Learning about Biomolecular Solvation from Water in Protein Crystals

Irem Altan,[†] Diana Fusco,[‡] Pavel V. Afonine,[¶] and Patrick Charbonneau[*,†,§]

Department of Chemistry, Duke University, Durham, NC 27708, USA, Department of Physics, University of California, Berkeley, Lawrence Berkeley National Laboratory, Berkeley CA 94720, USA, and Department of Physics, Duke University, Durham, NC 27708, USA

E-mail: patrick.charbonneau@duke.edu

[*]To whom correspondence should be addressed
[†]Department of Chemistry, Duke University, Durham, NC 27708, USA
[‡]Department of Physics, University of California, Berkeley
[¶]Lawrence Berkeley National Laboratory, Berkeley CA 94720, USA
[§]Department of Physics, Duke University, Durham, NC 27708, USA

Abstract

Water occupies typically 50% of a protein crystal, and thus significantly contributes to the diffraction signal in crystallography experiments. Separating its contribution from that of the protein is, however, challenging because most water molecules are not localized, and are thus difficult to assign to specific density peaks. The intricateness of the protein-water interface further compounds this difficulty. This information has, therefore, not often been used to study biomolecular solvation. Here, we develop a methodology to surmount in part this difficulty. More specifically, we compare the solvent structure obtained from diffraction data for which experimental phasing is available to that obtained from constrained molecular dynamics (MD) simulations. The resulting spatial density maps show that commonly used MD water models are only partially successful at reproducing the structural features of biomolecular solvation. The radial distribution of water is captured with only slightly higher accuracy than its angular distribution, and only a fraction of the water molecules assigned with high reliability to the crystal structure are recovered. These differences are likely due to shortcomings of both the water models and the protein force fields. Despite these limitations, we nevertheless achieve to infer protonation states of some of the side chains utilizing MD-derived densities.

1 Introduction

Water is not only a medium for biological processes, but an active participant.[1] It mediates interactions between proteins and small-molecule inhibitors,[2,3] and enables the enzymatic transfer of a proton to a protein residue.[4] Moreover ice-binding proteins alter the ordering of water around them, affecting ice nucleation.[5] A reliable physico-chemical description of water in the vicinity of biomolecules is thus needed both to properly solvate these complex objects and to comprehend their function. Yet, despite continued advances to our microscopic understanding of the properties of bulk water,[6–8] including its many phases[9–11] and

hydrophobicity,[12] our grasp of biomolecular solvation still markedly lags behind.[13,14] The intricate interplay between the mosaic of hydrophobic and hydrophilic surface residues, steric hindrance, and side-chain dynamics requires a careful balance of the various intermolecular interactions in order for a structurally accurate description of solvation to emerge. Standard water models, which are rigid, non-polarizable and parameterized to reproduce a standard set of bulk properties, attempt to do just that[8,15] (Figure 1), but it is unclear how they fare at solvating biomolecules.[16] The lack of reliable experimental information about solvation has thus far rendered this problem intractable.

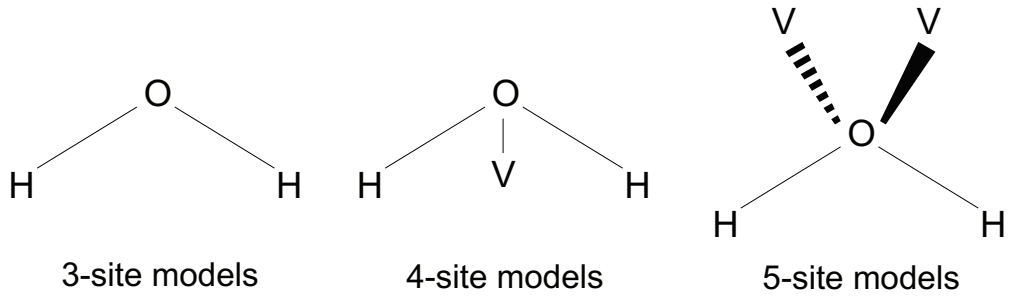

Figure 1: Typical water models used in biomolecular simulations vary mostly in the number of point charges they use to capture intermolecular interaction. All include charges at the hydrogen positions and a Lennard-Jones potential on the oxygen atom, but (a) three-site models contain an additional point charge on the oxygen atom (e.g., SPC[17] and SPC/E[18]), while (b) four-site models use a virtual site (V) (e.g., TIP4P,[19] with[20] and without Ewald summation, and TIP4P/2005[21]), and (c) five-site models split the charge between two virtual sites (e.g., TIP5P[22]). Although six-site models also exist, they are not commonly used.

A possible experimental headway into this problem comes from protein crystallography. Protein crystal unit cells contain a significant fraction of water (between 26% and 90% by volume with an average around 50%[23,24]), hence the inhomogeneous distribution of the solvent impacts the diffracting radiation – be it X-ray,[25–27] neutron,[28] or electron.[29,30] The phase problem of crystallography actually makes the accurate reconstruction of water density profile an essential component of most protein structure determinations.[27] Full structure factors – amplitude and phase – are needed to determine atomic densities within a unit cell, yet only amplitudes can typically be measured directly. Even when some of the phases can be gleaned from multiple intensity measurements or molecular replacement,[31] many phase

values can still go missing. Phases must thus be obtained by iteratively refining the unit cell description and the phase estimates. Because all atoms, i.e. both the macromolecule and the solvent, contribute to all structure factors, an accurate model of the solvent structure is required for this iterative refinement. Obtaining the structure of the protein chain therefore requires a careful treatment of water density fluctuations.

However, the description of the unit cell structure from refinement is far from perfect. The extent of the mismatch is commonly quantified by R-factors,

$$R = \frac{\sum_{\mathbf{k}\in S} |F_{\mathrm{obs}}(\mathbf{k}) - F_{\mathrm{model}}(\mathbf{k})|}{\sum_{\mathbf{k}\in S} F_{\mathrm{obs}}(\mathbf{k})}, \tag{1}$$

where $F_{\mathrm{obs}}(\mathbf{k})$ and $F_{\mathrm{model}}(\mathbf{k})$ are the experimentally observed and model structure factor amplitudes, respectively, from the set S of observed reflections $\mathbf{k}$, where $\mathbf{k} = 2\pi\mathbf{n}$ is a vector containing the Miller indices of the reflection *. Even for the highest-quality protein structures, R-factors average 15%, which is an order of magnitude larger than for small molecules.[33] Although part of the difference is attributable to experimental noise, the weaker agreement between model and experiment is more generally ascribed to the limited sophistication of the structural model of the unit cell content,[34] and especially of the solvent.[34,35] The water model used for structural refinement is assembled from the sum of (i) localized crystal water molecules and (ii) delocalized bulk water regions. A solvent model that improves the description of the water structure should increase agreement between model and data, and ultimately improve the quality of the protein structures obtained crystallographically. Whether molecular dynamics (MD) simulations, which allow a continuum of description between (i) and (ii), can complement the diffraction data[36] and thus improve the description

*If the set S contains all the measured structure factors, the resulting R-factor is R_{work}. As a measure of overfitting, crystallographers also calculate R_{free}[32] by choosing S to be a small set of structure factors that are not included in any stage of structure determination, including refinement. Note that in the crystallography literature the vector $\mathbf{k}$ is conventionally written without the factor of 2π, and an extra 2π then appears in the Fourier transform (see Eq. 2).

of biomolecular solvation is largely unexplored.

In this work, we make an attempt in this direction by comparing the MD and refinement-derived hydration structure of a single protein, a Yb^{+3}-substituted mannose binding protein (PDB ID: 1YTT).[34] Like Burling et al., who previously studied this protein to probe the surrounding water structure, we choose this system because its X-ray structure was determined from multi-wavelength anomalous diffraction (MAD) and a full set of experimental phases was experimentally extracted. This rare occurrence enables us to determine the experimental solvent density profile unbiased by the refinement process. This comparison also allows for benchmarking the water models used in MD simulations. From a methodological standpoint, our comparison relies on an ergodic-like hypothesis that the signal from diffraction techniques is spatially averaged over the configurations of water in the various unit cells, and thus can be recovered by averaging over water configurations obtained from long MD trajectories of a single unit cell. In the following, we first describe the test protein (Sec. 2.1), the water models used in the study (Sec. 2.2), as well as the MD simulation (Sec. 2.3), and comparison (Sec. 2.4), and protonation schemes (Sec. 2.5) before detailing the results of our analysis in Section 3.

2 Methods

This section presents the technical aspects of the experimental system and of the MD simulations as well as the solvent density analysis scheme.

2.1 Protein and Setup

We study the Yb^{+3}-substituted mannose binding protein (PDB ID: 1YTT) solved by MAD phasing up to a resolution of 1.8Å[34] from a crystal with space group symmetry $P2_12_12_1$.[37] The unit cell contains four protein dimers related by symmetry operations, and thus totals eight protein copies (see SI). The model deposited in the Protein Data Bank (PDB)

nearly two decades ago had $R_{\rm work} = 0.185$ and $R_{\rm free} = 0.206$,[38] but methodological advances achieved since by Phenix[39] (version phenix-dev-2405) have allowed us to make substantial improvements to the structural refinement and to update the assigned crystal waters. The biochemical reasonableness of the resulting structure was nevertheless verified by MolProbity.[40] No crystal waters were found to clash with protein atoms and all were at a reasonable hydrogen bonding distance from other crystal waters. Careful examination of the local difference density maps, however, led us to manually remove six water molecules that resulted in an excess electron density compared to the experimental data. Keeping the remaining 254 crystal waters per protein in place, an additional iteration of structural refinement gave $R_{\rm work} = 0.159$ and $R_{\rm free} = 0.183$. The robustness of this result to experimental and refinement noise indicates that the final structure is slightly overfitted but nevertheless a reasonable starting point for this study (see SI).

From the set of optimal structure factors obtained from the refinement process, the electron density at each point $\mathbf{r}$ within the unit cell can formally be computed as,

$$\rho(\mathbf{r}) = \frac{1}{v} \sum_{\mathbf{k} \in S} F_{\rm obs}(\mathbf{k})\ \mathrm{e}^{i[\varphi(\mathbf{k}) - \mathbf{k}\cdot\mathbf{r}]}, \tag{2}$$

where v is the volume of the unit cell. However, because $F(\mathbf{0})$ cannot be extracted experimentally – it is coincident with the transmitted beam – the density profile can only be determined up to an unknown constant, $\bar{\rho}$, and the sum is truncated at high $\mathbf{k}$ once experimental peaks become unresolvable (see SI[41]).

2.2 Water Models

The water models considered in molecular simulations are: (i) SPC,[17] (ii) SPC/E,[18] (iii) TIP3P,[19] (iv) TIP4P[19] with Ewald summation[20†], (v) TIP4P/2005,[21] and (vi) TIP5P[22] (see Fig. 1). The first five have three planar charges (TIP4P and TIP4P/2005 have a negative

[†]Throughout this paper, we refer to TIP4P with Ewald summation as TIP4P.

charge off the oxygen atom), while the sixth has four tetrahedrally-distributed charges. All overestimate the gas phase dipole moment of water, in order to treat some of the many-body contributions in condensed phases in an effective way.[8] SPC and SPC/E, unlike TIP3P, have an O-H bond length and a H-O-H bond angle that differs from the gas phase water geometry for a similar reason. The charge distribution in SPC/E also effectively takes into account the polarization correction to the energy.[18] Note that the only difference between TIP4P and TIP4P/2005 is that their parameters were optimized to match different sets of thermodynamic properties.

These models describe bulk water with varying degrees of success. For instance, TIP4P is better than SPC and TIP3P at reproducing the structure of the gas phase dimer as well as the water density, enthalpy of vaporization, and peak structure of the oxygen-oxygen radial distribution function.[19] TIP5P reproduces the oxygen-oxygen radial distribution function even better than TIP4P,[42] while TIP4P/2005 reproduces better the phase diagram of water than any other models of this type.[21] Although Vega and Abascal, judged TIP4P/2005 to be generally superior, their analysis mainly highlighted that all of such models result from compromises. Whether similar distinctions between these models exist for the structure of water near a protein surface, however, has not yet been tested.

2.3 Molecular Dynamics Simulations

The numerical solvent density profile was extracted from molecular dynamics simulations. Systems are initialized by first placing copies of the crystal structure (obtained in Sec. 2.1) of the protein following the crystal symmetry, within a simulation box that has the same dimensions as the crystal unit cell (see SI). Preserving the protein within its unit cell rather than solvating it within a larger simulation box more closely captures the confinement conditions within the crystal as well as the impact of protein-protein interfaces on water ordering. This choice, however, also introduces computational difficulties. In particular, sampling configurations near protein-protein interfaces can be sluggish, and tuning the water density in

confinement is nontrivial. Errors in the latter may result in a water activity that is quite different from that of a crystal grown in an experimental cocktail [‡]. In order to minimize the impact of both of these problems on the water density profile we run four simulations, each containing a different three-protein dimer copy subset of the unit cell. The absence of a protein dimer copy both accelerates sampling and endogenously introduces a reservoir of solvent that brings its activity near that of the bulk. Note that because only seven protein surface atoms (out of 1769) per chain lie at the interface of four protein dimers, the impact of this removal on the analysis of the solvent structure should be negligible. Water initialization is done by the solvate module in Gromacs, which results in a water density within the bulk region of the simulation box that deviates at most by 1% from its standard value, 1.00 g/mL, at temperature T=298K. Higo and Nakasako have found that the ionic strength does not noticeably affect the structure of water within the unit cell,[45] therefore our simulations use 0.05M NaCl, which is within the typical range of ionic strengths encountered in crystallization experiments.

The protonation states of side chains were at first automatically assigned by Gromacs[46] (version 5.1.2), based on the hydrogen-bonding network analysis of the software package.[47] In order to assess the impact of protonation on the surrounding water structure, we also generated variants with opposite protonation states for histidines, glutamates, lysines, and aspartates (see Sec. 2.5).

The protein chain was modeled using the Amber99sb biomolecular force field.[48] Parameters for Yb^{3+} ions, which are not defined in this force field, were constructed from the Lennard-Jones parameters for sodium ions, which has a similar ionic radius, but a charge of +3. Although this crude treatment cannot fully capture the rich coordination chemistry of a transition metal ion, only a small subset of nearby surface atoms are affected by this choice.

MD simulations were then run with various restraints. To minimize possible deviations from the experimentally-refined *protein* model, carbon and nitrogen atoms on the backbone

[‡]We assume that only water and small ions are present in the crystallization cocktail. In practice, other additives are often included.[43,44]

were kept immobile (restrained), while oxygens were allowed to move, as their position does not affect the overall protein backbone shape. Yb^{+3} ions were also restrained, in order not to bias the simulation results with the approximate parameters described above. In order to facilitate the sampling of water configurations near the protein surface, heavy atoms (i.e. all protein atoms except hydrogens) in the side chains as well as backbone oxygen atoms were constrained harmonically with a force constant of 1000 kJ nm^{-2} mol^{-1}, which is found to be weakest restraint that prevents side chains from changing conformation over the course of the simulations. Although these constraints slightly bias the final water density, they are required for a reliable comparison of the resulting MD water density with the experimentally observed density. Hydrogen atoms, water molecules and ions were allowed to move freely.

The simulations were thermostatted at 298 K. Although the crystallographic data was obtained at 110K upon flash freezing the crystal sample, we assume that the unit cell configuration at the crystallization temperature was preserved by this quench [§] and energy minimization has but a marginal impact on the structure. Amorphous water at that temperature in the protein crystal is indeed glassy.[9] The strong confinement experienced by water in the crystal is expected to leave water in a low-density amorphous (LDA) ice[49] with a structure similar to that of liquid water from which it was quenched. (Neutron diffraction[11] results suggest that the local spatial distribution around a given water molecule in LDA is closely related to that of the liquid phase.) We thus here assume that at distances comparable to the size of the solvent cavities in the protein unit cell, the amorphous structure of water closely resembles that of room-temperature liquid water. This temperature is also optimal for the protein force fields and water models that were used,[17–19,21,22] and facilitates the sampling of solvent configurations (see SI).

We optimize the sampling frequency and computational time by first equilibrating the systems for 30 ns, and then saving configuration snapshots every 3 ns. This provides a total

[§]The crystallization temperature for this protein was not reported but most structures deposited in the PDB are crystallized at room temperature,[38] and a lack of experimental details suggests that an atypical experimental procedure is unlikely.

of 40 fairly well decorrelated solvent configurations (see SI[50,51]). As a consistency check, we compare the water distribution surrounding a given protein atom with that of its symmetric counterparts by computing the real-space correlation coefficients around these atoms (as detailed in section 2.4). Less than 6% of the surface atoms were found to have a sampling error larger than 10%, which suggests that a thorough sampling was achieved.

2.4 Analyzing the Solvent Structure

Electron density maps are extracted from the MD simulations by averaging over the atomic densities obtained from the individual snapshots, and also from the diffraction dataset, using the Computational Crystallography Toolbox (CCTBX) library,[52] upon which Phenix is based. This algorithm uses a three-dimensional grid that spans the unit cell, with a grid spacing that is roughly one fourth of the maximum resolution of the dataset (see SI[53]). The contribution of water to the overall MD electron density, $\rho_{\text{solvent}}(\mathbf{r})$, is then estimated (using a standard Phenix routine), by centering an isotropic Gaussian on the positions of oxygen atoms, with a standard deviation determined by the given atomic B-factor (see SI).

In order to reconstruct the solvent density from the set of simulation boxes that contain only parts of the unit cell (see Sec. 2.3), we use the density information about protein-protein interfaces from the simulation box that contains the relevant copies of the protein dimer. In other words, we select the protein dimer copy that contains the given atom, and the two neighboring protein dimer copies that are closest to that atom. The densities are joined by first partitioning the unit cell, such that each grid point is assigned to the protein atom that is closest to it (considering the refined protein coordinates), and then by copying the density within the partition associated with each atom.

We compare the spatial distribution of water around protein atoms in both experimental and simulated systems using the grid described above. Because the nitrogens and carbons in the backbone are kept immobile, the protein structure in the various MD snapshots only differ from that of the refined structure in its side-chain positions. The radial distribution

functions (RDF), which capture the average solvent density profile as a function of distance from a protein heavy atom, offers the lowest-order correction to the bulk solvent description near an interface.[54] For a subset of atoms A and the grid described above, it is computed following a scheme similar to that of Lin and Pettitt. For an atom $i \in A$, let χ_i be the set of grid points assigned to that atom, and $X = \cup_{i \in A} \chi_i$ be the set of grid points assigned to atoms in A. Then,

$$g_A(r) = \frac{1}{\bar{\rho}_{\text{solvent}}} \frac{\sum_{i \in A} \sum_{p \in \chi_i'} \rho(p)\ \Theta[\rho(p)]}{\sum_{i \in A} \sum_{p \in \chi_i'} \Theta[\rho(p)]}, \tag{3}$$

where p is a grid point, $\rho(p)$ is the electron density at that grid point, χ_i' is the subset of χ_i that contains grid points $r \pm \Delta r$ away from atom i, Θ is the Heaviside function, and $\bar{\rho}_{\text{solvent}}$ is the average electron density in the solvent region. The chosen shell thickness, $\Delta r = 0.3$Å, is only slightly smaller than the grid spacing derived from the maximal resolution of the protein data, $d_{\text{min}} = 1.8$Å, which ensures that a statistically sufficiently number of grid points is captured within each shell, without overly coarsening the data.

RDFs are obtained both for separate sets of surface N, O, and C atoms and for individual surface atoms, in both cases considering only surface atoms that are well localized, i.e., with $|\chi_i| \geq 500$ and $B_i \leq 24$Å^2 (the B-factor of atom i), which roughly corresponds to a mean-squared displacement of 0.3Å, and follows the definition used in Ref. 34. We discard surface atoms that are within 6Å of Yb^{3+} ions due to the strong Fourier ripples that surround them (see SI). For the sets of surface N, O, and C atoms, an average radial correlation coefficient of the RDFs are computed for 2.4Å $< r <$ 6Å away from the protein atoms. For $r <$ 2.4Å it is not possible to deconvolute protein and solvent contributions to the observed electron density, whereas for $r >$ 6Å statistical noise and diffraction artifacts dominate because less than 2% of the grid points fall beyond that distance. The correspondence between the RDFs from experimental and MD-generated densities is assessed by the Pearson correlation coefficient.[56]

For individual surface atoms, we construct the set of RDFs, $\{(g_{i,\mathrm{MD}}(r), g_{i,\mathrm{obs}}(r))\}$ for all $i \in A$, and all radial bins. The Pearson correlation coefficient of this set of ordered pairs is also computed. In order to compare the radial position of a given peak in the RDFs, its 95% confidence interval is estimated by drawing 1000 perturbed RDFs according to the error margin in each radial bin.

Because RDFs average out information about the orientation of water molecules, we also consider angular distribution functions (ADFs), which depend on the hydrogen bond network in each configuration, and thus encode three-body and higher-order correlations. Only grid points within the first solvation shell, i.e., for 2.4Å $< r <$ 4.8Å, are considered for this computation. The heavy atom is placed at the origin, and then the orientation of each point around this atom is determined using spherical coordinates, (θ, ϕ), with the axis orientations following the PDB conventions,[57]

$$\gamma_i(\phi,\theta) = \frac{\sum\limits_{p\in\chi_i(I_\phi,I_\theta)} \rho(p)\ \Theta[\tilde{r}_p - r_1]\Theta[r_2 - \tilde{r}_p]}{\sum\limits_{p\in\chi_i(I_\phi,I_\theta)} \Theta[\tilde{r}_p - r_1]\Theta[r_2 - \tilde{r}_p]}, \tag{4}$$

where $\tilde{r}_p$ gives the distance from the grid point to the heavy atom, $\chi_i(I_\phi, I_\theta)$ is the set of grid points assigned to i and are oriented such that $\phi \in [\phi - \Delta\phi/2, \phi + \Delta\phi/2]$ and $\theta \in [\theta - \Delta\theta/2, \theta + \Delta\theta/2]$. We set $\Delta\phi = \Delta\theta = \pi/30$, which corresponds to an arc-length of 0.25Å at 2.4Å, and 0.5Å at 4.8Å, comparable to the radial binning of the RDF. The comparison between the angular distribution functions in experiments and simulations is also done using the Pearson correlation coefficient of $\Gamma_{\mathrm{obs}}(i,\phi,\theta) = \gamma_{i,\mathrm{obs}}(\phi,\theta)$ and $\Gamma_{\mathrm{MD}}(i,\phi,\theta) = \gamma_{i,\mathrm{MD}}(\phi,\theta)$, considering only cases in which both quantities are nonzero.

The real-space distribution of the water density combines information about both the radial and angular components. It thus provides an overall comparison of the solvent structure. Using the three-dimensional grid on which the electron density is calculated, we consider correlations between each grid point within 2.4Å $< r <$ 6Å of a surface atom. Because grid

points are roughly 0.4Å apart, the resulting coarsening is similar to that of both the RDF and the ADF, allowing for a meaningful comparison between their correlation coefficients. The discrepancy between the real-space simulation and experimental maps is also measured separately for the first solvation shell and for protein-protein contacts. Note that the latter are defined as the grid points at least 2.4Å and at most 3.0Å away from a pair of N or O atoms situated on different protein dimer copies.

Because the real-space distribution of water is calculated by averaging over exact electron densities calculated from MD snapshots, the peak shapes and locations are affected by the precise motion of the water molecules. In order to compare the water density peak locations, we eliminate the role of water density widths and shapes by selecting only peaks that appear above a given threshold density, ρ_{th}. We additionally deconstruct the solvent density by focusing exclusively on crystal waters, which by definition are associated with an observed local electron density well above experimental noise. This comparison thus deconvolutes the role of peak shape from that of peak location in assessing the density profile. Following Higo and Nakasako, we define a prediction $A_{\mathrm{pred}}(\rho_{\mathrm{th}})$ and a recall $A_{\mathrm{rec}}(\rho_{\mathrm{th}})$ score. The former yields the fraction of crystal waters that are within a distance smaller than the water radius, i.e., ∼1.4Å, of an MD peak above the threshold, while the latter gives the fraction of MD peaks above the threshold that are within ∼1.4Å of a crystal water,

$$\begin{aligned} A_{\mathrm{pred}}(\rho_{\mathrm{th}}) &= \frac{\sum\limits_{\alpha\in P_{\mathrm{MD}}} \Theta[\rho(\mathbf{r}_\alpha)-\rho_{\mathrm{th}}]\{1-\prod\limits_{\beta\in P_{\mathrm{CW}}}[1-w(|\mathbf{r}_\alpha-\mathbf{r}_\beta|)]\}}{\sum\limits_{\beta\in P_{\mathrm{MD}}} \Theta[\rho(\mathbf{r}_\alpha)-\rho_{\mathrm{th}}]}, \\ A_{\mathrm{rec}}(\rho_{\mathrm{th}}) &= \frac{\sum\limits_{\beta\in P_{\mathrm{CW}}} \{1-\prod\limits_{\beta\in P_{\mathrm{CW}}}[1-w(|\mathbf{r}_\alpha-\mathbf{r}_\beta|)]\theta[\rho(\mathbf{r}_\alpha)-\rho_{\mathrm{th}})]\}}{|P_{\mathrm{CW}}|}, \end{aligned} \tag{5}$$

where P_{MD} is the set of MD peaks, P_{CW} is the set of crystal waters, $\rho(\mathbf{r}_\alpha)$ is the density that corresponds to peak α, with $w(r) \equiv \Theta(1.4-r)$ the overlap function defined in terms of the Heaviside Θ function, $|\mathbf{r}_\alpha - \mathbf{r}_\beta|$ is the distance between the MD peak α and the crystal

water β, and $|P_{\mathrm{CW}}|$ is the total number of crystal waters in the refined protein structure. In other words, A_{pred} is the true positive rate, while A_{recall} is the true negative rate. Note that to assess the structural significance of the measured signal, we further compute these scores with a random distribution of crystal waters with the same number density in the solvent region. This null model results in a constant $A_{\mathrm{pred}}(\rho_{\mathrm{th}}) = 0.1$, and a $A_{\mathrm{rec}}(\rho_{\mathrm{th}})$ that steadily decays from 0.2 as ρ_{th} increases, both values being well below the level of the measured signal.

Finally, we compare the experimental and MD densities in reciprocal space by generating a model of the protein unit cell that combines the simulated density with the protein model (see SI[58]). Comparing the resulting R_{work} of this model with that of the original protein model determines whether or not the simulated densities improve the agreement with the experimental data. For this analysis, we estimate the error in the R_{work} values due to measurement errors to be one part in ten thousand (95% confidence interval, see SI). This analysis can also be performed by partitioning the set of reflections into different resolution bins and calculating R_{work} for each. Because higher resolution bins correspond to more structured parts of the unit cell, such as the protein atoms and ordered water molecules around the protein surface, while lower resolution bins correspond to regions with flatter electron density, such as the bulk solvent,[23] this analysis provides insight into the regions of MD-generated solvent density that better agree with experimental data.

2.5 Inferring Protonation States

The solvent distribution is a reflection of its environment. Given sufficiently accurate solvation information, it should thus be possible to determine the protonation state of a residue. To test this hypothesis, different MD simulations were run for alternative side-chain protonation states, and the resulting water density was compared with the experimental density. The default Gromacs protonation states for a subset of histidine, glutamate, aspartate and lysine residues were inverted in different simulations. The default and inverted protonation

states for the residue types we study are summarized in Table 1. For glutamates, aspartates and lysines, the residues to be (de)protonated were chosen, such that: (i) they have one surface side chain oxygen or nitrogen; (ii) they are at least 6Å away from another residue chosen for protonation analysis in the same simulation to avoid interference between the solvent distribution of one residue with the other; and (iii) do not neighbor a Yb^{3+} ion and thus are not affected by the approximations to its force field. We further verify that the water density in the vicinity of these examples is well sampled by making sure that all the surrounding water molecules decorrelate in at most $\sim$ 1 ns, and that observations are consistent for all four protein dimer copies. This whole set of simulations was run with the TIP4P water model.

Table 1: Default vs inverted protonation states

residue	default	inverted
histidine	$N_{\delta 1}$ protonated or $N_{\epsilon 2}$ protonated **charge**: 0	$N_{\epsilon 2}$ protonated or $N_{\delta 1}$ protonated **charge**: 0
lysine	N_{ζ} has 3 protons **charge**: +1	N_{ζ} has 2 protons **charge**: 0
aspartate	both $O_{\delta 1}$ and $O_{\delta 2}$ deprotonated **charge**: -1	either $O_{\delta 1}$ or $O_{\delta 2}$ has 1 proton **charge**: 0
glutamate	both $O_{\epsilon 1}$ and $O_{\epsilon 2}$ deprotonated **charge**: -1	either $O_{\epsilon 1}$ or $O_{\epsilon 2}$ has 1 proton **charge**: 0

3 Results and Discussion

In this section, the experimental and MD solvent information is used to assess the quality of the MD description first by comparing density profiles, and second by using standard crystallographic observables. The potential to infer the protonation state of residues from MD solvent density is also examined.

3.1 Real-Space Comparison of Water Densities

The RDF, which is a quintessential quantity in liquid state theory,[54] has been utilized as main observable by most prior studies of macromolecular solvation.[34,55,59,60] Some of these have even attempted to reconstruct protein hydration from RDFs alone.[55,59,60] It is therefore a natural starting point for our evaluation.

Comparing the RDF for different atom types and water models reveals that the various descriptions qualitatively agree with one another (Fig. 2). In particular, a clear first solvation shell is noted, and hints of a second shell can be gleaned, although the number of available grid points beyond 6Å is too small to obtain a reliable profile of that shell. Because of experimental noise and artifacts, such as Fourier ripples (see SI), it is difficult to determine whether the first peak position of the various simulation models match that of the experimental RDF. The first peak positions of all water models, however, agree with each other within the error margin, with the exception of TIP5P for surface oxygens, for which the peak is pushed further out. For nitrogens and oxygens, the peak amplitude is significantly higher in simulations than in experiment. One might be tempted to ascribe the sensitivity of this feature to the choice of B-factor for water. Some water molecules are indeed less localized than others, especially near fairly mobile surface protein atoms. Hence, no single B can reliably describe all water molecules. The fact that neither nitrogens ($B = 20.2\text{Å}^2$) nor oxygens ($B = 18.7\text{Å}^2$) have significantly higher average B-factors than carbons ($B = 19.3\text{Å}^2$) does not rule out this possibility, as B-factors are unreliable estimates of thermal motion in protein crystals.[61] It is also possible that the peak intensity could be weakened by experimental noise and artifacts (see SI).

The overall shape of the RDF should nevertheless be insensitive to these effects. The Pearson correlation coefficients between the averaged RDFs of surface N, O, and C atoms reveal that the water density in the vicinity of surface oxygens and carbons is more accurately reproduced than around nitrogens (dashed lines in Fig. 3a). However, the radial correlation coefficients of RDFs for individual atoms (solid lines in Fig. 3a) suggest that the

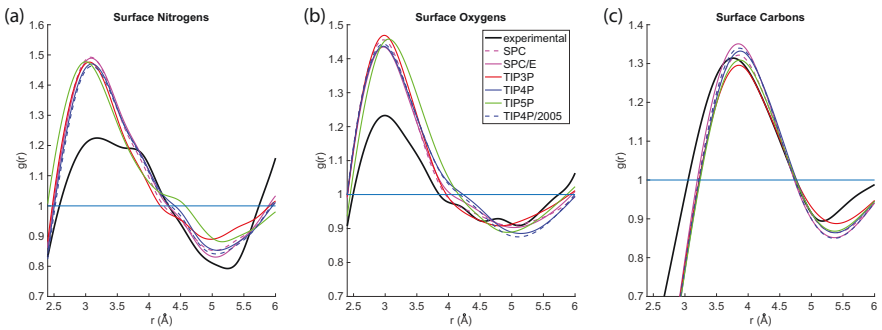

Figure 2: Averaged RDFs for surface (a) N, (b) O, and (c) C atoms, from different water models. Results obtained from different water models agree well with each other, as well as with the experimental RDFs.

distributions around oxygens are significantly worse. The radial distribution of water around individual oxygen atoms appears to depend more sensitively on the chemical environment than around nitrogens and carbons. We also conclude that the distribution of water around each atom is far from universal. Efforts to reconstruct water density using averaged radial distribution functions – as was previously attempted[59,60] – therefore have serious shortcomings. Interestingly, all water models perform identically within the estimated error, for both average and regular radial correlation coefficients. We get back to this point below.

Angular correlation coefficients are generally slightly lower than their radial counterparts. This effect is consistent with the latter being a higher-order structural feature. One might nonetheless expect that a model parameterized to more accurately reproduce the subtle orientational order of the various bulk water crystal phases,[21] such as TIP4P/2005, or a model like TIP5P, which explicitly treats tetrahedral point charges, to improve the description of ADFs. Neither TIP5P nor TIP4P/2005, however, perform significantly better than the other water models, including TIP4P.

Angular correlation coefficients are generally larger for nitrogens and oxygens than for carbons, which is particularly interesting. The orientation of water molecules around surface nitrogens and oxygens indeed mostly results from direct hydrogen bonding, while that of water molecules around carbons are affected by their interplay with the broader hydrogen-bond network and are thus less constrained by the protein force field. The resulting hydrophobic-

ity is structurally more subtle to capture, which likely explains why water models struggle to capture this effect (Fig. 3b). Water models that account more accurately for many-body correlations in water, such as E3B[62] and E3B2,[63] might improve the orientational description of water in these systems. Direct tests, however, are not immediately possible because these models have not yet been parameterized for macromolecular solvation.

Because it contains higher-order structural information, the real-space distribution generally gives rise to significantly lower correlations than either the radial or the angular correlation coefficients (Fig. 3). While water models capture the radial distribution of water around carbons equally well as around nitrogens, they rank last in spatial correlation coefficients. This is consistent with their poor performance describing angular correlations. Similarly, models reproduce the angular distribution around oxygens as well as around nitrogens, but perform worse for real-space correlations.

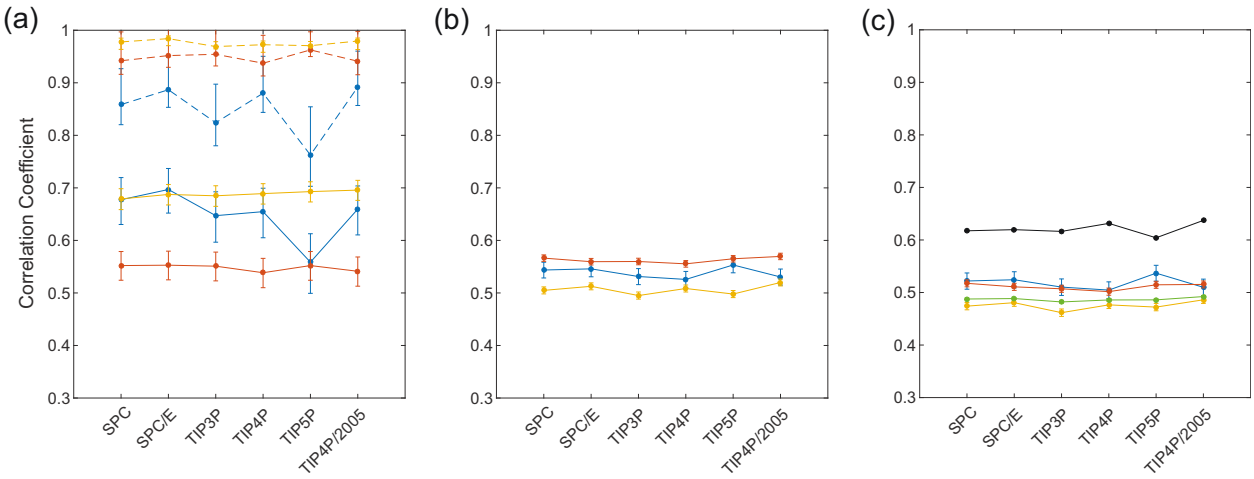

Figure 3: (a) Radial (solid) and averaged radial (dashed), (b) angular, and (c) real-space correlation coefficients for surface N (blue), O (red) and C (yellow) atoms. Real-space correlation coefficients for first-layer (green) and contact waters (black) are also given in (c). Error bars denote 95% confidence intervals. The lines connecting the data points are solely a guide for the eye and have no physical meaning.

To gain further insight into the aspects of water models that increase their propensity to capture water structure around biomolecules, we compare the spatial distribution of water in different regions of space. We first calculate real-space correlations separately for contact

and first-layer waters. Correlations for the first shell are consistent with the overall real-space correlations for surface N, O, and C. Beyond the first layer, errors get amplified by structural imprecisions in the first layer, a situation further worsened by the reduced number of grid points in that region of space. Protein-protein contacts, by contrast, show fairly good structural agreement. This likely results from the surface atoms in these regions being much less mobile than elsewhere, and from steric constraints there playing a larger role in dictating the solvent structure. The position and orientation of water molecules in protein contacts are thus likely less sensitive to water model and protein force field parametrizations.

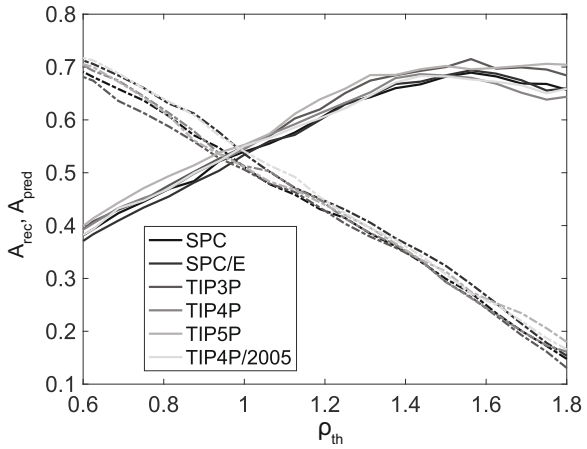

Figure 4: Prediction (solid) and recall scores (dashed), as defined in Eq. (5). At low threshold densities, too many MD peaks are identified, resulting in a high recall score but a low prediction score. As the threshold increases, MD peaks with stronger signals persist, which at high densities predict roughly 70% of the crystal waters. However, the recall scores fall as the density increases, suggesting that a significant fraction of crystal waters do not overlap with an MD peak.

We next consider the recall and prediction scores (Eq. (5)) of the MD peak locations with the assigned crystal waters. At low threshold densities many MD peaks are identified and a high fraction of crystal waters are recovered, although only a few of these MD peaks are near crystal waters. As ρ_{th} increases, the number of MD peaks decreases, but a greater fraction of the remaining ones overlap with crystal waters, decreasing the rate of false negatives. This encouragingly suggests that the strongest predictions (and interactions) of the MD model correlate with crystal waters with reasonably high accuracy (70%). The recall scores,

however, fall steadily with increasing ρ_{th}, and thus many crystal waters are not predicted by MD simulations. In other words, a low true positive rate is obtained. For all water models, the highest A_{pred} and A_{rec} is ~ 0.7. The discrepancy between MD and experiments is thus not purely due to imprecisions in the MD description of the shape and width of the density peaks, but also in their location. Some of this error is likely attributable to the protein force fields, as the location of high density peaks in the MD density are affected by the average positions of the nearby protein atoms throughout the simulation.

3.2 Reciprocal Space

The agreement between MD and experiments is assessed in reciprocal space by first combining MD densities with the refined PDB coordinates of the protein without the crystal waters. If MD simulations were to reproduce water densities reasonably well, the resulting R_{work} would be less than that of the refined PDB model. Yet for the best water model (SPC) we obtain $R_{\text{work}} = 0.208$, which is significantly higher than $R_{\text{work}} = 0.159$ obtained for the refined protein model (Fig. 5). The difference in R_{work} is also greater at higher resolution, suggesting that highly ordered solvent regions are not adequately captured. If we instead consider the entire solvent region to have a flat electron density, $R_{\text{work}} = 0.219$, which is about 1% worse than the best water model. Note that this increase is orders of magnitude larger than the estimated error in R_{work} (see SI). Hence, although the MD models contain some information about the water density within the unit cell, a significant fraction of it is inaccurate.

To check whether MD simulations capture solvent structure that is complementary to that of the crystal waters assigned from the experimental data, we combine MD densities with the refined PDB coordinates, including crystal waters. (The MD electron density of crystal waters is thus removed.) This strategy reduces R_{work} for water models to $R_{\text{work}} = 0.162$, which is an improvement over the previous scheme yet still appreciably higher than the refined model $R_{\text{work}} = 0.159$. The gap between R_{work} values at lower resolutions at least is

then closed.

It is important to note that R_{work} at high resolution is affected not only by the water density around each protein atom, but also by the average protein atom positions in the MD simulation being slightly different from that of the refined protein structure. Although refined coordinates are used for this analysis, the water density is affected by the slightly perturbed protein atom locations throughout the MD simulations, resulting in possible overlaps between the solvent density and refined atom positions.

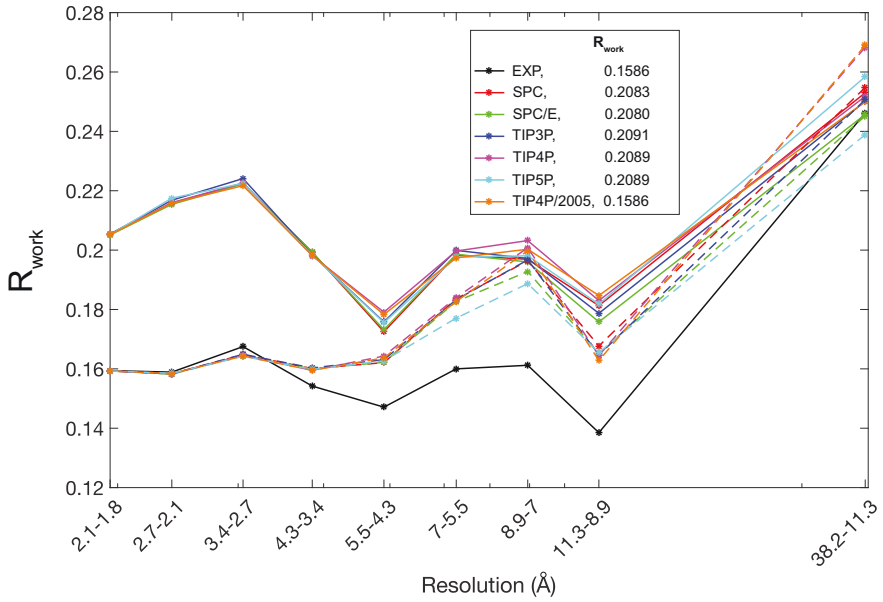

Figure 5: R_{work} in different resolution bins for the original model (EXP, black), and for models constructed by combining MD densities with the protein model. The overall R_{work} is as given in the legend. Dashed lines correspond to R_{work} for models with the refined crystal waters combined with the MD density.

Retaining crystal waters in the refined protein structure results in substantially lower R_{work} at high resolution than the MD density results alone. This strategy thus yields results comparable to the original protein model, but as resolution decreases, R_{work} becomes significantly worse than for the original protein model, which once again confirms that the refined model describes the electron density in the unit cell more accurately than the MD solvent density.

3.3 Inferring Protonation States

A complication that hinders the improvement of the solvent description in the analysis of X-ray diffraction experiments is that hydrogens, which are surrounded by relatively small electron clouds, cannot be detected unless a remarkably high diffraction resolution, i.e., better than roughly 0.7-1.0Å, is achieved. Although the position of many of the hydrogens on the protein chain can be inferred based on an elementary description of bonding (partly explaining the success of structure validation tools, such as MolProbity[40]), side chain protonation states can remain somewhat ambiguous, as do the positions of side chain hydrogens with a rotational degree of freedom. This problem is especially acute for side chains that contribute to an enzymatic pathway[4] or to protein-protein interactions,[64] such as salt-bridges.[65–67] Prediction servers have thus been developed to infer pK_a values and titration curves of individual side chains, based on the electrostatic properties of neighboring residues.[68,69] Other software packages rely on less involved algorithms to assign protonation states. For instance, MolProbity picks the most suitable protonation state and hydrogen atom position that minimizes clashes, while Gromacs[46] analyses the hydrogen bonding network around it.[47] Yet because the presence or absence of protons affects the solvent distribution around these sites, probing the solvent distribution around such residues should allow one to determine their protonation state more systematically.

The preceding analysis suggests that reconstructing the solvent density, and hence predicting every density peak, is not possible using existing water models. We are nevertheless encouraged by the fact that MD simulations reproduce a significant fraction of the high intensity peaks associated with crystal waters. It may thus be possible to infer protonation states by comparing the overlap between MD peaks and crystal water, if changing the protonation state of a residue gives rise to or eliminates such high intensity peaks in the MD solvent density.

In most cases considered here, either residues have insufficient solvent exposure to conduct the analysis, no significant difference in solvation is observed, or both sets of density

patterns are similarly mismatched with the refined structure. Despite this, a few examples for which inverting the protonation state of a residue significantly affects the surrounding water distribution can be found.

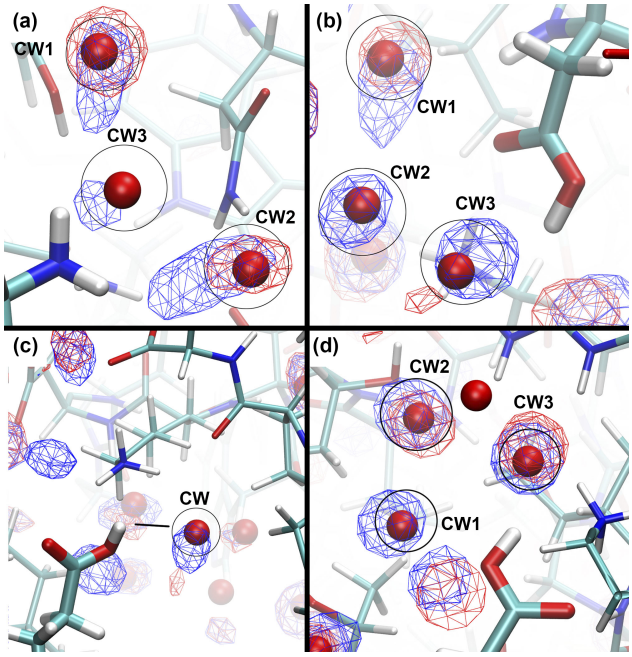

Figure 6: Comparison of water density distribution for simulations that contain different protonation states for (a) LYS 145 in chain A (the blue blob is behind the water and does not overlap it), (b) ASP 200 in chain A, (c) GLU 130 in chain B, and (d) GLU 218 in chain A. The water density from the default protonation state simulations are shown in blue wireframe, and the alternate protonation state simulations are shown in red wireframe. For all snapshots the isosurfaces are contoured at 0.88 $e^{-}/Å^{3}$. Crystal waters (CW) from the refined protein structure are denoted with red spheres.

Removing one of the three protons of the default +1 charged LYS 145 in chain B results in a slightly better overlap with two crystal waters (CW), labeled CW1 and CW2 in Figure 6a. There is, however, a third crystal water, CW3, within hydrogen bonding distance to the nitrogen that is unexplained by either protonation state. Although this lysine residue is relatively well localized and its average position does not deviate from that in the refined structure, MD models completely miss CW3. In addition, both protonation states result in an MD peak with the same orientation as the crystal waters, because removing a proton does not drastically change the geometry of the remaining two hydrogens. The MD peaks in

the deprotonated case are, however, pushed away from the protein, likely due to the altered charge distribution in the residue. We conclude that a neutral lysine with two protons at this position leads to a water density that is more consistent with the experimental density, although with caveats.

The case of ASP 200 in chain A is slightly more complicated. The MD peak resulting from the protonated case agrees better with CW1, compared to the peak resulting from the simulation in which the residue is unprotonated (Fig. 6b). However, two crystal waters (CW2 and CW3) overlap only with high intensity peaks in the density obtained with the standard protonation state. It is therefore likely that this residue is not protonated, but it is unclear why the protonated case better explains the CW1 peak.

For both GLU 130 in chain B and GLU 218 in chain A, the unprotonated case gives better agreement between MD peaks and crystal waters. Protonating the former results in a loss of an MD peak that overlaps the crystal water (Fig. 6c). Similarly, protonating GLU 218 in chain A results in the loss of an MD peak that overlaps CW1, but retains those on CW2 and CW3 (Fig. 6d). This is likely because CW3 is still within hydrogen bonding distance to the residue, and CW2 is within hydrogen bonding distance to CW3. However, the disappearance of the density peak on CW1 is unexpected because the protonated oxygen could still form a hydrogen bond to a water at that location.

While these results are encouraging, their robustness with respect to protein atom positions remains untested. The location of high density peaks in the MD density is likely affected by both the equilibrium position of protein atoms and the degree to which they are localized. In the case of CW1 near GLU 218 in chain A, for instance, the MD-density peak on CW1 might be missing in the protonated case (even though a water molecule at this location would be within hydrogen bonding distance), because other water molecules might be forming a more stable hydrogen bond network nearby. In addition, the success of these inferences ultimately depends on our ability to reliably reconstruct the water density around proteins by MD simulations. Using this method to detect protonation states thus relies on

being well above the noise inherent to the structural analysis.

4 Conclusion

Using a protein with a high-quality dataset from X-ray crystallography, we have attempted to extract complementary information about water structure in protein crystals from diffraction data and MD simulations. This work improved upon earlier efforts in a few different ways. (i) We used a simulation box equivalent to the protein unit cell, containing multiple protein copies in order to capture water structure in the protein-protein interfaces. (ii) We ran significantly longer simulations, which enabled the solvent configurations to be ergodically sampled. (iii) We used reference diffraction data,[34] for which experimental phases is available. Thanks to these advances, it was possible for us to compare more detailed aspects of the water structure and to explore the role solvent structure plays around titratable residues.

Comparison of experimental and MD densities in real space revealed that although water models are relatively good at capturing the radial distribution of water near the protein surface, they struggle to predict angular distributions and are somewhat deficient at reconstructing the overall water density. The relatively poor distribution of water around carbons, in particular, suggests that the structural consequences of hydrophobic effect are inadequately captured by these models. Remarkably, all water models we considered were found to behave rather similarly at the structural level.

Although MD water models are insufficient for reconstructing biomolecular hydration with a precision sufficient to conduct structural refinement, they nonetheless capture the position of a fraction of the crystal waters. In optimal hydration circumstances, these models may thus help assign protonation states to some of the protein side chains. The robustness of these predictions with respect to the choice of parameters, including the protein force field and the protonation state of the nearby residues, is untested, but these findings nevertheless

suggest MD simulations can provide information complementary to what is available from X-ray crystallography. It may further be possible to devise refinement schemes that utilize this information to improve biomolecule structure quality, similar to some already existing schemes.[59,70]

Our results suggest that it may be necessary to add more features to the common water models in order to reconstruct accurately the solvent structure around biomolecules. A re-parametrization of the existing models taking into account properties pertaining to protein-water interactions might improve the description of these interactions. Such a re-parametrization may not, however, adequately capture both bulk and interfacial water properties at once. A single, fixed dipole moment might indeed not be able to capture the behavior of water in both of these regions.[71] Considering more complex models that include polarizability[8] or include three-body interactions,[72] might provide a more robust starting point. To model a process that depends sensitively on the position of water molecules, it might be preferable to consider even higher-accuracy models of water that include *ab initio* descriptions. The use of models such as those based on quantum mechanics[73] may eventually become computationally tractable, allowing for a refinement process in which simulations are run at each refinement step.

Acknowledgement

PC, IA, and DF acknowledge support from National Science Foundation Grant no. NSF DMR-1055586. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562,[74] as well as the Duke Compute Cluster. PVA thanks the NIH (grant GM063210) and the PHENIX Industrial Consortium. We thank David and Jane Richardson, and Marat Mustyakimov for fruitful discussions. Data relevant to this work have been archived and can be accessed at `https://doi.org/10.7924/XXXXXXX`.

Supporting Information Available

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/XXXX. This material is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Ball, P. Water is an active matrix of life for cell and molecular biology. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, 201703781.

(2) Kim, E.; Baker, C.; Dwyer, M.; Murcko, M.; Rao, B.; Tung, R.; Navia, M. Crystal Structure of HIV-1 Protease in Complex with VX-478, a Potent and Orally Bioavailable Inhibitor of the Enzyme. *J. Am. Chem. Soc.* **1995**, *117*, 1181–1182.

(3) Balius, T. E.; Fischer, M.; Stein, R. M.; Adler, T. B.; Nguyen, C. N.; Cruz, A.; Gilson, M. K.; Kurtzman, T.; Shoichet, B. K. Testing Inhomogeneous Solvation Theory in Structure-Based Ligand Discovery. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, 201703287.

(4) Wan, Q.; Parks, J. M.; Hanson, B. L.; Fisher, S. Z.; Ostermann, A.; Schrader, T. E.; Graham, D. E.; Coates, L.; Langan, P.; Kovalevsky, A. Direct Determination of Protonation States and Visualization of Hydrogen Bonding in a Glycoside Hydrolase with Neutron Crystallography. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, 201504986.

(5) Voets, I. K. From Ice-Binding Proteins to Bio-Inspired Antifreeze Materials. *Soft Matter* **2017**, *13*, 4808–4832.

(6) Skinner, L. B.; Huang, C.; Schlesinger, D.; Pettersson, L. G.; Nilsson, A.; Benmore, C. J. Benchmark Oxygen-Oxygen Pair-Distribution Function of Ambient Water from X-Ray Diffraction Measurements with a Wide Q-Range. *J. Chem. Phys.* **2013**, *138*, 074506.

(7) Paesani, F.; Voth, G. A. The Properties of Water: Insights from Quantum Simulations. *J. Phys. Chem. B* **2009**, *113*, 5702–5719.

(8) Vega, C.; Abascal, J. L. Simulating Water with Rigid Non-Polarizable Models: a General Perspective. *Phys. Chem. Chem. Phys.* **2011**, *13*, 19663–19688.

(9) Kim, C. U.; Tate, M. W.; Gruner, S. M. Glass-to-Cryogenic-Liquid Transitions in Aqueous Solutions Suggested by Crack Healing. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 11765–11770.

(10) Sanz, E.; Vega, C.; Abascal, J.; MacDowell, L. Phase Diagram of Water from Computer Simulation. *Phys. Rev. Lett.* **2004**, *92*, 255701.

(11) Finney, J.; Hallbrucker, A.; Kohl, I.; Soper, A.; Bowron, D. Structures of High and Low Density Amorphous Ice by Neutron Diffraction. *Phys. Rev. Lett.* **2002**, *88*, 225503.

(12) Chandler, D. Interfaces and the Driving Force of Hydrophobic Assembly. *Nature* **2005**, *437*, 640–647.

(13) Ball, P. Water as an Active Constituent in Cell Biology. *Chem. Rev.* **2008**, *108*, 74–108.

(14) Macias-Romero, C.; Nahalka, I.; Okur, H. I.; Roke, S. Optical Imaging of Surface Chemistry and Dynamics in Confinement. *Science* **2017**, *357*, 784–788.

(15) Guillot, B. A Reappraisal of What We Have Learnt During Three Decades of Computer Simulations on Water. *J. Mol. Liq.* **2002**, *101*, 219–260.

(16) Smith, P. E.; Pettitt, B. M. Modeling solvent in biomolecular systems. *J. Phys. Chem.* **1994**, *98*, 9700–9711.

(17) Berendsen, H. J.; Postma, J. P.; van Gunsteren, W. F.; Hermans, J. *Intermolecular forces*; Springer, 1981; pp 331–342.

(18) Berendsen, H.; Grigera, J.; Straatsma, T. The missing term in effective pair potentials. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(19) 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*, 926–935.

(20) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an Improved Four-Site Water Model for Biomolecular Simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120*, 9665–9678.

(21) Abascal, J. L.; Vega, C. A General Purpose Model for the Condensed Phases of Water: TIP4P/2005. *J. Chem. Phys.* **2005**, *123*, 234505.

(22) Mahoney, M. W.; Jorgensen, W. L. A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid, Nonpolarizable Potential Functions. *J. Chem. Phys.* **2000**, *112*, 8910–8922.

(23) Weichenberger, C. X.; Afonine, P. V.; Kantardjieff, K.; Rupp, B. The Solvent Component of Macromolecular Crystals. *Acta Crystallogr. Sect. D* **2015**, *71*, 1023–1038.

(24) Weichenberger, C. X.; Rupp, B. Ten Years of Probabilistic Estimates of Biocrystal Solvent Content: New Insights via Nonparametric Kernel Density Estimate. *Acta Crystallogr. Sect. D* **2014**, *70*, 1579–1588.

(25) Jones, N. Atomic Secrets: 100 Years of Crystallography. *Nature* **2014**, *505*, 602–603.

(26) Chapman, H. N.; Fromme, P.; Barty, A.; White, T. A.; Kirian, R. A.; Aquila, A.; Hunter, M. S.; Schulz, J.; DePonte, D. P.; Weierstall, U. et al. Femtosecond X-Ray Protein Nanocrystallography. *Nature* **2011**, *470*, 73–77.

(27) Drenth, J. *Principles of Protein X-Ray Crystallography*; Springer Science & Business Media: New York, NY, USA, 2007.

(28) Myles, D. A. Neutron Protein Crystallography: Current Status and a Brighter Future. *Curr. Opin. Struct. Biol.* **2006**, *16*, 630–637.

(29) Nannenga, B. L.; Gonen, T. Protein structure determination by MicroED. *Curr. Opin. Struct. Biol.* **2014**, *27*, 24–31.

(30) Nannenga, B. L.; Shi, D.; Leslie, A. G.; Gonen, T. High-Resolution Structure Determination by Continuous-Rotation Data Collection in MicroED. *Nat. Methods* **2014**, *11*, 927–930.

(31) Taylor, G. L. Introduction to phasing. *Acta Crystallogr. Sect. D* **2010**, *66*, 325–338.

(32) Brünger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **1992**, *355*, 472–475.

(33) Wlodawer, A.; Minor, W.; Dauter, Z.; Jaskolski, M. Protein Crystallography for Non-Crystallographers, or How to Get the Best (But Not More) from Published Macromolecular Structures. *FEBS Journal* **2008**, *275*, 1–21.

(34) Burling, F. T.; Weis, W. I.; Flaherty, K. M.; Brünger, A. T. Direct Observation of Protein Solvation and Discrete Disorder with Experimental Crystallographic Phases. *Science* **1996**, *271*, pp. 72–77.

(35) Holton, J. M.; Classen, S.; Frankel, K. A.; Tainer, J. A. The R-Factor Gap in Macromolecular Crystallography: an Untapped Potential for Insights on Accurate Structures. *FEBS Journal* **2014**, *281*, 4046–4060.

(36) Jiang, J.-S.; Brünger, A. T. Protein hydration observed by X-ray diffraction: solvation properties of penicillopepsin and neuraminidase crystal structures. *J. Mol. Biol.* **1994**, *243*, 100–115.

(37) Brünger, A. T. Brünger Lab Web Site. `"http://atbweb.stanford.edu"`, Accessed 9-July-2017.

(38) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.;

Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(39) Adams, P. D.; Afonine, P. V.; Bunkóczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L.-W.; Kapral, G. J.; Grosse-Kunstleve, R. W. et al. *PHENIX*: a Comprehensive Python-Based System for Macromolecular Structure Solution. *Acta Crystallogr. Sect. D* **2010**, *66*, 213–221.

(40) Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. MolProbity: All-Atom Structure Validation for Macromolecular Crystallography. *Acta Crystallogr. Sect. D* **2010**, *66*, 12–21.

(41) Fokine, A.; Urzhumtsev, A. Flat Bulk-Solvent Model: Obtaining Optimal Parameters. *Acta Crystallogr. Sect. D* **2002**, *58*, 1387–1392.

(42) Vega, C.; McBride, C.; Sanz, E.; Abascal, J. L. Radial Distribution Functions and Densities for the SPC/E, TIP4P and TIP5P Models for Liquid Water and Ices I_h, I_c, II, III, IV, V, VI, VII, VIII, IX, XI and XII. *Phys. Chem. Chem. Phys.* **2005**, *7*, 1450–1456.

(43) Altan, I.; Charbonneau, P.; Snell, E. H. Computational Crystallization. *Arch. Biochem. Biophys.* **2016**, *602*, 12–20.

(44) Peat, T. S.; Christopher, J. A.; Newman, J. Tapping the Protein Data Bank for Crystallization Information. *Acta Crystallogr. Sect. D* **2005**, *61*, 1662–1669.

(45) Higo, J.; Nakasako, M. Hydration Structure of Human Lysozyme Investigated by Molecular Dynamics Simulation and Cryogenic X-Ray Crystal Structure Analyses: On the Correlation Between Crystal Water Sites, Solvent Density, and Solvent Dipole. *J. Comput. Chem.* **2002**, *23*, 1323–1336.

(46) Berendsen, H. J.; van der Spoel, D.; van Drunen, R. GROMACS: a Message-Passing Parallel Molecular Dynamics Implementation. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

(47) Lemkul, J. Personal exchange, 2016.

(48) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712–725.

(49) Bertrand, C. E.; Zhang, Y.; Chen, S.-H. Deeply-Cooled Water Under Strong Confinement: Neutron Scattering Investigations and the Liquid–Liquid Critical Point Hypothesis. *Phys. Chem. Chem. Phys.* **2013**, *15*, 721–745.

(50) Mukherjee, S.; Mondal, S.; Bagchi, B. Distinguishing Dynamical Features of Water Inside Protein Hydration Layer: Distribution Reveals What is Hidden Behind the Average. *J. Chem. Phys.* **2017**, *147*, 024901.

(51) Wang, J. H. Self-Diffusion Coefficients of Water. *J. Phys. Chem.* **1965**, *69*, 4412–4412.

(52) Grosse-Kunstleve, R. W.; Sauter, N. K.; Moriarty, N. W.; Adams, P. D. The Computational Crystallography Toolbox: Crystallographic Algorithms in a Reusable Software Framework. *J. Appl. Crystallogr.* **2002**, *35*, 126–136.

(53) Afonine, P.; Urzhumtsev, A. On a Fast Calculation of Structure Factors at a Subatomic Resolution. *Acta Crystallogr. Sect. A* **2004**, *60*, 19–32.

(54) Hansen, J.-P.; McDonald, I. R. *Theory of simple liquids*; Elsevier, 1990.

(55) Lin, B.; Pettitt, B. M. Note: On the Universality of Proximal Radial Distribution Functions of Proteins. *J. Chem. Phys.* **2011**, *134*, 106101.

(56) Riley, K. F.; Hobson, M. P.; Bence, S. J. *Mathematical Methods for Physics and Engineering*, 3rd ed.; Cambridge University Press: Cambridge, UK, 2006; p 1200.

(57) Protein Data Bank. Atomic Coordinate and Bibliographic Entry Format Description. `https://cdn.rcsb.org/wwpdb/docs/documentation/file-format/PDB_format_1992.pdf`, 1992; Accessed 21-09-2017.

(58) Afonine, P.; Grosse-Kunstleve, R.; Adams, P.; Urzhumtsev, A. Bulk-Solvent and Overall Scaling Revisited: Faster Calculations, Improved Results. *Acta Crystallogr. Sect. D* **2013**, *69*, 625–634.

(59) Lounnas, V.; Pettitt, B.; Phillips Jr, G. A Global Model of the Protein-Solvent Interface. *Biophys. J.* **1994**, *66*, 601–614.

(60) Virtanen, J. J.; Makowski, L.; Sosnick, T. R.; Freed, K. F. Modeling the Hydration Layer Around Proteins: HyPred. *Biophys. J.* **2010**, *99*, 1611–1619.

(61) Rupp, B. *Biomolecular crystallography*; Garland Science: New York, NY, USA, 2010; pp 635–636.

(62) Tainter, C.; Pieniazek, P.; Lin, Y.-S.; Skinner, J. Robust Three-Body Water Simulation Model. *J. Chem. Phys.* **2011**, *134*, 184501.

(63) Tainter, C. J.; Shi, L.; Skinner, J. L. Reparametrized E3B (Explicit Three-Body) Water Model Using the TIP4P/2005 Model as a Reference. *J. Chem. Theory Comput.* **2015**, *11*, 2268–2277.

(64) Wahle, C. W.; Martini, K. M.; Hollenbeck, D. M.; Langner, A.; Ross, D. S.; Hamilton, J. F.; Thurston, G. M. Model for Screened, Charge-Regulated Electrostatics of an Eye Lens Protein: Bovine gammaB-Crystallin. *Phys. Rev. E* **2017**, *96*, 032415.

(65) Anderson, D. E.; Becktel, W. J.; Dahlquist, F. W. pH-Induced Denaturation of Proteins: a Single Salt Bridge Contributes 3-5 kcal/mol to the Free Energy of Folding of T4 Lysozyme. *Biochemistry* **1990**, *29*, 2403–2408.

(66) Dey, M.; Cao, C.; Sicheri, F.; Dever, T. E. Conserved Intermolecular Salt Bridge Required for Activation of Protein Kinases PKR, GCN2, and PERK. *J. Biol. Chem.* **2007**, *282*, 6653–6660.

(67) Fusco, D.; Headd, J. J.; De Simone, A.; Wang, J.; Charbonneau, P. Characterizing Protein Crystal Contacts and Their Role in Crystallization: Rubredoxin as a Case Study. *Soft matter* **2014**, *10*, 290–302.

(68) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H++: a Server for Estimating pK_as and adding missing hydrogens to macromolecules. *Nucleic Acids Res.* **2005**, *33*, W368–W371.

(69) Rostkowski, M.; Olsson, M. H.; Søndergaard, C. R.; Jensen, J. H. Graphical Analysis of pH-Dependent Properties of Proteins Predicted Using PROPKA. *BMC Struct. Biol.* **2011**, *11*, 1.

(70) Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nat. Protoc.* **2007**, *2*, 2728–2733.

(71) Yu, H.; van Gunsteren, W. F. Accounting for Polarization in Molecular Simulation. *Comput. Phys. Commun.* **2005**, *172*, 69–85.

(72) Cisneros, G. A.; Wikfeldt, K. T.; Ojamae, L.; Lu, J.; Xu, Y.; Torabifard, H.; Bartok, A. P.; Csanyi, G.; Molinero, V.; Paesani, F. Modeling Molecular Interactions in Water: From Pairwise to Many-Body Potential Energy Functions. *Chem. Rev.* **2016**, *116*, 7501–7528.

(73) Zheng, M.; Reimers, J. R.; Waller, M. P.; Afonine, P. V. $Q|R$: Quantum-Based Refinement. *Acta Crystallogr. Sect. D* **2017**, *73*, 45–52.

(74) Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazle-

wood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D. et al. XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering* **2014**, *16*, 62–74.