(2)	0.5377(9)
$\mathrm{UC}_{\mathrm{Pol}}^{\mathrm{HNC}}$	1.04(1)	−0.26(3)	0.9471(3)	0.748(2)	0.4896(8)
$\mathrm{UCGF}_{\mathrm{Pol}}^{\mathrm{KH}}$	1.11(1)	−0.18(3)	0.9490(3)	0.962(2)	0.642(1)
$\mathrm{UCGF}_{\mathrm{Pol}}^{\mathrm{PSE3}}$	1.36(2)	−0.03(6)	0.9489(3)	1.946(3)	1.480(2)
$\mathrm{UCGF}_{\mathrm{Pol}}^{\mathrm{HNC}}$	1.38(2)	−0.14(5)	0.9234(6)	2.091(4)	1.590(2)
$\mathrm{NgB}_{\mathrm{Pol}}^{\mathrm{KH}}$	1.14(1)	−0.19(3)	0.9514(2)	1.067(2)	0.750(1)
$\mathrm{NgB}_{\mathrm{Pol}}^{\mathrm{PSE3}}$	1.16(1)	−0.16(4)	0.9531(2)	1.162(2)	0.833(1)
$\mathrm{NgB}_{\mathrm{Pol}}^{\mathrm{HNC}}$	1.14(1)	−0.23(3)	0.9535(3)	1.052(2)	0.774(1)
$\mathrm{ISc}_{\mathrm{Pol}}^{\mathrm{KH}}$	1.06(1)	−0.21(4)	0.9463(3)	0.820(2)	0.5240(9)
$\mathrm{ISc}_{\mathrm{Pol}}^{\mathrm{PSE3}}$	1.06(1)	−0.20(4)	0.9459(3)	0.837(2)	0.5330(9)
$\mathrm{ISc}_{\mathrm{Pol}}^{\mathrm{HNC}}$	1.04(1)	−0.26(3)	0.9472(3)	0.739(2)	0.4844(8)
$\mathrm{ISc}^{*\mathrm{KH}}_{\mathrm{Pol}}$	1.07(1)	−0.21(4)	0.9472(3)	0.845(2)	0.5435(9)
$\mathrm{ISc}^{*\mathrm{PSE3}}_{\mathrm{Pol}}$	1.08(1)	−0.19(4)	0.9475(3)	0.880(2)	0.5667(10)
$\mathrm{ISc}^{*\mathrm{HNC}}_{\mathrm{Pol}}$	1.05(1)	−0.26(3)	0.9487(3)	0.779(2)	0.5140(9)

Table 4.5: Bootstrap statistical comparison between predicted and molecular dynamics polar hydration free energies for neutral molecules (Mobley dataset).

	ΔG_{NP}				
	Slope	y-intercept	R^2	RMSE	MUE
KH	2.9(5)	20.5(9)	0.0822(8)	25.09(1)	24.09(1)
PSE-3	2.5(4)	16.8(7)	0.0967(8)	20.471(8)	19.687(8)
HNC	2.6(4)	16.2(7)	0.1022(8)	20.192(8)	19.417(8)
UC_{NP}^{KH}	0.69(3)	0.26(5)	0.586(1)	0.5729(6)	0.4522(5)
UC_{NP}^{PSE3}	0.64(3)	0.36(5)	0.542(1)	0.5856(6)	0.4614(5)
UC_{NP}^{HNC}	0.65(3)	0.26(5)	0.565(1)	0.6250(7)	0.5058(5)
$UCGF_{NP}^{KH}$	0.82(3)	0.11(6)	0.643(1)	0.5025(7)	0.3834(5)
$UCGF_{NP}^{PSE3}$	0.99(4)	0.58(7)	0.650(1)	0.7635(6)	0.6447(6)
$UCGF_{NP}^{HNC}$	0.97(4)	0.75(8)	0.621(1)	0.8641(7)	0.7526(6)
NgB_{NP}^{KH}	1.01(3)	−0.11(5)	0.7636(8)	0.4063(7)	0.2871(4)
NgB_{NP}^{PSE3}	0.99(2)	0.02(4)	0.7937(8)	0.3576(6)	0.2456(4)
NgB_{NP}^{HNC}	1.01(2)	−0.07(4)	0.7927(8)	0.3586(7)	0.2412(4)
ISc_{NP}^{KH}	0.50(2)	−1.32(4)	0.616(1)	2.3245(6)	2.2808(6)
ISc_{NP}^{PSE3}	0.56(2)	−0.01(3)	0.7284(9)	0.9318(6)	0.8528(5)
ISc_{NP}^{HNC}	0.59(2)	0.01(3)	0.7052(9)	0.8797(6)	0.7984(6)
$ISc*_{NP}^{KH}$	0.79(5)	1.32(9)	0.395(1)	1.1612(9)	0.9957(9)
$ISc*_{NP}^{PSE3}$	0.88(7)	2.7(1)	0.263(1)	2.655(1)	2.447(2)
$ISc*_{NP}^{HNC}$	0.92(7)	2.6(1)	0.264(1)	2.683(1)	2.472(2)

Table 4.6: Bootstrap statistical comparison between predicted and molecular dynamics non-polar hydration free energies for neutral molecules (Mobley dataset).

$\Delta H / \Delta \epsilon$

	Slope	y-intercept	R^2	RMSE	MUE
KH	1.01(6)	1.9(7)	0.802(1)	2.702(6)	2.334(6)
PSE-3	1.19(7)	1.2(8)	0.798(2)	2.79(1)	2.045(8)
HNC	1.19(8)	1.0(8)	0.802(2)	2.80(1)	2.041(8)
UC^{KH}	1.8(2)	−9(2)	0.565(3)	20.12(3)	18.82(3)
UC^{PSE3}	1.7(2)	−10(2)	0.568(3)	19.52(3)	18.32(3)
UC^{HNC}	1.7(2)	−10(2)	0.558(3)	19.52(3)	18.30(3)
UCT^{KH}	0.87(6)	−1.6(6)	0.794(1)	1.890(5)	1.541(5)
UCT^{PSE3}	0.89(6)	−1.4(6)	0.806(1)	1.849(5)	1.504(4)
UCT^{HNC}	0.87(5)	−1.6(6)	0.809(1)	1.823(5)	1.479(4)
$UCGF^{KH}$	1.6(2)	−6(2)	0.627(3)	14.36(3)	13.22(2)
$UCGF^{PSE3}$	1.8(1)	−2(1)	0.758(2)	12.13(2)	10.95(2)
$UCGF^{HNC}$	1.9(1)	−1(1)	0.774(2)	12.44(2)	11.20(2)
$UCTGF^{KH}$	0.90(6)	−1.3(6)	0.798(1)	1.897(5)	1.549(4)
$UCTGF^{PSE3}$	1.20(6)	2.4(7)	0.823(1)	2.442(8)	1.928(6)
$UCTGF^{HNC}$	1.29(7)	3.5(8)	0.814(1)	2.785(9)	2.183(7)
NgB^{KH}	1.01(6)	1.9(7)	0.803(1)	2.697(5)	2.330(5)
NgB^{PSE3}	1.19(7)	1.2(7)	0.802(1)	2.77(1)	2.030(8)
NgB^{HNC}	1.18(7)	1.0(8)	0.801(2)	2.81(1)	2.051(8)
$NgBT^{KH}$	1.01(6)	2.0(7)	0.804(1)	2.798(6)	2.424(6)
$NgBT^{PSE3}$	1.19(7)	1.1(8)	0.800(1)	2.88(1)	2.103(8)
$NgBT^{HNC}$	1.19(7)	0.9(8)	0.799(2)	2.92(1)	2.126(8)
ISc^{KH}	0.98(6)	0.0(7)	0.784(2)	2.130(6)	1.638(6)
ISc^{PSE3}	0.98(6)	0.3(7)	0.799(1)	2.117(6)	1.667(6)
ISc^{HNC}	0.97(6)	0.3(7)	0.800(1)	2.108(5)	1.656(5)
ISc^{*KH}	0.56(8)	−1.0(7)	0.497(3)	3.092(7)	2.734(6)
ISc^{*PSE3}	1.3(2)	−3(2)	0.466(3)	6.59(1)	5.79(1)
ISc^{*HNC}	1.4(2)	−3(2)	0.489(3)	6.79(1)	5.99(1)

Table 4.7: Bootstrap statistical comparison between predicted ΔH (all UC and NgB corrections) or $\Delta \epsilon$ (uncorrected 3D-RISM and parameter free corrections) and ΔH from experiment for neutral molecules (Abraham and Cabani datasets).

$\Delta H / \Delta \epsilon$

	Slope	y-intercept	R^2	RMSE	MUE
KH	0.84(9)	−13(9)	0.848(3)	8.29(5)	6.47(4)
PSE-3	0.88(9)	−12(8)	0.853(2)	8.16(6)	6.35(4)
HNC	0.91(8)	−9(8)	0.848(4)	7.67(8)	5.47(5)
UC^{KH}	0.66(7)	−36(6)	0.821(2)	10.97(5)	8.97(5)
UC^{PSE3}	0.64(7)	−37(6)	0.834(2)	10.95(5)	9.04(5)
UC^{HNC}	0.65(8)	−36(7)	0.827(4)	11.49(7)	9.23(7)
UCT^{KH}	0.80(9)	−19(9)	0.858(2)	8.09(6)	6.17(4)
UCT^{PSE3}	0.83(9)	−17(8)	0.867(2)	7.96(6)	6.12(4)
UCT^{HNC}	0.88(8)	−13(7)	0.872(4)	7.38(7)	5.43(5)
$UCGF^{KH}$	0.72(7)	−28(7)	0.837(2)	9.53(5)	7.77(4)
$UCGF^{PSE3}$	0.95(9)	−13(8)	0.853(2)	11.72(7)	8.60(6)
$UCGF^{HNC}$	0.94(10)	−14(9)	0.841(4)	11.83(10)	9.02(7)
$UCTGF^{KH}$	0.83(9)	−17(9)	0.857(2)	8.07(6)	6.17(4)
$UCTGF^{PSE3}$	1.1(1)	−2(1)	0.865(2)	11.95(8)	8.71(7)
$UCTGF^{HNC}$	1.1(1)	0.6(1)	0.850(4)	11.9(1)	8.53(8)
NgB^{KH}	0.85(9)	−12(9)	0.850(3)	8.23(5)	6.42(4)
NgB^{PSE3}	0.88(9)	−12(9)	0.850(3)	8.15(6)	6.34(4)
NgB^{HNC}	0.91(8)	−9(8)	0.848(4)	7.66(8)	5.46(5)
$NgBT^{KH}$	0.85(9)	−12(9)	0.850(3)	8.23(5)	6.42(4)
$NgBT^{PSE3}$	0.88(9)	−12(9)	0.849(3)	8.17(6)	6.36(4)
$NgBT^{HNC}$	0.91(8)	−9(8)	0.848(4)	7.69(8)	5.48(5)
ISc^{KH}	0.78(9)	−19(8)	0.856(2)	8.25(5)	6.16(5)
ISc^{PSE3}	0.80(8)	−17(8)	0.867(2)	7.89(5)	6.04(4)
ISc^{HNC}	0.85(7)	−13(7)	0.880(3)	6.89(5)	5.28(4)
ISc^{*KH}	0.53(9)	−1.4(7)	0.634(5)	2.614(10)	2.23(1)
ISc^{*PSE3}	0.6(2)	−4(2)	0.398(6)	2.39(1)	1.99(1)
ISc^{*HNC}	0.4(3)	−5(2)	0.238(6)	2.60(2)	2.20(1)

Table 4.8: Bootstrap statistical comparison between predicted ΔH (all UC and NgB corrections) or $\Delta \epsilon$ (uncorrected 3D-RISM and parameter free corrections) and ΔH from experiment for ions (Fawcett and Marcus datasets).

	$T\Delta S$				
	Slope	y-intercept	R^2	RMSE	MUE
KH	2.7(5)	−4(4)	0.432(3)	19.56(3)	18.10(3)
PSE-3	2.6(5)	−4(4)	0.439(3)	18.52(3)	17.13(3)
HNC	2.6(5)	−3(4)	0.460(3)	18.43(3)	17.01(3)
UC^{KH}	2.6(5)	−6(4)	0.435(3)	20.16(3)	18.85(3)
UC^{PSE3}	2.5(5)	−6(4)	0.443(3)	19.56(3)	18.35(3)
UC^{HNC}	2.6(5)	−5(4)	0.460(3)	19.49(3)	18.27(3)
UCT^{KH}	0.52(8)	−4.0(7)	0.489(3)	1.545(5)	1.205(4)
UCT^{PSE3}	0.56(8)	−3.7(7)	0.517(3)	1.497(5)	1.157(4)
UCT^{HNC}	0.58(7)	−3.5(6)	0.564(3)	1.427(5)	1.109(4)
$UCGF^{KH}$	2.1(4)	−4(3)	0.442(3)	14.27(2)	13.16(2)
$UCGF^{PSE3}$	1.9(3)	−3(3)	0.471(3)	11.71(2)	10.69(2)
$UCGF^{HNC}$	2.0(3)	−2(3)	0.502(3)	11.89(2)	10.87(2)
$UCTGF^{KH}$	0.49(8)	−4.2(7)	0.494(3)	1.531(6)	1.193(4)
$UCTGF^{PSE3}$	0.56(7)	−3.4(7)	0.483(3)	1.598(5)	1.205(4)
$UCTGF^{HNC}$	0.64(8)	−2.7(8)	0.492(4)	1.637(6)	1.214(4)
NgB^{KH}	0.67(1)	−0.9(8)	0.500(3)	2.418(6)	2.066(5)
NgB^{PSE3}	1.0(1)	−1(1)	0.495(3)	2.346(8)	1.807(6)
NgB^{HNC}	1.0(2)	−1(1)	0.518(3)	2.345(8)	1.800(6)
$NgBT^{KH}$	0.65(1)	−0.9(8)	0.502(3)	2.538(6)	2.182(6)
$NgBT^{PSE3}$	1.0(2)	−1(1)	0.493(3)	2.459(8)	1.895(7)
$NgBT^{HNC}$	1.0(2)	−1(1)	0.516(3)	2.466(8)	1.900(6)
ISc^{KH}	0.64(1)	−1.2(8)	0.492(3)	2.435(7)	2.065(5)
ISc^{PSE3}	0.7(1)	−1.3(9)	0.507(3)	1.930(6)	1.552(5)
ISc^{HNC}	0.8(1)	−1.0(9)	0.542(3)	1.887(6)	1.509(5)
ISc^{*KH}	0.83(5)	1.8(6)	0.787(1)	4.246(8)	3.806(8)
ISc^{*PSE3}	1.13(8)	−1.9(9)	0.749(2)	4.33(1)	3.48(1)
ISc^{*HNC}	1.13(9)	−2.0(9)	0.742(2)	4.48(1)	3.62(1)

Table 4.9: Bootstrap statistical comparison between predicted $T\Delta S_P$ (all UC and NgB corrections) or $T\Delta S_V$ (uncorrected 3D-RISM and parameter free corrections) and $T\Delta S_P$ from experiment for neutral molecules (Abraham and Cabani datasets).

$T\Delta S$					
	Slope	y-intercept	R^2	RMSE	MUE
KH	1.0(5)	−4(3)	0.295(6)	6.23(3)	5.16(3)
PSE-3	1.0(5)	−4(3)	0.316(6)	6.19(3)	5.16(3)
HNC	0.6(7)	−7(4)	0.183(5)	5.89(4)	4.77(3)
UC^{KH}	0.7(5)	−7(3)	0.195(5)	6.63(4)	5.51(3)
UC^{PSE3}	0.5(5)	−8(3)	0.140(4)	6.51(4)	5.32(3)
UC^{HNC}	0.4(9)	−9(5)	0.209(7)	7.27(5)	5.97(4)
UCT^{KH}	0.56(10)	−4.2(7)	0.604(5)	1.946(10)	1.577(9)
UCT^{HNC}	0.5(1)	−4.9(10)	0.479(6)	2.13(1)	1.81(1)
UCT^{PSE3}	0.7(1)	−3.9(10)	0.570(5)	2.21(1)	1.77(1)
$UCGF^{KH}$	0.7(4)	−5(3)	0.253(5)	4.48(3)	3.72(2)
$UCGF^{PSE3}$	0.9(4)	−4(3)	0.336(6)	4.70(2)	4.08(2)
$UCGF^{HNC}$	0.5(5)	−7(3)	0.195(6)	4.67(2)	4.08(2)
$UCTGF^{KH}$	0.6(1)	−4(1)	0.474(5)	2.27(1)	1.84(1)
$UCTGF^{PSE3}$	1.0(3)	−3(2)	0.383(6)	4.39(2)	3.34(2)
$UCTGF^{HNC}$	0.6(4)	−5(3)	0.231(7)	4.01(3)	3.15(2)
NgB^{KH}	0.7(1)	0.1(9)	0.730(4)	2.91(1)	2.57(1)
NgB^{PSE3}	0.8(1)	−0.3(9)	0.689(4)	1.91(1)	1.49(1)
NgB^{HNC}	0.7(2)	−1(1)	0.622(6)	1.88(1)	1.38(1)
$NgBT^{KH}$	0.7(1)	0.2(9)	0.727(4)	2.98(1)	2.64(1)
$NgBT^{PSE3}$	0.8(1)	−0.4(9)	0.687(4)	1.87(1)	1.43(1)
$NgBT^{HNC}$	0.7(2)	−1(1)	0.621(6)	1.84(1)	1.34(1)
ISc^{KH}	0.54(9)	−1.6(7)	0.646(5)	2.399(9)	2.06(1)
ISc^{PSE3}	0.6(1)	−1.7(9)	0.616(5)	2.004(7)	1.758(8)
ISc^{HNC}	0.45(9)	−2.5(7)	0.615(5)	1.926(9)	1.662(10)
ISc^{*KH}	0.80(9)	−16(9)	0.857(2)	8.37(6)	6.06(5)
ISc^{*PSE3}	0.77(8)	−21(7)	0.865(2)	8.12(5)	6.47(4)
ISc^{*HNC}	0.81(6)	−18(6)	0.876(3)	7.38(5)	5.85(4)

Table 4.10: Bootstrap statistical comparison between predicted $T\Delta S_P$ (all UC and NgB corrections) or $T\Delta S_V$ (uncorrected 3D-RISM and parameter free corrections) and $T\Delta S_P$ from experiment for ions (Fawcett and Marcus datasets).

4.1.6 k-fold Cross-Validation Statistics

Correction	a	a_0	a_1	b	b_0	b_1
UC^{KH}	$-0.1499(8)$	$0.009(7)$	$-0.00053(2)$	$-0.1(1)$	$-3.2(10)$	$0.011(3)$
UC^{PSE3}	$-0.1185(8)$	$0.033(7)$	$-0.00051(2)$	$-0.3(1)$	$-3.3(10)$	$0.010(3)$
UC^{HNC}	$-0.1186(8)$	$0.033(6)$	$-0.00051(2)$	$-0.2(1)$	$-3.3(9)$	$0.010(3)$
$UCGF^{KH}$	$-0.1044(9)$	$0.018(7)$	$-0.00041(2)$	$0.6(1)$	$-3.1(10)$	$0.013(3)$
$UCGF^{PSE3}$	$-0.042(1)$	$0.065(9)$	$-0.00036(3)$	$1.1(2)$	$-3(1)$	$0.013(4)$
$UCGF^{HNC}$	$-0.038(1)$	$0.069(10)$	$-0.00036(3)$	$1.1(2)$	$-3(1)$	$0.013(5)$

Correction	γ	γ_0	γ_1
NgB^{KH}	$0.333(1)$	$0.38(1)$	$-0.00015(4)$
NgB^{PSE3}	$0.366(1)$	$0.31(1)$	$0.00019(4)$
NgB^{HNC}	$0.364(1)$	$0.31(1)$	$0.00020(4)$

Table 4.11: Fit parameters for UC and NgB corrections using k-fold averaging. Standard error in the last digit is given in parentheses.

	ΔG				
	Slope	y-intercept	R^2	RMSE	MUE
KH	1.2(3)	25(1)	0.2234(10)	24.936(9)	24.016(9)
PSE-3	1.2(3)	20(1)	0.309(1)	20.258(7)	19.531(7)
HNC	1.2(3)	20(1)	0.305(1)	20.119(7)	19.398(8)
UC^{KH}	0.98(6)	−0.2(3)	0.8422(4)	1.261(2)	0.919(1)
UC^{PSE3}	0.97(6)	−0.2(2)	0.8420(4)	1.250(2)	0.918(1)
UC^{HNC}	0.93(6)	−0.3(2)	0.8620(4)	1.089(1)	0.8363(10)
$UCGF^{PSE3}$	1.31(8)	0.6(3)	0.8478(4)	1.882(4)	1.333(2)
$UCGF^{KH}$	1.04(6)	−0.0(3)	0.8444(4)	1.339(2)	0.965(1)
$UCGF^{HNC}$	1.30(8)	0.7(3)	0.8634(5)	1.721(3)	1.296(2)
NgB^{KH}	1.10(7)	0.1(3)	0.8522(4)	1.389(2)	1.015(1)
NgB^{PSE3}	1.12(7)	0.2(3)	0.8529(4)	1.425(2)	1.035(1)
NgB^{HNC}	1.08(6)	0.1(2)	0.8764(4)	1.184(2)	0.907(1)
ISc^{KH}	0.96(7)	−2.2(3)	0.8064(5)	2.482(2)	2.087(2)
ISc^{PSE3}	0.96(6)	−0.8(3)	0.8343(4)	1.416(2)	1.052(1)
ISc^{HNC}	0.92(6)	−0.7(2)	0.8578(4)	1.181(1)	0.919(1)
ISc^{*KH}	0.99(6)	1.1(3)	0.8386(4)	1.700(1)	1.453(1)
ISc^{*PSE3}	0.99(7)	2.6(3)	0.8167(5)	2.948(2)	2.690(2)
ISc^{*HNC}	0.96(7)	2.7(3)	0.8225(5)	3.057(2)	2.815(2)
MD	0.99(5)	0.6(2)	0.8865(3)	1.245(1)	1.0254(10)

Table 4.12: k-fold statistical comparison between predicted and empirical hydration free energies for neutral molecules (Mobley, Abagyan, Rizzo and Palmer datasets). As described in Methods, values are the mean of all resampled data. RMSE: root-mean-squared-error. MUE: mean unsigned error. Standard error in the last digit is given in parentheses.

	Slope	y-intercept	R^2	RMSE	MUE
KH	1.2(3)	22.4(215)	0.877(2)	12.38(3)	11.08(3)
PSE-3	1.2(3)	20.8(214)	0.884(2)	10.30(3)	9.27(3)
HNC	1.2(4)	17.0(285)	0.903(2)	9.72(3)	8.86(3)
UC^{KH}	1.0(2)	−3.9(164)	0.888(2)	6.12(2)	4.91(2)
UC^{PSE3}	1.0(2)	−2.8(170)	0.885(2)	6.09(3)	4.88(2)
UC^{HNC}	1.0(3)	−2.4(214)	0.917(2)	5.47(3)	4.49(2)
$UCGF^{PSE3}$	1.2(3)	5.8(210)	0.883(2)	9.03(4)	6.94(3)
$UCGF^{KH}$	1.0(2)	−3.3(168)	0.889(2)	6.19(3)	4.95(2)
$UCGF^{HNC}$	1.2(4)	6.0(289)	0.899(2)	9.45(5)	7.56(4)
NgB^{KH}	1.1(3)	3.4(200)	0.874(2)	7.34(3)	5.99(2)
NgB^{PSE3}	1.1(3)	5.3(206)	0.874(2)	7.62(3)	6.16(3)
NgB^{HNC}	1.1(3)	5.5(265)	0.907(2)	7.02(3)	5.82(3)
ISc^{KH}	0.9(2)	−5.9(165)	0.887(2)	6.25(3)	5.05(2)
ISc^{PSE3}	1.0(2)	−4.1(170)	0.884(2)	6.12(2)	4.90(2)
ISc^{HNC}	1.0(3)	−2.0(209)	0.915(2)	5.48(3)	4.49(2)
ISc^{*KH}	1.0(2)	−2.5(167)	0.889(2)	6.06(2)	4.89(2)
ISc^{*PSE3}	1.0(2)	−0.2(175)	0.886(2)	6.13(2)	5.03(2)
ISc^{*HNC}	1.0(3)	1.5(217)	0.915(2)	5.54(2)	4.73(2)

Table 4.13: k-foldstatistical comparison between predicted and empirical hydration free energies for ions (Rizzo dataset). Only the six Joung-Cheatham monovalent ions[39] are included for MD.

	ΔG				
	Slope	y-intercept	R^2	RMSE	MUE
KH	0.9(3)	23(1)	0.1576(9)	24.365(9)	23.343(9)
PSE-3	1.0(3)	18.8(10)	0.2383(10)	19.678(8)	18.858(8)
HNC	1.0(3)	18.7(10)	0.231(1)	19.451(8)	18.631(8)
UC^{KH}	0.98(4)	−0.9(1)	0.9338(2)	1.142(2)	0.886(1)
UC^{PSE3}	0.97(4)	−0.9(1)	0.9330(3)	1.138(2)	0.8860(10)
UC^{HNC}	0.95(4)	−0.9(1)	0.9317(3)	1.134(2)	0.9085(10)
$UCGF^{PSE3}$	1.31(5)	−0.2(2)	0.9348(3)	1.687(3)	1.170(2)
$UCGF^{KH}$	1.05(4)	−0.7(1)	0.9354(2)	1.200(2)	0.912(1)
$UCGF^{HNC}$	1.32(6)	−0.3(2)	0.9133(5)	1.744(4)	1.221(2)
NgB^{KH}	1.10(4)	−0.6(1)	0.9426(2)	1.219(2)	0.911(1)
NgB^{PSE3}	1.12(4)	−0.5(1)	0.9469(2)	1.215(2)	0.884(1)
NgB^{HNC}	1.11(4)	−0.6(1)	0.9459(2)	1.173(2)	0.883(1)
ISc^{KH}	0.98(4)	−2.8(1)	0.9351(3)	2.843(1)	2.726(1)
ISc^{PSE3}	0.97(3)	−1.4(1)	0.9407(3)	1.510(2)	1.318(1)
ISc^{HNC}	0.95(4)	−1.3(1)	0.9402(3)	1.394(1)	1.2142(10)
ISc^{*KH}	0.98(5)	0.4(2)	0.9073(3)	1.071(1)	0.8549(9)
ISc^{*PSE3}	0.97(6)	1.9(2)	0.8586(3)	2.283(1)	2.024(1)
ISc^{*HNC}	0.96(6)	1.9(2)	0.8380(4)	2.357(1)	2.094(2)

Table 4.14: k-fold statistical comparison between predicted and molecular dynamics hydration free energies for neutral molecules (Mobley dataset). As described in Methods, R^2 bootstrap is the mean of all resampled data and R^2 k-fold is the mean over all training sub-samples. RMSE: root-mean-squared-error. MUE: mean unsigned error.

	ΔG_{Pol}				
	Slope	y-intercept	R^2	RMSE	MUE
KH	1.13(4)	−0.2(2)	0.9524(2)	1.050(2)	0.749(1)
PSE-3	1.16(4)	−0.2(2)	0.9541(2)	1.145(2)	0.832(1)
HNC	1.14(4)	−0.2(2)	0.9549(3)	1.037(2)	0.774(1)
$\text{UC}_{\text{Pol}}^{\text{KH}}$	1.06(4)	−0.2(2)	0.9483(3)	0.813(2)	0.5351(9)
$\text{UC}_{\text{Pol}}^{\text{PSE3}}$	1.07(4)	−0.2(2)	0.9480(3)	0.821(2)	0.5387(9)
$\text{UC}_{\text{Pol}}^{\text{HNC}}$	1.04(4)	−0.3(2)	0.9489(3)	0.726(2)	0.4895(8)
$\text{UCGF}_{\text{Pol}}^{\text{KH}}$	1.10(4)	−0.2(2)	0.9499(2)	0.943(2)	0.6419(10)
$\text{UCGF}_{\text{Pol}}^{\text{PSE3}}$	1.35(4)	−0.0(2)	0.9512(3)	1.924(3)	1.477(2)
$\text{UCGF}_{\text{Pol}}^{\text{HNC}}$	1.38(5)	−0.1(3)	0.9286(6)	2.066(3)	1.590(2)
$\text{NgB}_{\text{Pol}}^{\text{KH}}$	1.13(4)	−0.2(2)	0.9523(2)	1.051(2)	0.749(1)
$\text{NgB}_{\text{Pol}}^{\text{PSE3}}$	1.16(4)	−0.2(2)	0.9542(2)	1.144(2)	0.832(1)
$\text{NgB}_{\text{Pol}}^{\text{HNC}}$	1.14(4)	−0.2(2)	0.9547(3)	1.038(2)	0.774(1)
$\text{ISc}_{\text{Pol}}^{\text{KH}}$	1.06(4)	−0.2(2)	0.9477(3)	0.798(2)	0.5231(8)
$\text{ISc}_{\text{Pol}}^{\text{PSE3}}$	1.06(4)	−0.2(2)	0.9475(3)	0.812(2)	0.5313(9)
$\text{ISc}_{\text{Pol}}^{\text{HNC}}$	1.04(4)	−0.3(2)	0.9487(3)	0.719(2)	0.4841(8)
$\text{ISc*}_{\text{Pol}}^{\text{KH}}$	1.07(4)	−0.2(2)	0.9484(3)	0.823(2)	0.5425(9)
$\text{ISc*}_{\text{Pol}}^{\text{PSE3}}$	1.08(4)	−0.2(2)	0.9490(3)	0.855(2)	0.5648(9)
$\text{ISc*}_{\text{Pol}}^{\text{HNC}}$	1.05(4)	−0.3(2)	0.9502(3)	0.759(2)	0.5136(8)

Table 4.15: k-fold statistical comparison between predicted and molecular dynamics polar hydration free energies for neutral molecules (Mobley dataset).

	ΔG_{NP}				
	Slope	y-intercept	R^2	RMSE	MUE
KH	3(1)	20(3)	0.0975(7)	25.057(9)	24.079(9)
PSE-3	3(1)	17(2)	0.1099(8)	20.455(7)	19.684(8)
HNC	3(1)	16(2)	0.1154(8)	20.161(8)	19.400(8)
UC_{NP}^{KH}	0.69(8)	0.3(2)	0.587(1)	0.5699(6)	0.4518(5)
UC_{NP}^{PSE3}	0.64(9)	0.4(2)	0.543(1)	0.5832(6)	0.4618(5)
UC_{NP}^{HNC}	0.65(9)	0.3(2)	0.5656(10)	0.6220(6)	0.5058(5)
$UCGF_{NP}^{KH}$	0.82(9)	0.1(2)	0.643(1)	0.4988(7)	0.3835(4)
$UCGF_{NP}^{PSE3}$	1.0(1)	0.6(2)	0.648(1)	0.7628(6)	0.6461(6)
$UCGF_{NP}^{HNC}$	1.0(1)	0.8(2)	0.619(1)	0.8624(6)	0.7532(6)
NgB_{NP}^{KH}	1.01(8)	−0.1(2)	0.7629(8)	0.4011(7)	0.2869(4)
NgB_{NP}^{PSE3}	0.99(7)	0.0(1)	0.7925(7)	0.3532(6)	0.2459(4)
NgB_{NP}^{HNC}	1.01(8)	−0.1(2)	0.7927(8)	0.3528(7)	0.2412(4)
ISc_{NP}^{KH}	0.50(6)	−1.3(1)	0.6141(10)	2.3227(6)	2.2797(6)
ISc_{NP}^{PSE3}	0.56(5)	−0.0(1)	0.7284(9)	0.9308(5)	0.8529(5)
ISc_{NP}^{HNC}	0.59(6)	0.0(1)	0.7056(9)	0.8783(5)	0.7984(5)
ISc^{*KH}_{NP}	0.8(1)	1.3(3)	0.398(1)	1.1585(9)	0.9958(8)
ISc^{*PSE3}_{NP}	0.9(2)	2.7(4)	0.270(1)	2.652(1)	2.448(1)
ISc^{*HNC}_{NP}	0.9(2)	2.6(5)	0.272(1)	2.681(1)	2.474(1)

Table 4.16: k-fold statistical comparison between predicted and molecular dynamics non-polar hydration free energies for neutral molecules (Mobley dataset).

$\Delta H / \Delta \epsilon$

	Slope	y-intercept	R^2	RMSE	MUE
KH	1.0(3)	2(4)	0.770(2)	2.649(6)	2.339(6)
PSE-3	1.2(4)	1(5)	0.773(2)	2.57(1)	2.034(8)
HNC	1.2(4)	1(5)	0.776(2)	2.57(1)	2.035(8)
UC^{KH}	1.7(9)	−10.5(1)	0.609(3)	19.87(3)	18.82(3)
UC^{PSE3}	1.6(9)	−11.1(1)	0.602(3)	19.33(3)	18.36(3)
UC^{HNC}	1.6(9)	−11.1(1)	0.604(3)	19.28(3)	18.30(3)
UCT^{KH}	0.9(3)	−2(4)	0.764(2)	1.825(5)	1.545(5)
UCT^{HNC}	0.9(3)	−2(3)	0.777(2)	1.751(5)	1.474(5)
UCT^{PSE3}	0.9(3)	−1(4)	0.773(2)	1.787(5)	1.510(5)
$UCGF^{KH}$	1.6(7)	−7(10)	0.649(3)	14.16(3)	13.23(2)
$UCGF^{PSE3}$	1.8(6)	−2(8)	0.750(2)	11.93(2)	10.97(2)
$UCGF^{HNC}$	1.9(6)	−1(8)	0.763(2)	12.21(2)	11.18(2)
$UCTGF^{KH}$	0.9(3)	−1(4)	0.768(2)	1.826(5)	1.547(5)
$UCTGF^{PSE3}$	1.2(3)	3(5)	0.804(2)	2.334(8)	1.937(7)
$UCTGF^{HNC}$	1.3(4)	4(5)	0.797(2)	2.662(9)	2.203(8)
NgB^{KH}	1.0(3)	2(4)	0.769(2)	2.640(6)	2.332(6)
NgB^{PSE3}	1.2(4)	1(5)	0.779(2)	2.58(1)	2.043(8)
NgB^{HNC}	1.2(4)	1(5)	0.774(2)	2.58(1)	2.044(8)
$NgBT^{KH}$	1.0(3)	2(4)	0.769(2)	2.739(6)	2.426(6)
$NgBT^{PSE3}$	1.2(4)	1(5)	0.778(2)	2.67(1)	2.116(9)
$NgBT^{HNC}$	1.2(4)	0(5)	0.773(2)	2.68(1)	2.118(9)
ISc^{KH}	1.0(3)	−0(4)	0.762(2)	2.019(7)	1.635(6)
ISc^{PSE3}	1.0(3)	0(4)	0.768(2)	2.037(6)	1.667(6)
ISc^{HNC}	1.0(3)	−0(4)	0.772(2)	2.028(6)	1.657(6)
ISc^{*KH}	0.6(3)	−1(3)	0.547(3)	3.034(7)	2.744(6)
ISc^{*PSE3}	1.5(8)	−2(7)	0.541(3)	6.47(1)	5.81(1)
ISc^{*HNC}	1.5(9)	−2(7)	0.562(3)	6.67(1)	6.00(1)

Table 4.17: k-fold statistical comparison between predicted ΔH (all UC and NgB corrections) or $\Delta\epsilon$ (uncorrected 3D-RISM and parameter free corrections) and ΔH from experiment for neutral molecules (Abraham and Cabani datasets).

$T\Delta S$					
	Slope	y-intercept	R^2	RMSE	MUE
KH	3(2)	−2.8(152)	0.519(3)	19.27(3)	18.07(3)
PSE-3	3(2)	−2.5(146)	0.527(3)	18.26(3)	17.11(3)
HNC	3(2)	−1.7(144)	0.542(3)	18.19(3)	17.03(3)
UC^{KH}	3(2)	−3.6(148)	0.528(3)	19.92(3)	18.85(3)
UC^{PSE3}	3(2)	−4.5(143)	0.525(3)	19.38(3)	18.38(3)
UC^{HNC}	3(2)	−3.7(142)	0.544(3)	19.27(3)	18.26(3)
UCT^{KH}	0.6(3)	−4(3)	0.546(3)	1.454(5)	1.200(4)
UCT^{PSE3}	0.6(3)	−4(3)	0.559(3)	1.420(5)	1.160(4)
UCT^{HNC}	0.6(3)	−3(3)	0.591(3)	1.355(5)	1.113(4)
$UCGF^{KH}$	2(1)	−2.7(117)	0.529(3)	14.10(2)	13.18(2)
$UCGF^{PSE3}$	2(1)	−2.0(101)	0.542(3)	11.55(2)	10.72(2)
$UCGF^{HNC}$	2(1)	−0.9(102)	0.574(3)	11.69(2)	10.85(2)
$UCTGF^{KH}$	0.5(3)	−4(2)	0.546(3)	1.445(5)	1.195(4)
$UCTGF^{PSE3}$	0.6(3)	−3(3)	0.550(3)	1.511(6)	1.207(5)
$UCTGF^{HNC}$	0.6(4)	−3(3)	0.565(3)	1.542(6)	1.224(5)
NgB^{KH}	0.7(4)	−1(3)	0.555(3)	2.338(6)	2.062(5)
NgB^{PSE3}	1.0(6)	−1(5)	0.553(3)	2.228(8)	1.813(7)
NgB^{HNC}	1.1(6)	−1(5)	0.577(3)	2.196(8)	1.789(7)
$NgBT^{KH}$	0.7(4)	−1(3)	0.556(3)	2.461(6)	2.177(6)
$NgBT^{PSE3}$	1.0(6)	−1(5)	0.552(3)	2.338(8)	1.901(7)
$NgBT^{HNC}$	1.1(6)	−1(5)	0.576(3)	2.313(9)	1.889(7)
ISc^{KH}	0.7(4)	−1(3)	0.545(3)	2.364(7)	2.071(6)
ISc^{PSE3}	0.8(4)	−1(4)	0.558(3)	1.849(6)	1.552(5)
ISc^{HNC}	0.8(4)	−1(4)	0.586(3)	1.817(6)	1.516(5)
ISc^{*KH}	0.8(3)	2(3)	0.759(2)	4.174(8)	3.813(8)
ISc^{*PSE3}	1.1(4)	−2(5)	0.737(3)	4.16(1)	3.50(1)
ISc^{*HNC}	1.1(4)	−2(5)	0.736(3)	4.29(1)	3.62(1)

Table 4.18: k-fold statistical comparison between predicted $T\Delta S_P$ (all UC and NgB corrections) or $T\Delta S_V$ (uncorrected 3D-RISM and parameter free corrections) and $T\Delta S_P$ from experiment for neutral molecules (Abraham and Cabani datasets).

References

[1] Human serum albumin. *Wikipedia, the free encyclopedia*, February 2016. Page Version ID: 704259799.

[2] Johan. Aaqvist. Ion-water interaction potentials derived from free energy perturbation simulations. *The Journal of Physical Chemistry*, 94(21):8021–8024, October 1990.

[3] G. M Abernethy and M. J Gillan. A new method of solving the hnc equation for ionic liquids. *Molecular Physics*, 39:839, Jan 1980.

[4] Michael H. Abraham, Gary S. Whiting, Richard Fuchs, and Eric J. Chambers. Thermodynamics of solute transfer from water to hexadecane. *Journal of the Chemical Society, Perkin Transactions 2*, (2):291, 1990.

[5] Nathan A. Baker, Donald Bashford, and David A. Case. Implicit Solvent Electrostatics in Biomolecular Simulation. In Benedict Leimkuhler, Christophe Chipot, Ron Elber, Aatto Laaksonen, Alan Mark, Tamar Schlick, Christoph Schütte, and Robert Skeel, editors, *New Algorithms for Macromolecular Simulation*, number 49 in Lecture Notes in Computational Science and Engineering, pages 263–295. Springer Berlin Heidelberg, 2006. DOI: 10.1007/3-540-31618-3_15.

[6] Dmitri Beglov and Benoît Roux. An integral equation to describe the solvation of polar molecules in liquid water. *Journal of Physical Chemistry B-Condensed Phase*, 101(39):7821–7826, 1997.

[7] Dmitrii Beglov and Benoît Roux. An Integral Equation To Describe the Solvation of Polar Molecules in Liquid Water. *The Journal of Physical Chemistry B*, 101(39):7821–7826, September 1997.

[8] L. Blum. Invariant Expansion. II. The Ornstein-Zernike Equation for Nonspherical Molecules and an Extended Solution to the Mean Spherical Model. *The Journal of Chemical Physics*, 57(5):1862–1869, September 1972.

[9] Juan M. Bollo-Rivas. Tutorial on the smooth particle-mesh Ewald algorithm, January 2015.

[10] A. J. Bordner, C. N. Cavasotto, and R. A. Abagyan. Accurate Transferable Model for Water, n-Octanol, and n-Hexadecane Solvation Free Energies. *J. Phys. Chem. B*, 106(42):11009–11015, 2002.

[11] Sergio Cabani, Paolo Gianni, Vincenzo Mollica, and Luciano Lepori. Group contributions to the thermodynamic properties of non-ionic organic solutes in dilute aqueous solution. *Journal of Solution Chemistry*, 10(8):563–595, August 1981.

[12] David Chandler and Hans C. Andersen. Optimized Cluster Expansions for Classical Fluids. II. Theory of Molecular Liquids. *The Journal of Chemical Physics*, 57(5):1930–1937, September 1972.

[13] David Chandler, Y Singh, and DM Richardson. Excess electrons in simple fluids. i. general equilibrium theory for classical hard sphere solvents. *J. Chem. Phys.*, 81:1975–1982, 1984.

[14] SH Chong and Fumio Hirata. Ion hydration: Thermodynamic and structural analysis with an integral equation theory of liquids. *Journal of Physical Chemistry B*, 101(16):3209–3220, Jan 1997.

[15] Song-Ho Chong and Sihyun Ham. Thermodynamic-Ensemble Independence of Solvation Free Energy. *Journal of Chemical Theory and Computation*, 11(2):378–380, February 2015.

[16] Gennady N Chuev, Maxim V Fedorov, Sandro Chiodo, Nino Russo, and Emilia Sicilia. Hydration of ionic species studied by the reference interaction site model with a repulsive bridge correction. *J. Comput. Chem.*, 29(14):2406–2415, Jan 2008.

[17] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald–an Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98:10089–10092, 1993.

[18] Jeremy L. England and Gilad Haran. Role of solvation effects in protein denaturation: from thermodynamics to single molecules and back. *Annual Review of Physical Chemistry*, 62:257–277, 2011.

[19] U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee, and L.G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103:8577–8593, 1995.

[20] BC Eu and K Rah. A closure for the ornstein-zernike relation that gives rise to the thermodynamic consistency. *Journal of Chemical Physics*, 111(8):3327–3338, Jan 1999.

[21] R. Evans. The nature of the liquid-vapour interface and other topics in the statistical mechanics of non-uniform, classical fluids. *Advances in Physics*, 28(2):143–200, April 1979.

[22] W. Ronald Fawcett. Thermodynamic Parameters for the Solvation of Monatomic Ions in Water. *The Journal of Physical Chemistry B*, 103(50):11181–11185, December 1999.

[23] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation, Second Edition: From Algorithms to Applications*. Academic Press, San Diego, 2 edition edition, November 2001.

[24] Samuel Genheden, Tyler Luchko, Sergey Gusarov, Andriy Kovalenko, and Ulf Ryde. An mm/3d-rism approach for ligand binding affinities. *Journal of Physical Chemistry B*, 114(25):8505–8516, Jan 2010.

[25] A.W. Götz, M.J. Williamson, D. Xu, D. Poole, S. Le Grand, and R.C. Walker. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.*, 8:1542–1555, 2012.

[26] B. Guillot, Y. Guissani, and S. Bratos. A computer-simulation study of hydrophobic hydration of rare gases and of methane. I. Thermodynamic and structural properties. *The Journal of Chemical Physics*, 95(5):3643–3648, September 1991.

[27] Sergey Gusarov, Bhalchandra S Pujari, and Andriy Kovalenko. Efficient treatment of solvation shells in 3d molecular theory of solvation. *J Comput Chem*, 33(17):1478–1494, Jun 2012.

[28] J.-P. Hansen and I. R. McDonald. *Theory of Simple Liquids*. Academic Press, London, Great Britain, second edition, 1990.

[29] Jean-Pierre Hansen and I. R. McDonald. *Theory of Simple Liquids*. Academic Press, February 2006.

[30] F. Hirata. Theory of molecular liquids. In F. Hirata, editor, *Molecular Theory of Solvation*, pages 1–60. Kluwer Acadamic Publishers, Dordrecht, 2003.

[31] Fumio Hirata, Peter J. Rossky, and B. Montgomery Pettitt. The interionic potential of mean force in a molecular polar solvent from an extended RISM equation. *The Journal of Chemical Physics*, 78(6):4133–4144, March 1983.

[32] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–1149, May 1995.

[33] Dominik Horinek, Shavkat I. Mamatkulov, and Roland R. Netz. Rational design of ion force fields based on thermodynamic solvation properties. *The Journal of Chemical Physics*, 130(12):124507, March 2009.

[34] WenJuan Huang, Nikolay Blinov, and Andriy Kovalenko. Octanol-Water Partition Coefficient from 3d-RISM-KH Molecular Theory of Solvation with Partial Molar Volume Correction. *The Journal of Physical Chemistry B*, April 2015.

[35] Philippe Hünenberger and Maria Reif. *Single-Ion Solvation: Experimental and Theoretical Approaches to Elusive Thermodynamic Quantities*. Royal Society of Chemistry, 2011.

[36] T Ichiye and David Chandler. Hypernetted chain closure reference interaction site method theory of structure and thermodynamics for alkanes in water. *J. Phys. Chem.*, 92(18):5257–5261, 1988.

[37] Mitsunori Ikeguchi and Junta Doi. Direct numerical solution of the Ornstein–Zernike integral equation and spatial distribution of water around hydrophobic molecules. *The Journal of Chemical Physics*, 103(12):5011–5017, September 1995.

[38] Ryosuke Ishizuka and Norio Yoshida. Extended molecular Ornstein-Zernike integral equation for fully anisotropic solute molecules: Formulation in a rectangular coordinate system. *The Journal of Chemical Physics*, 139(8):084119, August 2013.

[39] In Suk Joung, Tyler Luchko, and David A. Case. Simple electrolyte solutions: Comparison of DRISM and molecular dynamics results for alkali halide solutions. *The Journal of Chemical Physics*, 138(4):044103, January 2013.

[40] In Suk Joung, Tyler Luchko, and David A Case. Simple electrolyte solutions: comparison of drism and molecular dynamics results for alkali halide solutions. *J Chem Phys*, 138(4):044103, Jan 2013.

[41] Stefan M Kast and Thomas Kloss. Closed-form expressions of the chemical potential for integral equation closures with certain bridge functions. *J Chem Phys*, 129(23):236101, Jan 2008.

[42] Joseph W. Kaus, Levi T. Pierce, Ross C. Walker, and J. Andrew McCammon. Improving the Efficiency of Free Energy Calculations in the Amber Molecular Dynamics Package. *Journal of Chemical Theory and Computation*, 9(9):4131–4139, September 2013.

[43] M Kinoshita and Fumio Hirata. Application of the reference interaction site model theory to analysis on surface-induced structure of water. *Journal of Chemical Physics*, 104(21):8807–8815, Jan 1996.

[44] Georgios Kontogeorgis, Georgios Folas, Nuria Muro Sunè, Ferran Roca Leon, and Michael Locht Michelsen. Solvation phenomena in association theories with applications to oil & gas and chemical industries. *Oil & Gas Science & Technology*, 63(3):305–319, 2008.

[45] A. Kovalenko. Three-dimensional rism theory for molecular liquids and solid-liquid interfaces. In F. Hirata, editor, *Molecular theory of solvation*, chapter 4, pages 175–262. Kluwer Academic Publishers, 2003.

[46] Andriy Kovalenko and Fumio Hirata. Self-consistent description of a metal-water interface by the kohn-sham density functional theory and the three-dimensional reference interaction site model. *J Chem Phys*, 110(20):10095–10112, Jan 1999.

[47] Andriy Kovalenko and Fumio Hirata. Self-consistent description of a metal–water interface by the Kohn–Sham density functional theory and the three-dimensional reference interaction site model. *The Journal of Chemical Physics*, 110(20):10095–10112, May 1999.

[48] Andriy Kovalenko and Fumio Hirata. Potentials of mean force of simple ions in ambient aqueous solution. i. three-dimensional reference interaction site model approach. *J. Chem. Phys.*, 112(23):10391–10402, 2000.

[49] Andriy Kovalenko, Seiichiro Ten-no, and Fumio Hirata. Solution of three-dimensional reference interaction site model and hypernetted chain equations for simple point charge water by modified method of direct inversion in iterative subspace. *Journal of Computational Chemistry*, 20(9):928–936, July 1999.

[50] S. Le Grand, A.W. Goetz, and R.C. Walker. SPFP: Speed without compromise–A mixed precision model for GPU accelerated molecular dynamics simulations. *Comput. Phys. Commun.*, 184:374–380, 2013.

[51] Pil H Lee and Gerald M Maggiora. Solvation thermodynamics of polar molecules in aqueous solution by the xrism method. *J. Phys. Chem.*, 97(39):10175–10185, Sep 1993.

[52] Bo Li, Alexei V. Matveev, and Notker Rösch. Three-dimensional reference interaction site model solvent combined with a quantum mechanical treatment of the solute. *Computational and Theoretical Chemistry*, 1070:143–151, October 2015.

[53] Jiabo Li, Tianhai Zhu, Gregory D. Hawkins, Paul Winget, Daniel A. Liotard, Christopher J. Cramer, and Donald G. Truhlar. Extension of the platform of applicability of the SM5.42r universal solvation model. *Theoretical Chemistry Accounts*, 103(1):9–63, November 1999.

[54] Tyler Luchko, Sergey Gusarov, Daniel R Roe, Carlos Simmerling, David A Case, Jack Tuszynski, and Andriy Kovalenko. Three-dimensional molecular theory of solvation coupled with molecular dynamics in amber. *J Chem Theory Comput*, 6(3):607–624, Jan 2010.

[55] Y. Marcus and A. Loewenschuss. Chapter 4. Standard entropies of hydration of ions. *Annual Reports Section "C" (Physical Chemistry)*, 81(0):81–135, 1984.

[56] Yizhak Marcus. The thermodynamics of solvation of ions. Part 2.—The enthalpy of hydration at 298.15 K. *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases*, 83(2):339–349, 1987.

[57] D. A. McQuarrie. *Statistical Mechanics*. University Science Books, Sausalito, USA, 2000.

[58] W. Smith Md Cells. The Minimum Image Convention in Non-Cubic MD Cells. 1994.

[59] Paulius Mikulskis, Samuel Genheden, and Ulf Ryde. A Large-Scale Test of Free-Energy Simulation Estimates of Protein–Ligand Binding Affinities. *Journal of Chemical Information and Modeling*, 54(10):2794–2806, October 2014.

[60] Maksim Misin, Maxim V. Fedorov, and David S. Palmer. Communication: Accurate hydration free energies at a wide range of temperatures from 3d-RISM. *The Journal of Chemical Physics*, 142(9):091105, March 2015.

[61] David L. Mobley, Ken A. Dill, and John D. Chodera. Treating Entropy and Conformational Changes in Implicit Solvent Simulations of Small Molecules. *The Journal of Physical Chemistry B*, 112(3):938–946, January 2008.

[62] T Morita. Theory of classical fluids: Hyper-netted chain approximation, i — formulation for a one-component system—. *Progress of Theoretical Physics*, 20:920, Dec 1958.

[63] K Ng. Hypernetted chain solutions for the classical one-component plasma up to $\gamma = 7000$. *J. Chem. Phys.*, Jan 1974.

[64] David S. Palmer, Andrey I. Frolov, Ekaterina L. Ratkova, and Maxim V. Fedorov. Towards a universal method for calculating hydration free energies: a 3d reference interaction site model with partial molar volume correction. *Journal of Physics: Condensed Matter*, 22(49):492101, December 2010.

[65] David S Palmer, Andrey I Frolov, Ekaterina L Ratkova, and Maxim V Fedorov. Towards a universal method for calculating hydration free energies: a 3d reference interaction site model with partial molar volume correction. *J Phys Condens Matter*, 22(49):492101, Nov 2010.

[66] Luca Peliti. *Statistical Mechanics in a Nutshell*. Princeton University Press, Princeton N.J., y first english language edition edition edition, August 2011.

[67] John Perkyns and B. Montgomery Pettitt. A site–site theory for finite concentration saline solutions. *The Journal of chemical physics*, 97(10):7656–7666, 1992.

[68] Christine Peter, Chris Oostenbrink, Arthur van Dorp, and Wilfred F. van Gunsteren. Estimating entropies from molecular dynamics simulations. *The Journal of Chemical Physics*, 120(6):2652–2661, February 2004.

[69] Enrico Purisima and Traian Sulea. Solvation Models: Theory and Validation. *Current Pharmaceutical Design*, 20(20):3266–3280, May 2014.

[70] Ekaterina L. Ratkova, David S. Palmer, and Maxim V. Fedorov. Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy. *Chemical Reviews*, 115(13):6312–6356, July 2015.

[71] E.L. Ratkova, D.S. Palmer, and M.V. Fedorov. Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy. *Chem. Rev.*, 115:6312–6356, 2015.

[72] Robert C. Rizzo, Tiba Aynechi, David A. Case, and Irwin D. Kuntz. Estimation of Absolute Free Energies of Hydration Using Continuum Methods: Accuracy of Partial Charge Models and Optimization of Nonpolar Contributions. *Journal of Chemical Theory and Computation*, 2(1):128–139, January 2006.

[73] Gabriel J. Rocklin, David L. Mobley, Ken A. Dill, and Philippe H. Hünenberger. Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects. *The Journal of Chemical Physics*, 139(18):184103, November 2013.

[74] Benoit Roux, Hsiang Ai Yu, and Martin Karplus. Molecular basis for the Born model of ion solvation. *Journal of Physical Chemistry*, 94(11):4683–4688, 1990.

[75] J.-P. Ryckaert, G. Ciccotti, and H.J.C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23:327–341, 1977.

[76] R. Salomon-Ferrer, A.W. Götz, D. Poole, S. Le Grand, and R.C. Walker. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.*, 9:3878–3888, 2013.

[77] Roland Schmid, Arzu M. Miah, and Valentin N. Sapunov. A new table of the thermodynamic quantities of ionic hydration: values and some applications (enthalpy–entropy compensation and Born radii). *Physical Chemistry Chemical Physics*, 2(1):97–102, January 2000.

[78] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, pages 57–61, 2010.

[79] Volodymyr Sergiievskyi, Guillaume Jeanmairet, Maximilien Levesque, and Daniel Borgis. Solvation free-energy pressure corrections in the three dimensional reference interaction site model. *The Journal of Chemical Physics*, 143(18):184116, November 2015.

[80] Volodymyr P. Sergiievskyi, Guillaume Jeanmairet, Maximilien Levesque, and Daniel Borgis. Fast Computation of Solvation Free Energies with Molecular Density Functional Theory: Thermodynamic-Ensemble Partial Molar Volume Corrections. *The Journal of Physical Chemistry Letters*, pages 1935–1942, May 2014.

[81] SJ Singer and David Chandler. Free energy functions in the extended rism approximation (article) free energy functions in the extended rism approximation (article) free energy functions in the extended rism approximation (article) free energy functions in the extended rism approximation (article) free energy functions in the extended rism approximation (article) free energy functions in the extended rism approximation (article) free energy functions in the extended rism approximation. *Molecular Physics*, 55(3):621–625, Jan 1985.

[82] J. F Springer, M. A Pokrant, and F. A Stevens. Integral equation solutions for the classical electron gas. *J Chem Phys*, 58:4863, Jun 1973.

[83] W. Clark Still, Anna Tempczyk, Ronald C. Hawley, and Thomas Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society*, 112(16):6127–6129, August 1990.

[84] Jacopo Tomasi, Benedetta Mennucci, and Roberto Cammi. Quantum Mechanical Continuum Solvation Models. *Chemical Reviews*, 105(8):2999–3094, August 2005.

[85] Jean-François Truchon, B. Montgomery Pettitt, and Paul Labute. A Cavity Corrected 3d-RISM Functional for Accurate Solvation Free Energies. *Journal of Chemical Theory and Computation*, 10(3):934–941, March 2014.

[86] Vincent Vagenende and Bernhardt L. Trout. Quantitative Characterization of Local Protein Solvation To Predict Solvent Effects on Protein Structure. *Biophysical Journal*, 103(6):1354–1362, September 2012.

[87] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, July 2004.

[88] Lingle Wang, Yuqing Deng, Jennifer L. Knight, Yujie Wu, Byungchan Kim, Woody Sherman, John C. Shelley, Teng Lin, and Robert Abel. Modeling Local Structural Rearrangements Using FEP/REST: Application to Relative Binding Affinity Predictions of CDK2 Inhibitors. *Journal of Chemical Theory and Computation*, 9(2):1282–1293, February 2013.

[89] Takeshi Yamazaki, Nikolay Blinov, David Wishart, and Andriy Kovalenko. Hydration effects on the het-s prion and amyloid-beta fibrillous aggregates, studied with three-dimensional molecular theory of solvation. *Biophys J*, 95(10):4540–4548, Nov 2008.

[90] Takeshi Yamazaki, Andriy Kovalenko, Vladimir V Murashov, and Grenfell N Patey. Ion solvation in a water-urea mixture. *Journal of Physical Chemistry B*, 114(1):613–619, Jan 2010.

[91] HA Yu and M Karplus. A thermodynamic analysis of solvation. *J. Chem. Phys.*, 89:2366, 1988.

[92] HA Yu, Benoît Roux, and M Karplus. Solvation thermodynamics: An approach from analytic temperature derivatives. *J. Chem. Phys.*, 92:5020, 1990.

[93] Shuangliang Zhao, Rosa Ramirez, Rodolphe Vuilleumier, and Daniel Borgis. Molecular density functional theory of solvation: From polar solvents to water. *The Journal of Chemical Physics*, 134(19):194102, May 2011.