Learning about Biomolecular Solvation from Water in Protein Crystals

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

†*Department of Physics, University of California, Berkeley*

‡*Lawrence Berkeley National Laboratory, Berkeley CA 94720, USA*

¶*Department of Physics, Duke University, Durham, NC 27708, USA*

§*Department of Chemistry, Duke University, Durham, NC 27708, USA*

E-mail: patrick.charbonneau@duke.edu

Abstract

Water occupies typically 50% of a protein crystal, and thus significantly contributes to diffraction signals from crystallography experiments. Separating its contribution from that of the protein is challenging mainly because almost water molecules are mostly delocalized as a result of the probabilistic nature of solvation, and are thus are difficult to assign to specific density peaks. The intricate protein-water interface compounds this difficulty. Here, 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 capturing biomolecular solvation. In general, their radial distribution is captured with only slightly higher accuracy than their 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 this difficulty, we nevertheless attempt to infer protonation states of side chains

utilizing MD-derived densities, and observe some cases for which assignment is possible with reasonable confidence.

1 Introduction

Water is not only a medium for biological processes, but an active participant in them. It mediates interactions between proteins and small-molecule inhibitors,[1,2] and enables the enzymatic transfer of a proton to a protein residue.[3] One remarkable group of examples of such a water-biomolecule relationship is caused by ice-binding proteins, which can alter the ordering of water around them, with effects ranging from ice nucleation to antifreeze.[4] A reliable physico-chemical description of water in the vicinity of biomolecules is needed both to properly solvate these complex objects and to comprehend their function. Yet despite marked advances to our microscopic understanding of the properties of bulk water,[5–7] including its many phases[8–10] and the origine of hydrophobicity,[11] our grasp of biomolecular solvation still markedly lags behind.[12,13] The intricate interplay between the mosaic of hydrophobic and hydrophilic surface residues, steric hindrance, and side-chain dynamics indeed 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[7,14] (Figure 1), but it is unclear how they fare at solvating biomolecules. Reliable experimental information about solvation to assess their predictions is not forthcoming, hence the question largely remains unanswered.

A possible experimental headway into this problem comes from crystallography. Protein unit cells contain a significant fraction of water (between 26% and 90% by volume with an average of about 50%[21,22]), hence the solvent density fluctuations are captured by the diffracting radiation – be it X-ray,[23–25] neutron,[26] or electron.[27,28] The "phase problem" actually makes the accurate reconstruction of water density profile an essential component

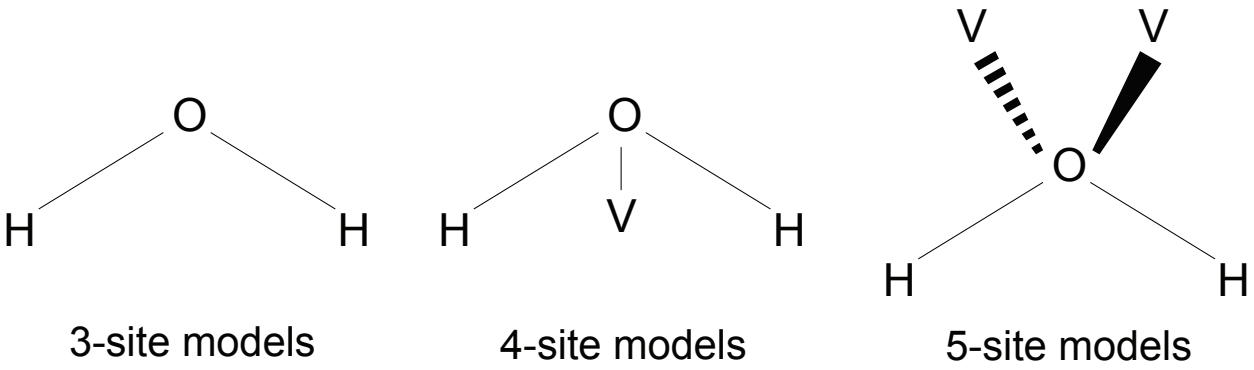

Figure 1: Typical water models used in biomolecular simulations vary mostly in the number of point charges they have. All include charges at the hydrogen positions and a Lennard-Jones potential on the oxygen atom, but (a) three-site models contain an additional point charge on the oxygen atom (e.g., SPC[15] and SPC/E[16]), while (b) four-site models use a virtual site (V) (e.g., TIP4P[17] (with Ewald summation[18]) and TIP4P/2005[19]), and (c) five-site models split the charge between two virtual sites (e.g., TIP5P[20]). Although six-site models also exist, they are not commonly used.

of most protein structure determinations.[25] Because full structure factors – amplitude and phase – are needed to determine atomic densities within a unit cell, but only amplitudes can typically be measured directly, phases must be obtained by iteratively refining the unit cell description and the phase estimates. Even when some of the phase values can be gleaned from multiple intensity measurements or molecular replacement,[29] iterative refinement is still needed to determine the remaining ones. Obtaining the structure of the protein chain therefore requires a careful treatment of water density fluctuations, which should is then available for detailed structural analysis.

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

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

where $F_{\mathrm{exp}}(\mathbf{k})$ and $F_{\mathrm{model}}(\mathbf{k})$ are the experimentally measured and model structure factor

amplitudes, respectively, from the set S of reflections $\mathbf{k}$, where $\mathbf{k} = 2\pi\mathbf{n}$ is a vector with $\mathbf{n} \in \mathbb{Z}^3$, *. is about 15% even for the highest-quality protein structures. Remarkably, this is an order of magnitude larger than for small molecules.[31] 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,[32] especially for the solvent.[32,33] In this context, the solvating water is indeed crudely decomposed between, on the one hand, localized crystal water molecules and, on the other, delocalized bulk water regions, with nothing in between. A solvent model that could improve water structure description would thus increase agreement between model and data, and ultimately improve the quality of protein structures obtained crystallographically. Could MD predictions help? Whether molecular dynamics (MD) or refined models of water better capture the solvation of biomolecules remains unclear.

In this work, we compare the MD and refinement-derived hydration structure of a single protein, a Yb^{+3}-substituted mannose binding protein (PDB ID: 1YTT).[32] Like by Burling *et al.*, we choose this system because its X-ray structure was determined from multi-wavelength anomalous diffraction (MAD) and a complete set of experimental phases is available. This rare occurrence enables us to extract an experimental solvent density profile unbiased by the refinement process. The 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), water models used in the study (Sec. **??**), the MD simulation scheme (Sec. 2.3), and the comparison scheme (Sec. 2.4,2.5), before detailing the results of our analysis in Section 3.

*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}[30] by choosing S to be a small set of structure factors that do not participate in any stage of structure determination, including refinement.

2 Methods

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

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[32] from a crystal with space group symmetry $P2_12_12_1$.[34] The unit cell containing four protein dimers related by symmetry operations, thus amounting to eight copies of the protein (SI). The model deposited in the Protein Data Bank (PDB) nearly two decades ago had $R_{\text{work}} = 0.206$ and $R_{\text{free}} = 0.185$,[35] but methodological advances achieved since enable Phenix[36] (version phenix-dev-2405) to make substantial improvements to the structural refinement and update the assigned crystal waters. The biochemical reasonableness of the resulting structure was nevertheless verified by MolProbity.[37] No crystals 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 remove six water molecules because they 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 $R_{\text{work}} = 0.1537$ and $R_{\text{free}} = 0.1804$.

In order to assess the robustness of these R-factors, estimates of the measurement errors were propagated to R_{work}. Because errors for the measured intensities were not reported, a typical experimental error, $\Delta F_{\text{exp}}/F_{\text{exp}} \approx 5\%$,[33] was used as estimate. Errors for the measured phases, $\Delta\varphi$, were obtained from the figure of merit, m, of each wavevector,

$$\Delta\varphi = \cos^{-1}(m). \tag{2}$$

Assuming a Gaussian noise with standard deviation $\Delta\varphi/2$ and $F_{\text{exp}}(\mathbf{k}) \times 0.05$, we generated

$N = 10,000$ perturbations, computing R_{work} for each. The resulting distribution of structure factors gives an estimate of $R_{\text{work}} = 0.1614(3)$, within the 95% confidence interval (Fig. 2). This analysis suggests that although the final refined model is slightly overfitted, it remains structurally quite reasonable.

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,

$$\rho(\mathbf{r}) = \frac{1}{v} \sum_{\mathbf{k}} F(\mathbf{k})\, \mathrm{e}^{i[\varphi(\mathbf{k}) - \mathbf{k}\cdot\mathbf{r}]}, \tag{3}$$

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, $\bar{\rho}$. It is here determined using the approach adopted by Lang *et al.*[38] (See Supporting Information).

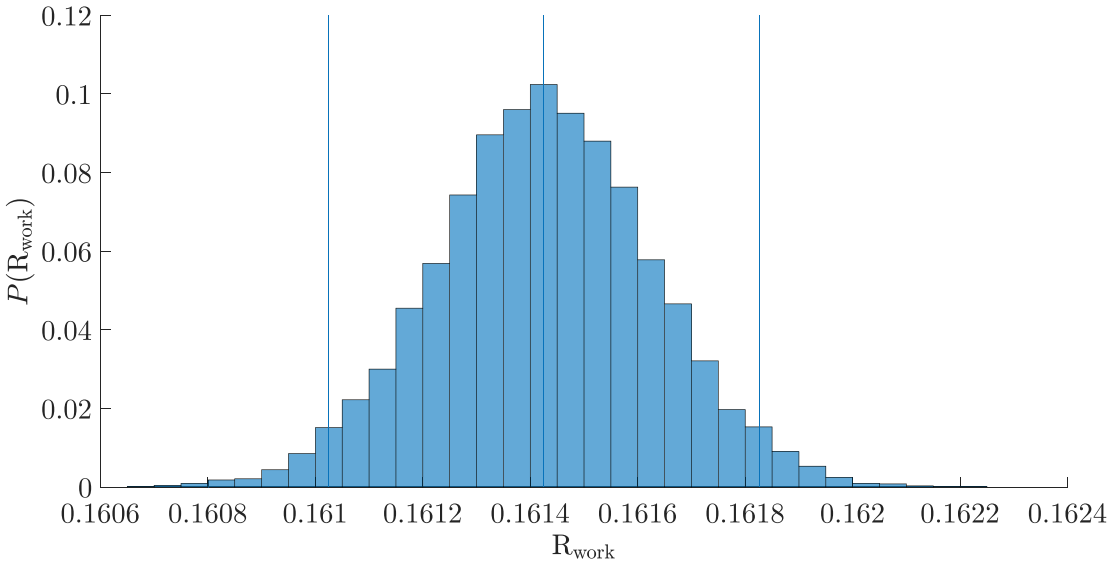

Figure 2: Probability distribution of R_{work} for structure factors perturbed by the estimated experimental errors as described in the text. From this analysis we get $R_{\text{work}} = 0.1614(3)$, which is slightly higher than the refined value. Because a small perturbations in the structure factors consistently lead to larger R_{work} values suggests that the data is overfitted slightly.

2.2 Water Models

The water models considered in molecular simulations are: (i) SPC,[15] (ii) SPC/E,[16] (iii) TIP3P,[17] (iv) TIP4P[17] with Ewald summation,[18] (v) TIP4P/2005,[19] and (vi) TIP5P[20] (see Fig. 1). The first five have three planar charges (TIP4P and TIP4P/2005 have a negative 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.[7] SPC and SPC/E, by contrast to TIP3P, have an O-H bond length and a H-O-H bond angle differing 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.[16] Note that the difference between TIP4P and TIP4P/2005 is that their parameters were optimized to match different sets of thermodynamic properties.

A number of differences between these models have been observed in bulk water. 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.[17] TIP5P reproduces the oxygen-oxygen radial distribution function even better than TIP4P,[39] while TIP4P/2005 reproduces better the phase diagram of water than any other models of this type.[19] Although Vega et al.,[7] judged TIP4P/2005 to be generally superior their analysis mainly highlight that all of such models result from compromises. Whether similar distinctions exist for the structure of water near the protein surface, however, has yet to be tested.

2.3 Molecular Dynamics Simulations

The numerical solvent density profile was extracted from molecular dynamics (MD) 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 (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. This choice, however, also introduces computational difficulties. In particular, sampling configurations near protein-protein interfaces can be sluggish, and selecting the water density in confinement is nontrivial. Errors in the latter may result in a water activity 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 copy both accelerates sampling and endogenously introduces a reservoir of solvent that brings its activity near the bulk value [†]. 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 is negligible. Water molecules are then inserted by tiling the whole unit cell with a disordered template configuration of 216 water molecules; molecules that clash with the protein chain–as determined by the sum of their van der Waals radii–are removed. This 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 a temperature of 298K.

In order to neutralize the net protein charge, 0.05 M of sodium and chloride ions are subsequently added by replacing some of the solvent molecules with ions. Higo *et al.* have found that the ionic strength does not noticeably affect the structure of water within the unit cell,[40] which we also verified by running test simulations with 0.05M, 0.1 M, 0.5 M and 1.0 M NaCl. The lowest salt concentration was thus used for the rest of the analysis.

The protonation states of side chains were at first automatically assigned by Gromacs, based on the hydrogen-bonding network analysis of the software package.[41] 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.

The protein chain was modeled using the Amber99sb biomolecular force field.[42] Param-

[†]This is an approximation in that it assumes that only water and small ions are present in the crystallization cocktail. Typically, other additives are included to enable crystallization.

eters 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.

Gromacs[43] (version 5.1.2) was used to run the MD simulations with various restraints imposed. To minimize possible deviations from the experimentally-refined *protein* model, carbon and nitrogen atoms on the protein backbone as well as the Yb^{+3} ions were kept immobile. In order to facilitate the sampling of water configurations near the protein surface, heavy atoms (all protein atoms but hydrogens) in the side chains as well as backbone oxygen atoms were also harmonically restrained with a force constant of 1000 kJ nm^{-2} mol^{-1}, which is the weakest restraint that prevented side chains from changing conformation over the course of the simulations. Hydrogen atoms, water molecules and ions were allowed to move freely.

The simulations were thermostatted at 298 K, which is the temperature at which the various water models used were parameterized.[15–17,19,20] Although the crystallization temperature for this protein was not specifically reported, it likely is around room temperature, because most structures deposited in the PDB are crystallized at room temperature.[35] The lack of experimental details indeed suggests that an atypical experimental procedure was unlikely. Conveniently, this temperature further allows an efficient sampling of solvent configurations (SI).

We optimize the sampling quality and computational time by first equilibrating the systems for 30 ns, and then saving configurational snapshots every 3 ns. This provides 40 fairly well decorrelated solvent configurations. 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 (detailed in section 2.5). This indicates that at most 6% of the surface atoms have a sampling error larger than 10%.

Because the experimental diffraction data with which we compare the simulation results were obtained at 110 K–after flash freezing in liquid nitrogen–the water within the protein crystal is expected to be glassy.[8] The strong confinement experienced by water in the crystal is expected to leave it in a low-density amorphous (LDA) ice[44] with a structure that is similar to the liquid water from which it came. It has indeed been established by neutron diffraction[10] that although the bulk structure of LDA ice more closely resembles that of crystalline water, its local spatial distribution around a given water molecule is closely related to that of the liquid phase. In this work, we thus assume that at distances comparable to the size of the cavities in the protein unit cell, the quenched structure water is closely reproduced by simulations at higher (liquid) temperatures, where sampling is ergodic.

2.4 Analyzing the Solvent Density

Electron density maps are extracted from the MD snapshots and the model unit cell using the Computational Crystallography Toolbox (CCTBX) library,[45] 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 (SI). The contribution of water to the overall electron density $\rho_{\text{oxygen}}(\mathbf{r})$ is then estimated by centering an isotropic Gaussian with a standard deviation determined by the given atomic B-factor, B, on the oxygen atom of each water molecule (SI).

In order to reconstruct the solvent density from combining simulation boxes that contain only parts of the unit cell (see Sec. 2.3), we use the water density at protein-protein interfaces from the simulation box that contains the relevant protein dimer copies. In other words, we select the protein copy that contains the given atom, and two neighboring protein dimer copies that are the closest to the atom. The densities are then joined by first partitioning the unit cell such that each grid point is assigned to the closest protein atom, and by then copying the density within the relevant partition of each atom. This partitioning uses protein atom positions averaged over the course of the MD simulation, as is also the case for the

analyses described in Sect. 2.5.

2.5 Water Structure Analysis

We compare the spatial distribution of water around protein atoms in both experimental and simulated systems, using the unit cell partitioning scheme described above and the refined protein atom positions as reference. Because the protein backbone is frozen, the protein structure throughout the MD simulation only differs from the refined structure in its side-chain configurations.

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.[46] For a subset of atoms A and the grid described above, it is computed as follows. For an atom $i \in A$, let χ_i be the set of grid points assigned to that atom, and define $X = \cup_{i \in A} \chi_i$ the set of grid points assigned to atoms in A. Then,

$$g_A(r) = \frac{1}{\rho_{\text{solvent}}} \frac{\sum_{i \in A} \sum_{p \in \chi_i'} \rho(p)\ \Theta(\rho(p))}{\sum_{i \in A} \sum_{p \in \chi_i'} \Theta(\rho(p))}, \tag{4}$$

where p is a grid point, $\rho(p)$ is the electron density at grid point p, χ_i' is a subset of χ_i that contains grid points that are $r \pm \Delta r$ away from atom i, and ρ_{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, $d_{\min} = 1.8$Å, which ensures that a statistically sufficiently number of grid points is captured within each shell, without overly coarsening the data. The normalization, ρ_{solvent}, is the average density computed over the solvent region in the unit cell.

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 $B_i \leq 24\text{Å}^2$. We discard surface atoms that are within 6Å of ytterbium ions

due to the strong Fourier ripples that surround these atoms (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. This particular range is chosen because for $r <$ 2.4Å it is not possible to deconvolute protein from solvent contributions to the experimental electron density, while for $r >$ 6Å statistical noise and diffraction artifacts dominate as $\lesssim 2\%$ 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.[47] For individual surface atoms, we construct the set of RDFs, $\{(g_{i,\mathrm{MD}}(r), g_{i,\mathrm{exp}}(r))\}$ 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 intervals is estimated by drawing 1000 perturbed RDFs according to the error margin in each radial bin.

Because RDFs lose information about the (relative) orientation of water molecules with respect to the protein and each other, we also consider the angular distribution function (ADF), which depends on the hydrogen bond network in each configuration, and thus encodes three-body and higher-order correlations. Only grid points within the first solvation shell, i.e., between 2.4Å to 4.8Å, are considered. Axis orientations follow the PDB conventions, placing the heavy atom at the origin, and then determining the orientation of each grid point around this atom in spherical coordinates, (θ, ϕ),

$$\gamma_i(\phi,\theta) = \frac{\sum\limits_{p\in\chi_i(I_\phi,I_\theta)} \rho(p)\ \Theta(\delta(p)-r_1)\Theta(r_2-\delta(p))}{\sum\limits_{p\in\chi_i(I_\phi,I_\theta)} \Theta(\delta(p)-r_1)\Theta(r_2-\delta(p))}, \tag{5}$$

where $\delta(p)$ 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, and is thus comparable in size to the radial binning used

for the RDF. The comparison between the angular distribution functions in experiments and simulations is also done using the Pearson correlation coefficient of $\Gamma_{\rm exp}(i,\phi,\theta) = \gamma_{i,\rm exp}(\phi,\theta)$ and $\Gamma_{\rm MD}(i,\phi,\theta) = \gamma_{i,\rm MD}(\phi,\theta)$, considering only cases in which both values are greater than zero.

The real-space distribution of the water density combines information about both the radial and angular components. It thus provides an overall comparison of solvation. Using the three-dimensional grid on which the electron density is calculated, we consider correlations between each grid point within $2.4 < r < 6$Å of a surface atom. Because grid points are roughly 0.4Å apart, the resulting coarsening is similar to that of the RDF and the ADF. The Pearson correlation coefficient for $\rho_{\rm MD}(p)$ and $\rho_{\rm exp}(p)$ computed for a set of grid points X can thus be meaningfully compared to that of the latter two. The discrepancy between the real-space simulation and experimental maps is further measured separately for the first solvation shell and for protein-protein contacts. 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.

In order to eliminate the role of the water density peak breadths and shapes from the structural analysis, water density peak locations are directly compared, selecting only peaks that appear above a threshold density, $\rho_{\rm th}$. We additionally deconstruct the solvent density by focusing exclusively on crystal waters, which by definition are associated with a local electron density well above experimental noise. This comparison also deconvolutes the role of peak shape from that of peak location in assessing the density profile. Following Higo *et al.*,[40] we define a prediction $A_{\rm pred}$ and a recall $A_{\rm rec}$ score as a function of a threshold electron density. 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 a threshold density, $\rho_{\rm th}$, while the latter gives the fraction of MD peaks above $\rho_{\rm th}$ that are within ∼1.4Å, of a crystal water,

$$A_{\mathrm{pred}}(\rho_{\mathrm{th}}) = \frac{\sum\limits_{p\in P_{\mathrm{MD}}} \Theta(\rho(p)-\rho_{\mathrm{th}})(1-\prod\limits_{w\in P_{\mathrm{CW}}}(1-w(|\mathbf{r}_w-\mathbf{r}_p|))}{\sum\limits_{p\in P_{\mathrm{MD}}} \Theta(\rho(p)-\rho_{\mathrm{th}})},$$
$$A_{\mathrm{rec}}(\rho_{\mathrm{th}}) = \frac{\sum\limits_{w\in P_{\mathrm{CW}}} (1-\prod\limits_{w\in P_{\mathrm{CW}}}(1-w(|\mathbf{r}_w-\mathbf{r}_p|))\theta(\rho(p)-\rho_{\mathrm{th}}))}{|P_{\mathrm{CW}}|}, \tag{6}$$

where P_{MD} is the set of MD peaks, P_{CW} is the set of crystal waters, $\rho(p)$ is the density that corresponds to peak p, with $w(r) \equiv \Theta(1.4-r)$ the overlap function defined in terms of the Heaviside Θ function, $|\mathbf{r}_{\mathrm{cw}}-\mathbf{r}_p|$ is the distance between the MD peak and the crystal water, and $|P_{\mathrm{CW}}|$ is the total number of crystal waters in the refined protein structure. Note that to assess the structural significant 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. The results give $A_{\mathrm{pred}} = 0.1$, and A_{rec} steadily decays from 0.2 with increasing ρ_{th}, which are both well below the level of the measured signal.

Finally, we compare the experimental densities with MD densities in reciprocal space by generating a model of the protein unit cell that combines the simulated density with the protein model. 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. 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,[21] this analysis provides some insight into the regions of MD-generated solvent density that better agree with experimental data.

2.6 Inferring Protonation States

The solvent distribution reflects the physical conditions of its environment. Given sufficiently accurate solvation information, it should thus be possible to determine the protonation state of some of the residues. 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 histidines, glutamates, aspartates, and lysines 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 side chain oxygen or nitrogen on the surface; (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 residue; and (iii) do not neighbor a ytterbium ionbecause the water density around ytterbium ions is 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. Note that the whole set of these simulations was run with the TIP4P water model.

3 Results and Discussion

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

Table 1: Default vs. inverted protonation states.

Residue	Default	Inverted
Histidine	$N_{\delta 1}$ protonated or $N_{\epsilon 2}$ protonated **charge**: +1	$N_{\epsilon 2}$ protonated or $N_{\delta 1}$ protonated **charge**: +1
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

3.1 Real-Space Comparison of Water Densities

The RDF, which is a quintessential component of liquid state theory,[46] has been utilized as main observable by most prior studies of macromolecular solvation.[32,48–50] Some of these have even attempted to reconstruct protein hydration from RDFs alone.[48–50] 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. 3). In particular, a clear first solvation shell is denoted, and hints of 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 (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 position 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 difference in mismatch between the peak height of various heavy atom types, however, does not support this hypothesis. Neither nitrogens ($B = 20.2\text{Å}^2$) nor oxygens ($B = 18.7\text{Å}^2$) have significantly higher average B-factors than carbons ($B = 19.3\text{Å}^2$). The peak intensity might thus simply be weakened by experimental noise and artifacts.

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, 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. 4a). However, when one considers the radial correlation coefficients for RDFs of individual atoms (solid lines in Fig. 4a), the distributions around oxygens are significantly worse. The radial distribution of water around individual oxygen atoms thus depends more sensitively on the chemical environment than around individual 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[48,50] – are therefore inherently flawed. 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.

Radial correlation coefficients are generally slightly higher than angular correlation coefficients. 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,[19] such as TIP4P/2005, or a model like TIP5P which explicitly treats tetrahedral point charges, to improve the angular description. Neither TIP5P nor TIP4P/2005, however, perform significantly better than other water models, including the original parameterization of TIP4P.

Angular correlation coefficients are also generally larger for nitrogens and oxygens than for carbons. This observation is particularly interesting. The orientation of water molecules around surface nitrogens and oxygens indeed mostly results from direct hydrogen bond-

ing, while that of water molecules around carbons are affected by their interplay with the broader hydrogen-bond network and are thus less constrained. The resulting hydrophobicity is structurally more subtle to capture, which likely explains why water models do a relatively poorer job at capturing this effect (Fig. 4b). Water models that account more accurately for many-body correlations in water, such as E3B[51] and E3B2,[52] might improve the orientational description in these systems, but direct tests of this hypothesis are not immediately possible as these models have not yet been parameterized for macromolecular solvation.

By getting rid of most of the spatial coarsening the real-space distribution includes structural correlations of all orders. It thus generally gives rise to significantly lower correlations than either the radial or the angular correlation coefficients (Fig. 4). 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 correlation coefficients. Similarly, models reproduce the angular distribution around oxygens as well as around nitrogens, but are worse for their real-space correlation coefficients.

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 that is worsened by the decreasing number of grid points in that region. Protein-protein contacts, by contrast, show a fairly good structural agreement. This likely results from the surface atoms in these regions being much less mobile than the other surface atoms, and from steric constraints there playing a larger role in dictating the solvent structure in this regime. The position and orientation of water molecules in protein contacts are then less sensitive to water model parametrization and protein force fields, although even there agreement is not perfect.

We next consider the recall and prediction scores (Eq. 6) 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 the threshold density increases, the number of MD peaks decreases, but a greater fraction of the remaining ones overlap with crystal waters. 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 threshold density, and thus many crystal waters are not predicted by MD simulations. For all water models, the highest A_{pred} and A_{rec} is ~ 0.7. From this analysis, we conclude that there is a small yet significant contribution to the error in the water structure that arises from the peak location. 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.

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 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 we obtain $R_{\mathrm{work}} = 0.203(5)$ (SPC), which is significantly higher than $R_{\mathrm{work}} = 0.155(2)$ obtained for the refined protein model (Fig. 6). The difference in R_{work} is also greater at higher resolution, suggesting that highly ordered solvent regions are not captured adequately. If we instead substitute a flat electron density for the solvent region, $R_{\mathrm{work}} = 0.212(5)$, which is about 1% higher than the result for the best water model. Note that this increase is orders of magnitude larger than the estimated error in R_{work}, 4×10^{-4}. Hence, although the MD model contains some information about the water density within the unit cell, a significant share of its contribution is still lacking.

To check whether MD simulations capture solvent structure that is complementary to crystal waters assigned from the experimental data, we combine MD densities with the refined PDB coordinates including crystal waters. (The MD electron density where crystal waters are found is thus removed.) This strategy reduces R_{work} for the best water model to $R_{\mathrm{work}} = 0.158$, while the others have $R_{\mathrm{work}} = 0.159$. This is clearly better than the previous scheme yet still appreciably higher than the refined model $R_{\mathrm{work}} = 0.1537$. The gap between R_{work} values at lower resolutions is nonetheless then closed.

It is important to note that the value of R_{work} at high resolution is affected not only by the water density around each protein atom, but also by the fact that the average protein atom positions in the MD simulation is slightly different from the refined protein structure. Although refined protein structure 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 protein atom positions, which contributes to the discrepancy at high resolution.

Retaining crystal waters in the refined protein structure results in substantially lower R_{work} in high resolution bins compared to using only the MD-generated density. As resolution decreases, R_{work} becomes substantially worse than that of the original protein model, which once again confirms that the original protein model describes the electron density in the unit cell more accurately than the model with the MD solvent density.

3.3 Inferring Protonation States

A complication hindering the improvement of the solvent description in the analysis of X-ray diffraction experiments is that hydrogen atoms, which are surrounded by relatively small electron clouds, cannot be distinguished unless one achieves a remarkably high diffraction resolution, i.e., 0.7-1.0Å. Although the position of many of the hydrogens can be inferred based on an elementary description of bonding (partly explaining the success of structure validation tools, such as MolProbity[37]), side chain protonation states can remain some-

what ambiguous. This problem is especially important for side chains that contribute to an enzymatic pathway[3] or to protein-protein interactions, such as salt-bridges.[53–55] 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.[56,57] 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[43] analyses the hydrogen bonding network.[41] 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 strong 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 strong peaks in the MD solvent density.

In most cases considered here, either residues have insufficient solvent exposure to conduct the analysis, either no significant difference in solvation is observed, or both sets of density patterns are similarly incompatible with the refined structure. However, a few successful examples could be found. In the following cases, inverting the protonation state of a residue significantly affects the water distribution around the residues.

For LYS 145 on chain B, where removing one of the three protons from the default +1 charged lysine results in a slightly better overlap with two crystal waters, labeled 1 and 3 in Figure 7a. There is, however, a third crystal water labeled 2 within hydrogen bonding distance to the nitrogen atom that is unexplained by either protonation state. (The blue blob is behind the water and does not overlap it.) 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 crystal water 2. 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 farther 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.

ASP 200 on chain A is slightly more complicated. The MD peak resulting from the protonated case agrees better with crystal water 1, compared to the peak resulting from the simulation in which the residue is not protonated (Fig. 7b). However, two crystal waters (2 and 3) are overlapped by blue blobs byt no red blobs. It is therefore more likely that this residue is unprotonated, but it is unclear why the protonated case explains the peak on crystal water 1 better.

Similarly for GLU 130 on 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. 7c). Similarly, protonating GLU 218 in chain A results in the loss of an MD peak that overlaps crystal water 1, but retains those on crystal waters 2 and 3 (Fig. 7d). This is likely because crystal water 3 is still in hydrogen bonding distance to the residue, and crystal water 2 is hydrogen bonding distance to crystal water 3. It is however unclear why the peak on crystal water 1 disappears, as 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. In addition, the success of these inferences ultimately depend on our ability to reliably reconstruct the water density around proteins. Using this method to detect protonation states thus ultimately depend on being well above the noise inherent to our structural analysis.

4 Conclusions

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. 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 atoms, in particular, suggests that the hydrophobic effect is inadequately captured by these models. Remarkably, all water models we considered were found to behave rather similarly at the structural level. The result of the comparison in reciprocal space suggests that despite the limitations, water models do capture structural information that is complementary to direct refinement of the X-ray data.

Although MD water models are insufficient for reconstructing biomolecular hydration with a precision sufficient to conduct structural refinement, they nonetheless capture a fraction of crystal waters. In optimal hydration circumstances, these models may thus suggest the assignment of a protonation states to some side chains. The robustness of these predictions with respect to the choice of variety of parameters, including the protein force field and the protonation state of the nearby residues, is untested. **Pavel: please add a sentence about how this work might serve as a source of inspiration for improving refinement in the future.**

Our results suggest that it may be necessary to add more features to the common water models in order to reconstruct accurately the water 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. Whether or not a re-parametrization can capture both bulk and interfacial water properties, however, relies on whether both these behaviors can be captured with a single, fixed dipole moment.[58] Considering more complex models that include polarizibility[7] or include three-body interac-

tions,[59] might thus provide a more robust starting point. To model a process that depends sensitively on the position of water molecules, it might thus be preferable to consider even higher-accuracy models of water that include *ab initio* descriptions. The use of such models might be sufficient to improve more directly the structural refinement process in the future.

References

(1) Kim, E.; Baker, C.; Dwyer, M.; Murcko, M.; Rao, B.; Tung, R.; Navia, M. *Journal of the American Chemical Society* **1995**, *117*, 1181–1182.

(2) Balius, T. E.; Fischer, M.; Stein, R. M.; Adler, T. B.; Nguyen, C. N.; Cruz, A.; Gilson, M. K.; Kurtzman, T.; Shoichet, B. K. *Proceedings of the National Academy of Sciences* **2017**, 201703287.

(3) 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. *Proceedings of the National Academy of Sciences* **2015**, 201504986.

(4) Voets, I. K. *Soft Matter* **2017**,

(5) Skinner, L. B.; Huang, C.; Schlesinger, D.; Pettersson, L. G.; Nilsson, A.; Benmore, C. J. *The Journal of chemical physics* **2013**, *138*, 074506.

(6) Paesani, F.; Voth, G. A. *The Journal of Physical Chemistry B* **2009**, *113*, 5702–5719.

(7) Vega, C.; Abascal, J. L. *Physical Chemistry Chemical Physics* **2011**, *13*, 19663–19688.

(8) Kim, C. U.; Tate, M. W.; Gruner, S. M. *Proceedings of the National Academy of Sciences* **2015**, *112*, 11765–11770.

(9) Sanz, E.; Vega, C.; Abascal, J.; MacDowell, L. *Physical review letters* **2004**, *92*, 255701.

(10) Finney, J.; Hallbrucker, A.; Kohl, I.; Soper, A.; Bowron, D. *Physical review letters* **2002**, *88*, 225503.

(11) Chandler, D. *Nature* **2005**, *437*, 640–647.

(12) Ball, P. *Chemical reviews* **2008**, *108*, 74–108.

(13) Macias-Romero, C.; Nahalka, I.; Okur, H. I.; Roke, S. *Science* **2017**, eaal4346.

(14) Guillot, B. *Journal of Molecular Liquids* **2002**, *101*, 219–260.

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

(16) Berendsen, H.; Grigera, J.; Straatsma, T. *Journal of Physical Chemistry* **1987**, *91*, 6269–6271.

(17) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *The Journal of chemical physics* **1983**, *79*, 926–935.

(18) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *The Journal of chemical physics* **2004**, *120*, 9665–9678.

(19) Abascal, J. L.; Vega, C. *The Journal of chemical physics* **2005**, *123*, 234505.

(20) Mahoney, M. W.; Jorgensen, W. L. *The Journal of Chemical Physics* **2000**, *112*, 8910–8922.

(21) Weichenberger, C. X.; Afonine, P. V.; Kantardjieff, K.; Rupp, B. *Biological Crystallography* **2015**, *71*, 1023–1038.

(22) Weichenberger, C. X.; Rupp, B. *Acta Crystallographica Section D: Biological Crystallography* **2014**, *70*, 1579–1588.

(23) Jones, N. *Nature* **2014**, *505*, 602–603.

(24) others,, et al. *Nature* **2011**, *470*, 73–77.

(25) Drenth, J. *Principles of protein X-ray crystallography*; Springer Science & Business Media, 2007.

(26) Myles, D. A. *Current opinion in structural biology* **2006**, *16*, 630–637.

(27) Nannenga, B. L.; Gonen, T. *Current opinion in structural biology* **2014**, *27*, 24–31.

(28) Nannenga, B. L.; Shi, D.; Leslie, A. G.; Gonen, T. *Nature Methods* **2014**, *11*, 927–930.

(29) Taylor, G. L. *Acta Crystallographica Section D: Biological Crystallography* **2010**, *66*, 325–338.

(30) Free, R. *Nature* **1992**, *355*, 472–5.

(31) Wlodawer, A.; Minor, W.; Dauter, Z.; Jaskolski, M. *Febs Journal* **2008**, *275*, 1–21.

(32) Burling, F. T.; Weis, W. I.; Flaherty, K. M.; Brünger, A. T. *Science* **1996**, *271*, pp. 72–77.

(33) Holton, J. M.; Classen, S.; Frankel, K. A.; Tainer, J. A. *FEBS Journal* **2014**, *281*, 4046–4060.

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

(35) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic acids research* **2000**, *28*, 235–242.

(36) others,, et al. *Acta Crystallographica Section D: Biological Crystallography* **2010**, *66*, 213–221.

(37) 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. *Acta Crystallographica Section D: Biological Crystallography* **2010**, *66*, 12–21.

(38) Lang, P. T.; Holton, J. M.; Fraser, J. S.; Alber, T. *Proceedings of the National Academy of Sciences* **2014**, *111*, 237–242.

(39) Vega, C.; McBride, C.; Sanz, E.; Abascal, J. L. *Physical Chemistry Chemical Physics* **2005**, *7*, 1450–1456.

(40) Higo, J.; Nakasako, M. *Journal of computational chemistry* **2002**, *23*, 1323–1336.

(41) Lemkul, J. personal exchange, 2016.

(42) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712–25.

(43) Berendsen, H. J.; van der Spoel, D.; van Drunen, R. *Computer Physics Communications* **1995**, *91*, 43–56.

(44) Bertrand, C. E.; Zhang, Y.; Chen, S.-H. *Physical Chemistry Chemical Physics* **2013**, *15*, 721–745.

(45) Grosse-Kunstleve, R. W.; Sauter, N. K.; Moriarty, N. W.; Adams, P. D. *Journal of Applied Crystallography* **2002**, *35*, 126–136.

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

(47) Riley, K. F.; Hobson, M. P.; Bence, S. J. *Mathematical Methods for Physics and Engineering*; Cambridge University Press, 2006; p 1200.

(48) Lounnas, V.; Pettitt, B.; Phillips Jr, G. *Biophysical journal* **1994**, *66*, 601.

(49) Lin, B.; Pettitt, B. M. *The Journal of chemical physics* **2011**, *134*, 106101.

(50) Virtanen, J. J.; Makowski, L.; Sosnick, T. R.; Freed, K. F. *Biophysical journal* **2010**, *99*, 1611–1619.

(51) Tainter, C.; Pieniazek, P.; Lin, Y.-S.; Skinner, J. *The Journal of chemical physics* **2011**, *134*, 184501.

(52) Tainter, C. J.; Shi, L.; Skinner, J. L. *Journal of chemical theory and computation* **2015**, *11*, 2268–2277.

(53) Anderson, D. E.; Becktel, W. J.; Dahlquist, F. W. *Biochemistry* **1990**, *29*, 2403–2408.

(54) Dey, M.; Cao, C.; Sicheri, F.; Dever, T. E. *Journal of Biological Chemistry* **2007**, *282*, 6653–6660.

(55) Fusco, D.; Headd, J. J.; De Simone, A.; Wang, J.; Charbonneau, P. *Soft matter* **2014**, *10*, 290–302.

(56) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. *Nucleic acids research* **2005**, *33*, W368–W371.

(57) Rostkowski, M.; Olsson, M. H.; Søndergaard, C. R.; Jensen, J. H. *BMC structural biology* **2011**, *11*, 1.

(58) Yu, H.; van Gunsteren, W. F. *Computer Physics Communications* **2005**, *172*, 69–85.

(59) Cisneros, G. A.; Wikfeldt, K. T.; Ojamae, L.; Lu, J.; Xu, Y.; Torabifard, H.; Bartok, A. P.; Csanyi, G.; Molinero, V.; Paesani, F. *Chemical reviews* **2016**, *116*, 7501–7528.

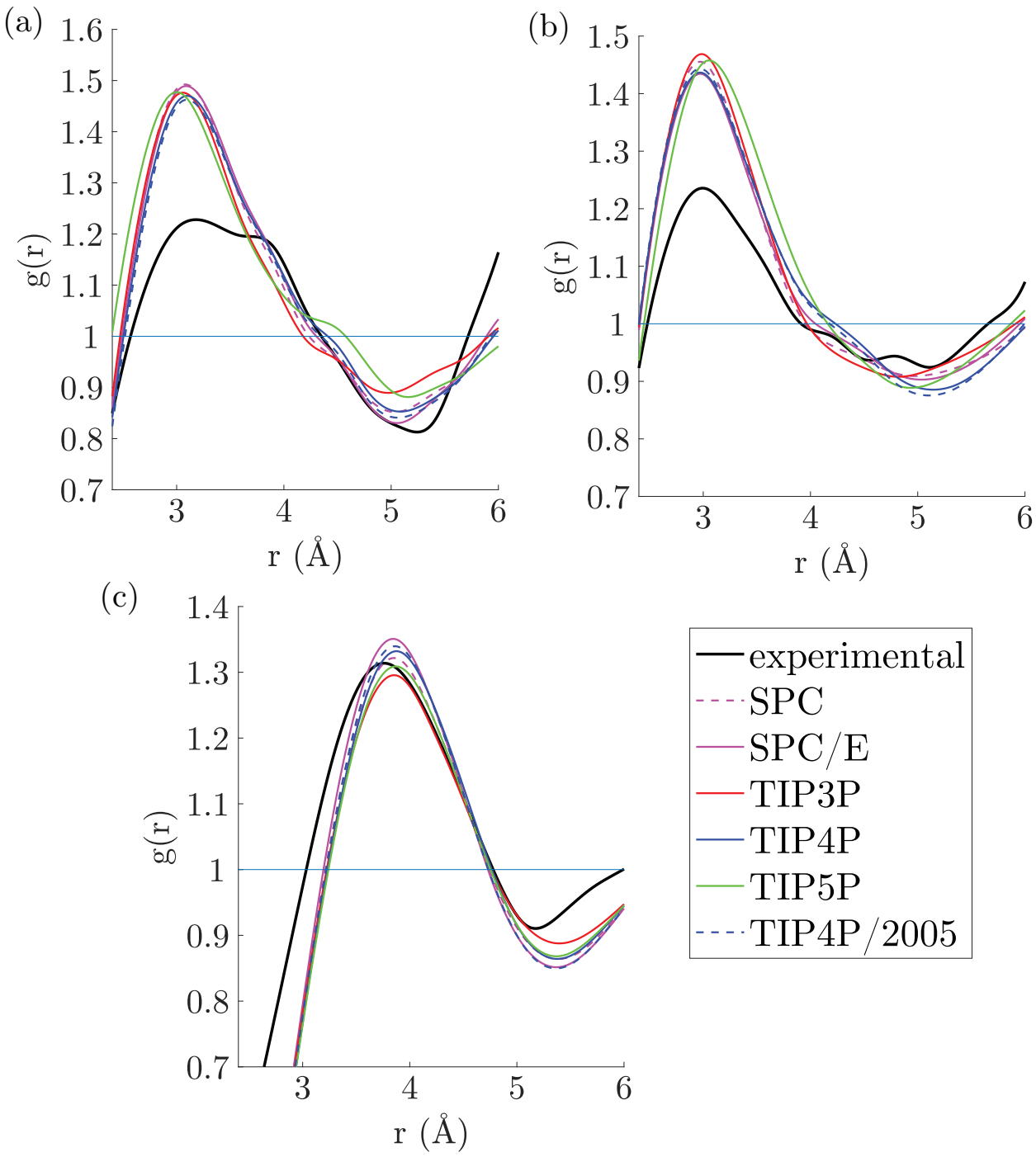

Figure 3: Averaged RDFs for surface (a) N, (b) O, and (c) C atoms, for different water models. Results obtained from different water models agree well with each other, as well as with the experimental RDFs.

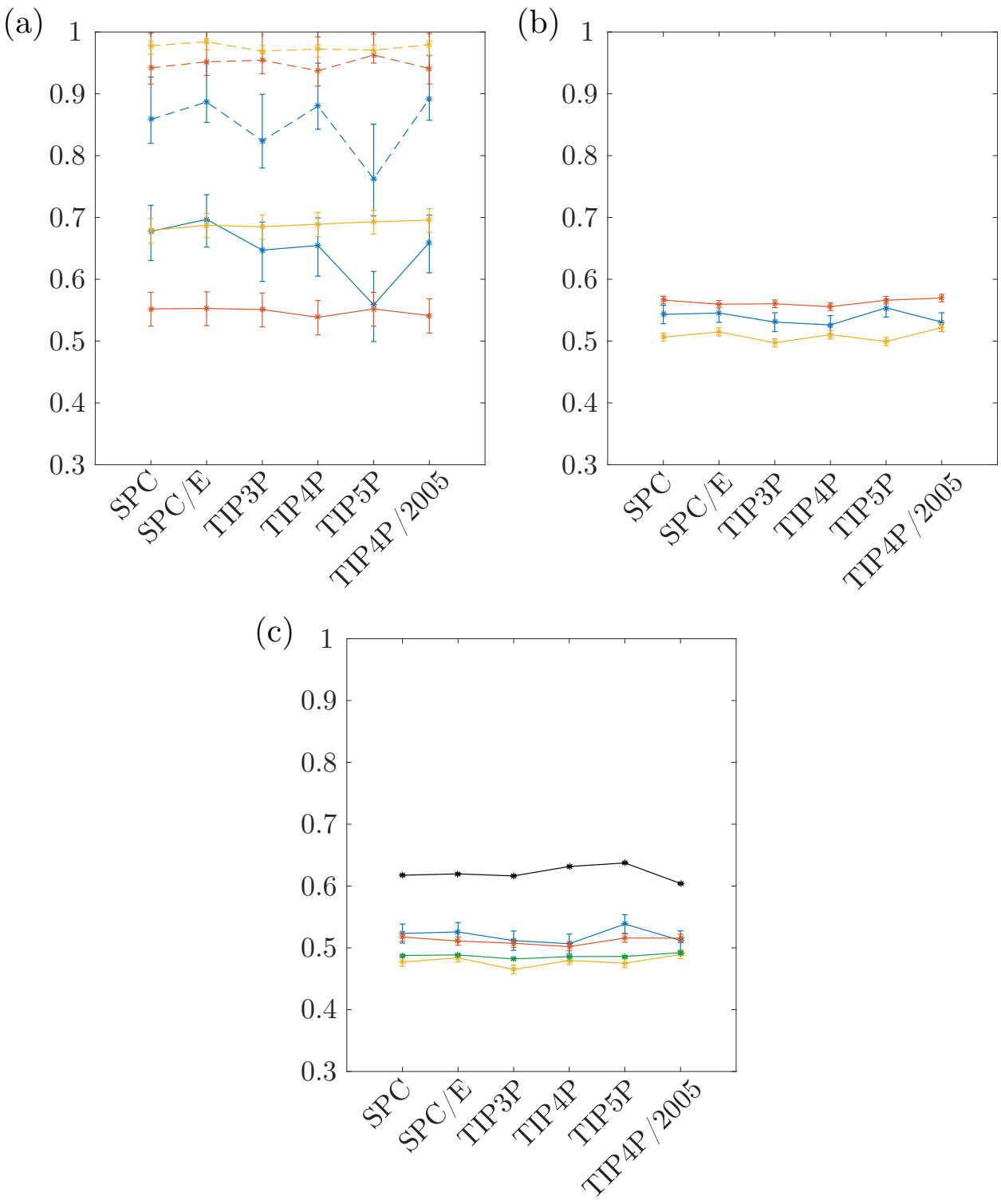

Figure 4: (a) Radial (solid) and averaged radial (dashed), (b) angular, and (c) spatial correlation coefficients for surface N (blue), O (red) and C (yellow) atoms. Real-space correlation coefficients for the first layer (green) and contact waters (black) are also given in (c). Error bars denote 95% confidence interval.

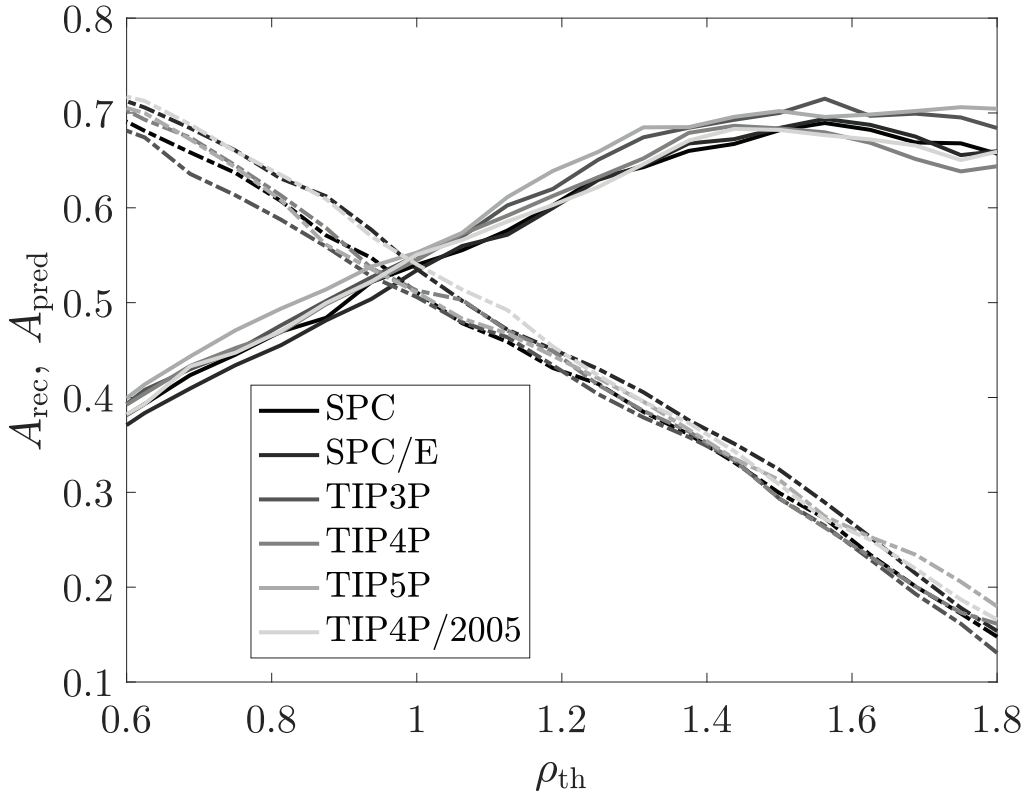

Figure 5: Prediction (solid) and recall scores (dashed), as defined in Eq. 6. At low threshold densities, too many MD peaks are identified, resulting in a high recall score but a low prediction score. As the threshold increases, MD peaks with stronger signals persist, which at high densities predict roughly 70% of the crystal waters. However, the recall scores fall as the density increases, suggesting that there is still a significant fraction of crystal waters that do not overlap with an MD peak.

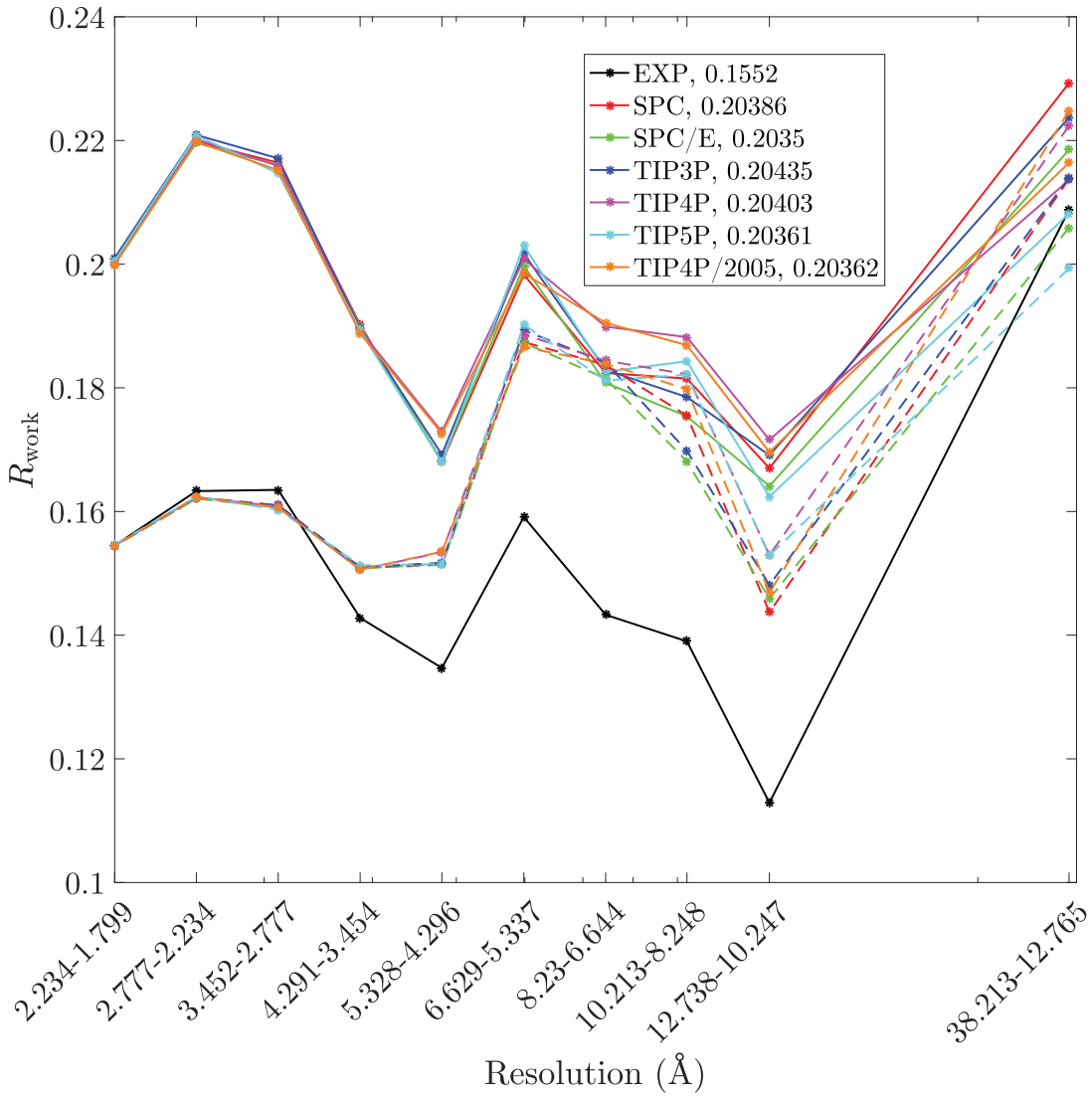

Figure 6: R_{work} in different resolution bins for the original model (EXP, black), and for models constructed by combining MD densities with the protein model. The overall R_{work} values are as given in the legend. Dashed lines show R_{work} when the crystal waters in the refined protein model are retained while combining it with the MD density.

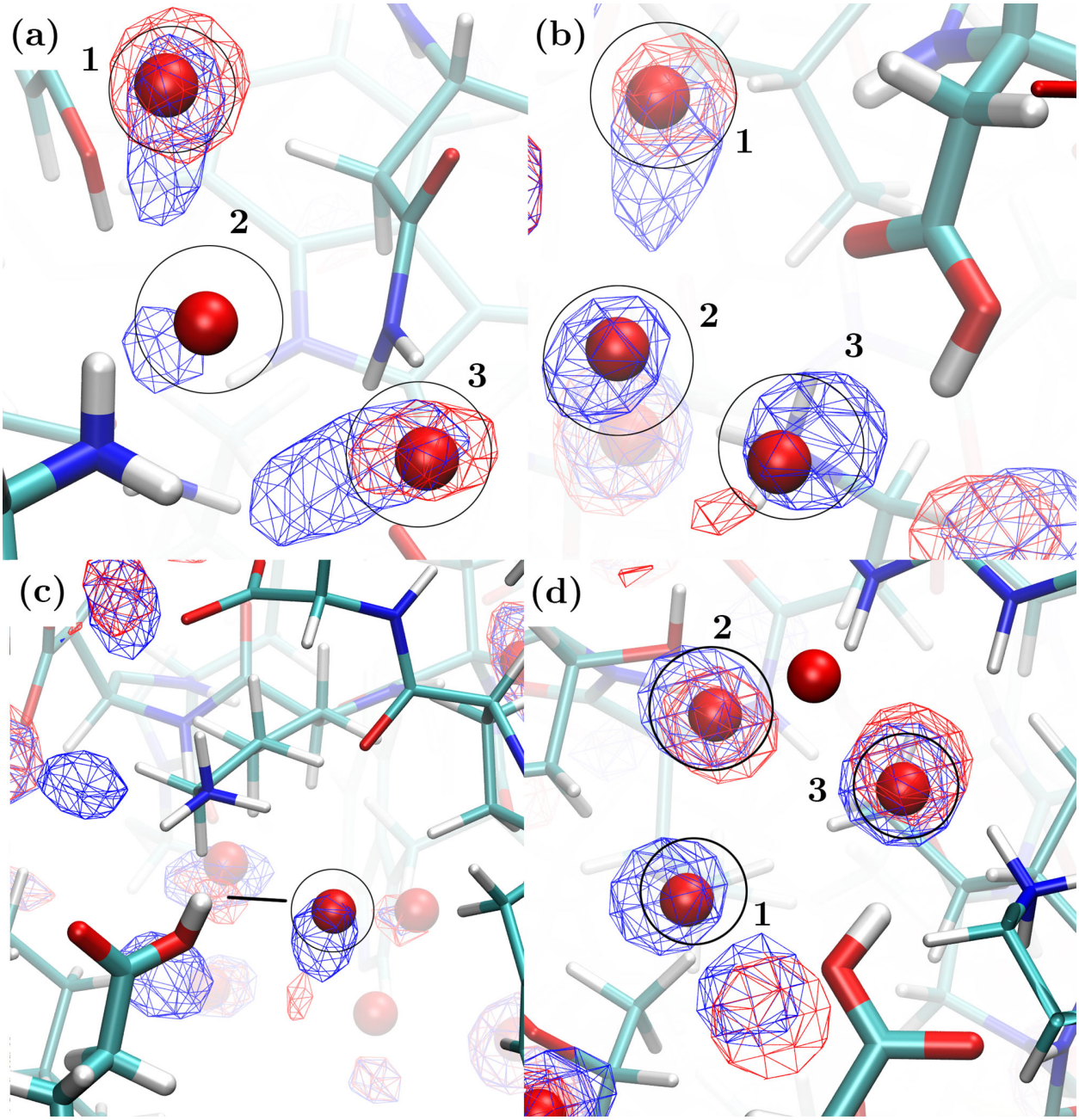

Figure 7: Comparison of water density distribution for simulations that contain different protonation states for (a) LYS 145 in chain A, (b) ASP 200 in chain A, (c) GLU 130 in chain B, and (d) GLU 218 in chain A. The water density from the default protonation state simulations are shown in blue wireframe, 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 from the refined protein structure are shown with red spheres.