Biophysical Journal
Article

Biophysical Society

LAWS: Local alignment for water sites—Tracking ordered water in simulations

Eugene Klyshko,[1,2] Justin Sung-Ho Kim,[1,2] and Sarah Rauscher[1,2,3,*]
[1]Department of Physics, University of Toronto, Toronto, Ontario, Canada; [2]Department of Chemical and Physical Sciences, University of Toronto Mississauga, Mississauga, Ontario, Canada; and [3]Department of Chemistry, University of Toronto, Toronto, Ontario, Canada

ABSTRACT Accurate modeling of protein-water interactions in molecular dynamics (MD) simulations is important for understanding the molecular basis of protein function. Data from x-ray crystallography can be useful in assessing the accuracy of MD simulations, in particular, the locations of crystallographic water sites (CWS) coordinated by the protein. Such a comparison requires special methodological considerations that take into account the dynamic nature of proteins. However, existing methods for analyzing CWS in MD simulations rely on global alignment of the protein onto the crystal structure, which introduces substantial errors in the case of significant structural deviations. Here, we propose a method called local alignment for water sites (LAWS), which is based on multilateration—an algorithm widely used in GPS tracking. LAWS considers the contacts formed by CWS and protein atoms in the crystal structure and uses these interaction distances to track CWS in a simulation. We apply our method to simulations of a protein crystal and to simulations of the same protein in solution. Compared with existing methods, LAWS defines CWS characterized by more prominent water density peaks and a less-perturbed protein environment. In the crystal, we find that all high-confidence crystallographic waters are preserved. Using LAWS, we demonstrate the importance of crystal packing for the stability of CWS in the unit cell. Simulations of the protein in solution and in the crystal share a common set of preserved CWS that are located in pockets and coordinated by residues of the same domain, which suggests that the LAWS algorithm will also be useful in studying ordered waters and water networks in general.

SIGNIFICANCE Protein-water interactions are fundamental to protein function. X-ray crystallography provides a high-resolution protein structure and positions of well-ordered water molecules that can be compared with time-resolved MD simulations. However, methods accounting for protein flexibility are needed to carry out a rigorous comparison between simulation and experiment. With these considerations, we developed an approach that tracks water sites relative to the local protein motion and determines if crystallographic waters are preserved in a simulation based on the local density of water. The LAWS algorithm is general and can be extended to tracking any arbitrary location in a simulation relative to reference points. Our approach could be applied to studying ordered water networks, which are important in the function of many proteins.

INTRODUCTION

Water is an essential component of biomolecular systems; it affects the structure and stability of biological machinery through the hydrophobic effect, hydrogen bonding, and polar interactions (1). Protein-water interactions are important driving forces in dynamic processes, such as protein folding, self-assembly, and binding (2,3). In addition, solvation is essential for protein function, since water networks have been shown to play a crucial role in the motion of protein domains (1,4–6). All-atom molecular dynamics (MD) simulations can provide information on the molecular basis for protein function, but require high-resolution structural information (1,7).

X-ray crystallography is the most commonly used experimental technique to obtain high-resolution protein structures found in the Protein Data Bank (PDB) (8). Crystal structures often contain crystallographic water sites (CWS), which are locations of high water density in the lattice. These CWS represent ordered water molecules, which are stabilized by noncovalent interactions with the protein, while bulk (unordered) water molecules comprise the remaining space between the protein chains. Past studies

Submitted May 16, 2022, and accepted for publication September 13, 2022.
*Correspondence: sarah.rauscher@utoronto.ca
Eugene Klyshko and Justin Sung-Ho Kim contributed equally to this work.
Editor: Lucie Delemotte.
https://doi.org/10.1016/j.bpj.2022.09.012

have used the preservation of CWS as a means to assess the accuracy of protein-water interactions in simulations (9–13). MD simulations can probe the dynamics of protein crystals because periodic boundary conditions mimic the periodic nature of crystal lattices (14–16). An accurate modeling of protein-water interactions implies that the protein structure along with the crystallographic waters should be well preserved in MD simulations of crystals. Since x-ray crystallography provides an ensemble-averaged conformation of the protein and CWS, a direct comparison between time-resolved MD trajectories and experiment requires special methodological considerations.

The analysis of CWS in MD simulations has been carried out using a variety of methods that can be broadly classified into two categories—density-based and coordinate-based methods (Fig. 1). In density-based methods (9,10,12,13,17–20), the time-averaged atomic density of the solvent molecules is computed, where high-density peaks correspond to CWS and are compared with the experimental CWS positions (Fig. 1, *A* and *C*). In contrast, the coordinate-based methods (11,13,18,21,22) analyze the explicit positions of the water molecules in each MD frame in the vicinity of CWS (Fig. 1, *B* and *D*). To identify if CWS are preserved in simulations, their locations are probed for the presence of water molecules. We refer to these locations in the MD trajectory as water sites (WS); they can be tracked in various ways depending on whether the protein conformation is globally or locally aligned to the crystal structure. When using global alignment (Fig. 1, *A* and *B*), the entire protein is superimposed onto the crystal structure and the experimental positions of the CWS represent the WS in a frame. In the local alignment approach, WS are

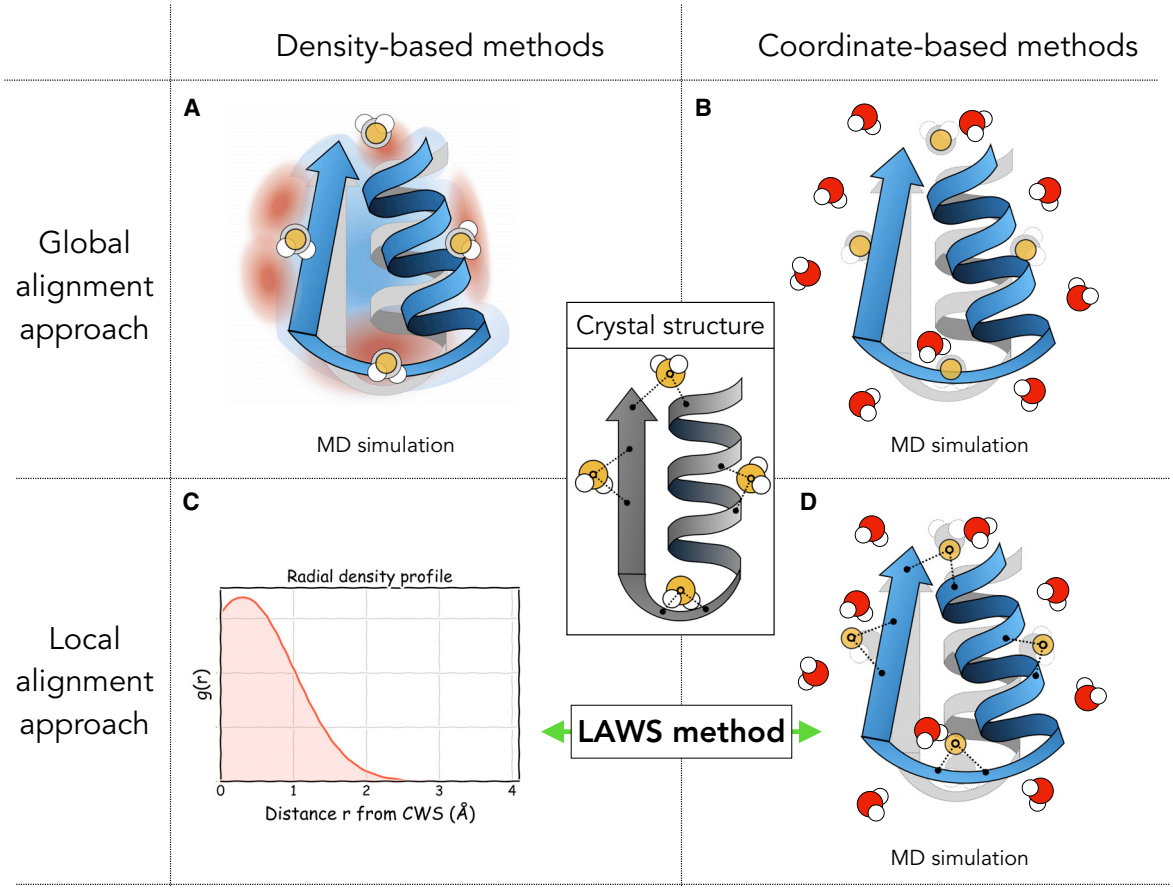

FIGURE 1 A schematic illustration of existing methods for the analysis of crystallographic water in MD simulations. These methods are based on either average MD density maps or explicit MD coordinates, and can rely on global (*A* and *B*) or local (*C* and *D*) alignment of the protein. The crystal structure (*central panel*) provides the experimental positions of CWS (*yellow spheres*). Density-based methods (*A* and *C*) analyze an average atomic density of water in an MD simulation and then compare CWS with the peaks in the density (*red*). Coordinate-based methods (*B* and *D*) analyze explicit positions of water molecules in every frame and measure the occupancy of each CWS, which can be done using either global (*B*) or local (*D*) alignment of the protein structure to the crystal structure. Global alignment approaches (*A* and *B*) superimpose the configuration of the protein from an MD simulation (*blue*) onto the static crystal structure (*gray*), which can result in clashes between CWS and protein atoms. Local alignment approaches (*D*) address this issue by only considering local protein regions in the vicinity of the CWS. The method proposed here—local alignment for water sites (LAWS)—is a coordinate-based algorithm relying on a local alignment approach, which additionally provides density-type data, including the local radial density profile (*C*) or density maps (Fig. S5). Therefore, LAWS can be classified as both a coordinate-based and a density-based method (*C* and *D*). To see this figure in color, go online.

defined using only a local region of the protein in the vicinity of the experimental CWS (Fig. 1 *D*).

In the analysis of CWS, a quantity of interest is the proportion of CWS preserved in a simulation, referred to as *recall statistics* (9–13,23). It has been demonstrated that density-based and coordinate-based methods produce comparable recall of CWS when global alignment is employed (13). However, the global alignment approach has one notable limitation; it is only suitable for analysis of MD simulations in which the protein does not deviate significantly from the crystal structure. Previous simulations were typically either short (tens of ns) or restrained (9–13). For example, Wall et al. showed that 93% of all CWS were preserved within 1.4 Å of their experimental positions in a simulation with position restraints, while only 46% of CWS were preserved when position restraints were removed, suggesting that the deviation of a protein from its crystal structure may contribute to the loss of CWS in simulations (12). It has been observed in simulations and experiments that, even in a crystal environment, proteins are flexible and can undergo conformational changes at sub-μs timescales (24–26). The dynamic nature of proteins motivates the local alignment approach.

To characterize the water structure in simulations to predict potential CWS, Henchman and McCammon developed a local alignment approach using backbone atoms (18). However, side-chain atoms also contribute to the coordination of crystallographic waters. Caldararu et al. proposed a method that includes local side-chain atoms in the alignment to track WS. Their method relies on clustering coordinates of the water molecules located near WS (13). However, clustering algorithms have their limitations as they are data driven and can be sensitive to the choice of parameters (27). A natural way to analyze CWS in MD simulations is by using a measure of density, as these data are probed in x-ray crystallography experiments.

Here, we propose a method called local alignment for water sites (LAWS) to assess the preservation of experimentally refined crystallographic waters in MD simulations. LAWS is based on a widely used algorithm in GPS navigation called multilateration (28) to track WS relative to nearby protein atoms in an MD trajectory. Instead of relying on absolute coordinates of CWS in the crystal structure, which is what conventional global alignment approaches do, LAWS defines WS at specified distances from nearby protein residues. The LAWS algorithm makes it possible to optimally compute such positions in space that would correspond to the positions seen in the crystal structure. First, we apply LAWS and the conventional global alignment approach to an MD simulation of a protein unit cell. We demonstrate that our approach characterizes CWS with a higher density of water compared with global alignment by reducing alignment errors. We also explore how various properties of the CWS, such as the experimental uncertainty (B-factors) and the location relative to the protein, affect the recall of CWS in the simulation. Finally, by analyzing the simulations of the same protein in solution, we discuss the contribution of interchain crystal contacts to the stability of CWS.

METHODS

The LAWS algorithm

In a crystal structure, each CWS can be characterized by a set of contacts with nearby protein atoms, referred to as "coordination." When similar interactions are maintained in a simulation, the position of the crystal water would be preserved over time. Therefore, with the LAWS algorithm, we track WS (corresponding to CWS in a crystal structure) in a simulation using these reference interaction distances and compute the density of water molecules around each WS (*yellow spheres* in Fig. 2 *A*).

Tracking WS in a simulation using LAWS

The first step is to find the atoms that coordinate each CWS in the crystal structure. We use the atomic coordinates from the crystal structure to determine a set of n protein atoms, A_i, with indices $i = 1, \ldots, n$, as well as the distances, $\widehat{d}_i$, between the CWS and these protein atoms A_i (dashed lines in the central panel of Fig. 1).

Now, knowing the coordination of each CWS, we can track the corresponding WS in each simulation frame. Let x, y, and z be the unknown coordinates of a WS in a simulation frame. The set of known distances $\widehat{d}_i$ to the nearby protein atoms can be used to find the unknown coordinates of the WS even when the protein atoms A_i change position x_i, y_i, z_i throughout the trajectory. In other words, for every frame, we are aiming to find the position x, y, z of the WS that satisfies the n distance equations:

$$\widehat{d}_i = \left((x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2\right)^{\frac{1}{2}} \quad (i = 1, 2, \ldots, n). \tag{1}$$

This problem is formulated in the literature as multilateration (28), which is commonly encountered in navigation and surveillance. For example, a GPS device calculates distances $\widehat{d}_i$ by measuring the times required for a signal to travel from a set of satellites. Since the position of each satellite x_i, y_i, z_i is known at any given time, the coordinates of the GPS device x, y, z can be determined by solving a system of Eq. (1). By analogy, the reference protein atoms act as the satellites and the GPS device represents a water site with a location that needs to be determined.

The system of nonlinear equations (1) has three unknowns and n equations. In theory, the exact solution of the equation can be found uniquely if $n = 4$ positions and distances are provided—analogous to identifying the point of intersection of four spheres with known radii and positions of the centers in 3D space. However, an ambiguity of the solution is possible when any two sphere centers and the unknown point are collinear. In that case, more spheres are required for a unique solution. On the other hand, Eq. 1 is not guaranteed to have a solution as the spheres may not intersect in every case. Thus, due to the stochastic nature of protein motions and the fact that it is not possible to find an exact solution in every instance, an *optimal* solution is required instead.

We define the LAWS error as the weighted sum of squares:

$$\mathrm{LAWS}(x, y, z) = \sum_{i=1}^{n} w_i^2 \Big(\left((x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2\right)^{\frac{1}{2}} - \widehat{d}_i \Big)^2, \tag{2}$$

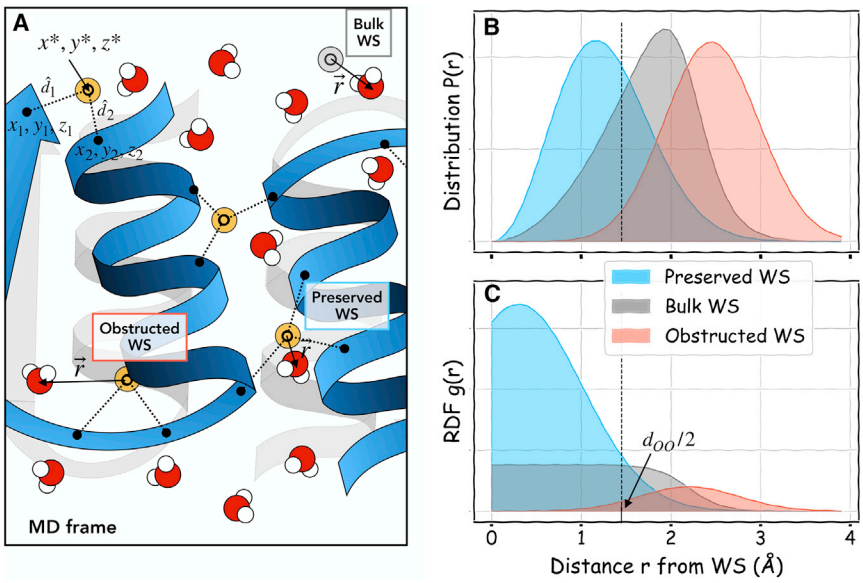

FIGURE 2 The LAWS pipeline. (*A*) A cartoon representation of one MD frame containing positions of protein and water molecules, overlapping with a crystal structure in gray. Yellow points represent WS that are tracked by the LAWS algorithm based on the distances $\hat{d}_i$ (obtained from the crystal structure) to reference protein atoms. Note that contacts can be formed by multiple protein chains. To check if a CWS is preserved in MD, we analyze how often the corresponding WS is occupied by water molecules. Bulk WS (shown in *gray*) are used as a control. To find the occupancy for each WS, we compute the vector $\vec{r}$ and distance $r = |\vec{r}|$ to the nearest water molecule at each frame and estimate the distribution $P(r)$ and the corresponding radial distribution function $g(r)$. Depending on $P(r)$ and $g(r)$, each WS can be classified into three groups. (*B*) An example of the distance distribution $P(r)$ for a preserved WS, bulk WS, and obstructed WS, shown in (*A*). (*C*) The corresponding radial distribution functions $g(r) = P(r)/r^2$. A radial density around a preserved WS has a peak close to zero. For a bulk WS the radial density must be uniform from zero to $d_{OO}/2$ (shown as a *dashed line*), where d_{OO} is the average distance between nearest neighbor oxygen atoms in bulk water. An obstructed WS has a radial density peak at $r > d_{OO}/2$. To see this figure in color, go online.

where w_i are the weights of each atom A_i coordinating the water site, such that $\sum_{i=1}^{n} w_i^2 = 1$. This function is then minimized to find the optimum position x^*, y^*, z^* of the water site in a given simulation frame:

$$x^*, y^*, z^* = \underset{x,y,z}{\operatorname{argmin}} \operatorname{LAWS}(x, \ y, \ z). \tag{3}$$

This procedure is the weighted least-squares problem for which we used a Python implementation of the Levenberg-Marquardt algorithm (29).

The motivation for using weights comes from the susceptibility of the least-squares method to outliers. Protein atoms closer to the CWS in the crystal have a higher contribution to the interaction energy than atoms further away. Hence, they have greater weights in the LAWS error (Eq. 2), which increases the robustness of the algorithm. Importantly, the value of the LAWS error at the optimum x^*, y^*, z^* shows a quantitative estimate of how much the local protein region is perturbed relative to the crystal structure. The exact solution will have a LAWS error of zero. Therefore, the LAWS error is a measure of the displacement of the WS from its ideal position observed in the crystal structure.

The application of this tracking algorithm can be extended beyond the scope of the current work, namely, tracking crystallographic water in simulations. It can be generalized to tracking various types of molecules relative to an arbitrary reference. When applied to a small molecule with a single functionally important atom or an ion, the approach is analogous to the one described here. The reference distances can be extracted from a crystal structure. However, applying LAWS to track larger molecules, such as a lipid, polyethylene glycol, or a molecule with multiple functional groups, requires additional considerations. When more than one atom is being tracked, the LAWS function (Eq. 2) will have additional terms corresponding to each atom, as well as distance constraints. The optimization problem will resolve multiple positions: (x^1, y^1, z^1), (x^2, y^2, z^2), and so on.

Parameters

When applying the LAWS algorithm to an MD trajectory, the number of protein atoms, n, coordinating the CWS and the weights of each atom, w_i, must be determined. These parameters are computed only once, at the start of the algorithm, and are based on the positions in the crystal structure.

There is a trade-off between computational efficiency and robustness of the numerical algorithm that depends on the number of protein atoms, n, used to coordinate each CWS. A larger n increases the computational cost of solving Eq. 3, while a smaller n makes the algorithm less robust to outliers. The value of n depends on the cutoff distance defining a contact. We use a cutoff distance of 4.5 Å for heavy protein atoms and set a maximum of $n = 10$. If there are fewer than four protein atoms in contact with the CWS, we increase the cutoff distance until we find at least $n = 4$. Therefore, n varies between 4 and 10 depending on the CWS. Importantly, we take into consideration protein atoms from multiple symmetrically related molecules if the CWS is located at the interface between molecules in the unit cell (Fig. 2 *A*). We also tested the LAWS algorithm with the same fixed number of heavy atoms n for all CWS in a range from 5 to 8 and the results were not significantly affected by this parameter choice. Once n reference protein atoms are determined, the distances $\hat{d}_i$ can be computed and used in Eq. 2.

The choice of weights w_i in Eq. 2 is motivated as follows: the protein atoms located closer to the WS contribute more to its coordination. The protein-water interaction energy depends on the distance, r, between atoms, where $\alpha \geq 1$ for various noncovalent interactions. We chose the weights w_i to be proportional to the interaction energy and hence inversely proportional to the distance between the CWS and a protein atom in the crystal structure ($\propto 1/\hat{d}_i$, i.e., $\alpha = 1$). We tested other values of α ($\alpha = 2, \ldots, 6$) and this did not affect the results significantly. The normalized weights are therefore given by:

$$w_i^2 = \frac{(1/\hat{d}_i)^2}{\sum_{i=1}^{n} (1/\hat{d}_i)^2}. \tag{4}$$

Computing density profiles of WS

To estimate how frequently a water molecule occupies any given WS, we compute the offset vector $\vec{r}$ and the distance r from a WS to the oxygen of the nearest water molecule at each frame. The distance distribution $P(r)$ (averaged over all symmetrically related copies) then provides a measure of the occupancy of the WS in the trajectory, as it shows the fraction of nearest neighbor water molecules found in the spherical shell with radius r

from the WS (Fig. 2 *B*). Since the probability to find water in a spherical shell with radius r is proportional to the area of a sphere, $P(r) \propto 4\pi r^2$, we normalize the distribution $P(r)$ by r^2 to define a radial distribution function (RDF). Then, $g(r) = P(r)/r^2$ provides a density profile as a function of the distance r from the WS (Fig. 2 *C*). As an alternative approach, offset vectors $\vec{r}$ around each WS can be used to generate local 3D density maps with, for example, GROma*ps* software (30). The details of this approach are presented in supporting material (Section S7).

Bulk WS as control

A CWS that is preserved in a simulation is expected to have on average a higher occupancy than a WS in a bulk water region, which can be treated as a reference for comparison (31). Bulk WS are defined as positions within the unit cell where water molecules are not significantly influenced by the interaction with the protein. Water was experimentally observed to have bulk-like properties (in terms of the lifetime of hydrogen bonds) at least 6 Å from the protein surface (32). Based on this observation, we checked if randomly sampled positions in a unit cell located at least 6 Å from the protein (*gray spheres* in Fig. 2 *A*) would display bulk water properties. We ran a 500 ns simulation of a box of water to model ideal bulk water behavior (Section S1). Analyzing the $P(r)$ distribution for randomly sampled locations in the box, we found that the bulk water $P(r)$ was statistically identical to the $P(r)$ sampled at 6 Å from the protein in our MD simulation of a unit cell (Fig. S1). Therefore, water molecules located >6 Å from the protein surface exhibit bulk behavior, and hence the $P(r)$ (and corresponding $g(r)$) from the unit cell simulation can be used as the control bulk distribution.

The radial density around a bulk WS should be uniform from zero to $d_{OO}/2$ (Fig. 2 *C*), where d_{OO} is the average interoxygen distance between nearest neighbor water molecules in liquid water. Hence, it is equally likely to find the nearest neighbor water molecule at any distance $r \leq d_{OO}/2$ from a bulk WS. When $r > d_{OO}/2$, the probability to observe a nearest neighbor water becomes negligible as r approaches d_{OO}. The experimental estimate of d_{OO} was reported to be in the range of 2.7–3.0 Å (33,34). We estimated a value of $d_{OO} = 2.8$ Å using the $g(r)$ from our simulations, consistent with previous MD studies (10,12).

Preserved, bulk-like, and obstructed WS in a simulation

Once the bulk control is established, a comparison can be made between $P(r)$ of each WS with the control bulk distribution (Fig. 2 *B*). There are three possible outcomes of this comparison. 1) If a WS is *preserved* in a simulation, the waters are distributed more closely to the WS, and the $P(r)$ is expected to be more left-shifted relative to the control (Fig. 2 *B*, *blue*). 2) Conversely, if the nearest waters are distributed away from the WS, $P(r)$ would be right-shifted and considered to be *obstructed* (Fig. 2 *B*, *red*). Obstructed WS occur when the protein or another cosolute occupies this space, obstructing access of water molecules. 3) Finally, if $P(r)$ is similar to the bulk water control, this WS would be considered *bulk-like* (Fig. 2 *B*, *gray*). Both obstructed and bulk-like WS correspond to CWS in the crystal structure that are not preserved (i.e., lost) in a simulation.

Here, we classify WS as preserved, bulk-like, or obstructed, using the RDF $g(r) = P(r)/r^2$. The radial density profile for a preserved WS is expected to have a higher peak, the radial density for a bulk-like WS should be uniform, while the radial density of an obstructed WS is less than the bulk control (Fig. 2 *C*). The metric we used to compare the RDF with the bulk control is the integral $\int_0^{d_{OO}/2} g(r)\mathrm{d}r$ of the RDF in the interval from 0 to $d_{OO}/2$. First, we computed this integral for randomly sampled bulk WS to estimate a range (I_{MIN}, I_{MAX}) for comparison. For our system, we sampled 120 different bulk water locations and found that the range of the $g(r)$ integrals was between $I_{\mathrm{MIN}} = 0.47\text{Å}^{-2}$ and $I_{\mathrm{MAX}} = 0.63\text{Å}^{-2}$. For a preserved WS, the RDF integral is greater than I_{MAX} (Fig. 2 *C*, *blue*), for a bulk-like WS the value of the integral is in the range from I_{MIN} to I_{MAX} (Fig. 2 *C*, *gray*), while an integral below I_{MIN} indicates an obstructed WS (Fig. 2 *C*, *red*).

The local density criteria are also applicable to the CWS that are represented by alternative conformations with partial occupancy in the crystal structure. From the experimental perspective, even an alternative conformation is represented by an electron density peak above the average bulk water density. This implies that a difference between the preserved WS and the bulk water should be distinguished by our density criteria.

Global alignment versus LAWS comparison

We compared global alignment and LAWS as two approaches for tracking WS in a simulation. In global alignment, for each frame of the trajectory, the protein structure was aligned to the crystal structure by minimizing root-mean-squared deviation (RMSD) using the MDAnalysis Python library (35). In this case, the positions of aligned crystallographic water oxygen atoms defined the WS. Once the WS are defined at each simulation frame (using either global alignment or LAWS), we can compute the preservation of these sites using the RDF approach described above.

Globally aligned WS positions are fixed relative to each other, whereas in LAWS they can move relative to each other due to changes in the protein structure. The reference points and distances used in LAWS are computed once, at the start, using only the experimental crystal structure, not a starting structure or any structure from the simulation. The same crystal structure is used for global alignment (aligning to this crystal structure at every frame). Therefore, LAWS does not have a bias or intrinsic advantage relative to global alignment since both approaches are supplied with the exact same information, namely the atomic positions in the experimental structure. While LAWS uses this information to extract distances (and hence, weights) to define WS in a trajectory, global alignment uses this information to carry out an RMSD superposition. Since the methods use exactly the same input information, a comparison of the two methods is fair. We note that, if there is no observed deviation from the crystal structure in a simulation (RMSD is negligible), then the results of global alignment and LAWS would be equivalent.

MD simulation details

Model building

The simulation system was constructed from the 95-residue-long second PDZ domain of the ligand of Numb protein X 2 ($\mathrm{LNX2}^{\mathrm{PDZ2}}$), which was obtained from PDB: 5E11 (24) with a resolution of 1.80 Å (Fig. 3). We chose this structure as an example of a room temperature protein crystal structure. Alternate conformations with the highest occupancy were chosen in model building. We used the CHARMM-GUI web-server to add atoms that were missing in the crystal structure (36). The positions of all hydrogen atoms in the original PDB file were ignored. Instead, hydrogen atoms were added to the initial structure using *pdb2gmx* in the GROMACS software package (37). Next, CHARMM-GUI (36) was used to reconstruct a triclinic unit cell (*C*121 space group, with parameters $a = 65.30$ Å, $b = 39.45$ Å, $c = 39.01$ Å, and $\alpha = \gamma = 90°$, $\beta = 117.54°$) with four symmetrically related proteins (Fig. 3). Since the conditions (salt concentration, pH, etc.) of the protein crystal are difficult to determine, these parameters were chosen to match the conditions of the crystallization buffer as closely as possible. Sodium chloride (NaCl) was added to neutralize the system and then to mimic the 35 mM concentration of NaH_2PO_4 found in the buffer. To mimic the pH of the crystallization buffer (pH 4.5), the residues were matched to their protonation states at this pH: N- and C-termini were kept charged and all of the histidine residues were protonated (24). All of the crystallographic water oxygen atoms (with the highest occupancy in case of alternate conformations) were included in the construction of the unit cell. Following the method of Cerutti and Case (15), additional water molecules were added to solvate the system and preserve the experimental volume of the unit cell in the NPT ensemble using the GROMACS utility *solvate* (37). The simulation

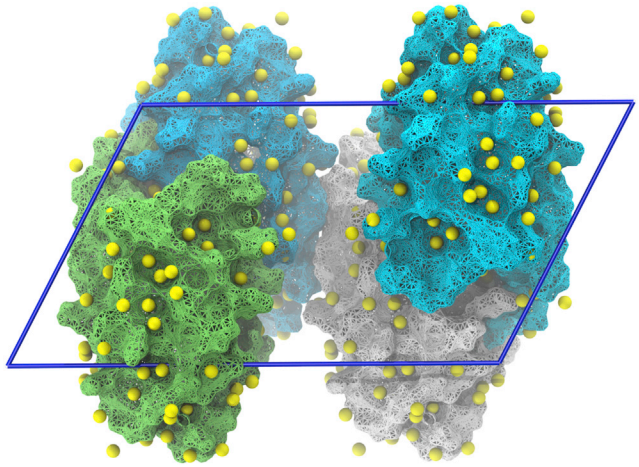

FIGURE 3 The simulation system. A single unit cell of the LNX2^{PDZ2} crystal with the *ac*-plane shown. There are 94 CWS (*yellow spheres*) around each of four symmetrically related protein chains (colored individually and shown in a surface representation). There are 376 CWS in total. To see this figure in color, go online.

system contained a total of 9650 atoms, with 6892 protein atoms, and 3750 water atoms, as well as 6 sodium ions and 2 chloride ions. To construct the system for simulating a protein in solution, a single PDZ domain was placed in a dodecahedral box. The same conditions (salt concentration, pH) were used as for the unit cell system. As a result, this system contained a total of 37,329 atoms, with 1723 protein atoms and 35,589 water atoms, as well as 9 sodium ions and 8 chloride ions.

CWS

In the crystal structure (PDB: 5E11), there are a total of 94 water oxygen atoms, of which 88 have occupancies of 100%, while 6 have alternative conformations with partial occupancies. For our analysis, we consider 94 CWS represented by the positions of these oxygen atoms, using the alternative conformation with the highest occupancy. Thus, we tracked 94 WS associated with each of the four individual PDZ domains in the unit cell, such that there are four symmetric copies of each CWS, providing a total of $94 \times 4 = 376$ WS for the entire unit cell (Fig. 3). A CWS was classified as intrachain if its coordination was limited to a single protein chain, whereas it was classified as interchain if it was coordinated by multiple distinct chains of the unit cell. Of the 94 CWS associated with each PDZ domain, 38 were found to be intrachain and 56 were found to be interchain.

Simulation

Simulations were carried out using GROMACS 2019.1 (37). The CHARMM36m force field (38) combined with the CHARMM-modified TIP3P water model (39) were used for the study. The time step of the simulation was 2 fs. The LINCS algorithm was used to constrain covalent bonds with hydrogen atoms (40). Short-range electrostatics and Lennard-Jones interactions were computed with a cutoff of 9.5 Å. Long-range electrostatics were computed using particle-mesh Ewald summation with a grid spacing of 1.2 Å with a fourth order interpolation (41). A compressibility of 2.5×10^{-5} bar^{-1} was used to mimic the compressibility of a protein crystal (42). The temperature of the simulation was kept constant at 298 K using the velocity rescaling thermostat (43) to match the experimental conditions. The pressure of the system was kept constant at 1 bar. Following energy minimization of the system using the steepest descent algorithm, 10 ns of position restrained simulation was performed, followed by equilibration in the NVT ensemble for 100 ns. Two types of NPT equilibration simulations were performed in succession: 1) 10 ns of isotropic Berendsen pressure coupling to quickly reach a pressure of 1 bar and 2) 100 ns of simulation using isotropic Parrinello-Rahman pressure coupling (44,45). The simulation was extended for an additional 1 μs, which was used for the analysis (100,000 frames with a 10-ps stride). The protein conformation in a unit cell reached an average heavy-atom RMSD of 1.8 Å relative to the crystal structure (Fig. S2). The experimental unit cell parameters were preserved in the simulation (Fig. S3). Two additional 1-μs independent simulations were carried out starting from the same initial coordinates and randomly assigned velocities.

The same simulation protocol was used for simulating a single PDZ domain in solution, except for the temperature (289 K) and the compressibility (4.5×10^{-5} bar^{-1}, corresponding to the value for water). A total of ten simulation replicas of 1 μs each were carried out. The protein conformation in the absence of crystal packing reached an average RMSD of 4.5 Å relative to the crystal structure (Fig. S4 *A*).

RESULTS AND DISCUSSION

Recall of CWS for a unit cell

We applied the LAWS pipeline (as described in Methods and summarized in Section S2) to the 1-μs simulation of the PDZ domain unit cell (Fig. 3), which contains four symmetrically related copies of the protein. In addition, we performed a similar analysis using the global alignment approach (Fig. 4 *A*). After we tracked the WS in the simulation (using either LAWS or global alignment), we classified each WS into one of three groups according its RDF: preserved, bulk-like, and obstructed (Fig. 4 *B*). An example radial density profile is shown for each type of WS (Fig. 4 *C*). Preserved WS are characterized by a high radial density of water in the range from 0 to 1.4 Å when compared with the control bulk distribution (Fig. 4 *C*, *top*). Bulk-like WS have an RDF similar to the control bulk water, which correspond to the regions of average solvent density (Fig. 4 *C*, *center*). Obstructed WS are located close to the protein (within the van der Waals radii of protein atoms), preventing water molecules from occupying these WS regions. This obstruction results in the RDF shifting to the right (Fig. 4 *C*, *bottom*).

We computed the number of WS classified as preserved, bulk-like, and obstructed in the simulation using both LAWS and global alignment (Table 1). Using LAWS to analyze the simulation, we find that 76% of the WS are preserved. The results with the global alignment approach shows a smaller CWS recall: 68% are identified in the simulation as preserved. Two additional simulation replicas show consistent CWS recall statistics (Table S1).

To test the robustness of our method, namely, estimating occupancy with radial density $g(r)$, we analyzed the 3D density peaks computed from the offset vectors $\vec{r}$ (Fig. S5). The CWS recall obtained using the radial density does not differ significantly from the results obtained using 3D density maps (Table S2). Irrespective of the method used to compute the WS occupancy, LAWS shows more WS to be preserved in the simulation when compared with global alignment.

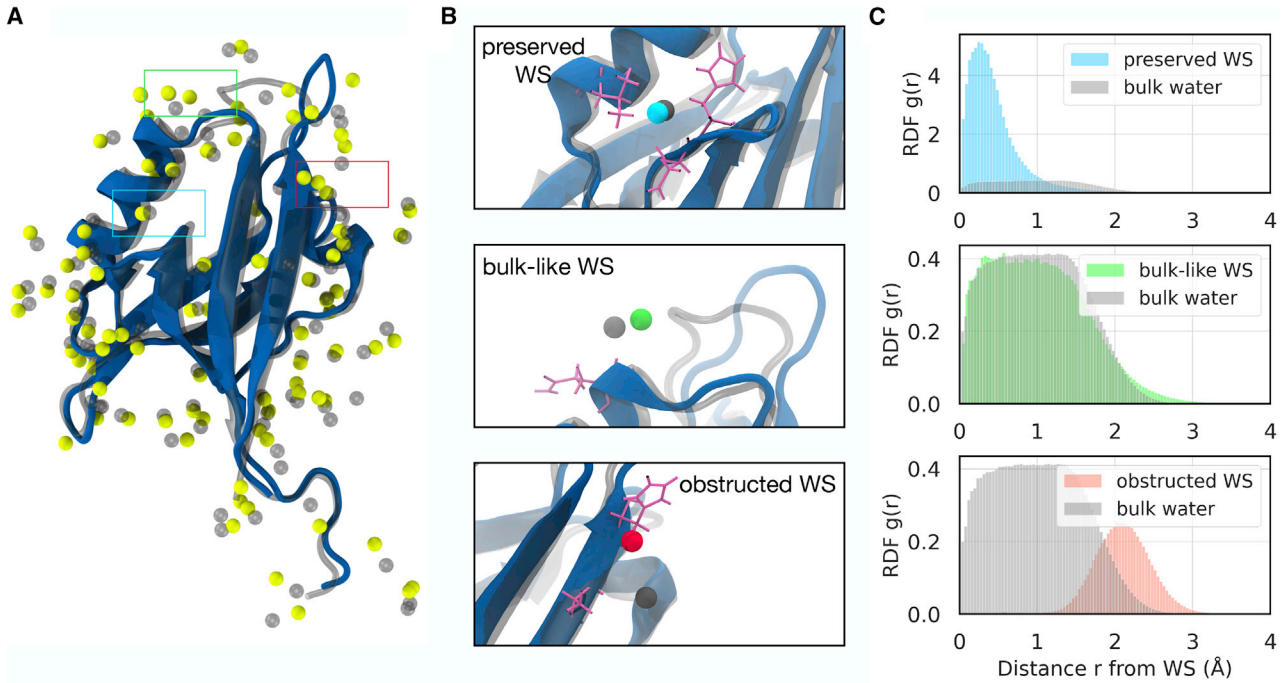

FIGURE 4 The LAWS method classifies WS into preserved, bulk-like, and obstructed. (*A*) A representative structure of the PDZ domain taken from the simulation (*blue*) is superimposed onto the crystal structure (*gray*). WS obtained by LAWS (*yellow*) are compared with the WS obtained by global alignment (*gray spheres*). Some of the LAWS-defined WS are coordinated by interactions with multiple protein chains in the unit cell. However, only a single protein is shown for clarity. (*B*) The local protein structure surrounding a representative preserved WS (*blue*), a bulk-like WS (*green*), and an obstructed WS (*red*). In each panel, the positions of globally aligned WS (*gray spheres*) and the coordinating side chains (*purple*) are shown. (*C*) Radial distribution functions for each representative water site in (*B*) compared with the control bulk water distribution are shown. To see this figure in color, go online.

LAWS detects CWS with higher density compared with global alignment

To compare the properties of the preserved WS identified with LAWS with those identified by the global alignment approach, we computed the combined distance distribution, $P(r)$, for preserved WS (Fig. 5 *A*), and the corresponding RDF $g(r)$ (Fig. 5 *B*). An apparent shift of the $P(r)$ distribution to smaller r is evident for the preserved WS identified by LAWS compared with global alignment (Fig. 5 *A*). This shift leads to a peak in the radial density that is significantly higher (Fig. 5 *B*). The analysis of 3D density maps demonstrates similar results; preserved WS defined by LAWS have higher 3D density peaks than those defined by global alignment (Fig. S6). Therefore, water is more localized around the WS tracked by the LAWS algorithm than by global alignment. Regarding the CWS with partial occupancy, all six CWS with partial occupancy are preserved in the simulation when analyzed using LAWS, while four out of six are preserved when analyzed using global alignment.

TABLE 1 The number of water sites and percentages classified as preserved, bulk-like, and obstructed using LAWS and global alignment methods

Method	Preserved WS	Bulk-like WS	Obstructed WS
LAWS	71 (76%)	10 (11%)	13 (14%)
Global alignment	64 (68%)	8 (9%)	22 (23%)

LAWS quantifies perturbation to the local protein structure

To determine how the deviation of the protein from the crystal structure contributes to the loss of crystal water, we analyzed the LAWS error for each group of WS. The LAWS error represents a quantitative measure of the perturbation to the protein region coordinating the WS in a given MD frame. The definition of the LAWS error (Eq. 2) does not consider the simulation water molecules, hence the LAWS error is only determined by the nearby atoms of the protein. A LAWS error of zero represents an unperturbed protein environment relative to the crystal structure. Well-preserved WS should be characterized by a low LAWS error. In contrast, a high LAWS error (>3 Å^2) represents a situation where the deviation of the protein structure is so considerable that the placement of the WS becomes meaningless in a given frame.

The distributions of LAWS errors for each of the three WS groups are presented as boxplots in Fig. 6. These distributions are exponential, with high variance indicated by the long whiskers. The preserved WS have a mean LAWS error ($\pm$ SD) of 0.7 ± 2.2 Å^2, the bulk-like WS have a mean value

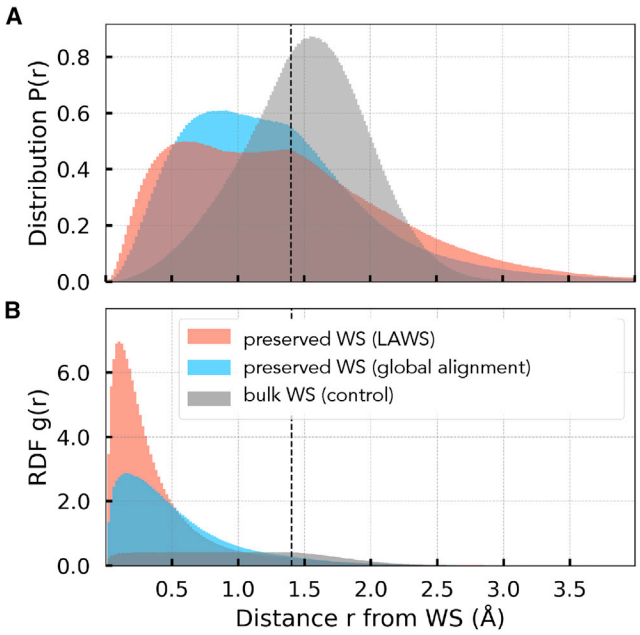

FIGURE 5 Preserved WS detected by LAWS exhibit higher water density compared with global alignment. (*A*) Distribution of distances, $P(r)$, between the nearest neighbor water molecule and (i) preserved WS tracked with LAWS (*red*), (ii) preserved WS tracked with global alignment (*blue*), and (iii) bulk WS (*gray*). (*B*) Corresponding RDF, $g(r)$, with the same color scheme as (*A*). For instance, it is $\sim 7\times$ more likely relative to bulk to find a water molecule within 0.2 Å from the WS as found by global alignment. In contrast, it is $\sim 17\times$ more likely as found by LAWS. The likelihood of finding a water molecule within 0.2 Å from the preserved WS relative to bulk is increased by $\sim 2.4\times$ by using LAWS versus global alignment. To see this figure in color, go online.

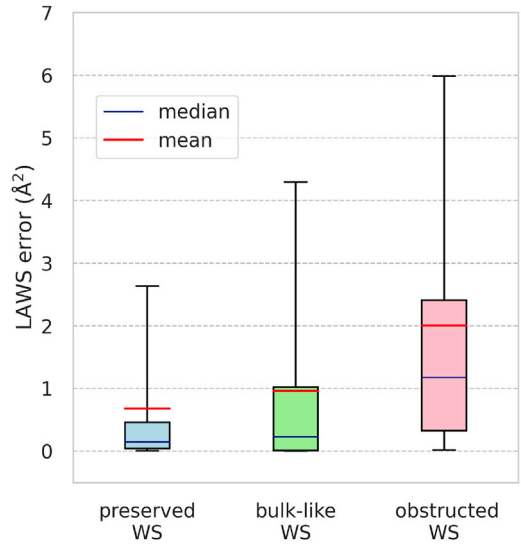

FIGURE 6 LAWS error for each category of WS. The boxplots show the distribution of LAWS error for preserved (*blue*), bulk-like (*green*), and obstructed (*red*) WS. The LAWS error represents the minimized value of Eq. 2 when the WS is optimally placed. In the plots, boxes represent the interquartile range (25th–75th percentile), while whiskers show 5th–95th percentile of the distribution. To see this figure in color, go online.

of 1.0 ± 2.2 Å^2, whereas the mean LAWS error for the obstructed WS is 1.8 ± 2.3 Å^2. As expected, preserved WS are characterized by relatively low LAWS errors (<1 Å^2) corresponding to minor deviations of the local structure compared with the crystal structure. In contrast, high LAWS errors are more commonly observed in obstructed and bulk-like WS, where the perturbations are more significant.

We note that a low LAWS error by itself is not a criterion for a preserved WS. To illustrate this, we remove the frames for which the LAWS error exceeds 1 Å^2 for each water site. By only analyzing the frames where WS are well coordinated (with a low LAWS error), we would expect perfect recall. Nevertheless, the observed recall with excluded frames is only slightly improved (79% as opposed to 76% for the entire trajectory). Therefore, a low LAWS error is a necessary, but not a sufficient, condition for a WS to be preserved.

Recall of CWS depends on the experimental B-factors

In diffraction experiments, B-factors are a measure of the spread of the electron density caused by atomic fluctuations in the crystal lattice. The experimental CWS (water oxygen atoms) have associated B-factors, which represent uncertainty in their positions. For this reason, we hypothesize that the experimental B-factor of a CWS should be related to the probability of preserving that CWS in the simulation.

To check whether this is the case, we analyzed CWS recall according to B-factor (Table 2). Using LAWS, 100% recall of CWS was obtained in the lowest B-factor range (<20 Å^2). For comparison, CWS in the middle range of 20–50 Å^2 have a recall of 74–81%, while CWS with high (>50 Å^2) B-factors are poorly preserved (0–67% recall). In addition, the LAWS error correlates with B-factor, suggesting that the local protein environment surrounding the low B-factor CWS is structurally less perturbed than for the higher B-factor CWS.

An advantage of simulating the entire unit cell is the independent information provided by the four symmetric copies of each CWS (Fig. 3), which can be used to estimate statistical errors. Making use of this information, we computed the number of copies (from 0 to 4) that were classified as preserved for each of the 94 CWS (Fig. 7). More

TABLE 2 Recall of CWS in the simulation according to the experimental B-factor using LAWS and global alignment

B-factor range (Å^2)	10–20	20–30	30–40	40–50	50–60	60–70
No. of CWS in range	12	20	26	23	9	4
Mean LAWS error (Å^2)	0.23	0.85	0.88	1.04	0.50	2.45
Percent preserved (LAWS)	100%	75%	81%	74%	67%	0%
Percent preserved (global alignment)	92%	65%	65%	61%	78%	50%

The 94 CWS were grouped according to B-factor in bins of 10 Å^2.

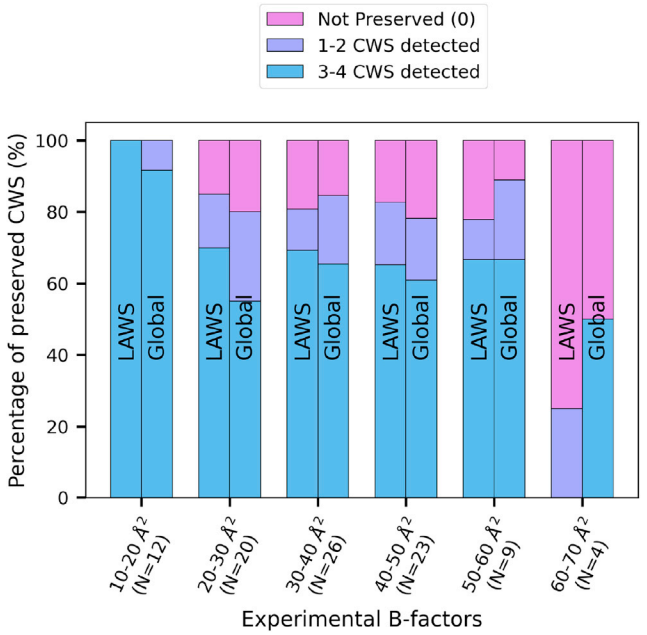

FIGURE 7 Experimental B-factors correlate with preserving CWS in the simulation. The bar chart shows the percentage of the CWS that are preserved in 0 (*pink*), 1–2 (*purple*), or 3–4 (*blue*) symmetric copies in the unit cell. Results are provided for both LAWS and the global alignment approach. The 94 CWS were grouped according to B-factor in bins of 10 Å^2, with the number of CWS in each bin, N, provided. The number of symmetric copies that are preserved provides a way to assess statistical uncertainty in each bin, with more copies representing a higher confidence in the result. To see this figure in color, go online.

copies of the CWS with low B-factors are found to be preserved in the simulation. Similar to Table 2, this trend is robust regardless of whether LAWS or global alignment is used to track the WS. Importantly, LAWS shows a higher consistency across symmetric copies than global alignment in the low B-factor range.

Taken together, this analysis reveals that the CWS with lower B-factors, i.e., lower uncertainty, are more likely to be preserved in the simulation, as opposed to the CWS with higher B-factors, and higher associated uncertainty in their positions. This trend is consistent with the findings of Sun et al., who investigated the preservation of binding site waters in simulations (22). Similarly, we analyzed local 3D density peaks for preserved WS, and a negative correlation between peak height and experimental B-factor is found (Fig. S6).

Interestingly, LAWS demonstrates a low recall of the highest uncertainty CWS, with B-factor >60 Å^2 (Table 2). These sites also have the highest LAWS error, which suggests that the surrounding protein structure is highly perturbed. It is unclear how many of these sites are expected to be captured by simulation, since B-factors of >60 Å^2 are close to the bulk threshold for this system. To understand these effects in more detail, we analyzed the locations of CWS that were lost (not preserved) in the simulation.

Lost CWS are coordinated by flexible regions of the protein

In the analysis carried out so far, we have used LAWS to quantify CWS recall. Next, we examine how the preservation of a water site is affected by the surrounding protein structure. The key similarity between lost CWS appears to be tied to their location relative to the protein. The lost CWS are mostly coordinated by the flexible structural elements, including the loops and the C-terminal tail (Table S3). These protein segments have a higher root-mean-squared fluctuation (RMSF) in the simulation (Fig. 8), indicating that the loss of CWS can be attributed to the higher mobility of the protein regions coordinating these WS. The relatively high LAWS error observed for bulk-like and obstructed WS (Fig. 6) suggests that the flexible regions of the protein also experience significant deviations relative to the crystal structure, leading to a loss of water coordination. To understand whether the loss of coordination is more pronounced at the protein-protein interfaces or in the regions within the protein, it is useful to consider the system where the protein is not constrained by crystal contacts.

Recall of CWS for a single PDZ domain in solution

In addition to simulating a crystal unit cell, which is the most realistic system for quantifying recall of CWS, we also carried out a set of simulations of the same PDZ domain in solution (10 replicas × 1 μs each, see Methods). Unsurprisingly, the protein in solution exhibits a much higher deviation from the crystal structure (RMSD = 4.5 Å, Fig. S4 *A*), compared with the crystal environment (RMSD = 1.8 Å, Fig. S2). In addition, the protein is more flexible in solution (Fig. S4 *B*) compared with the crystal environment (Fig. 8). Since both sets of simulations start from the same crystal structure, we

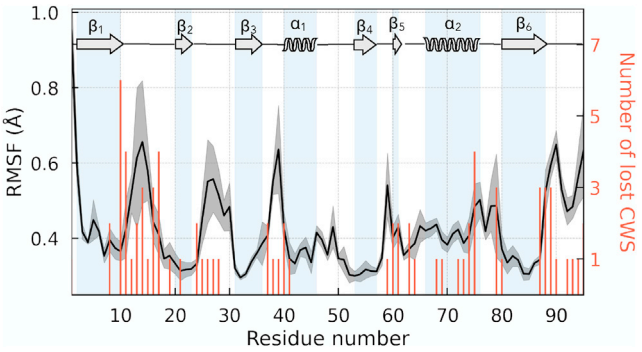

FIGURE 8 Lost CWS are coordinated by flexible regions of the protein. RMSF of Cα atoms represented by the mean value (solid line) over $N = 4$ protein copies in the unit cell. The shaded envelope shows standard deviation. There are 23 lost CWS which are each coordinated by multiple protein residues. The number of the lost CWS coordinated by each residue is shown with red bars. This figure illustrates the data provided in Table S2. To see this figure in color, go online.

investigated whether the recall of CWS decreases over time as the protein structure changes.

We analyzed the convergence of CWS recall with increasing simulation time for both systems (solution, Fig. 9 *A*, and crystal, Fig. 9 *B*). Using the global alignment approach, a significant decrease in CWS recall is observed during the course of the simulation for the protein in solution (Fig. 9 *A*). This decrease of CWS recall happens because the structure is drifting away from the initial crystal structure throughout the simulation (Fig. S4 *A*). In contrast, in the crystal environment, the RMSD curve is flat (Fig. S2) and no substantial change is observed in the CWS recall over time (Fig. 9 *B*). The CWS recall determined using LAWS is consistent over time for both solution and crystal simulations.

Using global alignment, the average recall of CWS is only 5% in the solution simulation; however, when using LAWS to track WS in solution, the time-averaged CWS recall is much higher—63% (*dashed lines* in Fig. 9 *A*). Together, these results demonstrate that the global alignment approach is strongly affected by changes in protein structure. Alignment errors cause WS to overlap with the protein or to be located in bulk water, resulting in them being classified as either obstructed or bulk-like. LAWS, on the other hand, is effective in eliminating these errors and shows a consistent CWS recall for both systems. The LAWS errors across all CWS (mean ± SD) are higher in solution (2.0 ± 9.9 Å^2) than in crystal (0.9 ± 2.2 Å^2), demonstrating lower coordination of CWS provided by a more perturbed protein environment in solution. Using LAWS, we find that more CWS are preserved in the crystal simulation compared with the protein in solution (76% vs. 63%, red dashed lines in Fig. 9, *A* and *B*). We hypothesized that many of the CWS are stabilized by contacts in the crystal lattice, which might affect the recall statistics.

To assess the contribution of crystal contacts, we considered two types of CWS in the crystal structure: 1) intrachain CWS, i.e., those making contacts with a single protein domain, and 2) interchain CWS found at the interface between symmetric copies of the proteins (Fig. 9 *C*). We compare the fraction of lost CWS for each group in the crystal and solution simulations. Since the experimental B-factors of these two groups do not differ significantly, with $B_{intra} = 36 \pm 13$ Å^2 $(n = 38)$ and $B_{inter} = 36 \pm 12$ Å^2 $(n = 56)$, we can eliminate the effect of

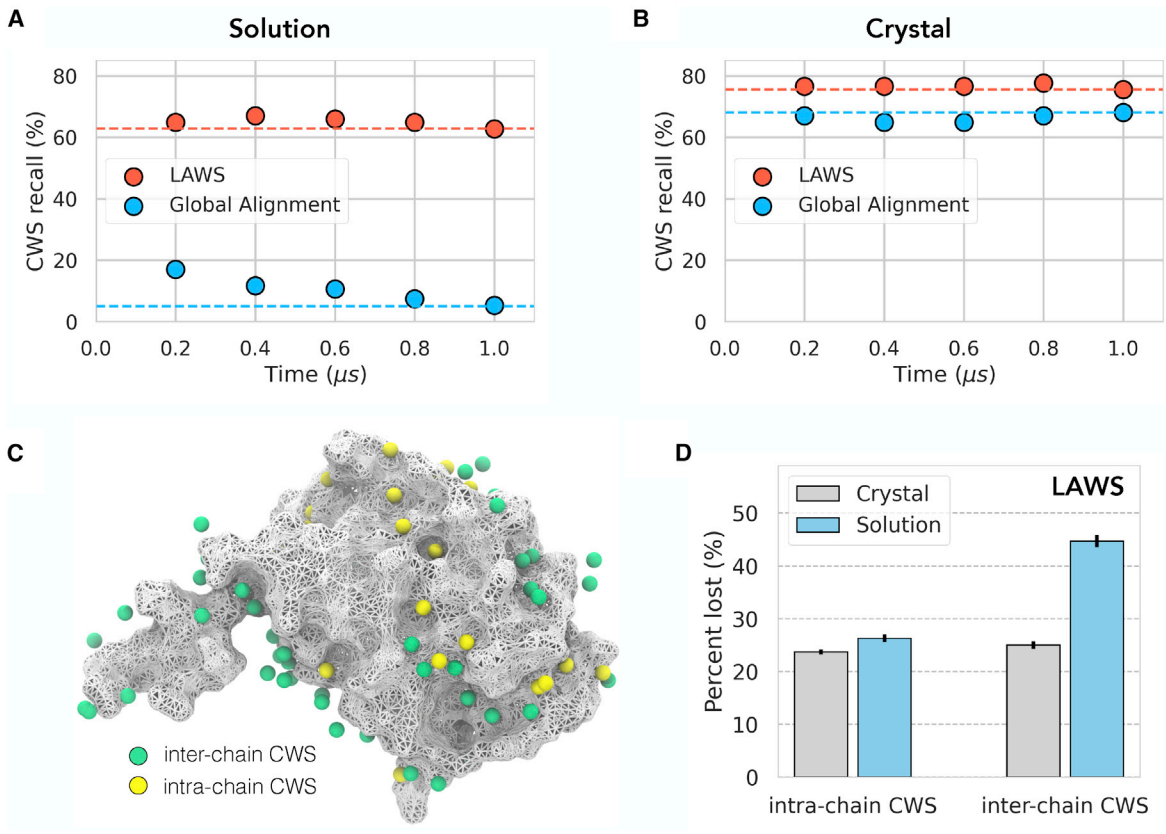

FIGURE 9 Recall of CWS in solution and crystal simulations. The CWS recall with increasing cumulative simulation time (0.2 μs, 0.4 μs, etc.) computed using LAWS and global alignment for the PDZ domain (*A*) in solution, (*B*) in a crystal unit cell. Dashed lines indicate average recall over the 1-μs trajectory. (*C*) The crystal structure of the PDZ domain in surface representation with locations of interchain (*green*) and intrachain CWS (*yellow*) shown. (*D*) The percent of lost CWS in the two groups (interchain and intrachain) as found in crystal and solution simulations (obtained using LAWS). Statistical errors are estimated via a block averaging method with a window size of 0.2 μs. There are 28 preserved intrachain CWS in solution, and 29 in a crystal, of which 25 are common. To see this figure in color, go online.

B-factor as a contribution to the CWS recall in this comparison. The fraction of lost intrachain CWS in the crystal (24%) is consistent with the solution (26%) simulations (Fig. 9 *D*). The set of preserved CWS in the crystal simulation is very similar to the set of preserved CWS in solution (29 preserved CWS in the crystal, 28 preserved CWS in solution, and 25 CWS are common between the two sets). While a comparable number of the interchain CWS are lost in crystal simulations (25%), significantly more of them are lost in solution (45%) (Fig. 9 *D*). These results suggest that the effect of the protein environment (crystal versus solution) is critical for the interchain CWS. The interchain coordination is specific to protein crystals and significantly contributes to the preservation of CWS in simulations. However, the preservation of intrachain CWS is not strongly affected by the protein environment. These results highlight a powerful application of the LAWS method—in particular, that LAWS can be applied to simulations of proteins in solution to analyze ordered waters located in pockets and coordinated by residues of the same domain.

CONCLUSIONS

In this study, we discuss the limitations of existing global alignment approaches for analyzing crystallographic water in MD simulations. Namely, these methods are only suitable for cases where the protein conformation remains close to the crystal structure. As the deviation from the crystal structure increases, global alignment introduces larger and larger errors in the positions of WS. To address this limitation, we developed the LAWS method, which becomes more advantageous when protein regions experience significant deviations from the crystal structure since this approach does not suffer from alignment errors. To the best of our knowledge, the only study that implemented a similar local alignment approach was Caldararu et al. (13). We build upon their approach by computing the radial density profile of water as the main criterion to determine if a water site is preserved (instead of relying on data-driven clustering). In addition, our method utilizes a reference (bulk WS) to decide whether or not a particular CWS is preserved, whereas their method does not. Unique to our method, the LAWS error serves as a measure of protein structural perturbation around each WS. Finally, compared with the previous local alignment method, LAWS takes into account the CWS coordination by multiple symmetrically related protein monomers in the crystal lattice (Fig. 2 *A*). This consideration is critical since crystal contacts contribute to the stability of many CWS in a protein crystal.

Applying the LAWS method to a 1-μs simulation of a protein crystal, we demonstrate that LAWS improves the recall of high-confidence (low B-factor) CWS and shows an increased density of water surrounding the CWS (Figs. 5 and S4) compared with global alignment. Previous studies reported the CWS recall as the fraction of water density peaks detected within 1.4 Å of their positions in the crystal structure (10–13). They found that CWS recall varied from 40 to 100% depending on the system, simulation setup, and the method used for analysis. To place our results in the context of these studies, we provide analogous recall statistics computed using 3D density maps: 70% with LAWS vs. 60% with global alignment (Table S2). This is broadly consistent with the earlier studies that did not restrain protein dynamics.

The question is: What recall of CWS is realistic, given the challenges of accurately replicating the experiments? The lack of consensus between independent experiments in the positions of CWS (46–49) suggests that complete recall of all CWS in simulations is most likely overly optimistic and should not be expected. The construction of the MD simulation system on its own has multiple challenges that can affect the accuracy of modeling crystallographic water. For example, the resolution of the starting structure plays a crucial role in the quality of simulations (50). In addition, the correct assignment of protonation states (51) and the inclusion of crowding agents in the crystal lattice (52) can be challenging. Finally, a limited sampling of protein conformational space can be an obstacle for simulations to provide equilibrium distributions. Despite all these difficulties, we observe 100% recall of the high confidence (low B-factor) CWS using LAWS in this study.

While the current work has focused on presenting the LAWS algorithm as a means to study how well crystallographic waters are preserved in simulation (e.g., for force field and model testing), our methodology has broader applications. The coordinate-based nature of LAWS allows analysis of dynamic information, such as residence times of individual water molecules located in WS. The foundation of LAWS is a tracking algorithm that can easily be generalized to track various ions, small molecules, and ligands. With the addition of further constraints, LAWS could be extended to track the interactions of proteins with larger molecules as well. Furthermore, an important application of LAWS may also be in the study of functional water networks in proteins. Experimentally resolved CWS have been used to study water networks, which have been found to be highly conserved and important to protein function (6,53). Perturbations to these networks have also been implicated in human disease (54). While we have found that many of the CWS are stabilized by interactions in the crystal lattice, a subset of the experimental CWS appears to be intrinsic to the protein structure, as demonstrated by the consistent preservation of a set of CWS in both crystal and solution simulations. What this suggests is that, by applying LAWS to study experimentally resolved waters of proteins in solution simulations, LAWS can be a powerful tool to study water networks that underlie protein function.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2022.09.012. Simulation data is available at https://doi.org/10.5281/zenodo.6478270. The implementation of the LAWS algorithm is available at https://github.com/rauscher-lab/LAWS.

AUTHOR CONTRIBUTIONS

E.K., J.S.K., and S.R. designed the research. E.K., J.S.K., and S.R. performed the research. E.K. developed the code and contributed analytic tools. J.S.K. and E.K. carried out the simulations. E.K., J.S.K., and S.R. analyzed the data. E.K., J.S.K., and S.R. wrote the manuscript.

ACKNOWLEDGMENTS

This research was supported by a Connaught New Researcher Award (to S.R.), a Natural Sciences and Engineering Research Council of Canada of Canada (NSERC) Discovery Grant and by Compute Canada. Computations were performed on the Niagara supercomputer at the SciNet HPC Consortium. SciNet is funded by: the Canada Foundation for Innovation; the Government of Ontario; Ontario Research Fund – Research Excellence; and the University of Toronto.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Helms, V. 2007. Protein dynamics tightly connected to the dynamics of surrounding and internal water molecules. *ChemPhysChem.* 8:23–33. https://doi.org/10.1002/cphc.200600298.
2. Bellissent-Funel, M.-C., A. Hassanali, …, A. E. Garcia. 2016. Water determines the structure and dynamics of proteins. *Chem. Rev.* 116:7673–7697. https://doi.org/10.1021/acs.chemrev.5b00664.
3. Levy, Y., and J. N. Onuchic. 2006. Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.* 35:389–415. https://doi.org/10.1146/annurev.biophys.35.040405.102134.
4. Schirò, G., Y. Fichou, …, M. Weik. 2015. Translational diffusion of hydration water correlates with functional motions in folded and intrinsically disordered proteins. *Nat. Commun.* 6:6490. https://doi.org/10.1038/ncomms7490.
5. Nakasako, M. 2021. Hydration Structures of Proteins. Springer Japan. https://doi.org/10.1007/978-4-431-56919-0.
6. Kim, T. H., P. Mehrabi, …, E. F. Pai. 2017. The role of dimer asymmetry and protomer dynamics in enzyme catalysis. *Science.* 355:eaag2355. https://doi.org/10.1126/science.aag2355.
7. Karplus, M., and J. Kuriyan. 2005. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. USA.* 102:6679–6685. https://doi.org/10.1073/pnas.0408930102.
8. Berman, H. M., J. Westbrook, …, P. E. Bourne. 2000. The protein Data Bank. *Nucleic Acids Res.* 28:235–242. https://doi.org/10.1093/nar/28.1.235.
9. Higo, J., and M. Nakasako. 2002. 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.* 23:1323–1336. https://doi.org/10.1002/jcc.10100.
10. Altan, I., D. Fusco, …, P. Charbonneau. 2018. Learning about biomolecular solvation from water in protein crystals. *J. Phys. Chem. B.* 122:2475–2486. https://doi.org/10.1021/acs.jpcb.7b09898.
11. Rudling, A., A. Orro, and J. Carlsson. 2018. Prediction of ordered water molecules in protein binding sites from molecular dynamics simulations: the impact of ligand binding on hydration networks. *J. Chem. Inf. Model.* 58:350–361. https://doi.org/10.1021/acs.jcim.7b00520.
12. Wall, M. E., G. Calabró, …, G. L. Warren. 2019. Biomolecular solvation structure revealed by molecular dynamics simulations. *J. Am. Chem. Soc.* 141:4711–4720. https://doi.org/10.1021/jacs.8b13613.
13. Caldararu, O., M. Misini Ignjatović, …, U. Ryde. 2020. Water structure in solution and crystal molecular dynamics simulations compared to protein crystal structures. *RSC Adv.* 10:8435–8443. https://doi.org/10.1039/c9ra09601a.
14. Cerutti, D. S., P. L. Freddolino, …, D. A. Case. 2010. Simulations of a protein crystal with a high resolution X-ray structure: evaluation of force fields and water models. *J. Phys. Chem. B.* 114:12811–12824. https://doi.org/10.1021/jp105813j.
15. Cerutti, D. S., and D. A. Case. 2019 Jul-Aug. Molecular dynamics simulations of macromolecular crystals. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 9:e1402. https://doi.org/10.1002/wcms.1402.
16. Case, D. A., T. E. Cheatham, …, R. J. Woods. 2005. The Amber biomolecular simulation programs. *J. Comput. Chem.* 26:1668–1688. https://doi.org/10.1002/jcc.20290.
17. Copie, G., F. Cleri, …, M. F. Lensink. 2016. On the ability of molecular dynamics simulation and continuum electrostatics to treat interfacial water molecules in protein-protein complexes. *Sci. Rep.* 6:38259. https://doi.org/10.1038/srep38259.
18. Henchman, R. H., and J. A. McCammon. 2002. Extracting hydration sites around proteins from explicit water simulations. *J. Comput. Chem.* 23:861–869. https://doi.org/10.1002/jcc.10074.
19. Feig, M., and B. M. Pettitt. 1998. Crystallographic water sites from a theoretical perspective. *Structure.* 6:1351–1354. https://doi.org/10.1016/s0969-2126(98)00135-x.
20. Steinbach, P. J., and B. R. Brooks. 1993. Protein hydration elucidated by molecular dynamics simulation. *Proc. Natl. Acad. Sci. USA.* 90:9135–9139. https://doi.org/10.1073/pnas.90.19.9135.
21. Young, T., R. Abel, …, R. A. Friesner. 2007. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc. Natl. Acad. Sci. USA.* 104:808–813. https://doi.org/10.1073/pnas.0610202104.
22. Sun, H., L. Zhao, …, N. Huang. 2014. Incorporating replacement free energy of binding-site waters in molecular docking. *Proteins.* 82:1765–1776. https://doi.org/10.1002/prot.24530.
23. van Gunsteren, W. F., H. J. Berendsen, …, J. P. Postma. 1983. Computer simulation of the dynamics of hydrated protein crystals and its comparison with x-ray data. *Proc. Natl. Acad. Sci. USA.* 80:4315–4319. https://doi.org/10.1073/pnas.80.14.4315.
24. Hekstra, D. R., K. I. White, …, R. Ranganathan. 2016. Electric-field-stimulated protein mechanics. *Nature.* 540:400–405. https://doi.org/10.1038/nature20571.
25. Janowski, P. A., D. S. Cerutti, …, D. A. Case. 2013. Peptide crystal simulations reveal hidden dynamics. *J. Am. Chem. Soc.* 135:7938–7948. https://doi.org/10.1021/ja401382y.
26. Janowski, P. A., C. Liu, …, D. A. Case. 2016. Molecular dynamics simulation of triclinic lysozyme in a crystal lattice. *Protein Sci.* 25:87–102. https://doi.org/10.1002/pro.2713.
27. Rodriguez, M. Z., C. H. Comin, …, F. A. Rodrigues. 2019. Clustering algorithms: a comparative approach. *PLoS One.* 14:e0210236. https://doi.org/10.1371/journal.pone.0210236.
28. Jaulin, L. 2015. Instantaneous Localization. Elsevier, pp. 171–196.. https://doi.org/10.1016/B978-1-78548-048-5.50005-X.
29. Moré, J. J. 1978. The Levenberg-Marquardt Algorithm: Implementation and Theory. Springer Berlin Heidelberg, pp. 105–116.. https://doi.org/10.1007/bfb0067700.
30. Briones, R., C. Blau, …, C. Aponte-Santamaría. 2019. GROmaps: a GROMACS-based toolset to analyze density maps derived from molecular dynamics simulations. *Biophys. J.* 116:4–11. https://doi.org/10.1016/j.bpj.2018.11.3126.

31. Chen, X., I. Weber, and R. W. Harrison. 2008. Hydration water and bulk water in proteins have distinct properties in radial distributions calculated from 105 atomic resolution crystal structures. *J. Phys. Chem. B.* 112:12073–12080. https://doi.org/10.1021/jp802795a.

32. Ebbinghaus, S., S. J. Kim, …, M. Havenith. 2007. An extended dynamical hydration shell around proteins. *Proc. Natl. Acad. Sci. USA.* 104:20749–20752.

33. Huang, Y., X. Zhang, …, C. Q. Sun. 2013. Size, separation, structural order and mass density of molecules packing in water and ice. *Sci. Rep.* 3:3005. https://doi.org/10.1038/srep03005.

34. Bergmann, U., A. Di Cicco, …, A. Nilsson. 2007. Nearest-neighbor oxygen distances in liquid water and ice observed by x-ray Raman based extended x-ray absorption fine structure. *J. Chem. Phys.* 127:174504. https://doi.org/10.1063/1.2784123.

35. Michaud-Agrawal, N., E. J. Denning, …, O. Beckstein. 2011. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* 32:2319–2327. https://doi.org/10.1002/jcc.21787.

36. Jo, S., T. Kim, …, W. Im. 2008. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* 29:1859–1865. https://doi.org/10.1002/jcc.20945.

37. Abraham, M. J., T. Murtola, …, E. Lindahl. 2015. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *Software*. 1-2:19–25. https://doi.org/10.1016/j.softx.2015.06.001.

38. Huang, J., S. Rauscher, …, A. D. MacKerell. 2017. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods.* 14:71–73. https://doi.org/10.1038/nmeth.4067.

39. MacKerell, A. D., D. Bashford, …, M. Karplus. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* 102:3586–3616. https://doi.org/10.1021/jp973084f.

40. Hess, B., H. Bekker, …, J. G. E. M. Fraaije. 1997. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* 18:1463–1472. https://doi.org/10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.0.co;2-h.

41. Essmann, U., L. Perera, …, L. G. Pedersen. 1995. A smooth particle mesh Ewald method. *J. Chem. Phys.* 103:8577–8593. https://doi.org/10.1063/1.470117.

42. Taulier, N., and T. V. Chalikian. 2002. Compressibility of protein transitions. *Biochim. Biophys. Acta.* 1595:48–70. https://doi.org/10.1016/s0167-4838(01)00334-x.

43. Bussi, G., D. Donadio, and M. Parrinello. 2007. Canonical sampling through velocity rescaling. *J. Chem. Phys.* 126:014101. https://doi.org/10.1063/1.2408420.

44. Berendsen, H. J. C., J. P. M. Postma, …, J. R. Haak. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690. https://doi.org/10.1063/1.448118.

45. Parrinello, M., and A. Rahman. 1981. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* 52:7182–7190. https://doi.org/10.1063/1.328693.

46. Levitt, M., and B. H. Park. 1993. Water: now you see it, now you don't. *Structure.* 1:223–226. https://doi.org/10.1016/0969-2126(93)90011-5.

47. Otting, G., E. Liepinsh, and K. Wüthrich. 1995. Protein Hydration in Aqueous Solution. World Scientific, pp. 632–638.. https://doi.org/10.1142/9789812795830_0057.

48. Zhang, X.-J., and B. W. Matthews. 1994. Conservation of solvent-binding sites in 10 crystal forms of T4 lysozyme. *Protein Sci.* 3:1031–1039. https://doi.org/10.1002/pro.5560030705.

49. Sanschagrin, P. C., and L. A. Kuhn. 1998. Cluster analysis of consensus water sites in thrombin and trypsin shows conservation between serine proteases and contributions to ligand specificity. *Protein Sci.* 7:2054–2064. https://doi.org/10.1002/pro.5560071002.

50. Garcia-Viloca, M., T. D. Poulsen, …, J. Gao. 2004. Sensitivity of molecular dynamics simulations to the choice of the X-ray structure used to model an enzymatic reaction. *Protein Sci.* 13:2341–2354. https://doi.org/10.1110/ps.03504104.

51. Socher, E., and H. Sticht. 2016. Mimicking titration experiments with MD simulations: a protocol for the investigation of pH-dependent effects on proteins. *Sci. Rep.* 6:22523. https://doi.org/10.1038/srep22523.

52. Cerutti, D. S., I. Le Trong, …, T. P. Lybrand. 2008. Simulations of a protein crystal: explicit treatment of crystallization conditions links theory and experiment in the streptavidin-biotin complex. *Biochemistry.* 47:12065–12077. https://doi.org/10.1021/bi800894u.

53. Venkatakrishnan, A. J., A. K. Ma, …, R. O. Dror. 2019. Diverse GPCRs exhibit conserved water networks for stabilization and activation. *Proc. Natl. Acad. Sci. USA.* 116:3288–3293. https://doi.org/10.1073/pnas.1809251116.

54. Rodríguez-Almazán, C., R. Arreola, …, A. Torres-Larios. 2008. Structural basis of human triosephosphate isomerase deficiency. *J. Biol. Chem.* 283:23254–23263. https://doi.org/10.1074/jbc.m802145200.