Molecular interpretation of preferential interactions in protein solvation: a solvent-shell perspective by means of minimum-distance distribution functions

Leandro Martínez*[1] and Seishi Shimizu[2]

[1]Institute of Chemistry and Center for Computational Engineering & Science, University of Campinas. Campinas, SP, Brazil.

[2]York Structural Biology Laboratory, Department of Chemistry, University of York. Heslington, York, UK.

November 8, 2017

Abstract

Preferential solvation is a fundamental parameter for the interpretation of solubility and solute structural stability. The molecular basis for solute-solvent interactions can be obtained through distribution functions, and the thermodynamic connection to experimental data depends on the computation of distribution integrals, specifically Kirkwood-Buff integrals for the determination of preferential interactions. Standard radial distribution functions, however, are not convenient for the study of the solvation of complex, non-spherical solutes, as proteins. Here we show that minimum-distance distribution functions can be used to compute KB integrals while at the same time providing an insightful view of solute-solvent interactions at the molecular level. We compute preferential solvation parameters for Ribonuclease T1 in aqueous solutions of urea and trimethylamine N-oxide (TMAO), and show that, while macroscopic solvation shows that urea is preferentially bound to the protein surface and TMAO is preferentially excluded, both display specific density augmentations at the protein surface in dilute solutions. Therefore, direct protein-osmolyte interactions can play a role in the stability and activity of the protein even for preferentially hydrated systems. The generality of the distribution function and its natural connection to thermodynamic data suggests that it will be useful in general for the study of solvation in mixtures of structurally complex solutes and solvents.

1 Introduction

Biomolecules perform their function in an environment comprised mainly of water molecules. In addition to water, there are many other small- and macro-molecules that influence the equilibrium and kinetics of binding, folding and conformational changes.[1] Especially important are the osmolytes, which are small, organic metabolites such as sugars and polyols, used by organisms to protect biomolecules under denaturation stress (due to heat, cold, pressure or chemical denaturants).[2,3] For an elucidation of biomolecular stability and function on a molecular scale, and towards the rational design and exploitation of enzymes, it is indispensable to understand how biomolecules in water interact with the third component.[4,5] (Throughout this paper, the third component is referred to as the cosolvent, which encompasses both protein stabilizers - osmolytes - and denaturants, both organic and inorganic, and both small- and macromolecules.)

The modulation of protein stability by cosolvents can be quantified via preferential solvation, i.e., the competition between protein-water and protein-cosolvent interactions for native and denatured states, which is accessible experimentally by membrane dialysis, scattering, and analytical ultracentrifugation.[6] Such experiments, when complemented by volumetric measurements,[7] can determine both the protein-water and protein-cosolvent interactions defined rigorously via statistical thermodynamics.[8–11] Such "interactions" are now defined rigorously in terms of the net excess or deficit of water and cosolvent concentrations from the bulk, referred to as the Kirkwood-Buff integrals (KBI).[8–11] KBIs serve not only as powerful tool for rationalizing and explaining experimental data on a variety of cosolvent-controlled effects (from biophysics, formulation science, pharmacy to food science)[5,12–15] but also as a benchmark for simulation and force field determination.[4,16]

KBIs are usually defined as the integrals of the solute-solvent distribution functions, most commonly the radial distribution functions (RDFs). Radial distribution functions are not convenient for the interpretation of the solvation of structurally complex solutes. This is because they are computed from the distances between the centers of masses, or specific atoms, of the solute and solvent molecules,[17] being highly dependent on the solvent and solute shapes. Thus, the necessity of more general distribution functions for the analysis of solutes of complex shapes has been recognized frequently.[18–21] In particular, the use of the distance of one solvent site (an atom or the center of mass) of the solvent molecule to the surface of the solute, or to the nearest solute atom, was proposed independently by different authors as an alternative to overcome the complexity of the solute shape.[18–20,22–27] This choice defines what has been called the "solvation-shell" distribution functions, $g_{ss}(r)$,[23,24] or proximal distribution functions, $g_{\perp}(r)$,[18,22,26,27] which appeal directly to the concept of Voronoi tesselation.[21,28] In all cases, the counting of nearest distances is straightforward from a simulation, but the normalization procedure leading to the distribution functions can be cumbersome.[20] When using Voronoi tesselation, the normalization might depend on the estimation of the volumes in space associated with each

reference atom or site.[21,26] The normalization of the solvation-shell distribution function, on the other hand, has employed a random distribution of solvent molecules.[23,25,29]

Independent computation of the KBIs for each solvent component and its comparison with a distribution function that reflects the solvent shell structure is important for understanding the molecular basis of solvation of complex solutes, biomolecules in particular. We demonstrate that these two aims can be fulfilled simultaneously by adopting an alternative distribution function, i.e., minimum-distance distribution function, which is better suited for characterizing solution structure according to the distance between the solute and solvent surfaces, which is in line with our classical view of the hydration shell of proteins. The resulting distribution functions are clear to interpret from the point of view of molecular interactions, and can be naturally decomposed into the contributions of each solvent atom.

Despite the powerfulness of KBIs as a bridge between the microscopic solution structure and macroscopic thermodynamics, KBI determination from simulation usually requires a very large simulation box to account for the long-ranged deviation of the radial distribution function (RDF) from 1 (i.e., the long-ranged deviation of the local concentration from the bulk). To put it precisely, small non-zero $g(r) - 1$ at large r is multiplied by the volume element, $4\pi r^2$ and contributes greatly to KBIs. Therefore, the use of RDFs as a route for calculation and interpretation of KBIs have been questioned,[21] and a number of proposals have been made to overcome this problem computationally, in order to facilitate the calculation of KBIs via RDFs. Because RDFs rely usually on the distance between the centres of mass, non-spherical molecules require longer distances for the RDF to converge to 1 due simply to their shape.[21] For ternary systems (solute, solvent and cosolvent) it is possible from KBIs to compute preferential interaction parameters, which are a measure of the excess number of cosolvent molecules in the domain of the solute, and can be determined experimentally.[4,19,30,31] Interestingly, it is easier to compute the preferential interaction parameter than KBIs.[4,32,33] This is because the preferential interaction parameter can be computed from the difference in the number of solvent and cosolvent molecules in the "solute domain", this being the volume in space for which the solute affects effectively the structure of the solvent.[4,19,34] The computation of the KBIs for each solvent component depends on the number molecules of that component on the protein domain relative to a reference state, the pure solvent with same density. The calculation of the number of molecules of the solvent in the solute domain depends, therefore, on the volume of the protein domain, which must be independently determined.

Using the minimum-distance framework, we study the preferential interactions of a model protein Ribonuclease T1 (RNaseT1) by urea and TMAO (trimethylamine-N-oxide), which are known experimentally to be preferentially attracted and excluded from the vicinities of the protein.[31] We show that simulations reproduce the experimental preferential solvation parameters, but that the molecular interpretation for the preferential

exclusion is counterintuitive at first sight. Indeed, both TMAO and urea exhibit local density augmentation at the protein surface, particularly at low concentrations. The two cosolvents, however, differ in their relative accumulation on the protein surface relative to that of water, explaining the overall preferential hydration or dehydration observed experimentally. The detailed interpretation of the density augmentation of the cosolvent molecules on the protein surface is beyond the reach of conventional RDFs. The generality of the approach presented here suggests that it can be useful for the study of solvation in mixtures of solutes and solvents of complex shapes, for which the definition of distribution functions to represent solvation interactions can be cumbersome.

2 Theory

2.1 General formalism

In this work we will use thermodynamic functions for the analysis of the solvation of complex solutes using minimum-distance computations. Throughout the paper, we will use the following notation: subscript u to refer to the solute (protein), s to any of the solvent componentes (water or cosolvent), w to water, and c to the cosolvent, either urea or TMAO.

Let $n_{us}(r)$ be ensemble average number density (which can be converted to molarity by multiplying by Avogadro's number) of solvent atoms which are minimum-distance atoms to the solute at r. We define $n^*_{us}(r)$ as the number of minimum-distances that would be observed at r if there were no solute-solvent interactions. Note that $n^*_{us}(r)$ represents the minimum distance distribution in the presence of the solute, but in the absence of solute-solvent interactions. Such a definition requires the presence of a "phantom" solute, which has no attractive nor repulsive interaction with the solvent molecules, but is present to define the position in the solution in the same way as when the solute-solvent interactions are present.

The ratio between $n_{us}(r)$ and $n^*_{us}(r)$ is the variation in density associated with the insertion of the solute, and we will call it $g^{md}_{us}(r)$,

$$g^{md}_{us}(r) \equiv \frac{n_{us}(r)}{n^*_{us}(r)} \tag{1}$$

$g^{md}_{us}(r)$ has a simple physical interpretation: $-RT \ln g^{md}_{us}(r)$ represents the change in the potential of mean force at the position r when the solute-solvent interactions are switched on. If the solvent has a single site, or if any single reference (center of mass or any atom) of the solvent is considered instead of the minimum distance, $g^{md}_{us}(r)$ reduces to a standard distribution function, because $n_{us}(r)$ becomes the density of solvent

molecules at r, and $n^*_{us}(r)$ is turns out to be constant and equal to the bulk density of the solvent.

The minimum-distance densities $n_{us}(r)$ and $n^*_{us}(r)$ differ from the densities of the solvent, because they are associated with the counting of any atom of the solvent molecule at each distance. For example, in the absence of solute-solvent interactions, the density of atoms of the solvent at short distances is associated with the probability of finding any atom of the solvent at that distance, which corresponds to the atomic density, not the molecular density of the solvent. The difference between the two densities will be discussed in further detail in Section 2.2.

The Kirkwood-Buff integral can be obtained from $n_{us}(r)$ and $n^*_{us}(r)$. To this end, let us define an integral up to a finite maximum distance R from the solute, as

$$G_{us}(R) = \frac{1}{\rho_s}\int_0^R [n_{us}(r) - n^*_{us}(r)]S(r)dr$$

where $S(r)$ is the surface defined by the minimum-distance r to any solute atom. By integration, we obtain

$$G_{us}(R) = \frac{1}{\rho_s}[N_{us}(R) - N^*_{us}(R)] \tag{2}$$

where $N_{us}(R)$ and $N^*_{us}(R)$ are, in this case, the number of solvent molecules with $at\ least$ an atom within R of the solute, in the presence of not of solute-solvent interactions, and ρ_s is the bulk density of this solvent component. The volume defined by distance R to the solute is, of course, dependent on the shape of the solute. These equation are generalizations of the standard equation to compute KB integrals from the standard radial distribution function, for which we have $S(r) = 4\pi r^2$ and $n^*_{us}(r) = \rho_s$ for all r, and are needed here because of the dependence of the minimum-distance densities on the shape of the solute and solvent molecules.

A brief justification concerning the convergence of $G_{us}(R)$ defined in Equation 2 to the actual KBI at the limit of $R \to \infty$ can be made in the following way. The equivalence of this definition of $G_{us}(R)$ at large R to the standard definition of KBIs requires that[4,35]

$$G_{us}(R) \xrightarrow[R\to\infty]{} V\frac{\langle N_s N_u\rangle}{\langle N_u\rangle\langle N_s\rangle} - V \tag{3}$$

This can be achieved most simply by noting that for each molecule the minimum-distance atom (i. e., the one which takes the shortest distance from the solute molecule) is uniquely defined (the configurational space in which there are two atoms at exactly the same distance is infinitesimal and hence can be neglected). Hence the $N^*_{us}(R)$ of Equation 2 yields the number of solvent molecules in the absence of solute-solvent interactions, which is $\langle N_s\rangle$; hence $\frac{N^*_{us}(R)}{\rho_s} \to V$. That the $\frac{N_{us}(R)}{\rho_s}$ term converges to the first term of Equation 3 can be

justified by the use of the inhomogeneous solvation theory, according to which [36]

$$V\frac{\langle N_s N_u\rangle}{\langle N_u\rangle\langle N_s\rangle} = V\frac{\langle N_s\rangle_u}{\langle N_s\rangle} = \frac{\langle N_s\rangle_u}{\rho_s}$$

where $\langle N_s\rangle_u$ is the number of solvent molecules in the inhomogeneous system, i. e., in presence of the solute at the fixed position in the origin. That the number of solvent molecules can be evaluated by enumerating the minimum-distance atoms again guarantees that $N_{us}(R)$ converges to $\langle N_s\rangle_u$, thereby justifying that the KB integrals can be calculated using Equation 2.

For small R, the use of the minimum distance counts for $N_{us}(R)$ and $N^*_{us}(R)$ associates the distance dependence of $G_{us}(R)$ with the $g^{md}_{us}(r)$ function of Equation 1, providing the connection between preferential solvation parameters and the molecular structure of the solvent in the solute solvation shell.

2.2 Interpretation of the $g^{md}(r)$ distribution in terms of atomic contributions

Let us consider the case of a single solute molecule and a single component of the solution, for simplicity of notation (the subscript us will be omitted here). Here we define r as the minimum-distance between any point in space and an atom of the solute. The minimum-distance distribution function $g^{md}(r)$ is defined in Equation 1, as a function of the density of minimum-distances at distance r, $n(r)$, and the density of minimum-distances at r in the absence of solute-solvent interactions, $n^*(r)$.

Given a solvent atom of type i ($i = 1, ...N$, where N is the number of atoms of the solvent molecule), we define $n_i(r)$ as the number density of atoms i at distance r from the solute, such that $n_i(r)dV(r)$ is the number of atoms of type i at within r and $r + dr$. Then, we define the probability of, given that the atom i is in volume element $dV(r)$, this atom being the closest atom to the solute. This probability is dependent non-trivially on the shape of the surface associated to the volume element at r and on the shape of the solvent molecule, and we will call it $w_i(r)$. The revolutions of a solvent molecule that are associated with an atom being the closest atom to the solute are illustrated in Figure 1.

It follows that the contribution of atoms of type i to the minimum-distance density at r will be $n_i(r)w_i(r)$. The total minimum-distance density at r is, therefore,

$$n(r) = \sum_{i=1}^{N} n_i(r)w_i(r) \tag{4}$$

The normalization of the $g^{md}(r)$ function of Equation 1 depends on the computation of the density of minimum-distances in the absence of solute-solvent interactions. However, for each r, the volume element

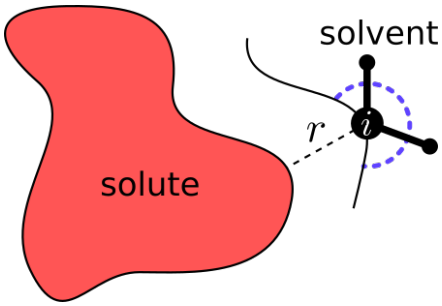

Figure 1: Fraction of revolutions of the solvent molecule (dashed blue line), with rotations centered at atom i that preserves the atom as being the closest to the non-interacting-solute.

associated with r is dependent on the structure of the solute. The intuitive picture we develop here is that of a "phantom"' solute molecule, immersed in a solvent with bulk properties, which will be useful to define the shape of the minimum-distance surfaces for every r.

If the density of the solvent was that of bulk and there were no solute-solvent interactions, the density of atoms of type i (or any other type) at every distance is simply $n_i^*(r) = \rho_s$, the solvent molecular density. However, the probability $w_i^*(r)$ that the atom is the closest atom to the solute is dependent on the shape of the surface defined by r and on the structure of the solvent molecule.

In general, the density of minimum-distances at r in bulk is, then,

$$n^*(r) = \rho_s \sum_{i=1}^{N} w_i^*(r) \tag{5}$$

Note that the minimum-distance density is in general larger than the bulk solvent density. In particular, if r is short enough, $n^*(r) = N\rho_s$, because $w_i^*(r) = 1$ for every i. In other words, the minimum-distance density at very short distances is the atomic density, not the molecular density of the solvent.

This result shows that the minimum-distance density differs from the bulk solvent density even in the absence of solute-solvent interactions. Each solvent atom contributes differently to the minimum-distance summation, according to the $w_i(r)$ parameter, which is the probability of it being the closest atom to the solute if it is at distance r. $w_i(r)$ is associated with the orientation of the solvent molecule relative to the solute, but also with the accessibility of each atom to the solvent, which might change if the solvent intramolecular structure is perturbed by the solute.

From Equations 4 and 5, it follows that the minimum distance distribution function can be written as

$$g^{md}(r) = \sum_{i=1}^{N} \frac{n_i(r)}{\rho_{bulk}} \left(\frac{w_i(r)}{\sum_{i=1}^{N} w_i^*(r)} \right) \tag{6}$$

which, as expected, is constant and equal to 1 in the absence of solute-solvent interactions, where $n_i(r) = \rho_{bulk}$ for every i (and $w_i(r) = w_i^*(r)$, by definition).

This definition provides a useful decomposition of $g^{md}(r)$ in terms of atomic contributions. $g^{md}(r)$ can be considered as the contributions of each atom of the solvent, with two weights: first, the density of each atom relative to bulk density at distance r. Second (in parenthesis in Equation 6), an orientational parameter associated with atom at r being the closest atom to the solute. An analysis of the distance dependence of the atomic contributions is available in Appendix A.1.

2.3 Computation of minimum-distance correlation functions

The process of obtaining the $g_{us}^{md}(r)$ (from Equation 1) and the corresponding KB integral (as defined in Equation 2) is illustrated in Figure 2. It relies on the construction of two histograms: 1) the histogram of $n_{us}(r)$, which contains the density of minimum-distance sites at a volume element in the vicinity of r in the actual solute-solvent simulation, and 2) the histogram of the normalization, $n_{us}^*(r)$, which must contain the density of minimum-distance sites in bulk solvent in the same volume. The computation of $n_{us}(r)$ is a straightforward site counting following directly from the simulation of the system of interest. The choice of the method to compute $n_{us}^*(r)$, however, deserves further justification.

In the absence of solute-solvent interactions (that is, in the presence of the "phantom" solute molecule), the density of any specific atom at r is the bulk density of the solvent, and thus the density of minimum-distances at r is given by Equation 5. It is dependent on the conditional probabilities $w_i(r)$, of each specific atom being the closest one to the solute if it is found at distance r. These probabilities are dependent on the solvent structure. For example, if the solvent contains a single atom $w_1 = 1$ and hence $n_{us}^*(r) = \rho_s$, which turns out to be equivalent to the normalization of standard radial distribution functions.

If the solvent has more than one atom, $n_{us}^*(r)$ must be obtained explicitly. The simplest conceptual alternative is to perform a simulation of the pure solvent with bulk concentration and compute the distribution by the insertion of a "phantom" solute. Nevertheless, we will argue that $n_{us}^*(r)$ can be obtained by a numerical procedure which avoids having to simulate the pure solvent.

First, we note that the probability, in bulk, of an atom at a distance r of the solute being the closest atom

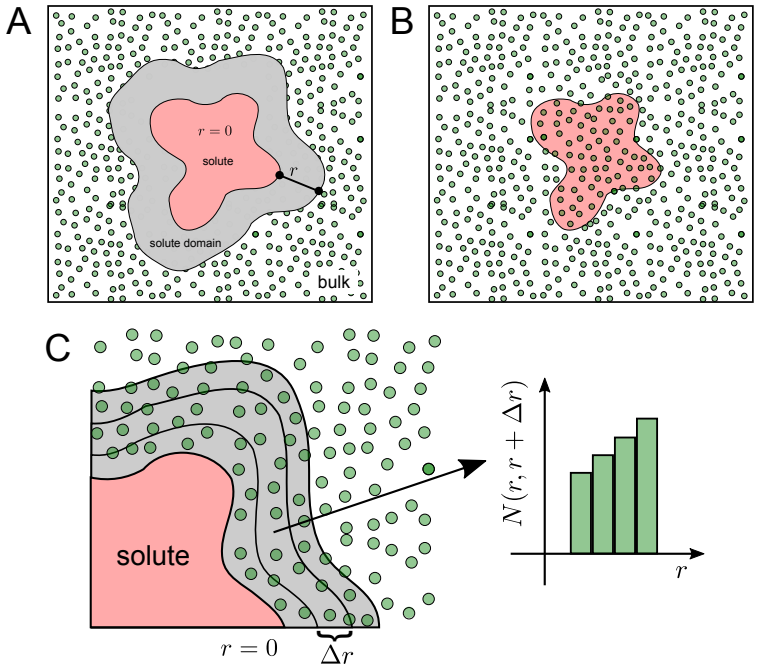

Figure 2: Illustration of the calculation of the minimum distance distributions. (A) A "solute domain" is defined as a region around the solute outside of which the effect of the solute on the solvent structure can be neglected for practical purposes. The volume of this region is determined by numerical integration. The bulk density of the solvent is then estimated from the number of solvent molecules outside the "solute domain" and the complementary volume of the box. (B) A random distribution of the solvent, with a density equal to the bulk density of the simulation, but occupying the whole box, is used to compute the minimum-distance density without solute-solvent interactions. (C) Histograms are computed for the minimum-distances counts obtained from the simulation and from the random solvent distribution, to compute the $g^{md}(r)$ distribution according to Equation 1.

depends only on the geometry of the solvent molecule and on the shape of the surface defined by r. It consists on the fraction of the revolutions of the molecule, with atom i as the rotation center, that preserves atom i as the closest atom to the solute, as illustrated in Figure 1.

If the solvent molecule is rigid, the probabilities $w_i^*(r)$ of Equation 5 can be computed by the numerical integration of the rotations of a solvent molecule at every distance. Equivalently, as discussed in Appendix A.2, they can be computed by simulating a set of non-interacting solvent molecules - with proper bulk density - in the presence of a "phantom solute", or simply by generating random positions and rotations for those molecules. These non-interacting solvent configurations could be used to compute the probabilities $w_i^*(r)$ to use Equation 5, or, more conveniently, to compute directly the minimum-distance count required for obtaining $n_{us}^*(r)$.

If the solvent molecules are not rigid, their conformational flexibility must be taken into account. The conformational flexibility in this case is that of the solvent molecule in bulk solvent. Therefore, it can be obtained by simulating the pure solvent, or by sampling solvent molecules from the bulk (large-distance) phase of the solute-solvent simulation. Given a properly-sampled set of solvent molecule conformations, the random solvent molecule distribution can be used to compute the histograms for $n_{us}^*(r)$.

Therefore, the algorithm we use here for computing $n_{us}^*(r)$ consists of:

1. Estimation of the bulk density of the solvent:
    1.1. Determine by numerical integration the volume of the "solute domain", $V_{solute-domain}$, defined by a distance to any atom of the solute considered large enough so that the solvent displays bulk properties outside this domain.
    1.2. Compute the volume of the bulk domain by $V_{box} - V_{solute\ domain}$, where V_{box} is the volume of the simulation box.
    1.3. Count the number of molecules of the solvent within the solute domain, to obtain by subtraction from the total number of molecules the number of molecules in the bulk domain.
    1.4. Using 1.2 and 1.3, estimate the bulk density of the solvent component.
2. Generation of a non-interacting solvent box:
    2.1. Choose a random molecule from the bulk region of the simulation.
    2.2. Generate a random position and orientation to the molecule within the simulation box, and compute the minimum distance to the solute, to add to the histogram of $n_{us}^*(r)$.

2.3. Repeat steps 2.1 and 2.2 until N^{rand} molecules with random positions were generated. N^{rand} must be large enough to allow a proper sampling, thus is it reasonable that it at least of the order of the number of solvent molecules in the simulation.

2.4. Scale the histogram of $n^*_{us}(r)$ by $\rho_s V_{box}/N^{rand}$, where ρ_s is the bulk density of the solvent and V_{box} is the total volume of the simulation box. That is, adjust the histogram to the proper bulk density.

With approximations of $n_{us}(r)$ and $n^*_{us}(r)$ determined with the procedures above, the correlation function $g^{md}_{us}(r)$ can be computed from Equation 1.

2.4 KB integrals and preferential interaction parameters

From the estimates of $n_{us}(r)$ and $n^*_{us}(r)$ obtained with the strategy summarized in the previous section, it is possible to compute the distance-dependent KBI for each of the solvent components independently using Equation 2. These KBIs can be compared with the corresponding $g^{md}_{us}(r)$ distributions for the molecular understanding of the interactions that build up the main characteristics of solvent accumulation or depletion from the surface of the solute. Additionally, the preferential interaction parameters in a ternary solution can be obtained from the KBIs using, for example, the following expression for the cosolvent,[4,11]

$$\Gamma_{uc}(R) \approx \rho_c[G_{uc}(R) - G_{uw}(R)], \tag{7}$$

where ρ_c is the bulk density of the cosolvent component of interest, and $G_{uc}(R)$ and $G_{uw}(R)$ are the KBIs for the cosolvent and water, respectively. Alternatively, the preferential interaction parameter can be computed directly from the solvent counts $n(r)$ for each component as described by Baynes and Trout,[19]

$$\Gamma_{uc} = N_c(R) - \frac{\rho_c}{\rho_w} N_w(R) \tag{8}$$

where $N_w(R)$ and $N_c(R)$ are the number of molecules of water and the cosolvent within R from the solute, and ρ_w and ρ_c are the corresponding bulk concentrations.

The distance R within which the number of molecules is integrated must be large enough such that the properties of the solution are those of bulk solvent, and $G_{us}(R)$ is converged, but minimizing the fluctuations of the $N_s(R)$ counts. Baynes and Trout estimated that the optimal R is the minimum distance for which the distributions are converged to one within the numerical errors.[19] The difference between our procedure and that of Baynes and Trout is that they use R as the distance between the center of mass of the solute and

the van der Waals surface of the protein, while we define it as the minimum-distance between any atom of the solute and any atom of the solvent. They find that 6Å is an adequate to obtain a qualitative estimate of Γ.[19] We use here $R = 8$Å, according to the observed convergence of the $g_{us}^{md}(r)$ distributions in our systems and the intention to compare the preferential solvation parameters quantitatively with the experimental data. The use of the minimum distance between the solute and the solvent provides a more precise definition of the distribution, and the molecular picture of the solvation shell arising from this distribution provides an intuitive representation of solute-solvent interactions.

3 Materials and Methods

Simulations of Ribonuclease T1 (RNaseT1) in solutions of TMAO and urea were performed as follows. The solution NMR structure of RNaseT1 (PDB id. 1YGW) was used.[37] It consists of 34 models, which were used as starting conformations for independent simulations. The models were solvated using the software Packmol[38,39] in pure water, in mixtures of water and urea, and in water and trimethylamine-N-oxide (TMAO), with added sodium and chloride ions for system neutrality. The TIP3P model was used for water,[40] CHARMM36[41] parameters were used to simulate the protein.

Urea was simulated either with the CHARMM General Force-Field (CGENFF)[42,43] or with the same force-field but substituting partial charges by those developed by Weerasinghe and Smith[44] to reproduce experimental KB integrals of solutions of urea in water (urea KBFF model). In this case we chose preserved other non-bonded parameters of the CGENFF model, particularly because the combination rules for atomic radii of the KBFF model are different. The results obtained supported this choice, and we will refer to this force-field as simply KBFF from now on. The details of the parameters used are reported in Supporting Information Table S2.

Three different force-fields were used to simulate TMAO molecules: The CGENFF model[42,43], and two models developed to reproduce experimental properties of aqueous TMAO solutions, which are improvements over the popular Kast model[45] of TMAO. The second model was developed to reproduce the osmotic pressures of TMAO solutions, and will be referred to as Osmotic model.[46] The third model was more recently developed by Scheck *et al.*[47] to reproduce experimental transfer free energies of peptides by varying the non-bonded parameters of the Kast model, and will be referred to as "Optimized Kast" model, or simply OptKast model.

Simulations were performed with the NAMD software,[48] with a 2 fs timestep, and were run at 298.15K and 1 atm, with 12 Å van der Waals interaction cutoff and Particle Mesh Ewald for evaluation of long range electrostatic potential. Constant temperature and pressure were maintained by a Langevin thermostat with

a damping coefficient of 5 ps^{-1} and Langevin Barostat with a piston period of 200 fs and a damping time scale of 100 fs. The systems' energies were minimized by 1000 steps of conjugated gradient method (CG) and equilibrated by 100 ps of constant-temperature and constant-pressure (NPT) MD with the protein fixed, followed by 100 CG steps and 100 ps NPT MD with the Cα atoms fixed, and finally 500 ps unrestrained NPT MD. Production simulations were then run for 10 ns. For each solvent concentration, 34 independent setups and simulations were performed starting with each of the NMR models, for a total of 340 ns of production MD. Simulations were performed for RNaseT1 in pure water, and in urea concentrations of 0.50, 1.00, and 2.00 mol$\times L^{-1}$, and for TMAO concentrations of 0.25, 0.50, and 1.00 mol$\times L^{-1}$ to reproduce the experimental conditions reported by Lin and Timasheff.[31] Initial volumes were obtained from experimental densities of the solvent mixtures.[49,50] A total of 5.1 μs of productive simulations were performed, using the high-performance computing environment of the Center of Computational Engineering & Science of the University of Campinas. The details of the systems simulated are reported in Supporting Information Table S1. Additional 400 ns simulations of systems composed only of one urea or TMAO molecule in 2000 water molecules were used for the examples discussed in Section 4. These simulations were performed with the same production protocol as described above.

The computation of the minimum distance distribution function was implemented as the `gmd` module of our MDAnalysis software suite, and is available at `http://leandro.iqm.unicamp.br/mdanalysis/gmd`.[51] A solvent molecule was considered to belong to the bulk domain if all of its atoms were farther from the protein than 8Å. The distribution functions and KBIs were observed to effectively converge at these distance. The volume associated with a minimum-distance of 8Å from the RNaseT1 surface contains a number of molecules equivalent to that of a sphere of about 25Å of pure water, such that the fluctuations expected are smaller than those of a conventional radial distribution function at much smaller distances. Independent distribution functions were computed for the simulations performed with each one of the 34 NMR models, from frames extracted from the simulations at every 10 ps. The fluctuations of the distribution functions within the simulations were negligible, and the fluctuations of the KBIs are reported as standard errors in Tables 1, S4, and S5, and in Figures 6, and S5. Finally, preferential interaction parameters were computed using Equation 7, with the standard error of the KB integrals being propagated to Γ.

4 Solvent-shell structure and Kirkwood-Buff integrals of irregular solutes: proof-of-concept example

To illustrate the effectiveness of minimum-distance distribution functions in representing the solvent-shell structure around irregularly-shaped solutes, we have performed simulations of single urea and TMAO molecules in pure water. We discuss here the results obtained for urea in water, with similar plots for TMAO available as Supporting Information (see Table S3 and Figure S2).

In Figure 3A, we illustrate the difference between the $g^{md}(r)$ and the standard radial distribution function, $g(r)$, in representing the water-urea interactions, for the KBFF urea model[44] solvated by water. The $g^{md}(r)$ displays a peak at $\sim$1.9Å, indicative of the formation of the expected hydrogen bonds. In this example, the $g(r)$ was computed between the water oxygen atom and the urea carbon atom, and displays a peak at $\sim$3.8Å. Obviously, this peak also results from water-urea hydrogen bonds, but the associated distance does not indicate that because of the references sites for the calculation of the $g(r)$. Naturally, this analysis could be, for a small molecule as urea, extended for the computation of the $g(r)$ between other pairs of sites and a more clear picture of the solvation structure could be obtained. However, the advantage of the $g^{md}(r)$ distribution is that it is properly defined for the representation of the solvation structure for solutes of any size and structural complexity.

In Figure 3B, we illustrate distance dependence of the KB-integrals computed from the integration of the radial distribution function, i. e.,

$$G_{us}(R) = 4\pi \int_0^R [g(r) - 1] r^2 dr, \tag{9}$$

or from the minimum distance count, using Equation 2. At large r, as expected, both integrals converge to the same value, validating the methods proposed here. The urea and water models in this case reproduce precisely the experimental apparent molar volume of urea.[52] The KB integral computed from the minimum distance count (blue curve) are determined by the excluded van der Waals volume of urea (which results to be of the order of -72 cm^3/mol (first dip), and the accumulation of water at the first solvation shell. The same integral computed from the $g(r)$ displays oscillations at larger distances, associated with the molecular volumes of urea and water, which turn out to perturb the integral up to distance of 8 to 10Å. The fact that the integration from the minimum-distances account for the molecular volumes at short distances is fundamental for the computation of KB integrals for larger solutes. Therefore, not only the convergence of the KB integral is favored by the use of the minimum-distance approach, but the molecular interpretation of the integral for variable r is clearer than that obtained from the standard radial distribution functions.

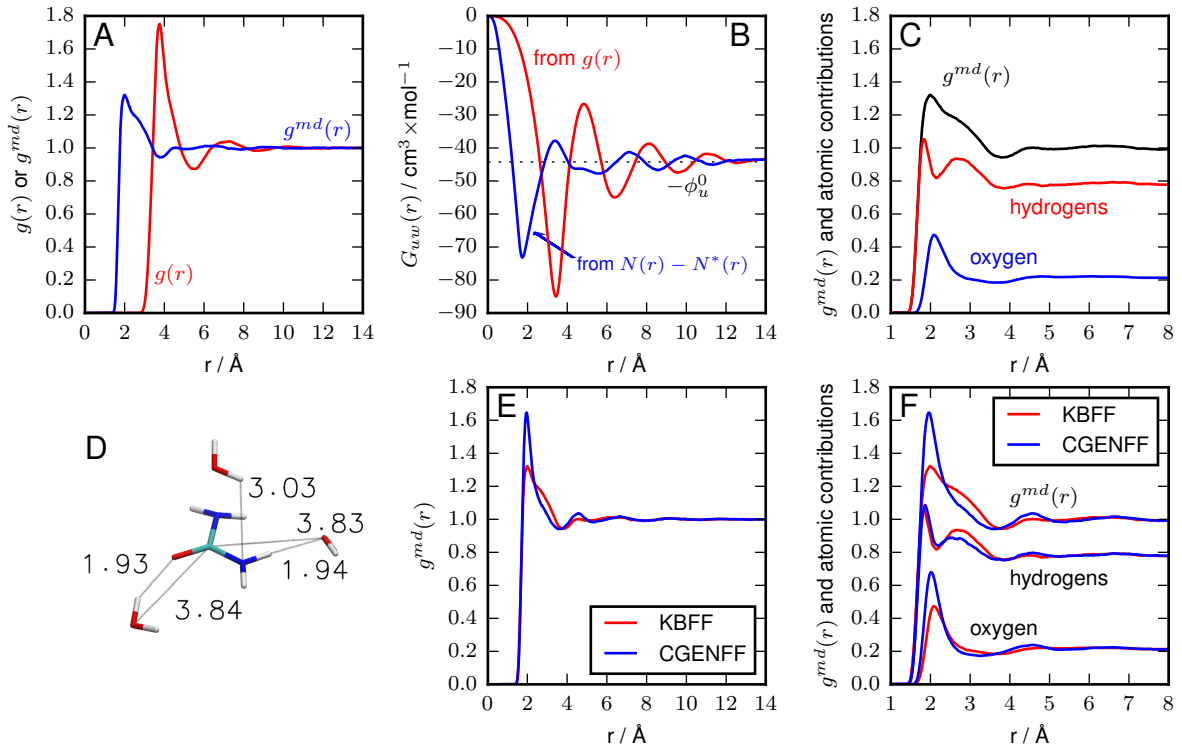

Figure 3: Comparison of the standard radial distribution function and the minimum-distance distribution function for the study of the solvation of a small irregular solute, as urea in water (using the KBFF urea model[44]). (A) Water-urea minimum distance distribution function compared to the radial distribution function (computed from water oxygen and urea carbon atoms). The distance of the peak of the $g(r)$ distribution, at ∼3.8Å, does not correspond to direct interactions, whereas the peak of the $g^{md}(r)$ distribution at ∼1.9Å clearly indicates the presence of hydrogen-bonds. (B) Distance-dependence of the KB-integrals computed by integrating the $g(r)$ (Eq. 9) or using the minimum distance count (Eq. 2). The integrals are equivalent for large r, but the use of the minimum distance count shows that the first solvation shell essentially determines the final $G_{uw}(r)$, and avoids fluctuations associated to the urea molecular volume. The urea model fits exactly the experimental apparent molar volume of urea (ϕ_u^0).[50] (C) The decomposition of the $g^{md}(r)$ distribution into atomic contributions shows that three types of interactions are important for the hydration shell of urea: with water donating or accepting hydrogen bonds, and secondary interactions with hydrogen atoms at ∼3.0Å. (D) Illustration of the interactions determinant for the distribution functions: the interactions of water hydrogen atoms with urea nitrogen atoms (at 3.03Å in the figure) can be associated to the shoulder of the $g^{md}(r)$ distribution. (E) and (F) Comparison of two urea force-fields. The KBFF model displays less hydrogen-bond capacity than the CGENFF model, and the difference can associated to the reduced bonding of urea to water oxygen atoms (panel F), which results from the reduced partial charges of urea hydrogen atoms (Supporting Information Table S2).

The $g^{md}(r)$ distribution can, additionally, be decomposed into atomic contributions, as we discussed in Section 2.2. This decomposition is shown in Figure 3C. The peak at $\sim$1.9Å results from the contributions of hydrogen bonds to water oxygen and hydrogen atoms. Additionally, there is a second peak on the hydrogen distribution which explains the shoulder of the complete $g^{md}(r)$ distribution. All these peaks, being associated to minimum-distances between water and urea, can clearly be linked to intermolecular configurations, as shown in Figure 3D. The hydrogen bonding are associated to urea-oxygen and urea-hydrogen bonding to water, and are expected. Additionally, the second peak of the hydrogen distribution can be clearly traced down to the interaction of the water hydrogen atoms with the urea nitrogen atoms, which is perpendicular to the urea molecular plane. The specificity of these interactions is hardly obtainable from the standard radial distribution functions, for which the peak at $\sim$3.8Å incorporates all types of contributions.

In Figures 3E and F we show that this precise molecular picture of solvation is useful for the interpretation, for example, of the differences between molecular models used. The $g^{md}(r)$ and atomic contributions computed from simulations using the KBFF model or the CGENFF model show that the KBFF leads to a lesser propensity of urea to form hydrogen bonds with water, and that this is particularly associated to hydrogen bonds with water oxygen atom. This is probably a direct result of the smallest partial charges of hydrogen atoms in the KBFF model relative to the CGENFF model. These smaller charges also favor the interactions of the water hydrogen atoms with the urea nitrogen atoms, increasing the second peak of the hydrogen distribution.

Finally, at long distances, the atomic contributions of water converge to ~ 0.21 for the oxygen atom and to ~ 0.79 for sum of the hydrogen contributions (Figures 3C and 3F). This is expected, and is characteristic of the shape of the water molecule and of the minimum-distance count at distances where the interactions with the solute are negligible. A theoretical analysis of these contributions is provided in Appendix A.1.

5 Reconciling preferential interaction parameters and the solvent-shell picture of protein solvation

Our discussion will focus in this section on the simulations performed with the KBFF[44] for urea, and with the Osmotic[46] model for TMAO, because the experimental preferential interaction parameters were better reproduced with these models, as it will be shown. Similar data for the other models simulated is provided as Supporting Information (Figures S3, S4, and S5, and Tables S4, and S5).

Figure 4 displays the minimum distance distribution functions [$g_{us}^{md}(r)$] computed for all systems simulated. As expected, $g_{us}^{md}(r)$ distributions provide a very clear picture of direct solute-solvent interactions. For example,

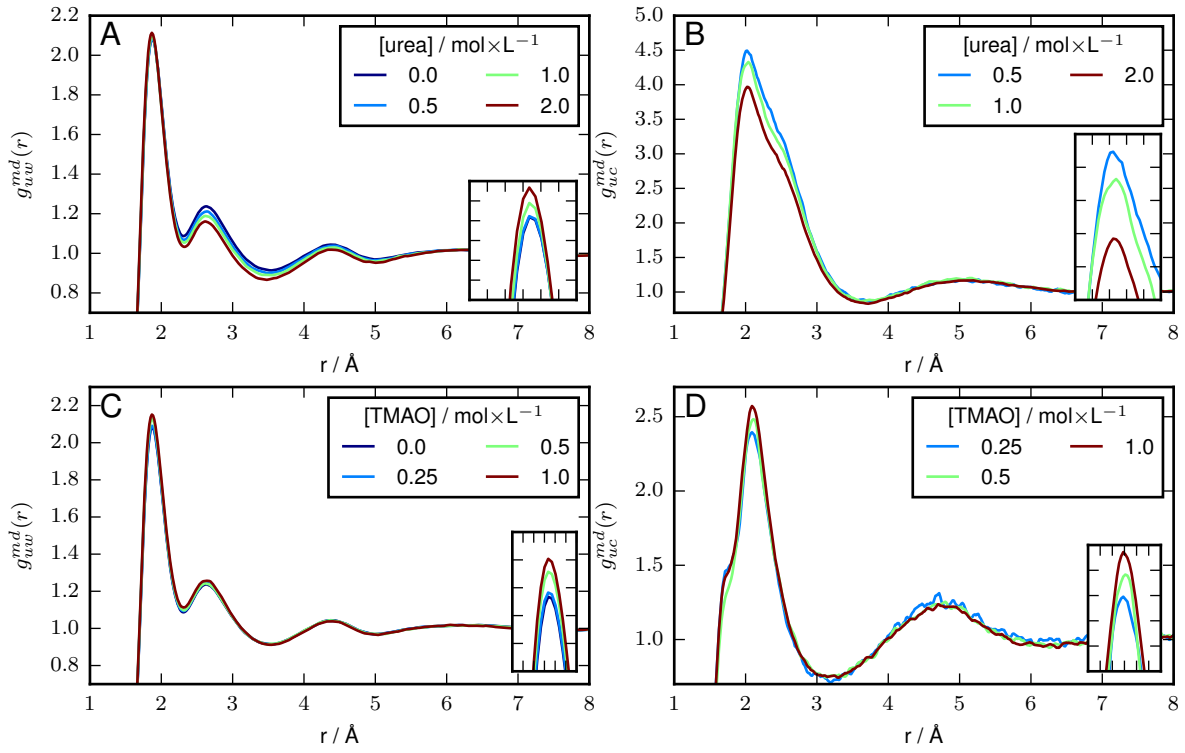

Figure 4: Minimum-distance distribution functions of (A) water in urea solutions, (B) urea, (C) water in TMAO solutions, (D) TMAO, relative to RNaseT1, for different cosolvent concentrations. The insets are augmented representations of the highest peak. These plots correspond to the use of the KBFF [44] model for urea and the Osmotic [46] model for TMAO. Similar plots for other solute models are provided as supplementary information.

Figure 4A displays the distribution of water around the protein. A distinct peak at ∼1.8Å indicates the formation of hydrogen bonds. A second peak at ∼2.7Å is associated to the second hydration shell. Water molecules therefore clearly form specific direct interactions with the protein, and display increased densities relative to bulk at short distances. The addition of urea decreases the relative water density at the surface of the protein, indicating that urea competes with water with apparent greater affinity (although there is a slight increase in the hydrogen-bonding peak of water, as shown in the inset of Figure 4A). Figure 4B shows the $g_{uc}^{md}(r)$ distribution of urea relative to the protein. The density of urea is 4 to 4.5 times greater at hydrogen bonding distances than in bulk. Therefore, urea forms a strongly-favorable interaction with the surface of protein. With increasing urea concentration, its local density augmentation decreases, indicating that the most urea-affine sites are occupied.

Interestingly, as shown in Figures 4C and D, the distributions of water and TMAO at the vicinity of the protein display similarities to those of urea: both water and TMAO display augmented relative densities at

short distances. The $g_{uw}^{md}(r)$ of water perturbed by TMAO at hydrogen-bonding distances as it was by urea. The density of water at larger distances appears to be mostly unaffected by TMAO. TMAO also accumulates on the protein surface, but the most important interactions occur at $\sim$2.3Å. A small shoulder at hydrogen-bonding distances is also visible in Figure 4D. Contrary to what is observed for urea, however, the relative density of TMAO increases at the highest peak with increasing TMAO concentration. This means that there might be some cooperative effect associated to TMAO binding to the protein surface.

Therefore, $g_{us}^{md}(r)$ correlation functions indicate that all molecules form specific and favorable interaction with the RNaseT1, leading to the augmentation of their densities in the proximity of the protein surface. This picture may, at first glance, seem at odds with the common interpretation that TMAO molecules are excluded from the protein surface. As we will see, in spite of this molecular picture, there is no contradiction when the preferential interaction parameters are computed, at least for dilute solutions.

As discussed in section 2.2, the $g_{us}^{md}(r)$ can be naturally decomposed into atomic contributions, allowing for a more detailed structural interpretation of the solvent interactions with the protein surface. Figures 5A and B show the contribution of each type of urea atom to the urea $g_{us}^{md}(r)$ distribution. Urea hydrogen and oxygen atoms account for the interactions at short distances, as expected. Minimum-distance interactions through the carbon or nitrogen atoms is infrequent, and occur only eventually at the second solvation shell.

The decomposition of the $g_{uc}^{md}(r)$ of TMAO into atomic contributions, shown in Figure 5C and D, is more interesting, because TMAO is an amphiphilic molecule. The small shoulder at hydrogen bonding distances is almost completely determined by the oxygen atom. At the same time, the significant density augmentation at $\sim$2.3Å results solely from methyl-hydrogen contributions, thus being hydrophobic in nature. These hydrophobic interactions are slightly strengthened with the increase of TMAO concentration. The stabilization of hydrophobic interactions is possibly the underlying cause of the apparent cooperative binding observed in the full $g_{uc}^{md}(r)$ distributions. At the same time, no particular trend can be discerned on the oxygen minimum distance distribution with varying TMAO concentration.

Figure 6 and Table 1 display the KB integrals computed for water, urea, and TMAO, in all systems. The KB integrals for water are around -8 L$\times$mol^{-1} in both cases (Figures 6A and C). Therefore, water is overall excluded from the protein domain. This exclusion results directly from the excluded protein volume, and it is only slightly compensated by the augmented density of water at the surface of the protein. The addition of urea decreases the water KBIs, thus inducing further water exclusion, while the addition of TMAO slightly increases the water KBIs, indicating water accumulation on the protein domain. Both effects are small relatively to the overall KBIs (Table 1). The value of the water-protein KB integral in pure water, -7,756 L$\times$mol^{-1}, agrees within 2.2% with the experimental apparent molar volume of RnaseT1[31] of 7.924 L$\times$mol^{-1}. This indicates

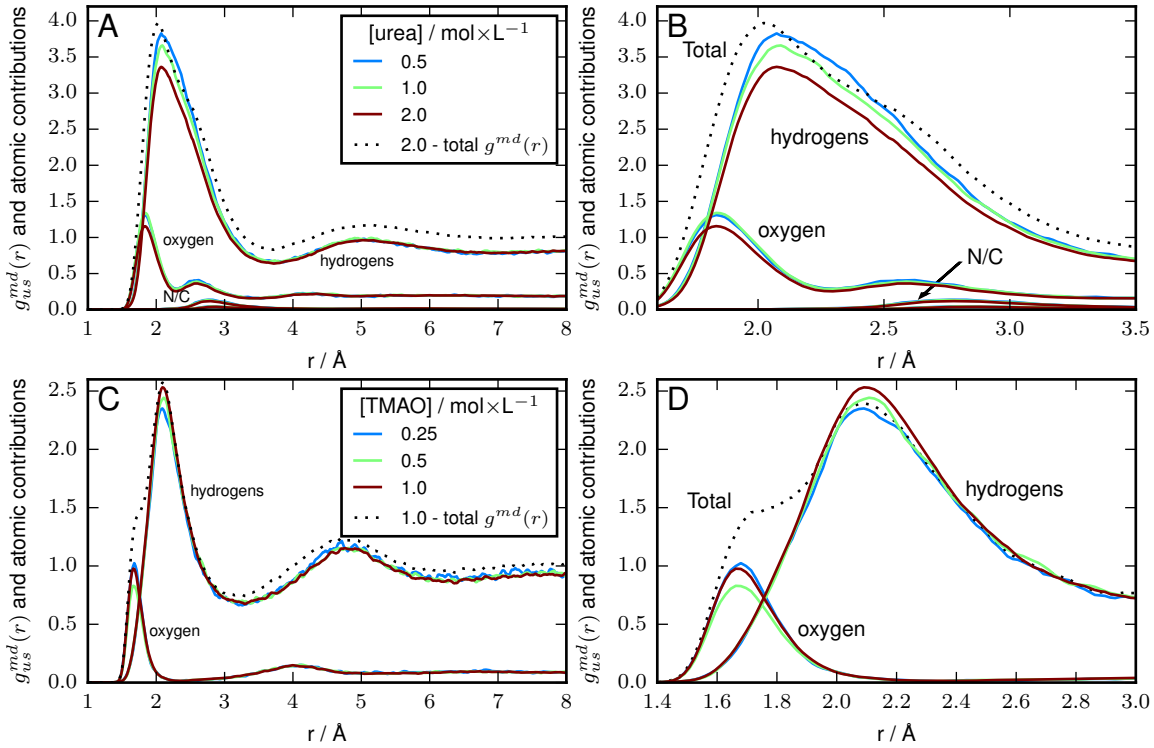

Figure 5: Decomposition of the $g^{md}_{us}(r)$ into atomic contributions. Plots (B) and (D) are insets of (A) and (C) focused on the first and second solvation layers. (A) and (B) Urea: the peak at ∼1.8Å is determined by hydrogen bonds of the protein with urea hydrogen or oxygen atoms. Urea only rarely interacts directly with the protein through its nitrogen or carbon atoms. The most important protein-urea interactions are associated with hydrogen bonds through the hydrogen atoms. (C) and (D) TMAO: hydrogen bonds of the protein with TMAO oxygen atom are associated with the shoulder on the $g^{md}(r)$ distribution at ∼1.7Å. The most prominent peak of the distribution is, however, determined by hydrophobic interactions of the protein with TMAO hydrogen atoms, which peak at ∼2.1Å. The stability of hydrophobic interactions through the hydrogen atoms increases with TMAO concentration. The contributions of TMAO nitrogen and carbon atoms for the $g^{md}(r)$ distributions are negligible.

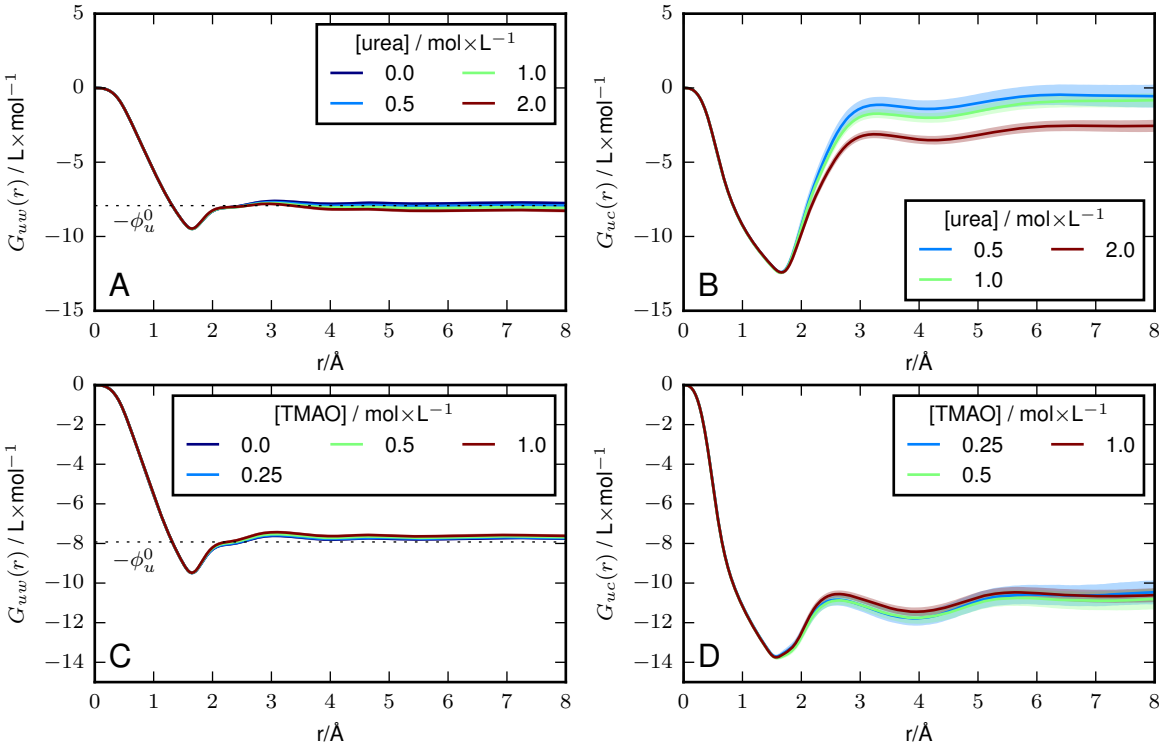

Figure 6: Kirkwood-Buff integrals of (A) water in urea solutions, (B) urea, (C) water in TMAO solutions, (D) TMAO, relative to RNaseT1, for different cosolvent concentrations. The simulations using the KBFF[44] and Osmotic[46] force-fields for urea and TMAO were used for these plots. The experimental apparent molar volume (ϕ_u^0) of RNaseT1[31] in water is indicated by dashed lines in (A) and (C), and agrees within 2.2% with the KB integral in pure water, indicating that the water and protein force-fields are adequate to study water-protein interactions. Similar plots using other force-fields and numerical values of the integrals are available as supplementary information.

that that the water-protein interactions are nicely modeled by with the current force-fields, and provides additional support for the methods proposed here.

Urea KBIs are numerically small, as shown in Figure 6B. This indicates that urea accumulation on the protein surface almost completely compensates the excluded protein volume (which contributes negatively to the KBI). The urea KBIs decrease with increasing urea concentration, indicating, as the distribution functions suggested, saturation of the urea interaction sites on the protein surface.

The TMAO KBIs are around -11 L×mol^{-1} thus smaller than those of water (Figure 6D). Therefore, TMAO is overall excluded from the protein domain. The TMAO KB integrals appear to become slightly more negative with increasing TMAO concentration (Table 1 - last column), but within numerical errors. The increase in the local density at hydrophobic interaction distances (∼2.1Å - Figure 4D) is reflected in the $G_{uc}(r)$

Table 1: Kirwood-Buff integrals computed from the simulations in L×mol^{-1} using the KBFF urea model[44] and the Osmotic TMAO model[46]. The deviations reported are the standard error of the means of the 34 simulations performed for each system. The experimental apparent molar volume of RnaseT1 in water[31] is ϕ_u^0=7.924 L×mol^{-1}, and differs by 2.2% from the predicted value from the simulations (7.756 L×mol^{-1}). Similar results using the other force-fields are provided in Supporting Information Tables S3 and S4.

	Urea		Trimethyl-N-Oxide	
C	G_{uw}	G_{uc}	G_{uw}	G_{uc}
0	-7.756±0.003	-	-7.756±0.003	-
0.25	-	-	-7.731±0.008	-10.46±0.53
0.50	-7.919±0.017	-0.55±0.68	-7.683±0.010	-10.81±0.45
1.00	-8.081±0.016	-0.83±0.38	-7.618±0.013	-10.61±0.28
2.00	-8.268±0.029	-2.55±0.31	-	-

integrals for up to about 4-5Å. The first solvation shell of TMAO roughly determines the final KBIs, but not to the precision required to distinguish a clear concentration dependent trend, as with urea.

The profiles of the KBIs shown in Figure 6 reveal how the accumulation or exclusion of the cosolvents are distance dependent. The negative character of most KB integrals is due to the rapid drop of $G_{us}(r)$ at short distances, and results directly from the excluded molecular volumes. In all cases, at roughly hydrogen bonding distances (∼1.8Å) there is reversal of the $G_{us}(r)$ drop, due to the accumulation of the solvents at the protein surface, as indicated in the $g_{us}^{md}(r)$ distributions. Only for urea this accumulation is enough to completely counteract the effect of the protein excluded volume. At about ∼6Å, all KB integrals are essentially converged. Therefore, the final KB integrals are determined mostly by the excluded volume of the protein and the accumulation of the solvents in the first solvation shell, characterized by protein-solvent minimum distances between ∼1.8Å and ∼3Å.

Finally, from the KB integrals, it is possible to compute the preferential interaction parameters to be compared with experimental data. Figure 7 shows the preferential interaction parameters computed from the simulations as a function of each cosolvent concentration, compared with the experimental parameters obtained by Timashef.[31] The results obtained for all solvent models are shown.

The simulations reproduce the qualitative dependence of the preferential interaction parameters for the osmolytes reasonably well. Preferential interaction parameters, computed using Equation 7, consist of the increase or decrease in the number of solvent molecules (water or osmolyte) on changing the concentration of the protein.[4] Therefore, for example, a negative value of Γ indicates that the increase in the concentration of the protein excludes solvent molecules.

Preferential interaction parameters for water, in urea or TMAO solutions, are shown in Figures 7A and B. Water preferential interaction parameters are deeply negative (Figure 7A), indicating there is preferential

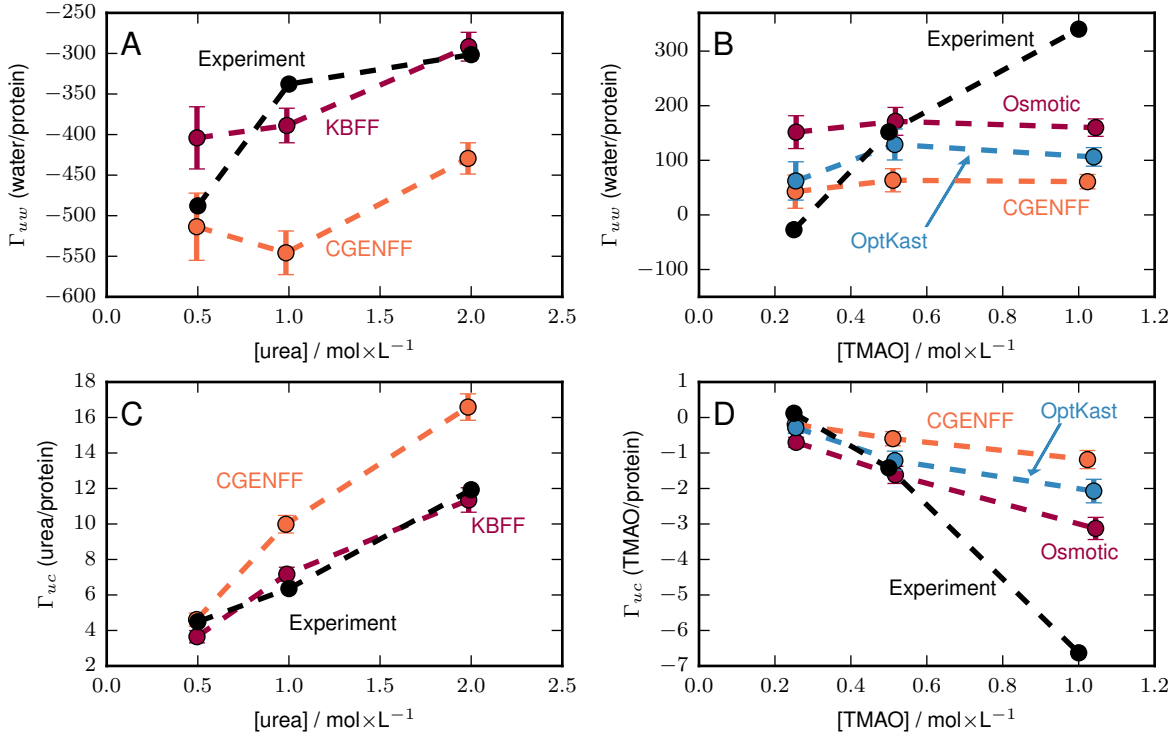

Figure 7: Preferential interaction parameters for (A) water in urea solutions, (B) water in TMAO solutions, (C) urea, (D) TMAO relative to RNaseT1 as a function of cosolvent concentrations. Experimental values were obtained from Lin and Timasheff,[31] with experimental errors not being significant for the purposes of this comparison. The error of the simulated data was computed by extrapolation of the standard error of the KBIs.

dehydration induced by urea. By contrast, water preferential interactions are positive in TMAO solutions. Therefore, the protein is preferentially dehydrated in urea solutions, and preferentially hydrated in TMAO solutions. However, experimentally, the preferential interactions with water increase with the increase in urea and TMAO concentrations, and the trends are not properly reproduced by the models, except perhaps for the KBFF urea force-field.

Figure 7C displays the preferential interaction parameters for urea. It is positive at all concentrations, indicating that urea is accumulated on the protein domain. The simulation results agree almost quantitatively with the experimental data in this case, specially when using the KBFF urea model.[44] With increasing urea concentration, the accumulation of the denaturant is increased. That is, a greater number of urea molecules accumulates on the protein domain. This occurs in spite of the fact that the free-energy of urea binding to the surface of the protein decreases, as indicated by the decrease of the $g_{uc}^{md}(r)$ peaks in Figure 4B at short distances. Therefore, the simulations and the experiments predict urea accumulation on the protein domain, associated with water exclusion.

As expected, the preferential interaction of TMAO displays the opposite trend, being negative at all concentrations (Figure 7D), and decreasing with the increase in TMAO concentration. This means that TMAO molecules are excluded from the protein domain, with qualitative agreement with the experimental result.[31] Therefore, more TMAO molecules are excluded from the protein domain with increasing TMAO concentration. Again, this occurs in spite of the fact that the free-energy of TMAO to the protein surface now increases with TMAO concentration, as indicated by the augmentation of the $g_{uc}^{md}(r)$ function at short distances (Figure 4D). The agreement of computed preferential interaction parameters with with the experimental values is quantitatively reasonable for the more dilute solutions. For the 1.0 mol$\times$L^{-1} solutions the exclusion of TMAO from the protein domain is underestimated by all models. Thus, the distribution functions of TMAO can be trusted for the two most dilute solutions, but additional improvements of the force-fields appear to be necessary to study protein-TMAO interactions at higher concentrations.

It is important to remark that the preferential accumulation (for urea) or preferential exclusion (for TMAO) are both associated with augmented local densities of the osmolytes at the surface of the protein, as indicated by minimum distance distribution functions (Figure 4). The difference between urea and TMAO is not qualitative in this sense, but associated with the greater urea accumulation, which is enough to completely counteract the effect of the excluded protein volume. The TMAO molecules also accumulate on the protein surface, mostly through hydrophobic interactions.

This picture of the interactions of each cosolvent, either an excluder or a crowder, with the protein surface, provides an additional level of understanding of protein-osmolyte interactions. Generally, protein stabilization

is associated with preferential hydration, and this indicates the exclusion exclusion of the osmolyte from the protein surface, as has been identified previously.[31] This picture, although being correct from the point of view of the preferential solvation parameters, might lead to a misleading interpretation of the microscopic interactions. Actually, both TMAO and urea (and it is likely to be the case for other osmolytes) do accumulate on the protein surface, and interact therewith, possibly playing additional roles in the protein stability and activity. This microscopic picture is not contradictory with the observed preferential interactions at low concentrations, which are determined by the relative weights of the excluded solute volume and solvent accumulation on the overall protein-solvent interactions. The use of minimum-distance distribution functions to visualize the solvent densities around the solutes provides intuitive pictures of these interactions, with easy atomic decomposition for structural interpretation.

6 Conclusions

The study of solvation of complex solutes, biomolecules in particular, is fundamental for the understanding of their stability in biologically or industrially important media. Distribution functions provide the means to connect the microscopic picture of solvation with experimental data, particularly preferential solvation, which is associated with protein stability. However, radial distribution functions are not appropriate for the representation of the solvation structure around complex solutes, and neither for the computation of the KBIs, which connect the molecular structure to thermodynamic data. The experimental observation of the preferential exclusion or accumulation of cosolvents often lead to the proposition of thermodynamic models of protein stabilization based on the hypothesis that some molecules avoid the interaction with the protein surface.[53–55]

In this work we have shown that using minimum-distance distribution functions both a molecular picture and the experimentally determined interaction parameters can be obtained from molecular dynamics simulations. The minimum-distance distribution functions are clearly associated with the solvation shell perspective which is often used intuitively to describe cosolvent interactions in terms of accumulation or exclusion.[55] However, while we confirm that urea, a denaturing osmolyte, accumulates preferentially on the protein surface and TMAO, a stabilizer, is preferentially excluded, both interact with the protein surface displaying density augmentation at short distances from the protein. The most determinant and universal factor associated with the exclusion of the solvent molecules from the protein domain is the protein molecular volume. The augmented densities of the solvent molecules at short distances may or may not be enough to lead to overall solvent accumulation on the protein domain, as measured by the Kirkwood-Buff integrals.

We conclude, therefore, that the observation of augmented densities of solvent components around proteins at a molecular level can occur even if the species is net-preferentially excluded from the protein domain, thereby acting as a stabilizer of the protein structure. This picture, which can be obtained by the use of minimum-distance distribution functions, provides an additional level of understanding of protein-solvent interactions and solvent induced stabilization. Finally, the distribution functions proposed here can be used for the study of solvation of solutes of any structural complexity without modification.

7 Associated Content

Supporting Information The Supporting Information is available free of charge on the ACS Publications website at DOI: XXXXX

Contains: Compositions of the solvents simulated (Table S1); details of the force-fields used (Tables S2 and S3); comparison of radial distribution and minimum distribution functions for urea and TMAO with different force-fields (Figures S1 and S2); protein-solvent minimum distance distribution functions for different force-fields and their atomic decompositions (Figures S3 and S4); KB integrals computed from minimum-distance distribution functions for all systems (Tables S4 and S5, and Figure S5).

8 Author Information

Corresponding author: *E-mail: leandro@iqm.unicamp.br

Notes

The authors declare no competing financial interest.

9 Acknowledgements

The authors thank FAPESP (Grants: 2010/16947-9, 2013/05475-7, 2013/08293-7) and CNPq (Grant: 470374/2013-6) for financial support.

A Appendix

A.1 Distance dependence of atomic contributions to $g^{md}(r)$

It is interesting to analyze the atomic contributions to the minimum-distance distribution function at very short and very long distances. Here we follow the notation used in Section 2.2.

Let us examine specific examples. For very short distances, $w_i(r) = 1$ for every i, thus

$$g^{md}(r \text{ short}) = \frac{1}{N} \sum_{i=1}^{N} \frac{n_i(r)}{\rho_{bulk}}$$

In the presence of solute-solvent interactions $n_i(r) = 0$ for every i at very short distances because of atomic volumes, thus $g^{md}(r) = 0$.

Far from the solute, $n_i(r) = \rho_s$ for all i, thus the $g^{md}(r)$ reduces to the sum of the orientational contribution of each atom, which sum up to one,

$$g^{md}(r \text{ large}) = \sum_{i=1}^{N} \frac{w_i(r)}{\sum_{i=1}^{N} w_i^*(r)} = \sum_{i=1}^{N} \frac{w_i^*(r)}{\sum_{i=1}^{N} w_i^*(r)} = 1.$$

The orientational contribution of each atom to this sum is dependent on the shape of the minimum-distance surface and on the solvent molecule structure. Let us analyze the properties of the orientational/conformational parameter w in more detail, at long distances. If the distance between the solute and the solvent molecule is large enough, the curvature of the minimum distance surface will be small relatively to the size of solvent molecule. Let us assume that this surface is flat.

The simplest case with some interest is that of a diatomic molecule, illustrated in Figure 8A. The parameter $w_i(r)$ will be, for each atom, the fraction of revolutions for which the other atom is farther from the solute. According to angles defined in Figure 8A, this fraction is

$$w(r) = \frac{\int_0^{\pi} d\theta \int_{-\pi/2}^{\pi/2} d\phi}{\int_0^{\pi} d\theta \int_0^{2\pi} d\phi} = \frac{1}{2}$$

A more interesting example is that of a triatomic molecule, as water. It is illustrated in Figure 8B. In this case, the reorientational parameter is dependent on the angle between the two covalent bonds. For the central

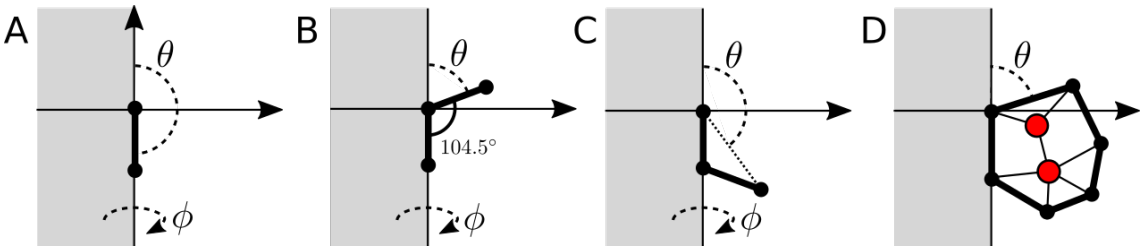

Figure 8: Orientational parameter at large distances: A) Diatomic molecule. B) Water molecule with oxygen as closest atom. C) Water molecule with hydrogen as closest atom. D) General molecule as a convex polyhedron.

(oxygen) atom, it is

$$w_O(r) = \frac{\int_0^{(180-104.5)^\circ} d\theta \int_{-\pi/2}^{\pi/2} d\phi}{\int_0^{\pi} d\theta \int_0^{2\pi} d\phi} \approx 0.210$$

and for the hydrogen atoms, illustrated in the second panel of Figure 8B,

$$w_H(r) = \frac{\int_0^{142.25^\circ} d\theta \int_{-\pi/2}^{\pi/2} d\phi}{\int_0^{\pi} d\theta \int_0^{2\pi} d\phi} \approx 0.395$$

These predictions can be numerically verified in the atomic decomposition water minimum-distance distribution functions around urea and TMAO, which are shown in Figures 3, S1, and S2.

At large r, $\sum_i^N w_i = 1$, because the molecule is for this computation a convex polyhedron, and the component of each atom consists on a fraction of the rotation of the polyhedron associated to each vertex. Many atoms might have zero contributions for being in the interior of the polyhedron, as illustrated in Figure 8D. The convergence of the $g^{md}(r)$ distribution to one at long distances corresponds to the convergence of this sum, as the atomic densities all converge to the bulk density. For the same reason, at long distances the $w(r)$ converge to the atomic contributions to $g^{md}(r)$.

A.2 Numerical integration schemes for the normalization of the minimum-distance distribution function

Here we will demonstrate that the normalization of the distribution functions by a random distribution of solvent molecules, considering the appropriate density and intramolecular flexibility, is equivalent to the normalization by solvent simulation in the presence of a non-interacting solute at infinite dilution.

First, let us consider distribution functions computed considering a single reference site at the solvent molecules (as conventional RDFs), usually the center of mass of the solvent molecule or one of its atoms. Let

ρ_s be the bulk density of the solvent. Without the solute, the ensemble average density of the solvent is also ρ_s everywhere. This is the standard normalization for distribution functions,

$$g(r) = \frac{\rho(r)}{\rho_s}$$

where $\rho(r)$ is the ensemble average density of solvent molecules at position r. The normalization by ρ_s can be considered equivalently as: 1) a constant density equal to that of bulk; 2) the average density of an ensemble of pure-solvent configurations around a non-interacting solute; 3) the average density of random solvent configurations with average bulk density (which is equivalent to an ideal-gas distribution of the solvent).

Now, we consider the distribution function computed from the minimum distance between any atom of the solvent and any atom of the solute. Initially, we will consider the case of a rigid solvent molecule. Let us suppose that a simulation of the pure solvent was performed. The average density of any specific atom at any position in space is ρ_s. A non-interacting, "phantom", solute is added to the solution, with no effect on the structure of the solvent. Therefore, the atomic densities remain unchanged. However, to compute the minimum-distance distribution, we need to know which is the probability that each atom, if found at a given position in space, is the closest atom to the "phantom" solute, in order to compute the normalization of the minimum-distance distribution using Equation 5.

At this point, we note that all configurations of the solvent which differ only by rotation around the atom of interest are thermodynamically equivalent, because there are no solute-solvent interactions. We illustrate these configurations which differ by rotation in Figure 9. Therefore, the probability that the reference atom is the closest atom to the solute is the same as the fraction of these rotations that maintain the reference atom as the closest atom. This is dependent, exclusively, on the shape of the solute and on the geometry of the individual solvent molecule. In particular, it is not dependent on the structure of the solvent.

It follows, then, that the minimum-distance density at each distance from the non-interacting solute can be computed by purely geometrical parameters of the solute and an individual solvent molecule, as described for long distances in the previous section. Explicitly, it can be computed by generating at each position r a solvent molecule centered at one of its atoms, and computing by numerical integration the fraction of configurations that preserve the reference atom as the closet atom to the solute ($w^*(r)$ in Equation 5). The procedure is repeated for each atom of the solvent molecule at the same position, and the atomic contributions $w^*(r)$ are summed up to compute the $n^*(r)$ minimum-distance density of Equation 5.

The numerical integration described above is equivalent to a Monte-Carlo integration by generation of many random positions and rotations for solvent molecules around a non-interacting solute. If the number of

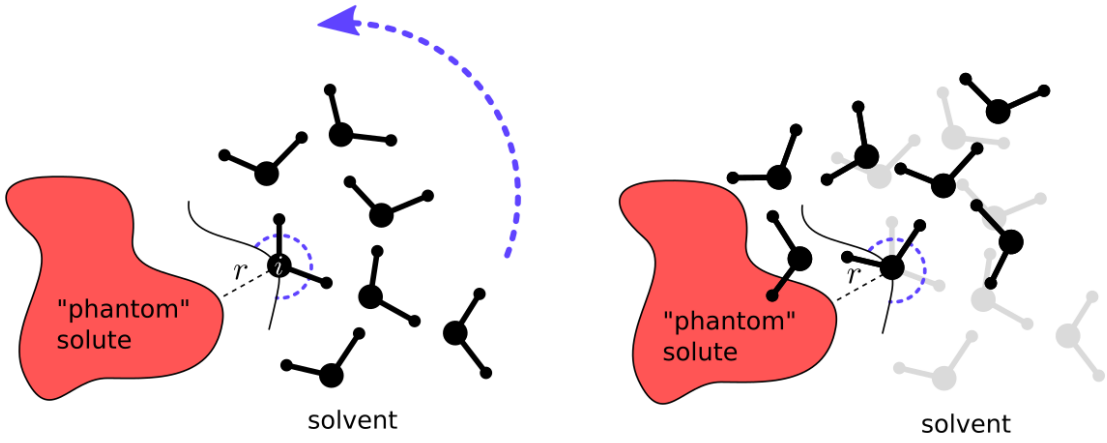

Figure 9: In the presence of a non-interacting solute, all configurations of the solvent associated to rotations around a specific atom i are equivalent. The probability of a solvent atom being the closest atom to the solute is then solely dependent on the shape of the solute and on the geometry of the solvent molecule.

random molecules is large enough, this procedure can be used to approximate the fraction of configurations at each position in space that display a specific atom closer to the solute. However, the explicit computation of the probabilities $w^*(r)$ are not necessary in this case, since the $n^*(r)$ minimum-distance counting can be explicitly computed, with the only condition that the final density of randomly generated molecules is adjusted to the desired solvent density, as we describe in the procedure of Section 2.3.

The simulation of the pure solvent and the generation of a random solvent configuration for the computation of $n^*(r)$ can differ, however, if the solvent molecules are not rigid. In this case, the interactions within solvent molecules might alter their geometry, thus directly affecting the probability of each atom being the closest atom at each r. In this case, the ensemble of the solvent molecules must reflect the conformational variability of the solvent molecules in bulk solvent. It is not possible to obtain that ensemble without simulating bulk solvent. On the other side, if a bulk solvent simulation is available, this problem is solved by randomly picking solvent molecules from this simulation to numerically compute $n^*(r)$ by the numerical procedure described. Here, the bulk solvent configurations are obtained directly from the solvent molecules at large distances from the solute.

Finally, let us remark that the solute conformations that are considered here are those of the solute in the presence of the solvent. The normalization is thought to evaluate how these conformations affect the solvent structure and local density. Thus, the minimum-distance distribution around the solute in the presence of solute-solvent interactions, $n(r)$, or without them, $n^*(r)$, are computed from each solute conformation

obtained in the solute-solvent simulations independently, the ensemble averages being the final reported result.

References

[1] Zhou, H.-X.; Rivas, G.; Minton, A. P. Macromolecular Crowding and Confinement: Biochemical, Biophysical, and Potential Physiological Consequences. *Annu. Rev. Biophys.* **2008**, *37*, 375–397.

[2] Holthauzen, L. M. F.; Auton, M.; Sinev, M.; Rösgen, J. Protein Stability in the Presence of Cosolutes. *Methods Enzymol.* **2011**, *492*, 61–125.

[3] Canchi, D. R.; García, A. E. Cosolvent Effects on Protein Stability. *Annu. Rev. Phys. Chem.* **2013**, *64*, 273–293.

[4] Pierce, V.; Kang, M.; Aburi, M.; Weerasinghe, S.; Smith, P. E. Recent Applications of Kirkwood-Buff Theory to Biological Systems. *Cell Biochem. Biophys.* **2008**, *50* (1), 1–22.

[5] Shimizu, S.; Stenner, R.; Matubayasi, N. Gastrophysics: Statistical Thermodynamics of Biomolecular Denaturation and Gelation from the Kirkwood-Buff Theory towards the Understanding of Tofu. *Food Hydrocoll.* **2017**, *62*, 128–139.

[6] Timasheff, S. N. Control of Protein Stability and Reactions by Weakly Interacting Cosolvents: The Simplicity of the Complicated. *Adv. Protein Chem.* **1998**, *51*, 355–432.

[7] Chalikian, T. V. Volumetric Properties of Proteins. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 207–235.

[8] Shimizu, S. Estimating Hydration Changes upon Biomolecular Reactions from Osmotic Stress, High Pressure, and Preferential Hydration Experiments. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (5), 1195–1199.

[9] Shimizu, S.; Boon, C. L. The Kirkwood-Buff Theory and the Effect of Cosolvents on Biochemical Reactions. *J. Chem. Phys.* **2004**, *121* (18), 9147–9155.

[10] Smith, P. E. Cosolvent Interactions with Biomolecules: Relating Computer Simulation Data to Experimental Thermodynamic Data. *J. Phys. Chem. B* **2004**, *108* (48), 18716–18724.

[11] Shulgin, I. L.; Ruckenstein, E. A Protein Molecule in an Aqueous Mixed Solvent: Fluctuation Theory Outlook. *J. Chem. Phys.* **2005**, *123* (5), 054909.

[12] Abbott, S.; Booth, J. J.; Shimizu, S. Practical Molecular Thermodynamics for Greener Solution Chemistry. *Green Chem.* **2017**, *19* (1), 68–75.

[13] Booth, J. J.; Omar, M.; Abbott, S.; Shimizu, S. Hydrotrope Accumulation around the Drug: The Driving Force for Solubilization and Minimum Hydrotrope Concentration for Nicotinamide and Urea. *Phys. Chem. Chem. Phys.* **2015**, *17* (12), 8028–8037.

[14] Reid, J. E. S. J.; Walker, A. J.; Shimizu, S. Residual Water in Ionic Liquids: Clustered or Dissociated? *Phys. Chem. Chem. Phys.* **2015**, *17* (22), 14710–14718.

[15] Nicol, T. W. J.; Matubayasi, N.; Shimizu, S. Origin of Non-Linearity in Phase Solubility: Solubilisation by Cyclodextrin beyond Stoichiometric Complexation. *Phys. Chem. Chem. Phys.* **2016**, *18* (22), 15205–15217.

[16] Ploetz, E. A.; Rustenburg, A. S.; Geerke, D. P.; Smith, P. E. To Polarize or Not to Polarize? Charge-on-Spring versus KBFF Models for Water and Methanol Bulk and Vapor–Liquid Interfacial Mixtures. *J. Chem. Theory Comput.* **2016**, *12* (5), 2373–2387.

[17] Section 2.6: Structural Quantities; In: Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press, Oxford, UK, 1989.

[18] Mehrotra, P. K.; Beveridge, D. L. Structural Analysis of Molecular Solutions Based on Quasi-Component Distribution Functions. Application to [H2CO]aq at 25.degree.C. *J. Am. Chem. Soc.* **1980**, *102* (13), 4287–4294.

[19] Baynes, B. M.; Trout, B. L. Proteins in Mixed Solvents: A Molecular-Level Perspective. *J. Phys. Chem. B* **2003**, *107* (50), 14058–14067.

[20] Coutinho, K.; Rivelino, R.; Georg, H. C.; Canuto, S. The Sequential QM/MM Method and Its Applications to Solvent Effects in Electronic and Structural Properties of Solutes. In *Solvation Effects on Molecules and Biomolecules*; S. Canuto (Ed.), Springer Netherlands, 2008; pp 159–189.

[21] Zeindlhofer, V.; Khlan, D.; Bica, K.; Schröder, C. Computational Analysis of the Solvation of Coffee Ingredients in Aqueous Ionic Liquid Mixtures. *RSC Adv.* **2017**, *7* (6), 3495–3504.

[22] Mezei, M. Modified Proximity Criteria for the Analysis of the Solvation of a Polyfunctional Solute. *Mol. Simul.* **1988**, *1* (5), 327–332.

[23] Song, W.; Biswas, R.; Maroncelli, M. Intermolecular Interactions and Local Density Augmentation in Supercritical Solvation: A Survey of Simulation and Experimental Results. *J. Phys. Chem. A* **2000**, *104* (30), 6924–6939.

[24] Furlan, A. C.; Fávero, F. W.; Rodriguez, J.; Laria, D.; Skaf, M. S. Solvation in Supercritical Fluids. In *Solvation Effects on Molecules and Biomolecules*; S. Canuto (Ed.), Springer Netherlands, 2008; pp 433–453.

[25] Oliveira, I. P.; Martínez, L. Molecular Basis for Competitive Solvation of the *Burkholderia Cepacia* Lipase by Sorbitol and Urea. *Phys. Chem. Chem. Phys.* **2016**, *18* (31), 21797–21808.

[26] Ou, S.-C.; Pettitt, B. M. Solute-Solvent Energetics Based on Proximal Distribution Functions. *J. Phys. Chem. B* **2016**, *120* (33), 8230–8237.

[27] Ou, S.-C.; Drake, J. A.; Pettitt, B. M. Nonpolar Solvation Free Energy from Proximal Distribution Functions. *J. Phys. Chem. B* **2017**, *121* (15), 3555–3564.

[28] Haberier, M.; Schröder, C.; Steinhauser, O. Hydrated Ionic Liquids with and without Solute: The Influence of Water Content and Protein Solutes. *J. Chem. Theory Comput.* **2012**, *8*(10), 3911–3928.

[29] Patel, N.; Biswas, R.; Maroncelli, M. Solvation and Friction in Supercritical Fluids: Simulation−Experiment Comparisons in Diphenyl Polyene/CO_2 Systems. *J. Phys. Chem. B* **2002**, *106* (28), 7096–7114.

[30] Gekko, K.; Timasheff, S. N. Mechanism of Protein Stabilization by Glycerol: Preferential Hydration in Glycerol-Water Mixtures. *Biochemistry* **1981**, *20* (16), 4667–4676.

[31] Lin, T. Y.; Timasheff, S. N. Why Do Some Organisms Use a Urea-Methylamine Mixture as Osmolyte? Thermodynamic Compensation of Urea and Trimethylamine N-Oxide Interactions with Protein. *Biochemistry* **1994**, *33* (42), 12695–12701.

[32] Schroer, M. A.; Michalowsky, J.; Birgit, F.; Smiatek, J.; Grübel, G. Stabilizing Effect of TMAO on Globular PNIPAM States: Preferential Attraction Induces Preferential Hydration. *Phys. Chem. Chem. Phys.* **2016**, *18*, 31459–31470.

[33] Diddens, D.; Volker, L.; Heuer, A.; Smiatek, J. Aqueous Ionic Liquids and Their Influence on Peptide Conformations: Denaturation and Dehydration Mechanisms. *Phys. Chem. Chem. Phys.* **2017**, *19*, 20430-20440.

[34] Smiatek, J. Aqueous ionic liquids and their effects on protein structures: an overview on recent theoretical and experimental results. *J. Phys.: Condens. Matter* **2017**, *29*, 233001.

[35] Kirkwood, J. G.; Buff, F. P. The Statistical Mechanical Theory of Solutions. I. *J. Chem. Phys.* **1951**, *19* (6), 774–777.

[36] Shimizu, S.; Matubayasi, N. Hydrotropy: Monomer–Micelle Equilibrium and Minimum Hydrotrope Concentration. *J. Phys. Chem. B* **2014**, *118* (35), 10515–10524.

[37] Pfeiffer, S.; Karimi-Nejad, Y.; Rüterjans, H. Limits of NMR Structure Determination Using Variable Target Function Calculations: Ribonuclease T1, a Case Study. *J. Mol. Biol.* **1997**, *266* (2), 400–423.

[38] Martínez, J. M.; Martínez, L. Packing Optimization for Automated Generation of Complex System's Initial Configurations for Molecular Dynamics and Docking. *J. Comput. Chem.* **2003**, *24* (7), 819–825.

[39] Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A Package for Building Initial Configurations for Molecular Dynamics Simulations. *J. Comput. Chem.* **2009**, *30* (13), 2157–2164.

[40] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935.

[41] MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616.

[42] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2009**, NA – NA.

[43] Yu, W.; He, X.; Vanommeslaeghe, K.; MacKerell, A. D. Extension of the CHARMM General Force Field to Sulfonyl-Containing Compounds and Its Utility in Biomolecular Simulations. *J. Comput. Chem.* **2012**, *33* (31), 2451–2468.

[44] Weerasinghe, S.; Smith, P. E. A Kirkwood-Buff Derived Force Field for Mixtures of Urea and Water. *J. Phys. Chem. B* **2003**, *107*(16), 3891–3898.

[45] Kast, K. M.; Brickman, J; Kast, S. M.; Berry, S. Binary Phases of Aliphatic N-Oxides and Water: Force Field Development and Molecular Dynamics Simulation *J. Phys. Chem. A* **2003**, *107*(27), 5342-5351.

[46] Canchi, D. R.; Jayasimha, P.; Rau, D. C.; Makhatadze, G. I.; Garcia, A. E. Molecular Mechanism for the Preferential Exclusion of TMAO from Protein Surfaces. *J. Phys. Chem. B* **2012**, *116*(40), 12095-12104.

[47] Scheck, E.; Horinek, D.; Netz, R. R. Insight into the Molecular Mechanisms of Protein Stabilizing Osmolytes from Global Force-Field Variations. *J. Phys. Chem. B* **2013**, *117*(28), 8310-8321.

[48] Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802.

[49] Makarov, D. M.; Egorov, G. I.; Kolker, A. M. Density and Volumetric Properties of Aqueous Solutions of Trimethylamine N -Oxide in the Temperature Range from (278.15 to 323.15) K and at Pressures up to 100 MPa. *J. Chem. Eng. Data* **2015**, *60* (5), 1291–1299.

[50] Egan, E. P.; Luff, B. B. Heat of Solution, Heat Capacity, and Density of Aqueous Urea Solutions at 25°C. *J. Chem. Eng. Data* **1966**, *11* (2), 192–194.

[51] Martínez, L. MDAnalysis. Version 17.224, http://leandro.iqm.unicamp.br/mdanalysis. Institute of Chemistry - University of Campinas, 2017.

[52] Hamilton, D.; Stokes, R. H. D. Hamilton, R. H. Stokes, Apparent Molar Volumes of Urea in Several Solvents as Functions of Temperature and Concentration. *J. Solut. Chem.* **1972**, *1*, 213–221.

[53] Kumar, A.; Attri, P.; Venkatesu, P. Effect of Polyols on the Native Structure of α-Chymotrypsin: A Comparable Study. *Thermochim. Acta* **2012**, *536*, 55–62.

[54] Arakawa, T.; Ejima, D.; Kita, Y.; Tsumoto, K. Small Molecule Pharmacological Chaperones: From Thermodynamic Stabilization to Pharmaceutical Drugs. *Biochim. Biophys. Acta* **2006**, *1764* (11), 1677–1687.

[55] Yancey, P. H.; Siebenaller, J. F. Co-Evolution of Proteins and Solutions: Protein Adaptation versus Cytoprotective Micromolecules and Their Roles in Marine Organisms. *J. Exp. Biol.* **2015**, *218* (Pt 12), 1880–1896.

Graphical abstract