Proximal distributions from angular correlations: A measure of the onset of coarse-graining

Kippi M. Dyer and B. Montgomery Pettitt^{a)}

Department of Biochemistry and Molecular Biology, Sealy Center for Structural Biology, University of Texas Medical Branch, Galveston, Texas 77555-0304, USA

(Received 31 August 2013; accepted 7 November 2013; published online 4 December 2013)

In this work we examine and extend the theory of proximal radial distribution functions for molecules in solution. We point out two formal extensions, the first of which generalizes the proximal distribution function hierarchy approach to the complete, angularly dependent molecular pair distribution function. Second, we generalize from the traditional right-handed solute-solvent proximal distribution functions to the left-handed distributions. The resulting neighbor hierarchy convergence is shown to provide a measure of the coarse-graining of the internal solute sites with respect to the solvent. Simulation of the test case of a deca-alanine peptide shows that this coarse-graining measure converges at a length scale of approximately 5 amino acids for the system considered.

© 2013 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4832895>]

I. INTRODUCTION

Since their introduction,^{1–3} so-called proximal distribution functions have been utilized to reduce the complexity of the molecular pair distribution functions of macromolecular systems³ to a set of conditional pair functions. In principle, the proximal distributions admit an approximate connection to the more general pair correlations of complex fluids while using a simpler, more easily understood description of the liquid structure around solutes, especially convenient for macromolecules.²

The proximal functions are based on a conditional property of the standard pair distribution functions in the liquid state. They are defined within a decomposition of the averages which build up the radial pair distribution functions via a sum over a hierarchy of neighbors, such that the radial distribution function is a direct sum over the nearest neighbor, next-nearest neighbor, etc., conditional probability distributions. The decomposition is exact, and finite where the number of atomic (or otherwise) sites within a given pair of molecules is finite. Taken as a whole, the neighbor hierarchy is an alternate, and equivalent, mathematical description of the radial distribution functions of the fluid.²

The existence of a series decomposition of the fundamental structural description of a fluid system leads naturally to the question of whether or how the series can be truncated to construct a reduced description of a fluid system which still accurately measures the thermodynamics. To this point, it has been shown^{4–8} that the truncation of the neighbor hierarchy of the average solvent distribution around a solute at the nearest neighbor approximation, with suitable renormalization, is a reasonable approximation for the density distribution and thus the electrostatic densities and energies of small biologically interesting molecules. The resulting electrostatic free energies of solvation of the approximate density distributions

computed with linear response are within a chemically acceptable error and generally better than continuum solvent approaches.⁷

In this paper we consider the generalization of proximal distributions to full angular dependence. This leads us to address two questions. First, at the level of the radial distribution functions, we consider the convergence of solute length scales as measured by the neighbor hierarchy. We demonstrate that series convergence is slow for molecules of the same size, where the potential functions act over distances for which more than one element of the neighbor hierarchy is meaningful to the immediate structure. For the van der Waals interactions of molecules of comparable size, many members of the hierarchy are essential to the complete description of the system energy. As we demonstrate, this is due to the fact that many of the terms in the hierarchy of the neighbor decomposition of the radial distribution functions are of the same relative magnitude for similar size molecules, and thus, the hierarchy is instead more usefully considered to be a finite basis set expansion of the radial distribution function in which all members of the hierarchy may contribute to the sum. We and others⁸ have successfully used truncation of one of the series for molecules which are very different in size. Here we use a slightly different set of functions than have previously been studied^{3,4,6–8} to measure coarse-graining for large solutes with respect to the smaller solvent molecule interactions.

Previously the proximal functions have been used only as a decomposition of the radial distribution of solvent around a solute.³ They are, however, not limited to the angle-averaged radial distribution functions, and can be generalized to apply to the complete pair distribution functions, including both distance and angular correlations between molecular pairs. Once this generalization is made, then, the question we address is whether and how the complete set of proximal functions can be utilized to describe the structure and thermodynamics of solvation of a complex molecule in a way which keeps the

^{a)}Electronic mail: mpettitt@utmb.edu

spirit of the simpler structure offered by the original proximal distribution function methods.

We apply the current method to consider the aqueous solvation of deca-alanine. We wish to consider the general construction of the proximal distributions. We conclude with the idea of using the proximal functions as physical or environmental criteria measures for coarse graining of all-atom potentials for certain purposes.

II. THEORY

Here, we review the proximal distribution function theory^{2,3} and present a generalization of the theory for the full angular-dependent molecular pair distribution functions.

Consider a fluid system of molecules whose interaction potential is defined by a site-site model, such that, for a given pair of molecules labeled 1, 2, the total potential $U(1, 2)$ between the molecules is

$$U(1, 2) = \sum_{i=1}^m \sum_{j=1}^n u_{ij}(r_{ij}), \quad (1)$$

where r_{ij} is the distance between site i on molecule 1 and site j on molecule 2, m is the number of sites on molecule 1, n is the number of sites on molecule 2, and u_{ij} is the potential between the sites i, j . Here and throughout, we assume that $u_{ij}(r_{ij})$ is a radially symmetric function of the scalar distance between two sites, and that the number of sites m, n are finite. For most of this work, these assumptions do not represent a loss of generality, though we will point out where they are a necessary condition of certain results. Finally, the site-site notation is usually simplified⁹ from $u_{ij}(r_{ij})$ to $u_{ij}(r)$ where it is understood that the scalar distance r implicitly refers to the distance between sites i and j . We will follow this convention, except where there is a possibility for confusion.

If we have N such molecules at a given thermodynamic phase point, then one of the standard descriptions of the structure of the fluid is given by the set of site-site pair distribution functions,¹⁰

$$\rho_i g_{ij}(\mathbf{r}) \rho_j = \frac{1}{V} \langle N(N-1) \delta(\mathbf{r}_j - \mathbf{r}_i - \mathbf{r}) \rangle, \quad (2)$$

where the angle brackets here represent the ensemble average, N is the number of molecules, V is the volume, the $\delta(\mathbf{r}_j - \mathbf{r}_i - \mathbf{r})$ functions are the operators restricting the average to the vector $\mathbf{r}_{ij}$ between site i on molecule 1 and site j on molecule 2, ρ_i is the total number density of site i , and we assume a homogeneous, non-reactive system where the site density $\rho_i = \rho_1$, ρ_1 is the total number density of molecular species 1, and similarly for ρ_j for species 2. With this standard definition, for cases where the site-site potential terms $u_{ij}(r)$ are radially symmetric functions of the distance between sites alone and there are no external fields, then the $\mathbf{r}$ -vector dependence of the $g_{ij}(\mathbf{r})$ functions reduces to the radially symmetric case, so that

$$4\pi r^2 g_{ij}(r) \rho_j dr \quad (3)$$

is the probability of finding site j of molecule 2 within the distance $r + dr$ given that site i of molecule 1 is at the origin.

For thermodynamic quantities whose operators are functions of only the distance between sites, such that

$$O = \langle \hat{O}(1, 2) \rangle = \sum_i \sum_j \langle \hat{o}_{ij}(r) \rangle \quad (4)$$

for O the thermodynamic average of the molecular operators $\hat{O}$ and site-site operators $\hat{o}$, such as the excess energy, then the radial distribution functions $g_{ij}(r)$ are a complete description of the thermodynamics of interaction of the molecular system, since in this case

$$O = \sum_{ij} \int \hat{o}(r_{ij}) g_{ij}(r) d\mathbf{r}. \quad (5)$$

However, for many important cases, the radial site-site functions must be generalized to take into account the orientational correlations between molecules.¹¹

The potential between pairs of arbitrary molecules with fixed intramolecular distributions is usually written as¹¹

$$U(1, 2) = U(\mathbf{r}, \Omega_1, \Omega_2), \quad (6)$$

where $\mathbf{r}$ is the vector between arbitrary points in molecule 1 and molecule 2, Ω_1 is the orientation of molecule 1 with respect to an arbitrary origin, and similarly for Ω_2 for molecule 2. If we choose the origin of $\mathbf{r}$ coincident with site i on molecule 1 and pointing to site j on molecule 2, labeled as $\mathbf{r}_{ij}$, then the radial site-site distribution functions are angular averages of the molecular pair distribution functions,

$$g_{ij}(r) = \langle g(r_{ij}, \Omega_1, \Omega_2) \rangle_{\Omega_1, \Omega_2}, \quad (7)$$

with $g(r_{ij}, \Omega_1, \Omega_2)$ the molecular pair distribution function fully dependent on the distance and orientations of the pair of molecules, and the $\langle \rangle_{\Omega_1, \Omega_2}$ notation here indicates a normalized average over the angular coordinates.¹¹ This notation usually assumes fixed internal molecular coordinates, but can be understood to apply in the case with internal degrees of freedom if the Ω -weighted ensemble average is understood to include as well the average over all of the respective intramolecular degrees of freedom of each molecule in the pair terms. The notation here is that of Gray and Gubbins,¹¹ and Hansen and McDonald,⁹ and is standard in liquid state theory. The site-site labels $g_{ij}(r)$ are discussed above. The ij labels inside of $g(\mathbf{r}_{ij}, \Omega_1, \Omega_2)$ mean simply that the origin of the coordinates is chosen to coincide with sites i and j , and is thus a simple coordinate change of the general $g(\mathbf{r}, \Omega_1, \Omega_2)$, without loss of generality. This is the standard convention when discussing the correspondence between site-site distribution functions and the general molecular pair distribution functions, and we follow it here.

To define the usual proximal distribution functions, we consider the coordinate geometry in a given pair function. From the definition of the site-site potential in Eq. (1), each point $(\mathbf{r}_{ij}, \Omega_1, \Omega_2)$ is equivalent to defining the set of distance vectors (or distance matrix) from m sites i on molecule 1 to n sites j on molecule 2, $[\mathbf{r}_{ij}]_{m,n} \equiv [\mathbf{r}_{11}, \mathbf{r}_{12}, \dots, \mathbf{r}_{mn}]$. This collection of distance vectors can be well-ordered, up to degeneracies, by scalar magnitude. For example, if we pick a particular site i on molecule 1, and label the set of distance vectors from site i to the n sites j of molecule 2 by their relative magnitude, $[r_{ij}]_n$ as $[r_i^{(m)}]$ where $r_i^{(1)}$ is the shortest

distance of the set $[r_{ij}]_n, r_i^{(2)}$ is the second-shortest, and so on, then we have the ordered set $[r_{ij}]_n \equiv [r_i^{(1)} \leq r_i^{(2)} \leq \dots \leq r_i^{(n)}]$ for the n distances between any single site i on molecule 1 and the n sites in molecule 2 for a particular configuration $(\mathbf{r}_{ij}, \Omega_1, \Omega_2)$. Given this ordering, we define the conditional radial distribution function

$$\rho_i g_{ij}(r, n) \rho_j = \frac{1}{V} \langle N(N-1) \delta(r_i^{(n)} - r) \rangle \quad (8)$$

as the probability density of site j on molecule 2 around site i on molecule 1 given that site j is the n th closest site on molecule 2 to site i on molecule 1. By definition, the complete set of conditional probabilities sums to the radial site-site distribution function:

$$g_{ij}(r) = \sum_n g_{ij}(r, n). \quad (9)$$

As a specific example, using any standard 3-site model of water, $g_{\text{OH}}(r)$ the radial site-site distribution between an oxygen site on molecule 1 and a specific hydrogen site on molecule 2, then the site-site distribution is given as the sum of the series

$$g_{\text{OH}}(r) = g_{\text{OH}}(r, 1) + g_{\text{OH}}(r, 2) + g_{\text{OH}}(r, 3), \quad (10)$$

and is constructed to be the sum of probabilities for $g_{\text{OH}}(r, 1)$ the radial distribution of hydrogen sites on molecule 2 with respect to the oxygen site on molecule 1 where the r_{OH} distance is the shortest distance between the oxygen on molecule 1 and the 3 sites of molecule 2, $g_{\text{OH}}(r, 2)$ is the distribution where r_{OH} is the second shortest distance, and $g_{\text{OH}}(r, 3)$ is the distribution for r_{OH} the third shortest distance. Here we have considered water as a solute in water for an illustration.

These are the most familiar, and basic, form of the proximal radial distribution functions. However, there is an interesting degree of freedom left to examine. The development of the series expansion in this form is chiefly aimed at understanding the solvation density around a given solute. Thus, the proximal expansions of the site-site radial pair functions $g_{ij}(r, n)$ are taken from the complete mutual orientational averages of the full $g(\mathbf{r}, \Omega_1, \Omega_2)$. There are two principal intermediate averages over the orientations of molecule 1 or 2 separately,

$$\begin{aligned} g_{ij}(\mathbf{r}_{ij}, \Omega_2) &= \langle g(\mathbf{r}_{ij}, \Omega_1, \Omega_2) \rangle_{\Omega_1}, \\ g_{ij}(\mathbf{r}_{ij}, \Omega_1) &= \langle g(\mathbf{r}_{ij}, \Omega_1, \Omega_2) \rangle_{\Omega_2}, \end{aligned} \quad (11)$$

which we call^{12–14} the *right*- and *left*-handed angular correlation functions, respectively. For clarity, we point out that the *left*-hand correlation function is generated from the complete pair function by a *right*-hand average, and *vice versa* for the *right*-hand correlation function. Since, by definition,

$$\begin{aligned} g_{ij}(r) &= \langle g_{ij}(\mathbf{r}_{ij}, \Omega_2) \rangle_{\Omega_2} \\ &= \langle g_{ij}(\mathbf{r}_{ij}, \Omega_1) \rangle_{\Omega_1}, \end{aligned} \quad (12)$$

the site-site distribution functions $g_{ij}(r)$ are independent of the order in which the angular averages are taken.

For the proximal functions, instead, since the definition in Eq. (8) is for the n sites on molecule 2 relative to a given site i on molecule 1, there is an additional set of functions

$g_{ij}(r, m)$,

$$\rho_i g_{ij}(r, m) \rho_j = \frac{1}{V} \langle N(N-1) \delta(r_i^{(m)} - r) \rangle, \quad (13)$$

where $4\pi g_{ij}(r, m) \rho_j r^2 dr$ is the probability of finding site j on molecule 2 at a distance r from site i on molecule 1, given that site i is the m th closest site of molecule 1 to site j on molecule 2, and $\sum_m g_{ij}(r, m) = g_{ij}(r)$. Just as for the angular functions, there is a *left*-handed and a corresponding *right*-handed sense to the proximal functions. They are formally distinguished by recognizing which angular averages are unrestricted in the construction. For the $g_{ij}(r, n)$ functions of Eq. (8), the angular average is unrestricted for molecule 1, and thus the $g_{ij}(r, n)$ are conditional averages of the *right*-handed $g_{ij}(\mathbf{r}_{ij}, \Omega_2)$ angular correlation functions. Similarly, for the $g_{ij}(r, m)$ functions in Eq. (13), the averages over molecule 2 are unrestricted, and the $g_{ij}(r, m)$ are functions of the *left*-handed $g_{ij}(\mathbf{r}_{ij}, \Omega_1)$ angular correlation functions. We therefore label the usual proximal distribution function hierarchy set $g_{ij}(r, n)$ from Eq. (8) as the *right*-handed proximal distribution functions, and the complementary set $g_{ij}(r, m)$ from Eq. (13) as the *left*-handed proximal distribution functions.

The relation of the *left* and *right* intermediate proximal radial distribution functions to their respective intermediate angular averages of the pair distribution function $g(\mathbf{r}, \Omega_1, \Omega_2)$ indicates that there exists an even more general set of functions that correspond to the full pair distribution function. Rather than the m sites of molecule 1 with respect to site j of molecule 2, or the n sites of molecule 2 with respect to a site i of molecule 1, if we instead take the complete set of $\eta = m \cdot n$ site-site distances between the pair of molecules 1 and 2 ordered from least to most, $[r^{(\eta)}] \doteq [r^{(1)} \leq r^{(2)} \leq \dots \leq r^{(\eta)}]$, we then have the set of functions

$$\rho_i g_{ij}(r, \eta) \rho_j = \frac{1}{V} \langle N(N-1) \delta(r_i^{(\eta)} - r) \rangle, \quad (14)$$

where $4\pi g_{ij}(r, \eta) \rho_j r^2 dr$ is the probability of finding site j of molecule 2 at distance r from site i of molecule 1 given that r_{ij} is the η th closest of the $m \cdot n$ site-site distances between molecules 1 and 2. As before,

$$g_{ij}(r) = \sum_{\eta} g_{ij}(r, \eta). \quad (15)$$

For molecular pair functions, this is the most general set of proximal distribution functions that can be defined using only the set of site-site distances between molecules, and has the same formal relationship to the left and right functions $g_{ij}(r, m)$ and $g_{ij}(r, n)$ as does the pair function $g(\mathbf{r}_{ij}, \Omega_1, \Omega_2)$ to the left and right angularly dependent correlation functions $g_{ij}(\mathbf{r}_{ij}, \Omega_1)$ and $g_{ij}(\mathbf{r}_{ij}, \Omega_2)$.

To summarize, the proximal radial distribution functions as previously constructed^{2,3} are one set of a general class of proximal distribution functions, each member of the class being delineated by its construction from one of the various angle averages of molecular pair correlation functions. Additionally, each of the classes of proximal distribution functions, as a complete series, is a set of distributions intermediate between the given angular pair function and the radially symmetric site-site distribution functions. In particular to

this point, the standard proximal radial distribution functions $g_{ij}(r, n)$ are intermediate between the right-handed $g_{ij}(\mathbf{r}_{ij}, \Omega_2)$ angular correlation function and $g_{ij}(r)$, the $g_{ij}(r, m)$ functions intermediate to the left-handed $g_{ij}(\mathbf{r}_{ij}, \Omega_1)$ angular correlation function, and the *both* $g_{ij}(r, \eta)$ functions are intermediate to the complete $g(\mathbf{r}_{ij}, \Omega_1, \Omega_2)$ pair correlation function. In strict terms of angular information content, the set of proximal functions are intermediate with relation to the radial and the angular pair correlations of the fluid.

Finally, we can always in principle re-normalize the conditional functions $g_{ij}(r, \eta)$, $g_{ij}(r, m)$ and $g_{ij}(r, n)$ such that they approach 1 at long distance when using a partial or quasi-component set. In such case we will denote the re-normalized functions as the perpendicular distribution functions, $g_{ij}^\perp(r, \eta)$, $g_{ij}^\perp(r, m)$, and $g_{ij}^\perp(r, n)$.

With these definitions of the proximal distribution functions, both with respect to the site-site radial distribution functions and the molecular pair distribution functions, it is reasonable to ask whether the series representations that the proximal functions provide of the different pair functions allow for a truncation of the series that still predicts the thermodynamic averages of the system with reasonable accuracy. To demonstrate, from the definitions of the thermodynamic averages, the excess internal energy per particle of a pure site-site model system is given by⁹

$$\begin{aligned} \frac{\beta U^{ex}}{N} &= \frac{1}{2} \sum_{ij} \int \beta u_{ij}(r) g_{ij}(r) \rho_j d\mathbf{r} \\ &= \frac{1}{2} \sum_{ij} \sum_n \int \beta u_{ij}(r) g_{ij}(r, n) \rho_j d\mathbf{r} \\ &\approx \frac{1}{2} \sum_{ij} \int \beta u_{ij}(r) g_{ij}^\perp(r, 1) \rho_j d\mathbf{r}, \end{aligned} \quad (16)$$

where $g_{ij}^\perp(r, 1)$ is the nearest neighbor proximal distribution normalized to 1 at large distances, $\beta^{-1} = k_b T$ with k_b Boltzmann's constant, T the absolute temperature in units of Kelvin, and N the total number of molecules in the pure fluid. Since $g_{ij}(r) = \sum_n g_{ij}(r, n)$ for the right-handed proximal distribution functions, we can approximate the total $\beta U^{ex}/N$ using the truncated series in the last line after appropriate normalization. Previous work⁷ has shown that this truncation works reasonably well for the Coulomb term, $u_{ij}^c(r)$, in potentials of the form

$$\begin{aligned} u_{ij}(r) &= \frac{q_i q_j}{r} + 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right] \\ &= u_{ij}^c(r) + u_{ij}^{lj}(r), \end{aligned} \quad (17)$$

when used in conjunction with a Voronoi-like spacial (volume) normalization⁶ to the distribution's first terms, $g_{ij}(r, 1) \rightarrow 1$ as $r \rightarrow \infty$, so that the approximation $g_{ij}(r) \approx g_{ij}^\perp(r, 1)$ holds at long distance.

We note that this can be done from a right or a left handed angular set. In past work we have used the right-handed $g_{ij}(r, n)$ proximal decomposition of the solvent density around a solute. In results given below, we will discuss the numerical results of this truncation for the distributions and poten-

tial energy terms considering the solute angular average using the left-handed decomposition $g_{ij}(r, m)$. In considering the truncation of the series in this way, however, we find that there is another utility to the proximal distribution function set that can be used to directly measure the onset of the coarse-graining convergence of the solution properties of a macromolecular solute.

Coarse-grained averaging as a concept in the physics of atomic systems has a long history. As an illustration, consider that the Lennard-Jones potential is itself a coarse-graining of the non-bonded interactions of the electronic distributions of atoms and molecules. We have in mind here the current and ongoing work¹⁵ that seeks to generate nano and mesoscale potential surface models of molecular systems in order to accurately project the underlying classical Hamiltonians of molecular mechanics models, for instance, the CHARMM¹⁶ and AMBER¹⁷ force fields, to longer length or time scales. The principal idea is that, rather than having macromolecules composed of hundred of thousands of sites, we use models composed of ten to a hundred times fewer sites, each such site representing some unique subset of the atomic sites of the generating potential surface. If these units can be rigorously parameterized,¹⁸ then in principle we could have a transferable set of nano and mesoscale models applicable to larger length and time scales than are generally accessible via all-atom molecular dynamics and Monte Carlo simulations. Our purpose here is not to produce such a model of the solute but to consider length scales where the properties as seen by solvent converge.

Formally, we seek to reduce the total number of site-site calculations necessary to generate the effective solvent-solute potential, so that

$$\sum_{i=1}^m \sum_{j=1}^n u_{ij}(r_{ij}) \cong \sum_{i=1}^m \sum_{J=1}^v U(\mathbf{r}_{iJ}, \Omega_2) \quad (18)$$

where the sites labeled by i, j are for the all-atom potential, $v \ll n$, J labels the sites of the coarse-grained model molecule 2, and $\sum_J U(\mathbf{r}_{iJ}, \Omega_2)$ is an arbitrary set of functions of many fewer sites which can be parameterized to make the equality approximately true for all corresponding configurations. In principle, the set of v sites can be defined arbitrarily for any given solvent-solute system. If we take a hint from generalized nonlocal grid techniques, the v sites can be defined dynamically so as to parameterize the potential surface according to mathematical and computational convenience. Given the ideal goal of transferability of any given parameterization, however, then a strictly utilitarian argument would be for the v sites to be associated with a particular defined subset of the sites in a given molecule.

For instance, we might restrict the v sites to represent a given number of adjacent base pairs in a DNA sequence, or amino acids in a sequence in a protein, averaged such that the parameters roughly represent the sequence potential regardless of the particular macromolecule modeled. This localization of the v sites also has the virtue of representing the same step across hierarchies of physical subunits as, for instance, the electronic structure to Lennard-Jones plus charges potential for small molecules and atomic species, as well as keeping

the principal physical meaning of coarse-graining in the sense of summing over j in small numbers of sites that are adjacent to each other *and* which make up a physically and chemically useful subunit of the macromolecule being modeled. We note that this type of coarse-graining will be most useful for studying macromolecular systems at finite concentration, where the sheer number of all-atom sites of any polymer is a significant hurdle, for example in polymer melts or cellular crowding.

If we want to then understand the general idea of coarse-graining over a well-defined molecular length scale, we need to account for, and measure, the results of all possible ways that one can partition the atomic sites of a molecule, given that the admissible subgroups are restricted to those which are close to each other (proximal) relative to an arbitrary solvent molecule. The proximal distribution functions provide precisely this measure. To see this, we consider the excess energy per molecule for species α in a mixture of site-site model fluids:^{9,10}

$$\begin{aligned} \frac{\beta U^{\text{ex}}}{N_\alpha} &= \frac{1}{2} \rho_\gamma \sum_{\alpha\gamma} \sum_{ij} \int d\mathbf{r} g_{\alpha_i\gamma_j}(\mathbf{r}) \beta u_{\alpha_i\gamma_j}(\mathbf{r}) \\ &= \frac{1}{2} \rho_\gamma \sum_{\alpha\gamma} \sum_{ij} \sum_m \int d\mathbf{r} g_{\alpha_i\gamma_j}(\mathbf{r}, m) \beta u_{\alpha_i\gamma_j}(\mathbf{r}), \quad (19) \end{aligned}$$

where α and γ are species labels, i and j are the site labels within species α and γ , respectively, and for this case, we use the left-handed $g_{ij}(\mathbf{r}, m)$ functions.⁴ Now, for the left-handed proximal distributions of the macromolecule-water distributions, the m sites are many more than if the right-handed $g_{ij}(\mathbf{r}, n)$ functions were used, since $n \ll m$ for water, as in the more common convention.^{3,4,6-8} Using the larger number of macromolecular sites to expand the neighbor hierarchy in allows for the fine-grained functionalization of the excess energy per molecule over the m -member hierarchy:

$$\frac{\beta U^{\text{ex}}(\nu)}{N_\alpha} = \frac{1}{2} \rho_\gamma \sum_{m=1}^{\nu} \sum_{\alpha\gamma} \sum_{ij} \int d\mathbf{r} g_{\alpha_i\gamma_j}(\mathbf{r}, m) \beta u_{\alpha_i\gamma_j}(\mathbf{r}), \quad (20)$$

where $\nu \in [1, \dots, m]$ for the m sites of the α molecule.

This parameterization of $\beta U^{\text{ex}}(\nu)$ with respect to ν measures the solute potential energy as a function of all subsets of ν sites within molecule 1 nearest to a test molecule 2. For a descriptive geometry, consider a linear polymer of arbitrary (large) length. A schematic of this idealized case is given in Figure 1. Now, consider an infinite radius disc perpendicular to the long axis of the polymer molecule, containing a single site of the polymer and a test site of a water molecule in the bulk. This disc is the restricted geometric region for which the solute site included is the nearest solute site to the test site of the solvent molecule, and thus is the region of $g(\mathbf{r}, \Omega_w, \Omega_p)$ which generates $g(r, 1)$, the nearest neighbor proximal distribution function for the polymer-water pair. The higher order members of the proximal distribution function hierarchy will have similar closed, but increasingly complex, geometric regions of $g(\mathbf{r}, \Omega_1, \Omega_2)$ as generators. In total, the potential $U^{\text{ex}}(\nu)$ is that of the ν sites closest to the solvent molecule, summed through all ν measuring geometries centered to the individual solute sites and then summed over all sites of the solute. That is, the first approximation,

FIG. 1. Schematic for the coordinates of the solute-solvent distributions $g(\mathbf{r}, \Omega_w, \Omega_p)$ and proximal components discussed in the text. The black outlined large spheres are the sites on a linear polymer solute with orientation Ω_p , with connecting bonds. The dark blue sphere is the test solvent site. The blue line indicates the nearest neighbor site of the solute to the test solvent site within the blue averaging disc, the green line is for the next nearest neighbor(s), and the red line is for the next-next nearest neighbor.

the potential from nearest neighbors, contains the contributions from only the solute site connected to the solvent via the corresponding blue line averaged over the blue disc region, then summed over each change of register which moves the blue measuring disc and line to the next site and corresponding test solvent molecule in the chain. The next approximation, nearest neighbor plus next-nearest neighbor, contains the blue line/disc contribution plus the green line contribution in its (more complex) next-nearest measuring geometry, and the total sum over solute sites sums these contributions for all changes of register along the solute. The next approximation would include as well the red line contributions, and so on. Further, the construction of the $g_{ij}(\mathbf{r}, m)$ functions insures that all possible Boltzmann-weighted orientations of the test disc are also taken into account. Or, for solutes with arbitrary internal geometry, that all possible localized subsets of the ν -nearest sites to a solvent molecule are accounted for.

We point out that the choice of the solute energy is not a unique measure of coarse graining. There are a variety of possible thermodynamic measures which can be used as the proxy by which the parameters of a given coarse-grained model are chosen, though minimization of the free energy functional is probably the choice most commonly considered formally.^{19,20} The choice for a given system can be formally justified in a variety of ways, and all of the thermodynamic measures that we are aware of have a range of problems for which the given construction is appropriate.²¹⁻²³ The energy here is chosen both for reasons of simplicity, as well as generality, since the integral over the distribution functions would be the key element for, e.g., virial pressure or, through the Kirkwood-Buff approach,²⁴ any of a variety of thermodynamic variables. In application, for example, we have shown that the free energy via the linear response approximation is also accessible from the proximal functions.⁷ Additionally, recent work by Voth and co-workers²⁵ has been concerned with χ^2 fitness measures for coarse-grained models, and in particular for our purposes they have investigated the properties of $\chi^2(n)$, that is, the fitness of a given model as a function of the number of coarse-grained sites used to construct the potential surface. Our simple energy measure is meant to be

TABLE I. Excess internal energies for SPC water system at 298.15 K and $\rho = 0.03346 \text{ N}/\text{\AA}^3$.

Generating function	Total $\frac{\beta U^{(ex)}}{N}$	Coulomb	Lennard-Jones	Running i, j total over n
Total over all sites $\sum_{ij} g_{ij}(r) = -15.2$				
$g_{OO}(r) = \sum_n g_{OO}(r, n)$	624.8	622.0	2.8	...
$g_{OO}(r, 1)$	171.9	170.4	1.5	171.9
$g_{OO}(r, 2)$	365.7	364.2	1.5	537.6
$g_{OO}(r, 3)$	87.2	87.4	-0.2	624.8
$g_{OH}(r) = \sum_n g_{OH}(r, n)$	-319.2	-319.2	0.0	...
$g_{OH}(r, 1)$	-163.2	-163.2	0.0	-163.2
$g_{OH}(r, 2)$	-60.8	-60.8	0.0	-224.0
$g_{OH}(r, 3)$	-95.2	-95.2	0.0	-319.2
$g_{HO}(r) = \sum_n g_{HO}(r, n)$	-319.0	-319.0	0.0	...
$g_{HO}(r, 1)$	-99.2	-99.2	0.0	-99.2
$g_{HO}(r, 2)$	-176.0	-176.0	0.0	-275.2
$g_{HO}(r, 3)$	-43.8	-43.8	0.0	-319
$g_{HH}(r) = \sum_n g_{HH}(r, n)$	159.0	159.0	0.0	...
$g_{HH}(r, 1)$	77.6	77.6	0.0	77.6
$g_{HH}(r, 2)$	32.8	32.8	0.0	110.4
$g_{HH}(r, 3)$	48.5	48.5	0.0	159.0

representative of this sort of χ^2 fitness measure. The proximal convergence of the energy integral, or any other thermodynamic integral over the distribution functions can be viewed as physical or environmental criteria/measures for coarse graining of the all-atom system. As such, the method developed here is meant as a tool for analysis of all-atom systems in order to complement the existing tools for building coarse-grained *models* from them.

III. RESULTS

Here we present numerical results for two systems. Note, throughout this work, we do not renormalize any of the functions in the various proximal distribution function series, so that the identity $g_{ij}(r) = \sum_n g_{ij}(r, n)$ holds rigorously throughout. For our first example, in order to illustrate the fundamental properties of both sets of proximal functions, the right-handed set $g(r, n)$ and the complete set $g(r, \eta)$, we use results from number, volume and temperature of the simple point charge water model (SPC) water model; that is, we consider water about water. All results for pure water were derived from a system of 1024 water molecules, at 300 K and 0.03346 molecules per $\text{\AA}^3$. We use the full Ewald sum, the Nosé thermostat, the Verlet algorithm, and otherwise standard methods of molecular dynamics simulation.²⁶ In our second example, we will use two previously determined⁷ fixed configurations of a deca-alanine peptide in solution using the Amber model¹⁷ with 2200 TIP3P water molecules.

A. Pure water

We first examine the excess internal energy integrals over the set of right-handed proximal distribution functions for the pure water system. In Table I, we break out the results of the integrals in Eq. (19) into their Lennard-Jones and Coulomb contributions term-by-term with the sum over n for $g(r, n)$, in comparison to the total sum, and the direct integral over $g_{ij}(r)$ calculated in the standard way. The trivial result is that the

total sum of energy contributions is identical to the $g_{ij}(r)$ integrals. The more illustrative result is the total (Lennard-Jones plus Coulomb) potential for each site-site pair as given as a running total (i.e., the sum as each additional term is added to the total) over n in Table I. For this case of a very small solute molecule, it is clear that each term of the series over n , for either the Lennard-Jones contributions or the Coulomb case, makes a substantial contribution to the individual site terms. However, for the Coulomb potential, the total site-site Coulomb sum for $n = 1$ is $U^{C; total} = -14.4$, while for the full $n = 3$ the result is $U^{C; total} = -18.4$. The molecular Coulomb potential (i.e., the total Coulomb contribution to the molecular potential) remains of the same magnitude across the series. Truncation of n for the Lennard-Jones potential for the very small molecule case considered would lead to large errors for any set of functions chosen, while the Coulomb results are less sensitive as seen in previous work.^{3,4,6,7} We note that, because the $n = 2, 3$ terms oscillate around this result, the conclusion that can be drawn from all of these various results together is that the fundamental conservation of charge rescues the partial sums of the Coulomb potential for the total molecular potential, but as is normal and expected the term-by-term convergence of a Coulomb series is not generally so well behaved.

For the water around water case, we plot the $g(r, n)$ series for each site-site pair in Figures 2–5. Notice due to the handedness of the angular averages $OH \neq HO$. All of the terms, nearest neighbor, next nearest, and next-next nearest, are of the same relative order of magnitude over the full range of the functions. While there is a clear ordering of the probabilities over any given i, j pair, that ordering is not consistent across pairs. Taken together, this means that no truncation of the series over n is a representation or good approximation of the relative probabilities for the averages considered for small molecules and one should take caution to examine any series.

For molecules with tight angular coupling over longer ranges, correlations between pairs of molecules are non-trivial. Since the $g_{ij}(r, n)$ set corresponds to $g(r, \Omega_2)$, the

FIG. 2. The oxygen-oxygen $g_{OO}(r, n)$ functions for the SPC model water system discussed in the text, together with the total site-site distribution function, $g_{OO}(r) = \sum_n g_{OO}(r, n)$. The blue line is the total $g_{OO}(r)$, the black line is the nearest neighbor $g_{OO}(r, 1)$, the red line is the next-nearest neighbor $g_{OO}(r, 2)$, and the green line is the next-next nearest neighbor $g_{OO}(r, 3)$.

averaging which generates the $g_{ij}(r, n)$ functions is not given by a strict series expansion, in the sense of n being a strict order parameter. As such, while the sub-averages which generate each member $g(r, n)$ are well-defined, direct measurement shows that there is no *a priori* reason to assign an ordering weight to the sequence $1, 2, \dots, n$. By considering the correlations of the averaged solute, in some cases, the nearest neighbor sites of a small molecule solvent may be less important than the next nearest or next-next nearest, at any

FIG. 4. The hydrogen-oxygen $g_{HO}(r, n)$ functions for the SPC model water system discussed in the text, together with the total site-site distribution function, $g_{HO}(r) = \sum_n g_{HO}(r, n)$. The line color conventions are as for Figure 1, blue for the total $g_{HO}(r)$, and black, red, and green for $g_{HO}(r, 1)$, $g_{HO}(r, 2)$, and $g_{HO}(r, 3)$, respectively.

given particular distance, at least as far as which probabilities are important to the whole. Here, since there are only a small number of possible site neighbors for a given solvent there is no reason not to include them all. This is opposite to the case where we consider the angularly averaged solvent around a solute composed of a large number of sites. There, the neighbor hierarchy provides considerable computational convenience with more controllable precision⁵ because proximity then correlates with influence, i.e., a solute site on one

FIG. 3. The oxygen-hydrogen $g_{OH}(r, n)$ functions for the SPC model water system discussed in the text, together with the total site-site distribution function, $g_{OH}(r) = \sum_n g_{OH}(r, n)$. The line color conventions are as for Figure 1, blue for the total $g_{OH}(r)$, and black, red, and green for $g_{OH}(r, 1)$, $g_{OH}(r, 2)$, and $g_{OH}(r, 3)$, respectively.

FIG. 5. The hydrogen-hydrogen $g_{HH}(r, n)$ functions for the SPC model water system discussed in the text, together with the total site-site distribution function, $g_{HH}(r) = \sum_n g_{HH}(r, n)$. The line color conventions are as for Figure 1, blue for the total $g_{HH}(r)$, and black, red, and green for $g_{HH}(r, 1)$, $g_{HH}(r, 2)$, and $g_{HH}(r, 3)$, respectively.

FIG. 6. The oxygen-nearest neighbor $g_{Oj}(r, 1)$ functions for the SPC model water system discussed in the text. The solid red line is the oxygen-oxygen nearest neighbor function $g_{OO}(r, 1)$ and the dotted-dashed red line is the oxygen-hydrogen nearest neighbor function $g_{OH}(r, 1)$.

side of a macromolecule has little influence on the solvent on the opposite side.

The difference considering small molecules is striking. For water as a solute in a solvent of water the results of breaking r, Ω_2 out into $r_{ij}^{(n)}$ in their thermodynamic contributions indicate that, at distances beyond the immediate contact distance between sites, we cannot look at a given orientation and assign the equivalent thermodynamic weight to the bare potential between sites. In other words, since $g_{ij}(r, n)$ and $g(\mathbf{r}_{ij}, \Omega_2)$ average over the orientations of molecule 1 and correlate with all other molecules 2 in the system, we cannot any longer make that direct prediction.

To illustrate, we plot the nearest neighbor functions for SPC water in Figures 6 and 7, for oxygen and hydrogen, respectively. In both cases, at any distance greater than the ~ 3 Å Lennard-Jones contact distance between oxygen sites, the most probable nearest site neighbor j on molecule 2 for site i (whether oxygen or hydrogen) when averaged over the angles of molecule 1 is the hydrogen site on molecule 2. Naively, taking only the distances and bare potential of two SPC water molecules, this might seem counter-intuitive, since the hydrogen-hydrogen interaction is repulsive. However, if we have a site of molecule 1 at the origin and average over the orientations of molecule 1, then the contributions leading to both oxygen and hydrogen of molecule 1 more favorably interacting with the hydrogens of molecule 2 arise from two distinct physical paths. For an oxygen on molecule 1, the hydrogen on molecule 2 should be a *direct* interaction, while for a hydrogen on molecule 1, the interaction with a hydrogen on molecule 2 is *indirect*, mediated by the companion oxygen site of molecule 1 in its interaction with the hydrogen of molecule 2. It is more complicated to use distances between molecular surfaces rather than sites. This appears to alleviate some of the problems of widely disparate atomic site

FIG. 7. The hydrogen-nearest neighbor $g_{Hj}(r, 1)$ functions for the SPC model water system discussed in the text. The solid red line is the hydrogen-oxygen nearest neighbor function $g_{HO}(r, 1)$ and the dotted-dashed line is the hydrogen-hydrogen nearest neighbor function $g_{HH}(r, 1)$.

diameters,^{4,27} and for molecules with a wide range of site charges that provide an additional challenge coupled to the one of sizes.

B. Ala₁₀

For the more complicated deca-alanine solute, the right-handed angle averaged solvent distributions have been discussed elsewhere.⁷ Here in this work we consider the left-handed distributions. The excess energy per solute results $\beta U^{ex}(\nu)/N_{solute}$ are presented in Figures 8 and 9, for the alpha-helix and the extended structures previously identified. These calculations were taken from simulations run expressly for this work. The total Lennard-Jones potentials, -5.8 kcal/mol for the alpha-helix structure and -1.8 kcal/mol for the extended structure, and the total Coulomb potentials, -57.2 kcal/mol for the alpha-helix and -99.2 kcal/mol for the extended, are consistent with the previously reported results for these systems.

There are two noteworthy features of the Lennard-Jones energy results. First, the convergence with respect to the ν nearest solute sites to a solvent molecule, and second the positive shoulder at shorter length scale. We identify the first as the onset of solute coarse-graining in the system, due to the construction of $U^{ex}(\nu)$ as discussed above. Inspection indicates that the coarse-graining convergence in this particular system takes effect at approximately the length of 4 or 5 amino-acids in the solute (assuming that each interior amino acid has ~ 10 sites). We identify the second feature, the positive shoulder, as related to the repulsion/exclusion work, or cavity work, necessary to insert the solute into the solvent. That this must be the cavity work is due to the form of the Lennard-Jones potential: the positive energy contribution is from the repulsive r^{-12} term alone. The effect for the linear

FIG. 8. The excess Lennard-Jones solvent-solute potential for deca-alanine in water, as a function of the ν nearest solute sites to a solvent molecule. The black line is for the alpha-helix structure and the red line is for the extended structure. The energy units here are in kcal/mol.

polymer chosen here is that the cavity work is measurable by the solvent at the scale of 1 or 2 amino acids, while the attractive r^{-6} term dominates the total at larger scales. Finally, we note that the physical scaling of the Lennard-Jones potential of the all-atom system appears to follow a similar decay as the $\chi^2(n)$ functions used to measure the accuracy of coarse-grained models as a function of the number of sites used to represent the potential model.²⁵ Our results for the scaling of the different extended and helical structures also agree with the results²⁵ showing that the number of sites necessary to converge a coarse-grained macromolecule model may be conformationally dependent, due to the fluctuation in structures.

FIG. 9. The excess Coulomb solvent-solute potential for deca-alanine in water, as a function of the ν nearest solute sites to a solvent molecule. The line types and units are the same as for Figure 3.

FIG. 10. The excess solute site-water potential for the extended deca-alanine configuration in water, as a function of ν . The black line is the total for all hydrogen sites in the solute, the red line is for all carbons, the blue line is for all nitrogens, and the green line is for all oxygens.

For the Coulomb surface, the convergence is dominated by cancellation of charges. This is guaranteed by the sum over sites in $U^{\text{ex}}(\nu)$, and the result is a function that oscillates around its ultimate result. The oscillations are on the scale of the contributions from the partially charged sites of the individual amino acids, and are a representation of the fact that the ν parameterization is with respect to any ν -nearest neighbors of a solvent molecule. As with the Lennard-Jones potential, the fluctuations converge at approximately the scale of 4 or 5 amino acids (less for the extended system, more for the alpha-helix).

A direct display of the data analogous for deca-alanine to that of the water as in Table I would contain too many entries to be immediately useful. However, if we take the total for each site type within the solute, the sum over j water sites of $U_{ij}(\nu)$ for the various (hydrogen, carbon, nitrogen, oxygen) site types i of the solute summed over the total number (of carbon, oxygen, etc.) of each type present in the solute, gives a manageable graphical presentation of the information for the solute, as well as a solute site decomposition of the total energies given in Figures 8 and 9. The total is presented in Figure 10 for the extended structure. The major feature, other than the expected convergence scale, is that the total solute-water potential is dominated by the oxygen and carbon terms, with the hydrogen and nitrogen terms being small due to the summation over water sites.

In the context of using the proximal distribution function methods outlined above to help generate coarse-grained models of all-atom site-site model molecules, we see the utility of the method given here as follows. We would start by running a test polypeptide simulation using an all-atom potential model. Then, one would use the proximal functions to measure the all-atom simulation as above, and find the natural scaling of the system, which here occurs at approximately 5

amino acids for the solute-solvent interaction. Assuming that further all-atom test cases found a similar scaling, then we would conclude that the potential model chosen to represent that basic unit, regardless of the particular functional form, could be optimally parameterized so as to reproduce the energy of interaction of the all-atom sequence dependent test cases. That is, the proximal functions are a tool for *a priori* measuring the convergence of coarse-graining at the all-atom level. One might then build a coarse-grained model according to the required solute-solute interactions, efficiency, and so on.

We seek convergence, as a function of the coarse-graining parameter ν , to hold over a range of oligomeric macromolecules. For molecular mechanics potentials, it is usually the case that a given polymer residue or subunit has the same potential parameters (within a given model set) regardless of the macromolecule investigated. Meaning, e.g., a glycine is modeled the same whether it is used in tyrosine or myoglobin. Though chemically this is only approximately true, in molecular mechanics computations it is a common compromise across a huge variety of different competing necessities. Since the potentials are the same, then it is also reasonable to assume that the approximate solution structures of the resulting subunits will, on average, have features that are similar. As an example, the average solution structure around an arbitrary glycine in a globular protein should be broadly similar to any other glycine in any other globular protein, given its proximity to the solvent in the context of condition of it in the protein geometry. While any particular subunit may have a local solution structure that deviates, the average structure of all sub-units should be similar in different protein cases simply due to the fact that the same basic physical properties hold for each system in the class. This suggests that, just as molecular mechanics potential models are constructed from a broad sample of the electronic structures of a chemical species across many states, a distribution model can be constructed in the liquid state.

Consider a general set of model solvation distributions which reproduces on average the solvation of an arbitrary polymer. If there is a ν' which holds for a sequence specific subset of polymers, then there is a functional description of the solvation of the members of the subset that allow a coarse-graining length related to ν' , and the functions which describe it are

$$G_{ij}(r, \nu') = \sum_{n=1}^{\nu'} g_{ij}(r, n), \quad (21)$$

where i is the solvent label, j is the sites of the individual sub-units of the polymers in consideration (i.e., base pairs, amino acids, etc.), and the $g_{ij}(r, n)$ functions in the collection are averaged over many examples of the solute class in question. Using our example above, one would have $G_{ij}(r, 50)$ for an individual alanine amino acid, averaged over a representative sample of alanine-containing proteins in a class for which $\nu' \approx 50$ is found to be representative as a measure of the coarse graining, and the general model of the average solvation structure of an arbitrary protein in the class is that constructed from the G functions for all of the amino acids in

the protein sequence. We have only considered here the solvation and considerable effort is required for the intra-solute or solute-solute model aspects.

IV. CONCLUSIONS

The purpose for this work was to examine the generalization of proximal pair distribution functions to consider the solute coarse grained properties as opposed to the solvent. We considered how the contributions of the effective potential energy surface of a polymer in aqueous solution could be meaningfully represented using a one sided, angularly averaged proximal radial distribution function methods. We have formally sketched the four constructed proximal alternatives, angularly averaging on either the left or the right side, for either the solute or the solvent. Previous work in this lab showed that the most well-known such proximal distribution functions, the nearest-neighbor or 1st-order terms of the solute-solvent site-site proximal radial distribution function hierarchy, were sufficient as an approximation to the solvation structure of the deca-alanine peptide to predict the effective Coulomb contribution to the potential energy with acceptable accuracy. For this work we showed that we could also use the solute averaged distributions to determine convergence criteria for coarse graining. For 3-site water models considered as both solvent and solute, acceptable convergence of the energies for the individual site-site terms of the total potential requires all terms in the proximal distribution function expansion of the site-site distribution functions. Yet, it is still the case that the total molecular Coulomb potential approximately converges at the first term in the proximal series n . The charge density zeroth moment condition aids convergence of the terms of the partial sums. For the Lennard-Jones terms, of course, there are no similar fundamental relations. Convergence in this case requires adding sufficient proximal terms to obtain the usual expected distance criteria used in cutoffs and other potential truncation schemes.

We also derived several formal generalizations of the basic proximal distribution theory. The first result was that the distance criterion of the neighbor hierarchy generalizes to all collections of site-site distances within the molecule. As such, there is a general set of site-generated proximal distribution functions that follow the same *left*, *right*, and *both* hierarchy of terms as do the full, angularly dependent molecular pair distribution functions.¹⁴ In turn, each type of generalized proximal distribution function is uniquely generated from an average of its related angular correlation function over the familiar r , Ω_1 , Ω_2 molecular coordinates within the proximal ordering of the complete set of $\mathbf{r}_{ij}$ site-site vectors unique to a given configuration of a pair of molecules.

Consideration of which side to angularly average and which group to leave as sites allows us to generate $U^{ex}(\nu)$, where when $\nu = n$, $U^{ex}(\nu) = U^{ex}$, the complete potential energy, in this case, of the alanine solute in water. A result of this paper was that this functionalization gives a measure of the length scale, in terms of the number of sites ν in the solute molecule, at which coarse-graining of the internal sites of the solute molecule with respect to the solvent sets in. Using this result, we generated the total set of right-handed

solvent-solute proximal distribution functions of an extended and an alpha-helix conformation of deca-alanine in water, and then showed the functionalization of the total $U^{ex}(\nu)$ for each conformation. Both conformations showed convergence of the solute potential energy at a ν value on the scale of approximately 4 or 5 amino acids in length. Additionally, while the extended structure converged more smoothly at that length scale, the alpha-helical structure demonstrates a more persistent length correlation, indicative of a different local solvation structure.

We note that both of these results are consistent with the $\chi^2(n)$ fitness measures developed to study the accuracy of the representation of coarse-grained models.²⁵ Overall, these results are consistent with a coarse-grained picture in which the length scale of the polymer that a proximal solvent water molecule samples is of the order of 5 amino acids for this poly-peptide. This is also consistent with previous and ongoing free-energy calculations on this and similar systems in this lab, and with a standard picture of the solvation mechanism of large molecules.

These formal and numerical results lead us to the idea that, for broad classes of molecules, similar average scaling results should hold for the solvation potential functions $U^{ex}(\nu)$ for the members of a class. The functions which may form a transferable solvation structure consistent with the measured solvation potentials of the members of the class are, or can be constructed from, the average $g_{ij}(r, \nu)$ sets of proximal distribution functions for the whole class. Specifically, suppose that, for some set of polypeptides on the scale of the deca-alanine studied here, the solvation energies $U^{ex}(\nu)$ all converge at roughly the 5 amino acid scale. Then, for that set of polypeptides, the set of transferable solvation distribution functions broadly describing the solvation of the members of the set and consistent with the measured scale ν' are the sets of $G_{ij}(r, \nu') = \sum_{n=1}^{\nu'} g_{ij}(r, n)$ functions for each amino acid averaged over the set, i.e., $G_{ij}(r, \nu')$ for alanine, glycine, and so on. The remaining question for practical implementation requires consideration of whether a ν' exists for such classes of polymer as well as sequence variations.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support of the National Institutes of Health (GM 037657), the Na-

tional Science Foundation (CHE-1152876), and the Robert A. Welch Foundation (H-0037). A portion of the research was performed at EMSL, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory. Additional high performance computing work was carried out through the resources from Xsede via the Texas Advanced Computing Center (TACC) at The University of Texas at Austin.

- ¹A. Ben-Naim, *Water and Aqueous Solutions* (Plenum Press, New York, 1992).
- ²P. K. Mehrotra and D. L. Beveridge, *J. Am. Chem. Soc.* **102**, 4287 (1980).
- ³M. Mezei and D. L. Beveridge, *Methods Enzymol.* **127**, 21 (1986).
- ⁴V. Lounnas, B. M. Pettitt, and G. N. Phillips, Jr., *Biophys. J.* **66**, 601 (1994).
- ⁵W. R. Rudnicki and B. M. Pettitt, *Biopolymers* **41**, 107 (1997).
- ⁶V. A. Makarov, B. K. Andrews, and B. M. Pettitt, *Biopolymers* **45**, 469 (1998).
- ⁷B. Lin, K.-Y. Wong, H. Kokubo, and B. M. Pettitt, *J. Phys. Chem. Lett.* **2**, 1626 (2011).
- ⁸J. J. Virtanen, L. Makowski, T. R. Sosnick, and K. F. Freed, *Biophys. J.* **99**, 1611 (2010).
- ⁹J. P. Hansen and I. R. McDonald, *Theory of Simple Liquids*, 2nd ed. (Academic Press Inc., London, 1986).
- ¹⁰D. Chandler, in *The Liquid State of Matter: Fluids, Simple and Complex*, edited by E. W. Montroll and J. L. Lebowitz (North Holland Pub. Co., Amsterdam, 1982), p. 275.
- ¹¹C. G. Gray and K. E. Gubbins, *Theory of Molecular Fluids* (Clarendon Press, Oxford, 1984), Vol. 1.
- ¹²B. M. Ladanyi and D. Chandler, *J. Chem. Phys.* **62**, 4308 (1975).
- ¹³D. Chandler, R. Silbey, and B. M. Ladanyi, *Mol. Phys.* **46**, 1335 (1982).
- ¹⁴K. M. Dyer, J. S. Perkyns, and B. M. Pettitt, *J. Chem. Phys.* **127**, 194506 (2007).
- ¹⁵E. M. Curtis and C. Hall, *J. Phys. Chem. B* **117**, 5019 (2013).
- ¹⁶N. Foloppe and A. D. MacKerell, Jr., *J. Comput. Chem.* **21**, 86 (2000).
- ¹⁷J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, *J. Comput. Chem.* **25**, 1157 (2004).
- ¹⁸W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, *J. Chem. Phys.* **128**, 244114 (2008).
- ¹⁹C. G. Gray, K. E. Gubbins, and C. H. Twu, *J. Chem. Phys.* **69**, 182 (1978).
- ²⁰J.-P. Hansen, C. I. Addison, and A. A. Louis, *J. Phys.: Condens. Matter* **17**, S3185 (2005).
- ²¹T. Head-Gordon and F. H. Stillinger, *J. Chem. Phys.* **98**, 3313 (1993).
- ²²G. Toth, *J. Phys.: Condens. Matter* **19**, 335222 (2007).
- ²³W. Li and S. Takada, *Biophys. J.* **99**, 3029 (2010).
- ²⁴J. G. Kirkwood and F. P. Buff, *J. Chem. Phys.* **19**, 774 (1951).
- ²⁵A. V. Sinititskiy, M. G. Saunders, and G. A. Voth, *J. Phys. Chem. B* **116**, 8363 (2012).
- ²⁶M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, Oxford, 1987).
- ²⁷V. Lounnas and B. M. Pettitt, *Proteins* **18**, 133 (1994).