Functional Protein Dynamics in a Crystal

Eugene Klyshko,[†,‡,#] Justin Sung-Ho Kim,[†,‡,#] Lauren McGough,[¶] Victoria Valeeva,[‡] Ethan Lee,[‡,§] Rama Ranganathan,[‖,⊥] and Sarah Rauscher[*,†,‡,§]

†*Department of Physics, University of Toronto, Toronto, ON, Canada*

‡*Department of Chemical and Physical Sciences, University of Toronto Mississauga, Mississauga, ON, Canada*

¶*Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA*

§*Department of Chemistry, University of Toronto, Toronto, ON, Canada*

‖*Center for Physics of Evolving Systems and Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL, USA*

⊥*Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL, USA*

#*contributed equally*

E-mail: sarah.rauscher@utoronto.ca

Abstract

Proteins are molecular machines and to understand how they work, we need to understand how they move. New pump-probe time-resolved X-ray diffraction methods open up ways to initiate and observe protein motions with atomistic detail in crystals on biologically relevant timescales. However, practical limitations of these experiments demands parallel development of effective molecular dynamics approaches to accelerate progress and extract meaning. Here, we establish robust and accurate methods for simulating dynamics in protein crystals, a nontrivial process requiring careful attention to equilibration, environmental composition, and choice of force fields. With more than seven milliseconds of sampling of a single chain, we identify critical factors

controlling agreement between simulation and experiments and show that simulated motions recapitulate ligand-induced conformational changes. This work enables a virtuous cycle between simulation and experiments for visualizing and understanding the basic functional motions of proteins.

Introduction

Protein functions in the cell, such as enzymatic activity, signalling, and transport, are driven by conformational changes between multiple states.[1–3] To understand biological processes at the molecular level, we therefore require a precise description of protein dynamics.[1] Pump-probe time-resolved x-ray methods have made it possible to observe functionally relevant motions in a crystal environment in atomistic detail.[4–7] In these experiments, the protein's motion is a non-equilibrium response to an external perturbation, such as reaction initiation,[5] temperature-jump,[4] or the application of an electric field.[6] The latter, known as electric field-stimulated x-ray crystallography (EFX) can resolve protein dynamics on a sub-μs timescale by providing structural snapshots separated by 50 to 100 ns time intervals.[6] These snapshots represent ensemble averages, as proteins may adopt diverse conformations even within the crowded environment of a crystal.[8,9] Because of this conformational averaging, it is challenging to interpret the dynamical information obtained in the EFX experiment.[6]

Molecular dynamics (MD) simulations can explicitly probe conformational heterogeneity, describing protein motions with a high temporal resolution (femtoseconds) on timescales relevant to pump-probe experiments (ns to μs). Therefore, MD simulations are capable of bridging the gap between structural snapshots obtained in the EFX experiment in order to provide a more complete description of protein motions. At the same time, high-resolution crystallographic measurements obtained in EFX may be used to test the performance of simulation models and force fields. Such benchmarking demands an accurate representation of experimental conditions in simulations, including explicit modeling of the crystal environment, temperature and the magnitude of the applied electric field.

Although MD simulations are typically used to model protein dynamics in solution, simulating proteins in the crystalline state has been the focus of dozens of studies.[10–29] The exponential growth of the total available sampling in crystal simulations (quantified by the system size in number of atoms multiplied by the simulation length) follows a type of "Moore's law", with the total sampling increasing tenfold every five years (Fig. S1). Increases in the length

of protein crystal simulations (from ps[10] to μs[21]) and the system size (from a single unit cell[21] to a supercell[25,28]) has led to improved agreement with experimental data.[21,25,28] In particular, protein crystal simulations can now be used to assist in the interpretation of ambiguous electron density and refine protein crystal structures.[29] Furthermore, protein conformational ensembles from crystal MD simulations can be directly compared to multiconformer ensemble models obtained from x-ray diffraction data.[9,30] Despite these methodological advances, there are challenges in modeling non-equilibrium experiments.

In order to simulate the EFX experiment, the protein crystal must fully relax to an equilibrium state before the dynamic response to an electric field can be investigated. Starting a simulation from the experimentally resolved crystal structure placed in a lattice has been observed to lead to μs-long relaxation,[24,25] as the initial configuration of the lattice might not correspond to a free energy minimum for the force field and the conditions simulated.

In this work, we use MD simulations to capture the equilibrium state of a protein crystal — an initial step towards simulating the EFX experiment. As a model system, we chose a human PDZ domain (LNX2^{PDZ2}, Fig. 1a), which was studied in the first EFX experiment.[6] PDZ domains bind the C-terminal residues of partner proteins, resulting in the assembly of large intracellular protein complexes involved in a variety of cellular processes.[31] Many pathogenic viruses produce PDZ ligands that disrupt the assembly of these complexes in the host organism. For example, the LNX2^{PDZ2} domain has been shown to interact with the E protein of SARS-CoV-2.[32] Because PDZ domains exhibit local and non-local conformational changes upon ligand binding,[33–35] they are an ideal model system to investigate functional motions. Furthermore, PDZ domains have been extensively studied in the context of single domain protein allostery.[34–37]

Here, using the most extensive simulations of a protein crystal to date (Fig. S1), we first identify critical factors controlling the agreement between the simulations and experiment. The conformational ensembles obtained using different force fields (Amber ff14SB[38] and CHARMM36m[39]) are found to be distinct and non-overlapping, which can be traced to

a difference in the population of specific side-chain rotameric states. We show that this discrepancy is likely to be general, as it is observed for other protein systems. Importantly, we establish that the simulations using the Amber ff14SB force field most accurately reproduce the crystal structure. Then, we combine this data set with equilibrium simulations of the PDZ domain in solution (apo and ligand-bound) to describe the effects of ligand binding on the free energy landscape. We find that the structural changes of the protein in the crystal resemble ligand-induced conformational changes in other related PDZ domains, which suggests that the motions observed in the crystal are functionally relevant. This work lays the foundation for a virtuous cycle between simulations and the new pump-probe experiments: experimental data can be used to identify potential areas for force field improvement, while simulations enable us to visualize and understand functional motions of proteins.

Results

Optimizing the model of the protein crystal

The first aim of this study is to determine the simulation setup that provides an accurate representation of the protein crystal at equilibrium. To model a crystal lattice, we constructed a supercell with a $3 \times 3 \times 3$ unit-cell layout (Fig. 1a). This layout prevents self-interaction of the unit cells across periodic boundaries, allowing each unit cell to be surrounded by independent (non-periodic) neighbours. The supercell arrangement has the added benefit of increased conformational sampling, as it contains 108 individual protein chains. We used three different force fields: Amber ff14SB[38] (ff14SB), CHARMM36m[39] (C36m), and Amber ff94[40] (ff94) to determine the one providing the highest simulation accuracy. Additionally, we considered two ways to model the solvent inside the protein crystal, either in a simplified way (water and counter-ions) or by including crowding agents present in the crystallization buffer. Detailed protocols and descriptions of the systems studied are provided in Table 1, Methods and section 4.1 of Supplementary Information (SI).

Before assessing accuracy, we must first ensure that the simulations have reached equilibrium. Here, we specify two necessary conditions for the system to be at equilibrium: (i) average structural observables must converge for each simulation replica, and (ii) multiple replicas starting from different initial conditions must become indistinguishable. To establish condition (ii), we simulated three replicas of each crystal system (Table 1).

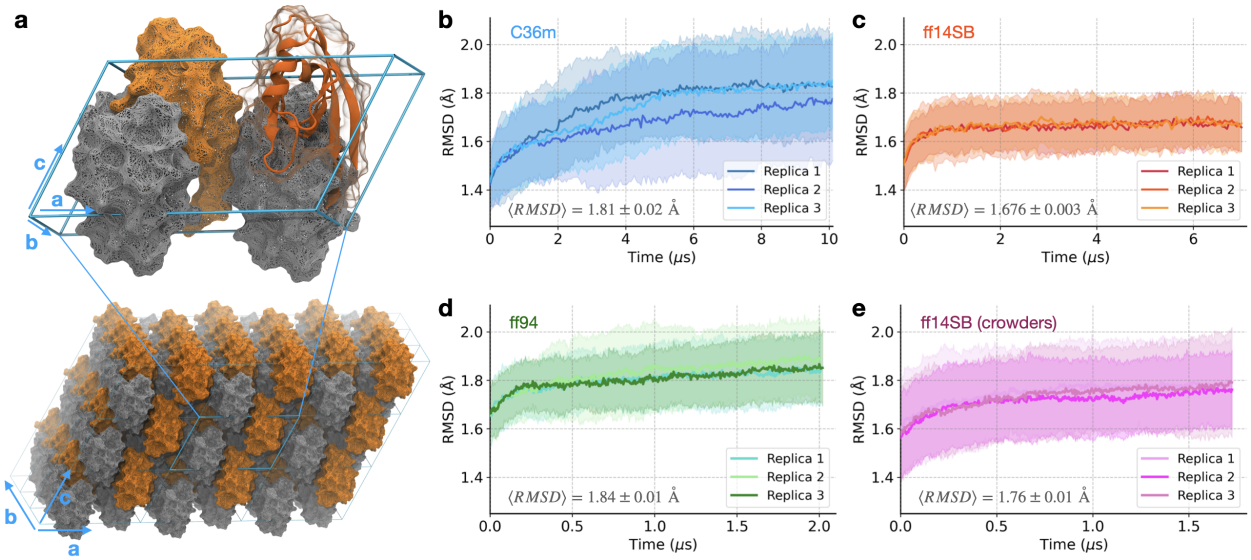

Figure 1: **Equilibrating the PDZ domain crystal.** (**a**) The simulated system is a supercell (108 protein chains) in a $3 \times 3 \times 3$ unit-cell arrangement, where each unit cell contains four symmetrically-related copies. The second PDZ domain of the human E3 ubiquitin ligase LNX2 ($LNX2^{PDZ2}$, PDB ID: 5E11) was used as the model system. (**b-e**), The average RMSD with respect to the crystal structure (computed using heavy atoms only) is shown for the simplified crystal environment using force fields (**b**) C36m, (**c**) ff14SB, (**d**) ff94, and (**e**) for the system with crowders using ff14SB. Standard deviation ($n = 108$) for each replica is represented by a shaded envelope. Each plot shows the $\langle RMSD \rangle$ ($\pm$ standard error for $n = 3$ replicas) computed for the last 1 μs of simulations. Note that the simulations with ff94 (**d**) and ff14SB with crowders (**e**) have not reached equilibrium.

Simplified crystal environment (water and ions). Using the two conditions for equilibrium, we analyzed the deviation of the protein from the crystal structure based on the root-mean-squared deviation (RMSD) and the fraction of preserved native contacts (Q). In the ff14SB simulations, both RMSD and Q are consistent between replicas and reach a plateau after 1.5 μs of simulation (Fig. 1c and Fig. S2b), satisfying both conditions for equilibrium. In the C36m simulations, both RMSD and Q exhibit a considerably slower

relaxation (Fig. 1b and Fig. S2a). Even after 10 μs, these simulations fail to satisfy both conditions for equilibrium, as one out of three replicas did not reach a plateau. Previous MD studies of protein crystals have also shown that the length of the equilibration period differs between force fields.[19,24] However, it is quite surprising that the relaxation of the crystal in our simulations takes microseconds.

We wanted to understand whether this slow relaxation could be due to conformational heterogeneity resulting from the parametrization of modern protein force fields. Specifically, disordered and partially folded states of proteins have been included in the development of recent force fields,[38,39] while older force fields were optimized to accurately describe folded states. To address this question, we simulated the same crystal with a much older force field, ff94, which predates this type of parametrization,[40] and was previously reported to provide fast (nanosecond) equilibration of protein crystals.[41,42] We find that ff94 also exhibits slow relaxation ($>$ 2 μs), but with significantly lower accuracy compared to ff14SB and C36m (see mean RMSD values in Fig. 1b-d). Thus, we discontinued the ff94 simulations without reaching equilibrium (Fig. 1d). Since the crystal simulated with ff94 also exhibits a slow relaxation, we conclude that it is not the approach to optimization of modern force fields (specifically, C36m and ff14SB) that leads to the slow relaxation of the crystal.

Prior crystal simulation studies showed equilibration times of tens to hundreds of nanoseconds[24] (Table S1), while in our study, microseconds are required to reach equilibrium for both ff14SB and C36m. The longer equilibration times observed here are likely due to the large system size (a supercell). Indeed, in our previous simulation study of the same PDZ domain crystal as a single unit cell (using C36m), the relaxation time was found to be 600 ns,[43] which is comparable to relaxation times observed for systems of a similar size.[24] A smaller system should reach equilibrium more quickly than a larger system because there are fewer copies of the protein and fewer protein-protein interfaces.

Crystal environment with crowding agents. So far, we have modeled the protein crystal using a simplified crystal environment (including only water and ions). Next, we

investigated if this model could be further improved by considering more realistic solution conditions. This is motivated by the study of Cerutti et al.,[44] who found that including crowders improved the accuracy of their crystal simulations. Using ff14SB, we simulated the supercell with the crowders found in the crystallization buffer (PEG and glycerol, see Methods) to determine if this setup could improve the accuracy. However, we find that the protein exhibits a higher deviation from the crystal structure in the system with crowders (Fig. 1e) compared to a simplified environment (Fig. 1c). In contrast to the work of Cerutti et al.,[44] explicit modeling of crowding agents does not improve (and slightly worsens) the agreement with the crystal structure in our simulations. For this reason, we discontinued these simulations after 1.7 μs without reaching equilibrium.

In setting up the simulations with explicit crowders, we assumed that the molarity of the crowding agents in the lattice is the same as the crystallization buffer. However, the actual concentrations of crowders in the crystal might be different from the concentration in the buffer because the crystal lattice may favor the inclusion of certain molecules more than others.[45,46] We also lack experimental information on the location of crowders. In addition, when adding PEG to the simulation system, there is a bias for conformations of PEG that are overly compact because extended conformations will not fit in the interstices between protein chains. Due to these challenges in modeling, adding explicit crowders to the system will not necessarily result in improved agreement with experiment.

Accuracy of the simulated protein crystal

After optimizing the simulation setup to model the protein crystal (four systems shown in Fig. 1), we assessed the agreement between simulation and experiment across additional observables. Since the ff14SB and C36m simulations sample a conformational ensemble closest to the equilibrium state, we focus on these two force fields. For completeness, analysis of the other systems, ff94 and ff14SB with crowders, is presented in SI sections 3.3, 3.4 and 3.7.

Protein structure. To evaluate how well the crystal structure is preserved in the simulation, we computed the mean squared deviation (MSD) using the final 1 μs of simulation (Fig. 2a). For both ff14SB and C36m, the structure is well-preserved overall, with the loop regions and C-terminal tail exhibiting the highest deviation from the crystal structure. To visualize these structural differences, we computed the ensemble-averaged structure in each force field (Fig. 2b). The RMSD between the average structure and the experimental structure is 1.28 Å and 1.58 Å for ff14SB and C36m, respectively. Because this degree of structural divergence is comparable to deviations between different crystal structures of this PDZ domain (Fig. S3), these results indicate that the average structure is well-captured in the simulations. When comparing the average MD structures to each other (Fig. 2a, gray), we find a smaller RMSD (1.02 Å). These results suggest that the ensembles sampled by these two force fields are more similar to one another than to the real crystal, at least according to this structural metric.

B-factors. The spread of electron density in the crystal is characterized by B-factors, which include contributions from protein structural mobility and lattice disorder.[19,47] In simulations, B-factors can be computed from the root-mean-squared fluctuation (RMSF) of atomic positions using the equation $B = 8\pi^2/3\ RMSF^2$. Here, we consider two types of B-factors, $B_{lattice}$ and B_{chain}, which depend on the degrees of freedom contributing to the fluctuations.[19] $B_{lattice}$ represents the total atomic fluctuations within the crystal lattice (visualized in Supplementary Videos S5 and S6), including both atomic fluctuations within the protein chain and motions of the protein chain as a whole within the lattice. B_{chain} captures only the atomic fluctuations within a protein chain, ignoring the rigid-body motions of the individual chains with respect to each other (see SI section 4.4).

The overall profile of the B-factors from both the ff14SB and C36m simulations are consistent with the experimental B-factor profile (Fig. 2c). The correlation with the experimental B-factors is high for both $B_{lattice}$ (Pearson $r = 0.70/0.76$ for ff14SB/C36m) and B_{chain} (Pearson $r = 0.66/0.70$ for ff14SB/C36m). Both force fields accurately describe the

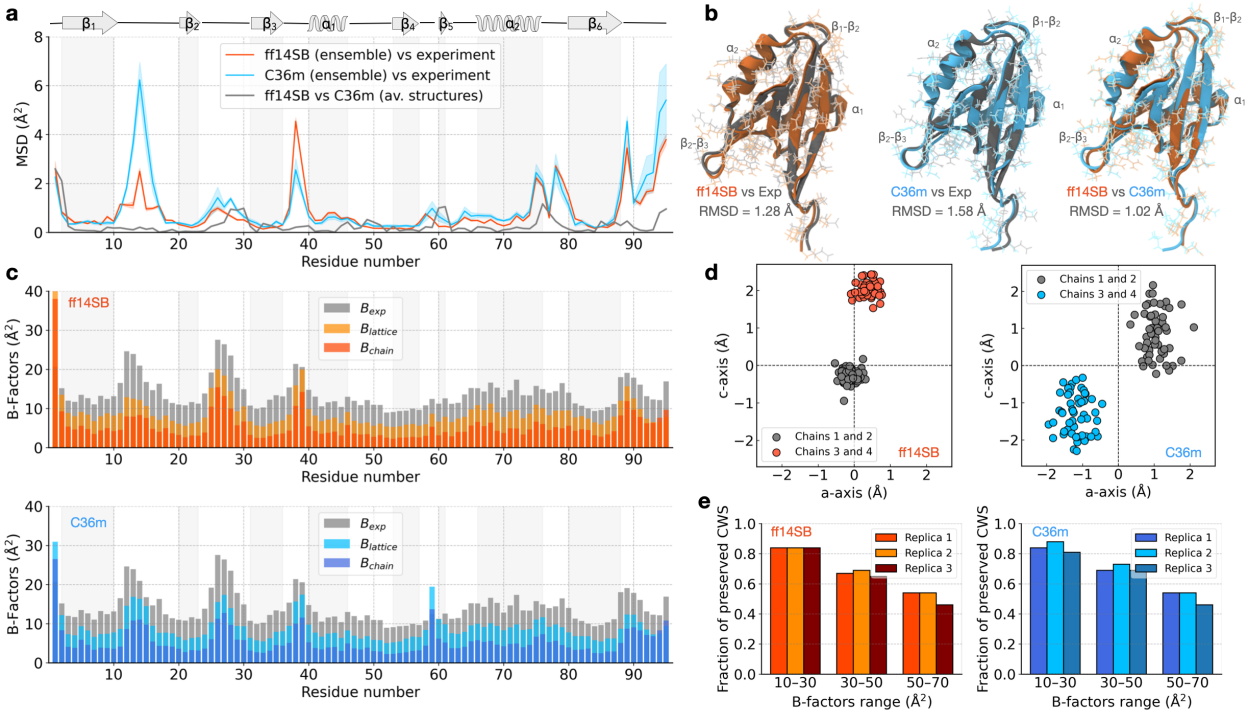

Figure 2: **Assessing the accuracy of the ff14SB and C36m crystal simulations.** (**a**) The average MSD of protein Cα positions relative to the crystal structure, with the shaded envelope representing the mean +/- standard deviation ($n = 324$ chains $= 3$ replicas $\times$ 108 copies); the MSD between average MD structures (ff14SB vs C36m) is shown in gray. (**b**) Comparison between the crystal structure (dark gray) and the average MD structures computed from the simulation ensembles. RMSD (including all heavy atoms) is indicated for each pair of structures. (**c**) Comparison between simulation ($B_{lattice}$ and B_{chain}) and experimental (B_{exp}) B-factors computed for Cα atoms. Note that bars are overlaid, not stacked. (**d**) Projection of protein centers-of-mass on the ac unit cell plane via inverse crystal transformations, where the origin represents positions in an undistorted lattice. A representative snapshot of the supercell at $t = 5$ μs of simulation replica 1 was used. See SI section 3.4 for details. (**e**) The fraction of preserved crystallographic water sites in simulations grouped by experimental B-factors.

relative mobility of the protein loop regions compared to the rest of the protein. Previous MD simulation studies of protein crystals reported similarly high correlations of computed and experimental B-factors.[24,25] Lattice B-factors ($B_{lattice}$) are on average $\sim$5 Å^2 lower than the experimental values for both force fields. This underestimation could be attributed to the idealized model of the lattice (a $3 \times 3 \times 3$ supercell), which is likely insufficient to fully capture the lattice disorder observed in real crystals caused by impurities.[48] The average difference, $B_{lattice} - B_{chain} = 4$ Å^2, represents the contribution of rigid body motions of individual proteins to the overall atomic fluctuations. This value corresponds to lattice vi-

brations of 0.5 Å in amplitude, which is consistent between force fields and is typical for MD simulations of protein crystals.[24,49]

Crystal symmetry and contacts. To assess how well the properties of the crystal lattice are preserved, inverse crystallographic transformations were applied to the positions of each of the 108 protein chains in the supercell simulations (see SI section 3.4). In the ideal case of an undistorted lattice (from which the simulations started, Fig. 1a), all points would remain at the origin. In simulations, due to thermal motion, lattice vibrations occur and the positions of the center of mass of individual chains scatter around their undistorted position (corresponding to the origin). The amplitude of these deviations from the origin indicates the level of disorder in the lattice (Table S2). For the ff14SB simulations, the average amplitude is 1.6 Å, while for the C36m simulations, it is higher, 1.8 Å. To visualize the lattice disorder, we projected transformed chain positions onto the **ac** crystallographic plane (Supplementary Videos S5 and S6). Representative frames are shown in Fig. 2d. In the unit cell, two protein chains (oriented in the same direction, Fig. 1a) are shifted by 2-4 Å relative to the other two, oriented in the opposite direction. This shift, which we refer to as crystal symmetry melting, is caused by anisotropic pressure coupling, which scales the simulation box vectors and reshapes the lattice (see SI section 3.5).

To further investigate this change in the crystal lattice, we analyzed changes in the protein-protein interfaces during the simulation (Fig. S8, Table S3). In the undistorted crystal, 69 out of 95 protein residues are involved in inter-molecular contacts. During the simulations, this number increases to 73 for ff14SB and 74 for C36m. The crystal lattice adopts a more favorable conformation with additional inter-protein contacts compared to the initial structure due to melting symmetry. We located specific regions of the protein where inter-molecular interactions are reformed (Table S3); the same regions also show the highest deviation from the crystal structure (Fig. 2a). These results demonstrate the coupling between protein structure and crystal lattice geometry. In particular, perturbations of the protein structure, followed by changes to the protein-protein interfaces, lead to the distortion

of the crystal lattice, and vice versa.

Crystallographic water. Crystallographic water sites (CWS) are the positions of ordered water molecules that are resolved in a crystal structure. These locations can be detected in MD trajectories and compared to the crystal structure to benchmark the accuracy of modeling crystals in simulations.[43,50,51] We recently developed a method called local alignment for water sites (LAWS) to conduct this type of analysis by explicitly accounting for protein motion in the lattice.[43] The LAWS algorithm determines if CWS are preserved in a simulation based on the local water density around these sites.

Using LAWS, we find that the proportion of preserved CWS is 71% for both the ff14SB and C36m force fields (Table S4), which is consistent with our study of the same PDZ domain crystal modeled as a single unit cell.[43] For both force fields, the CWS with higher confidence (smaller experimental B-factors) are found to be more frequently preserved in simulations than the CWS with lower confidence (higher B-factors), as shown in Fig. 2e. Analysis of the water sites that are not preserved in simulations demonstrates that they are nearly all coordinated by flexible protein regions at perturbed protein-protein interfaces (Tables S3 and S5). These observations suggest that ordered water molecules in the crystal are strongly affected by protein dynamics and changes in the crystal geometry.

Elucidating force field differences using dimensionality reduction

While ff14SB and C36m have similar accuracy in capturing various properties of the protein crystal, there are notable differences; namely, in the average protein structure (Fig. 2b) and at least a ten-fold difference in relaxation time (Fig. 1b and c). To investigate the cause of these differences, we analyzed the conformational ensembles of the PDZ domain obtained in all of the simulations carried out in this study (Table 1). We carried out principal component analysis (PCA) using a set of dihedral angles (Ramachandran ϕ/ψ and Janin χ_1/χ_2 for each residue) as a feature vector. The PCA projection separates the ensembles generated using different force fields (ff14SB, C36m, and ff94) into distinct, non-overlapping basins (Fig. 3a),

indicating that the protein conformational space strongly depends on the force field. We note that the separation between force fields is observed not only for the crystal but also for the ensembles obtained in solution (Fig. 3a). The region of conformational space sampled by the simulations predominantly depends on the force field, irrespective of the protein environment. Interestingly, we would be able to deduce which force field was used to generate a simulated protein conformation with near certainty by projecting it onto this PCA space. While, in general, it is reasonable to expect that force fields will generate ensembles that differ, the extent of these differences (i.e. the fact that the ensembles are non-overlapping) is surprising.

The crystal structure of the PDZ domain (black star, Fig. 3a), from which all simulations were initialized, is found in the ff14SB region. This result suggests that the ff14SB ensemble more accurately represents the crystal structure compared to the other force fields, consistent with the previous analysis of simulation accuracy (Figs. 1 and 2). For the ff94 and C36m simulations, the transition from the initial crystal structure to the corresponding force field basin in the PCA projection occurs within the first 1 ns. Such rapid transitions suggest that local, rather than global, conformational changes are involved.

To understand these local conformational changes, we analyzed the coefficients of the first principal component (PC 1), which identifies residues responsible for separating the force field ensembles (Fig. S10). When considering the backbone dihedral angles, the residues that contribute most to PC 1 are located in the protein regions that show significant deviation between the average MD structures (Fig. S10a and Fig. 2a). When considering the side-chain dihedral angles, we find that all five glutamine residues have a high contribution to PC 1 (Fig. S10b), meaning that glutamine side chains are a primary determinant separating the ff14SB and C36m ensembles. Indeed, when analyzing glutamine χ_1 and χ_2 distributions, we find significant differences between the force fields in terms of the populations of rotameric states (Fig. 3c). As a control, the distribution of χ_1 and χ_2 rotamers for the residues that contribute the least to PC 1 are consistent between ff14SB and C36m (Fig. S11), establishing that the differences between force fields observed for glutamine residues are meaningful.

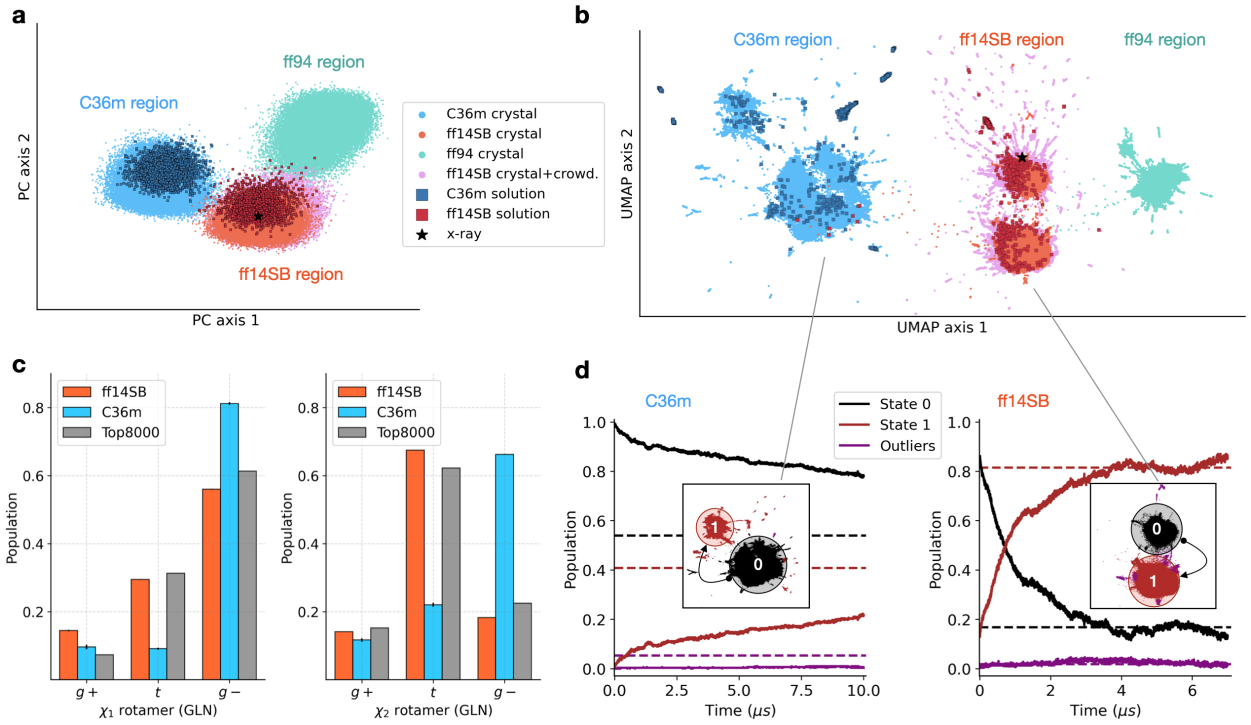

Figure 3: **Force field differences elucidated with two-dimensional projection of the protein conformational space and Markov state models**. (**a**) Two-dimensional PCA. Dihedral angles (Ramachandran and Janin) were used as features. The first two principal components account for 14% of the total variance in the data set. Each point represents the structure of the PDZ domain sampled from the crystal or solution simulations using one of the force fields: ff14SB, C36m or ff94. (**b**) Two-dimensional UMAP projection of the same space with the same legend as in (**a**). (**c**) Distributions of χ_1 and χ_2 rotameric states for glutamine residues in the ff14SB and C36m simulations, as well as the population of these rotameric states in the Top8000 dataset.[52] Error bars represent standard error over $n = 3$ replicas. (**d**) The population dynamics of a three-state Markov model for the C36m and ff14SB crystal simulations, respectively. The state population is defined as the number of individual chain trajectories occupying the state at a given time divided by 324 (3 replicas $\times$ 108 chains). Dashed lines show equilibrium populations computed from the MSM transition matrix. The insets show the corresponding force field region from the UMAP projection in (**b**) colored by the state. The structural differences between the dominant states (states 0 and 1) in each force field are provided in Fig. S14.

In order to determine if the difference in glutamine rotameric states is unique to this PDZ domain, or a more general discrepancy between ff14SB and C36m, we extended this analysis to MD simulations of two other protein systems (SI section 3.10). Similar to the results for the PDZ domain, these simulations also show significant differences between force fields for the glutamine rotamers (Fig. S12), demonstrating that these differences are not unique to either protein crystal simulations or this PDZ domain. To uncover which of the two force fields

more accurately captures glutamine rotameric states, we carried out the same analysis on the Top8000 dataset,[52] which includes high-resolution crystal structures of diverse proteins from the Protein Data Bank (PDB). The ff14SB force field is more consistent with the Top8000 dataset compared to C36m (Figs. 3c and S12). We note that this analysis is limited, as the backbone dependence of rotameric states is not taken into account here. Nevertheless, the comparison to the Top8000 dataset strongly suggests that ff14SB more accurately represents the populations of the side-chain rotameric states of glutamine residues.

Accurately capturing the correct populations of rotameric states is important when using MD simulations to understand the functional mechanism of proteins. For example, the rotameric state distribution of a specific glutamic acid residue in the potassium channel MthK was found to differ between ff14SB and C36m.[53] In our analysis, several (but not all) glutamic acid residues contribute significantly to PC 1 (Fig. S10b), indicating their contribution to the difference between force fields. Taken together, these results suggest that the C36m dihedral energy parameters may need to be reevaluated to more accurately represent populations of rotameric states (see SI section 3.11 for additional analysis).

To identify the conformational changes that are responsible for the slow relaxation observed for both force fields, we employed a Markov state model (MSM) approach. MSMs can be used for identifying slow conformational transitions, estimating the timescales of these transitions, and predicting equilibrium state populations.[54] To define conformational states, we utilized universal manifold approximation and projection (UMAP, Fig. 3b), which has previously been applied to construct MSMs for long timescale protein simulations.[55] Consistent with PCA (Fig. 3a), the UMAP projection separates the conformational ensembles into distinct regions based on force field. UMAP axis 1 corresponds to rapid side-chain rearrangements, which is similar to PC 1. We constructed MSMs for both the ff14SB and C36m ensembles using the UMAP projection to classify structures into states (see Methods). Based on analysis of the MSMs (Figs. 3d and S14), slow conformational transitions involving backbone structural changes occur along UMAP axis 2. In the ff14SB simulations, these slow

transitions involve changes in the conformation of the C-terminal tail (Fig. S14a), while in C36m, the transitions occur mainly in the β_1-β_2 loop and the α_2 helix (Fig. S14b). The population of the two dominant conformational states converges exponentially with characteristic timescales of 1.1 μs for ff14SB and 16.8 μs for C36m (Fig. 3d). These timescales are consistent with the timescales estimated from structural analysis (Figs. 1 and S2), confirming that the crystal simulations using ff14SB, and not C36m, have reached equilibrium.

Taken together, the results shown in Fig. 3 provide a detailed account of the conformational changes leading to the equilibration of the PDZ domain crystal. Starting from the crystal structure, the proteins in the lattice undergo fast (ns-timescale) conformational changes involving reorientation of the side-chains. These fast conformational changes result in the separation of the ensembles into distinct force field regions in both the PCA and UMAP projections, and they are followed by slow (μs-timescale) rearrangements of the backbone structure. It is the slow backbone conformational changes that correspond to the long timescales needed to reach equilibrium.

Functional relevance of the protein motions in the crystal

Based on the comparison to experimental data and the structural analysis presented above, the supercell simulated with ff14SB (Fig. 1) provides the most suitable model for the crystal at equilibrium. Using this simulation setup, we have accumulated nearly 2 ms of equilibrium sampling for a single protein chain (Table 1), which can be used to study conformational heterogeneity and motions in the crystal. To determine the motions with the largest amplitudes, we analyzed the pairwise Cα distances in the equilibrium portion of the trajectories using PCA (Fig. 4a). Based on the coefficients of the first PC, the dominant conformational change corresponds to the motion of the β_1-β_2 loop (Fig. S15a), while the second PC represents the much smaller motion of the β_2-β_3 loop (Fig. S15b). The relative position of the β_1-β_2 loop can be described by a single distance between two residues, Ser14 and Gly78, spanning a range from 4 to 15 Å (indicated in Fig. 4b,c). The large range of motions for this

loop (10 Å) is remarkable given the tightly packed environment of the crystal. Such a high conformational heterogeneity cannot be inferred from the crystallographic data alone.[56,57]

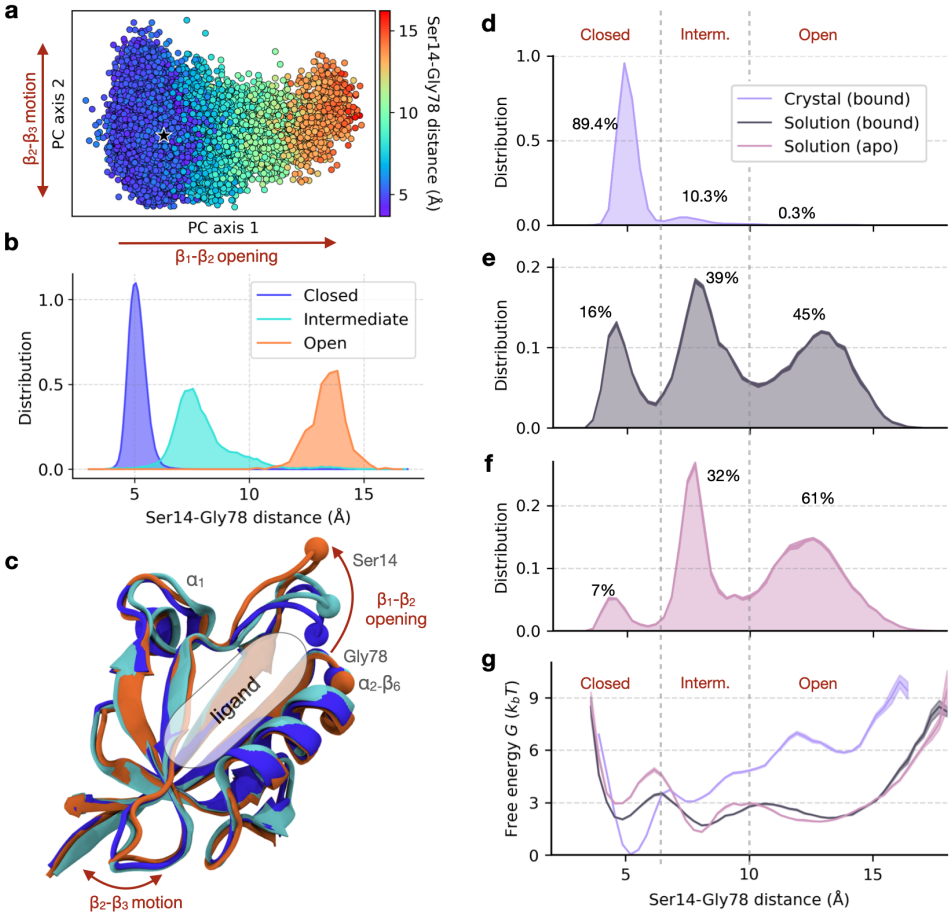

Figure 4: **Effect of environment and ligation state on the β_1-β_2 clamping motion.** (**a**) A two-dimensional PCA projection (using pairwise Cα distances) of the protein conformational space in the ff14SB crystal simulations, colored by the Ser14-Gly78 distance. The first principal component (15% of the variance) represents the motion of the β_1-β_2 loop, while the second principal component (7% of the variance) primarily captures the fluctuations of the β_2-β_3 loop. The crystal structure is shown with a black star. (**b**) The distribution of the Ser14-Gly78 distance in the crystal for each of the three states: open, intermediate, and closed. (**c**) Representative structures of the three loop states, colored as in **b**. The ligand-binding site (indicated) is occupied by the C-terminal tail of the neighbouring chain. (**d–g**) The distribution of the Ser14-Gly78 distance and free energy profile in each simulation ensemble: crystal (ligand-bound, **d**), solution (ligand-bound,**e**) and solution (ligand-free, **f**). The percent population of each state is indicated on each plot (**d–f**). (**g**) Free energy is computed as $G = -k_B T \log P(x)$, where $P(x)$ is the probability distribution for the Ser14-Gly78 distance. The shaded envelope represents standard error computed from bootstrapping with $n = 5$.

In the equilibrium ensemble, we identified three conformational states (Fig. S16), which are well-described by the position of the β_1-β_2 loop (Fig. 4b,c). The distributions of the

Ser14-Gly78 distance indicate that these states correspond to *closed*, *intermediate*, and *open* loop conformations (Fig. 4b). While the open state is characterized by stable contacts between the loop and the α_1 region, the closed state instead forms contacts between the loop and the α_2-β_6 region (Figs. 4c, S17).

Our next goal was to understand whether the three conformational states found in the crystal are related to the functional activity of the PDZ domain — namely, binding of a ligand, which is known to induce conformational changes in PDZ domains.[37,58] The functional relevance of these conformational states can be addressed by comparing ligand-bound and ligand-free ensembles of the protein. Two additional systems of the PDZ domain in solution were simulated, one in which the PDZ is ligand-bound and the other which is ligand-free. Note that the supercell simulations represent a ligand-bound form, as the C-terminal tail of each chain is bound to its neighbour's ligand binding site (see Methods, Fig. 1a). With these three simulation systems, we can study the effect of environment (crystal vs. solution) in addition to the effect of ligand binding on the conformation of the β_1-β_2 loop. Representative simulations for each of the three systems are shown in Supplementary Videos S1, S3 and S4.

From the distributions of the Ser14-Gly78 distance in the three simulation systems (Fig. 4d-f), we can determine the population of the β_1-β_2 loop states. In the crystal lattice (ligand-bound), the closed state dominates the ensemble, while the open state is rarely observed (Fig. 4d). In contrast, in solution (both the ligand-bound and ligand-free simulations), the open state has the highest population (Fig. 4e-f). The conformational ensemble in the crystal has a much lower structural diversity than in solution, as demonstrated by conformational entropy: $S_{cryst} = -\Sigma_i p_i \ln(p_i) \approx 3$ vs. $S_{sol} \approx 7$. The crowded crystal environment stabilizes a more compact state of the protein, with a closed β_1-β_2 loop, due to inter-protein contacts.

Next, we address the effects of ligand binding on the conformational ensemble of the PDZ domain in solution. The presence of a ligand in the active site results in a two-fold increase in the population of the closed state (Fig. 4e-f). The free energy profiles show that

the ligand decreases the free energy of the closed state by $\sim$1 k_BT and lowers the free energy barrier separating the closed and intermediate states by $\sim$2 k_BT (Fig. 4g). Therefore, ligand binding stabilizes the closed state, suggesting that the β_1-β_2 loop motion is important in the functional activity of the PDZ domain. Indeed, the "clamping" of the β_1-β_2 loop upon ligand binding has been observed in other PDZ domains,[58–61] supporting the functional relevance of these states.

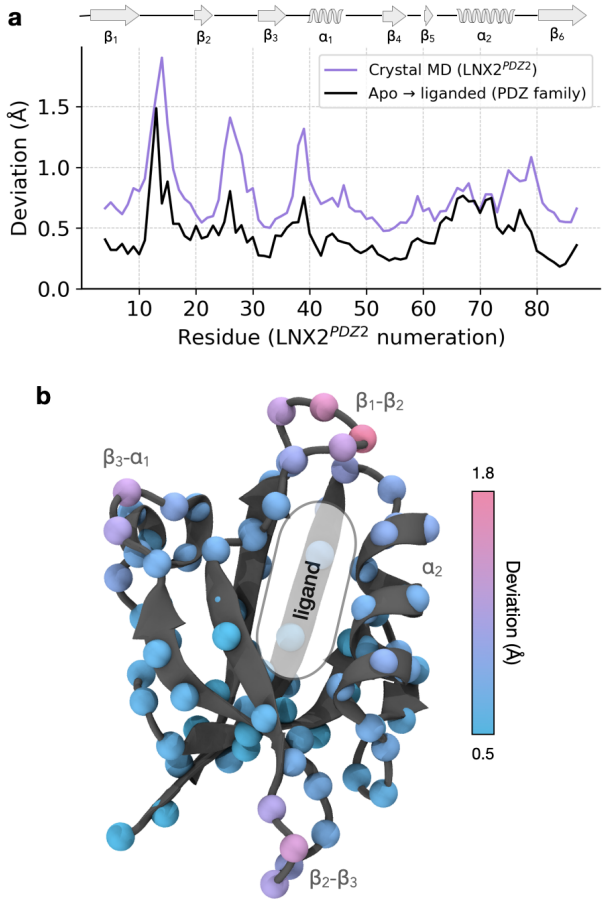

Figure 5: **The relationship between equilibrium motions in the crystal and PDZ domain function.** (**a**) The median deviation profile for the equilibrium ensemble in the crystal (purple) and ligand-induced deviation of backbone atoms per residue (LNX2^{PDZ2} numbering) in the ensemble of 10 PDZ homologues (black). (**b**) Deviations due to thermal fluctuations in the crystal mapped onto the Cα atoms of the LNX2^{PDZ2} crystal structure.

We have shown that the dominant motion in the crystal (β_1-β_2 loop motion) is functionally relevant. We extended our analysis to probe the functional relevance of the other

equilibrium motions in the crystal. In particular, we assessed the similarity between the conformational changes occurring in the crystal simulations and those induced by ligand binding across crystal structures of the PDZ domain family. The overall profile of observed atomic deviations in the equilibrium ensemble correlates strongly with the pattern of ligand-induced structural changes in other PDZ domains (Fig. 5a; Pearson $r = 0.7$ for $n = 84$ residues, P<0.0001). We find that the conformational changes in our simulations are the most pronounced in the β_1-β_2 and β_2-β_3 loops, as well as the α_1 region, consistent with crystal structures of PDZ domains (Fig. 5a-b). We note that the motions of the loops were also captured by the PCA analysis (Fig. 4a). The same regions are identified using an analysis of strain energy (Fig. S18), which is a measure that does not require explicit structural alignment of the protein conformations.

Together, these results suggest that the global pattern of equilibrium protein motions in the crystal is consistent with both local and non-local (allosteric) conformational changes induced by ligand binding. These changes occur in conserved functional sites of the PDZ domain family (the α_1 region and the β_1-β_2 loop[59]) — the same regions where conformational changes were induced by an electric field in the EFX experiment.[6] Our findings suggest that regardless of the type of perturbation, whether caused by ligand binding, electric fields, or thermal fluctuations, the dynamic response of the protein structure appears to remain the same. Therefore, from the equilibrium simulations alone, it may be possible to study the functionally relevant conformational changes that are probed in the non-equilibrium pump-probe experiments.

Discussion

In this work, we have studied the equilibrium dynamics of a PDZ domain in a crystal supercell with a total sampling time of over 7 ms. We address two major methodological challenges: (1) accurate modeling of the crystal environment, and (2) the long timescales required for

equilibration. In determining the most accurate simulation setup for the protein crystal, we found that structural ensembles generated by two widely used force fields (Amber ff14SB and CHARMM36m) occupy distinct regions of conformational space due to differences in side-chain dihedral angle distributions. Using experimental data, we identify potential parameters for improving protein force fields, but also provide insights into protein motions unattainable by experimental approaches alone. In proteins, local reorientation of side-chains happens quickly on a nanosecond timescale, while it can take several μs for collective side-chain and backbone configurations to stabilize. Importantly, the motions observed in our simulations were found to recapitulate the functional motions of PDZ domains,[6] validating their use for probing intramolecular mechanisms.

It is worth pointing out the critical importance of sub-μs motions in understanding protein function, both in terms of the dynamics defining the reaction coordinate and associated with allosteric regulation. Access to protein dynamics on this time-scale and with atomic resolution has been the motivating rationale for the new pump-probe time-resolved diffraction experiments.[6] While effective, the practical complexities of these experiments demand a parallel validated computational strategy to make predictions and to reduce the data such that intuition about protein mechanism can emerge. The present study provides the technical foundation for this process - a virtuous cycle between molecular simulations and experiments to systematically analyze the dynamics associated with protein function.

Methods

Model building. The second PDZ domain of the human E3 ubiquitin ligase Ligand-of-Numb protein X2 ($LNX2^{PDZ2}$) was used as a simulation system. The protein structure (PDB ID: 5E11)[6] with the highest occupancy was used among alternative conformations. The CHARMM-GUI web-server[62] was used to add atoms that were missing in the crystal structure and to reconstruct a triclinic unit cell of the $C121$ space group with four symmetrically-

related copies of the protein (Fig. 1a). The unit cell parameters were $a = 65.30$ Å, $b = 39.45$ Å, $c = 39.01$ Å and $\alpha = \gamma = 90°$, $\beta = 117.54°$. For all crystal simulations, a supercell layout was used, which consisted of a $3 \times 3 \times 3$ layout of unit cells (Fig. 1a). Two types of solution simulations were carried out: ligand-bound and ligand-free. The ligand is a four-residue peptide (Glu-Ile-Glu-Leu), as in the crystal structure.[6] The solution simulation system consisted of a single protein chain in a rhombic dodecahedron simulation box, with the diameter of the system based on the maximum distance between protein atoms plus 30 Å. The protonation states of the protein residues were determined using PROPKA[63] at pH 4.5, consistent with the conditions of the EFX experiment.[6] Na^+ and Cl^- ions were used to neutralize each system at a salt concentration of 35 mM. To solvate the crystal lattice, we used a simplified crystal environment (water) and an environment with crowders to match the experimental conditions of the crystallization buffer[6] (see SI section 4.1 for details). Three force field/water model combinations were used in simulations: (i) Amber ff14SB[38] + TIP3P water model,[64] (ii) Amber ff94[65] + TIP3P water model,[64] and (iii) CHARMM36m + CHARMM-modified TIP3P water model.[48] The total number of atoms in each system is provided in Table S6.

Simulation protocol. All simulations were conducted using GROMACS version 2019.1.[66] Periodic boundary conditions were employed. The time step of the simulation was 2 fs. The LINCS algorithm was used to constrain the covalent bonds involving hydrogen atoms.[67] The cut-off for short-range electrostatics and Van der Waals interactions was 9.5 Å. The smooth particle mesh Ewald (PME) method with a Fourier spacing of 1.2 Å and a fourth-order interpolation[68] was utilized. The velocity rescaling thermostat[69] was employed for constant temperature simulations ($T = 289$ K). Compressibility values of 2.5×10^{-5} bar^{-1} and 4.5×10^{-5} bar^{-1} were used for the supercell and solution simulations, respectively.[70] For the supercell simulations, the following simulation protocol was used. Following energy minimization, simulations with position restraints on all protein atoms (1000 kJ/mol nm^2) were carried out, followed by a simulation in the NVT ensemble for 10 ns. Two types of NPT sim-

ulations were then performed: (i) 10 ns of simulation using an isotropic Berendsen barostat[71] at P = 1 bar, followed by (ii) a 40 ns simulation using the anisotropic Parrinello-Rahman barostat.[72] These simulations were continued for production. For the solution simulations, the same simulation protocol was used except that isotropic pressure coupling was used in the Parrinello-Rahman barostat rather than anisotropic. Three replicas were run for each supercell system, 10 replicas for each apo solution simulation, and 9 replicas for ligand-bound simulations in solution. All of these systems were initialized from the same atomic coordinates and randomly seeded velocities. The total simulation time and combined sampling for an individual PDZ domain are listed in Table 1. Additional details on model building and simulations are provided in Supplementary Methods (SI section 4).

Table 1: MD simulations of a PDZ domain in the study.

#	Simulation setup	Force Field / Water Model	No. of replicas × sim. time	Total sampling for a single chain
1	crystal (water) $3 \times 3 \times 3$ supercell	C36m / TIP3P*	3×10 μs	3300 μs
2	crystal (water) $3 \times 3 \times 3$ supercell	ff14SB / TIP3P	3×7 μs	2300 μs
3	crystal (water) $3 \times 3 \times 3$ supercell	ff94 / TIP3P	3×2 μs	600 μs
4	crystal (water + crowders) $3 \times 3 \times 3$ supercell	ff14SB / TIP3P	3×1.7 μs	540 μs
5	solution (water) single PDZ domain	C36m / TIP3P*	10×4.5 μs	45 μs
6	solution (water) single PDZ domain	ff14SB / TIP3P	10×3 μs	30 μs
7	solution (water) single PDZ domain + ligand	ff14SB / TIP3P	9×2 μs	18 μs

* CHARMM-modified TIP3P

Analysis. The structural analysis of the MD trajectories was carried out using the MDAnalysis[73] package for Python. The detailed algorithms for computing RMSD, MSD, B-factors, residue contacts, and motivation for the choice of parameters are provided in SI section 4.4. Visual Molecular Dynamics (VMD) 1.9.4[74] and ChimeraX[75] were used for visualizing structures. The LAWS algorithm was used to determine if CWS were preserved in the supercell simulations.[43]

Average MD structures. The average MD structure for each force field was computed from the supercell simulations (3 replicas × 108 chains per replica for a total of 324 individual trajectories). The structure at each frame was aligned to the crystal structure using backbone

atoms. Using these trajectories, 324 time-averaged structures were computed, which were then averaged to produce a single average structure. It should be noted that the resulting average structure is non-physical, because realistic atomic geometries were not enforced. To produce a "physical" average structure (shown in Fig. 2b), the entire simulation ensemble was then scanned to identify the conformation that is the most similar to the average, i.e. the structure with the minimum heavy-atom RMSD (0.70 Å for C36m and 0.72 Å for ff14SB).

Conformational space. Normalized from -1 to 1 by sine and cosine transformations, 93 Ramachandran (ψ, ϕ) and 60 Janin (χ_1, χ_2) pairs of angles produce a $(93+60) \times 2 \times 2 = 612$ dimensional feature vector. PCA[76,77] and UMAP[78] were then applied to the combined ensemble sampled in crystal and solution simulations (Fig. 3a,b). Protein conformations were sampled every 10 ns from all simulation trajectories (Table 1), resulting in 726,912 protein conformations in total. The *sklearn 1.2.0* Python package[79] was used to perform PCA. The *umap-learn 0.4.6* package with parameters $min\_dist = 0.1$ and $n\_neighbors = 25$ was used to compute the UMAP projection. The resulting embeddings were not sensitive to the variation of these parameters.

Markov state models. Markov state models[54] for the crystal systems were constructed from the trajectories of isolated protein chains projected on the UMAP space. For the purpose of constructing MSMs, all protein chains in the supercell are considered to be independent. Although this assumption is not strictly true due to interaction between chains, it is a reasonable approximation as all chains experience nearly identical environments. For each force field (ff14SB and C36m), three-state MSMs were constructed. In these models, state 0 represents the initial state occupied by all protein chains in the lattice at the start of the simulation. State 1 is the state to which the majority of the population gradually transitions (i.e. the state with the highest probability inflow). All other structurally diverse states with low populations were aggregated into a single state called *outliers*, which has a total combined population of $< 5\%$. States 0 and 1 describe the majority of the conformational ensemble, as they have much larger populations than the outliers. Consequently, the

transition from state 0 to state 1 represents a global conformational change, which is the dominant contribution to equilibration. Therefore, the equilibration period of the crystal lattice can be estimated from the characteristic timescale of this transition provided by the MSM.

MSMs were constructed using MSMBuilder.[80] For each force field, a total of 324 individual trajectories (3 replicas $\times$108 chains) with 4096 frames per trajectory were used. Initial geometric clustering with the Mini Batch K Means[81] algorithm ($n_clust = 200$) was used to define microstates which were joined into three macrostates with the PCCA+ algorithm[82] such that the states with insignificant equilibrium population ($< 5\%$) were merged together. The lag time of 600 ns was chosen according to the convergence of the microstates transition timescales vs. lag time graph[83] (Fig. S22). Transition timescales for the macrostates were computed from the eigenvalues of the transition matrix.[54]

Equilibrium protein motions. Only the equilibrium portion ($t > 1.5$ μs) of the supercell simulations using the ff14SB force field was used in the analysis of equilibrium protein motions (Figs. 4 and 5). PCA was carried out using pairwise distances between Cα atoms as a feature vector in order to identify the dominant motions in the system. We excluded the C-terminal tail (residues 89–95) from this analysis for two reasons: (i) the C-terminal tail is not part of the native protein and was included to promote crystallization by serving as a self-binding motif extension,[6,84] and (ii) this tail is shown in the MSM (Fig. S14a) as being involved in a long timescale conformational relaxation. The conformational space was projected onto the two-dimensional PCA (Fig. 4a) and UMAP (using parameters $min\_dist = 0.0$ and $n\_neighbors = 20$, Fig. S16).

Structural analysis of related PDZ domains. Ten pairs of high-resolution (< 2 Å) crystal structures of PDZ domains with and without ligand were used: E3-LNX^{PDZ}: 3VQG vs. 3VQF; PSD-95^{PDZ3}: 1BFE and 6QJJ vs. 1BE9; PTP-1E^{PDZ2}: 3LNX vs. 3LNY; GRIP-1^{PDZ6}: 1N7E vs. 1N7F; SAP97^{PDZ2}: 2AWU vs. 2AWW; PALS-1^{PDZ}: 4UU6 vs. 4UU5; Erbin^{PDZ}: 2H3L vs. 1MFG; Dishevelled^{PDZ}: 2F0A vs. 1L6O; PDZK-1^{PDZ3}: 3R68 vs.

3R69. These PDZ domains were chosen because they have the highest sequence similarity to LNX2^{PDZ2} and they have both apo and ligand-bound crystal structures available. Apo and liganded structures were first aligned to each other in PyMOL[85] using *align* for Cα atoms, then aligned to the crystal structure of LNX2^{PDZ2} (PDB ID 5E11) using *super*. The median (over PDZ domains) ligand-induced deviation of $C\alpha$ atoms (Fig. 5a, black) were computed as described in Fig. 5e of Hekstra et al.[6] Only residues with matching positions in LNX2^{PDZ2} for each pair of apo and bound structures were included in the analysis. The deviation profile for the crystal ensemble (Fig. 5a, purple) is defined as the mean per-residue deviation between $C\alpha$ atoms in $N = 2000$ randomly sampled pairs of structures from the equilibrium ensemble in ff14SB.

Author Contributions

E.K., J.S.K., L.M., R.R., and S.R. conceptualized the research; J.S.K, E.K., L.M., and E.L. carried out the simulations; E.K. developed analytic tools and designed the figures; E.K., J.S.K., V.V., and S.R. analyzed the data; all authors wrote and revised the manuscript.

Declaration of Interests

The authors declare no competing interests.

Acknowledgement

This research was supported by a Connaught New Researcher Award and a Natural Science and Engineering Research Council of Canada (NSERC) Discovery Grant to S.R. This research was also enabled in part by support provided by the Digital Research Alliance of Canada (alliancecan.ca). 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. L.M. was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number F32GM134721. We thank Doeke Hekstra and members of the Ranganathan Lab for discussions. R.R. acknowledges support from NIH grants RO1GM12345, RO1GM141697, and P41GM118217 from the National Institute for General Medical Sciences.

Supporting Information Available

Supplementary Information document includes Figures S1–S22, Tables S1–S6, and description of Videos S1–S6. It is organized into four sections:

1. Prior MD simulation studies of protein crystals
2. Supplementary Videos
3. Supplementary Results
4. Supplementary Methods

The data underlying Figs. 1–5 as well as the initial and final conformations of the MD simulations for all systems have been deposited to the Zenodo repository and are available from `https://doi.org/10.5281/zenodo.7987473`.

References

(1) Henzler-Wildman, K.; Kern, D. Dynamic personalities of proteins. *Nature* **2007**, *450*, 964–972, DOI: `10.1038/nature06522`.

(2) Miller, M. D.; Phillips, G. N. Moving beyond static snapshots: Protein dynamics and the Protein Data Bank. *Journal of Biological Chemistry* **2021**, *296*, 100749, DOI: `10.1016/j.jbc.2021.100749`.

(3) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The Energy Landscapes and Motions of Proteins. *Science* **1991**, *254*, 1598–1603, DOI: 10.1126/science.1749933.

(4) Wolff, A. M.; Nango, E.; Young, I. D.; Brewster, A. S.; Kubo, M.; Nomura, T.; Sugahara, M.; Owada, S.; Barad, B. A.; Ito, K.; Bhowmick, A.; Carbajo, S.; Hino, T.; Holton, J. M.; Im, D.; O'Riordan, L. J.; Tanaka, T.; Tanaka, R.; Sierra, R. G.; Yumoto, F.; Tono, K.; Iwata, S.; Sauter, N. K.; Fraser, J. S.; Thompson, M. C. Mapping Protein Dynamics at High-Resolution with Temperature-Jump X-ray Crystallography. *bioRxiv* **2022**, DOI: 10.1101/2022.06.10.495662.

(5) Mehrabi, P.; Schulz, E. C.; Dsouza, R.; Müller-Werkmeister, H. M.; Tellkamp, F.; Miller, R. J. D.; Pai, E. F. Time-resolved crystallography reveals allosteric communication aligned with molecular breathing. *Science* **2019**, *365*, 1167–1170, DOI: 10.1126/science.aaw9904.

(6) Hekstra, D. R.; White, K. I.; Socolich, M. A.; Henning, R. W.; Šrajer, V.; Ranganathan, R. Electric-field-stimulated protein mechanics. *Nature* **2016**, *540*, 400–405, DOI: 10.1038/nature20571.

(7) Gruhl, T.; Weinert, T.; Rodrigues, M. J.; Milne, C. J.; Ortolani, G.; Nass, K.; Nango, E.; Sen, S.; Johnson, P. J. M.; Cirelli, C.; Furrer, A.; Mous, S.; Skopintsev, P.; James, D.; Dworkowski, F.; Båth, P.; Kekilli, D.; Ozerov, D.; Tanaka, R.; Glover, H.; Bacellar, C.; Brünle, S.; Casadei, C. M.; Diethelm, A. D.; Gashi, D.; Gotthard, G.; Guixà-González, R.; Joti, Y.; Kabanova, V.; Knopp, G.; Lesca, E.; Ma, P.; Martiel, I.; Mühle, J.; Owada, S.; Pamula, F.; Sarabi, D.; Tejero, O.; Tsai, C.-J.; Varma, N.; Wach, A.; Boutet, S.; Tono, K.; Nogly, P.; Deupi, X.; Iwata, S.; Neutze, R.; Standfuss, J.; Schertler, G.; Panneels, V. Ultrafast structural changes direct the first molecular events of vision. *Nature* **2023**, *615*, 939–944, DOI: 10.1038/s41586-023-05863-6.

(8) Fraser, J. S.; van den Bedem, H.; Samelson, A. J.; Lang, P. T.; Holton, J. M.; Echols, N.;

Alber, T. Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proceedings of the National Academy of Sciences* **2011**, *108*, 16247–16252, DOI: 10.1073/pnas.1111325108.

(9) Du, S.; Wankowicz, S. A.; Yabukarski, F.; Doukov, T.; Herschlag, D.; Fraser, J. S. Refinement of Multiconformer Ensemble Models from Multi-temperature X-ray Diffraction Data. *bioRxiv* **2023**, DOI: 10.1101/2023.05.05.539620.

(10) Gunsteren, W. F. V.; Karplus, M. Protein dynamics in solution and in a crystalline environment: a molecular dynamics study. *Biochemistry* **1982**, *21*, 2259–2274, DOI: 10.1021/bi00539a001.

(11) van Gunsteren, W. F.; Berendsen, H. J.; Hermans, J.; Hol, W. G.; Postma, J. P. Computer simulation of the dynamics of hydrated protein crystals and its comparison with x-ray data. *Proceedings of the National Academy of Sciences* **1983**, *80*, 4315–4319, DOI: 10.1073/pnas.80.14.4315.

(12) Berendsen, H.; van Gunsteren, W.; Zwinderman, H.; Geurtsen, R. Simulations of Proteins in Water. *Annals of the New York Academy of Sciences* **1986**, *482*, 269–286, DOI: 10.1111/j.1749-6632.1986.tb20961.x.

(13) Avbelj, F.; Moult, J.; Kitson, D. H.; James, M. N. G.; Hagler, A. T. Molecular dynamics study of the structure and dynamics of a protein molecule in a crystalline ionic environment, Streptomyces griseus protease A. *Biochemistry* **1990**, *29*, 8658–8676, DOI: 10.1021/bi00489a023.

(14) Héry, S.; Genest, D.; Smith, J. C. Fluctuation and Correlation in Crystalline Lysozyme. *Journal of Chemical Information and Computer Sciences* **1997**, *37*, 1011–1017, DOI: 10.1021/ci970234a.

(15) Stocker, U.; Spiegel, K.; van Gunsteren, W. On the similarity of properties in solution

or in the crystalline state: A molecular dynamics study of hen lysozyme. *Journal of Biomolecular NMR* **2000**, *18*, 1–12, DOI: 10.1023/a:1008379605403.

(16) Meinhold, L.; Smith, J. C. Fluctuations and Correlations in Crystalline Protein Dynamics: A Simulation Analysis of Staphylococcal Nuclease. *Biophysical Journal* **2005**, *88*, 2554–2563, DOI: 10.1529/biophysj.104.056101.

(17) Joti, Y.; Nakagawa, H.; Kataoka, M.; Kitao, A. Hydration-Dependent Protein Dynamics Revealed by Molecular Dynamics Simulation of Crystalline Staphylococcal Nuclease. *The Journal of Physical Chemistry B* **2008**, *112*, 3522–3528, DOI: 10.1021/jp710039p.

(18) Hu, Z.; Jiang, J. Assessment of biomolecular force fields for molecular dynamics simulations in a protein crystal. *Journal of Computational Chemistry* **2010**, *31*, 371–380, DOI: 10.1002/jcc.21330.

(19) Cerutti, D. S.; Freddolino, P. L.; Duke, R. E.; Case, D. A. Simulations of a Protein Crystal with a High Resolution X-ray Structure: Evaluation of Force Fields and Water Models. *The Journal of Physical Chemistry B* **2010**, *114*, 12811–12824, DOI: 10.1021/jp105813j.

(20) Janowski, P. A.; Cerutti, D. S.; Holton, J.; Case, D. A. Peptide Crystal Simulations Reveal Hidden Dynamics. *Journal of the American Chemical Society* **2013**, *135*, 7938–7948, DOI: 10.1021/ja401382y.

(21) Wall, M. E.; Benschoten, A. H. V.; Sauter, N. K.; Adams, P. D.; Fraser, J. S.; Terwilliger, T. C. Conformational dynamics of a crystalline protein from microsecond-scale molecular dynamics simulations and diffuse X-ray scattering. *Proceedings of the National Academy of Sciences* **2014**, *111*, 17887–17892, DOI: 10.1073/pnas.1416744111.

(22) Xue, Y.; Skrynnikov, N. R. Ensemble MD simulations restrained via crystallographic data: Accurate structure leads to accurate dynamics. *Protein Science* **2014**, *23*, 488–507, DOI: 10.1002/pro.2433.

(23) Li, Y.; Zhang, J. Z. H.; Mei, Y. Molecular Dynamics Simulation of Protein Crystal with Polarized Protein-Specific Force Field. *The Journal of Physical Chemistry B* **2014**, *118*, 12326–12335, DOI: 10.1021/jp503972j.

(24) Janowski, P. A.; Liu, C.; Deckman, J.; Case, D. A. Molecular dynamics simulation of triclinic lysozyme in a crystal lattice. *Protein Science* **2015**, *25*, 87–102, DOI: 10.1002/pro.2713.

(25) Wall, M. E. Internal protein motions in molecular-dynamics simulations of Bragg and diffuse X-ray scattering. *IUCrJ* **2018**, *5*, 172–181, DOI: 10.1107/s2052252518000519.

(26) Zhu, T.; Wu, C.; Song, J.; Reimers, J. R.; Li, Y. Polarization effect within a protein crystal: A molecular dynamics simulation study. *Chemical Physics Letters* **2018**, *706*, 303–307, DOI: 10.1016/j.cplett.2018.06.018.

(27) Wych, D. C.; Fraser, J. S.; Mobley, D. L.; Wall, M. E. Liquid-like and rigid-body motions in molecular-dynamics simulations of a crystalline protein. *Structural Dynamics* **2019**, *6*, 064704, DOI: 10.1063/1.5132692.

(28) Meisburger, S. P.; Case, D. A.; Ando, N. Diffuse X-ray scattering from correlated motions in a protein crystal. *Nature Communications* **2020**, *11*, 1271, DOI: 10.1038/s41467-020-14933-6.

(29) Wych, D. C.; Aoto, P. C.; Vu, L.; Wolff, A. M.; Mobley, D. L.; Fraser, J. S.; Taylor, S. S.; Wall, M. E. Molecular-dynamics simulation methods for macromolecular crystallography. *Acta Crystallographica Section D Structural Biology* **2023**, *79*, 50–65, DOI: 10.1107/s2059798322011871.

(30) Ploscariu, N.; Burnley, T.; Gros, P.; Pearce, N. M. Improving sampling of crystallographic disorder in ensemble refinement. *Acta Crystallographica Section D Structural Biology* **2021**, *77*, 1357–1364, DOI: 10.1107/s2059798321010044.

(31) Tonikian, R.; Zhang, Y.; Sazinsky, S. L.; Currell, B.; Yeh, J.-H.; Reva, B.; Held, H. A.; Appleton, B. A.; Evangelista, M.; Wu, Y.; Xin, X.; Chan, A. C.; Seshagiri, S.; Lasky, L. A.; Sander, C.; Boone, C.; Bader, G. D.; Sidhu, S. S. A Specificity Map for the PDZ Domain Family. *PLoS Biology* **2008**, *6*, e239, DOI: 10.1371/journal.pbio.0060239.

(32) Zhu, Y.; Alvarez, F.; Wolff, N.; Mechaly, A.; Brûlé, S.; Neitthoffer, B.; Etienne-Manneville, S.; Haouz, A.; Boda, B.; Caillet-Saguy, C. Interactions of Severe Acute Respiratory Syndrome Coronavirus 2 Protein E With Cell Junctions and Polarity PSD-95/Dlg/ZO-1-Containing Proteins. *Frontiers in Microbiology* **2022**, *13*, DOI: 10.3389/fmicb.2022.829094.

(33) Ivarsson, Y. Plasticity of PDZ domains in ligand recognition and signaling. *FEBS Letters* **2012**, *586*, 2638–2647, DOI: 10.1016/j.febslet.2012.04.015.

(34) Gerek, Z. N.; Keskin, O.; Ozkan, S. B. Identification of specificity and promiscuity of PDZ domain interactions through their dynamic behavior. *Proteins: Structure, Function, and Bioinformatics* **2009**, *77*, 796–811, DOI: 10.1002/prot.22492.

(35) Ashkinadze, D.; Kadavath, H.; Pokharna, A.; Chi, C. N.; Friedmann, M.; Strotz, D.; Kumari, P.; Minges, M.; Cadalbert, R.; Königl, S.; Güntert, P.; Vögeli, B.; Riek, R. Atomic resolution protein allostery from the multi-state structure of a PDZ domain. *Nature Communications* **2022**, *13*, 6232, DOI: 10.1038/s41467-022-33687-x.

(36) Lockless, S. W.; Ranganathan, R. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science* **1999**, *286*, 295–299, DOI: 10.1126/science.286.5438.295.

(37) Stevens, A. O.; He, Y. Allosterism in the PDZ Family. *International Journal of Molecular Sciences* **2022**, *23*, 1454, DOI: 10.3390/ijms23031454.

(38) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* **2015**, *11*, 3696–3713, DOI: 10.1021/acs.jctc.5b00255.

(39) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods* **2017**, *14*, 71–73, DOI: 10.1038/nmeth.4067.

(40) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **1995**, *117*, 5179–5197, DOI: 10.1021/ja00124a002.

(41) Ceccarelli, M.; Marchi, M. Simulation of a Protein Crystal at Constant Pressure. *The Journal of Physical Chemistry B* **1997**, *101*, 2105–2108, DOI: 10.1021/jp9701810.

(42) García, A. E.; Blumenfeld, R.; Hummer, G.; Krumhansl, J. A. Multi-basin dynamics of a protein in a crystal environment. *Physica D: Nonlinear Phenomena* **1997**, *107*, 225–239, DOI: 10.1016/s0167-2789(97)00090-0.

(43) Klyshko, E.; Kim, J. S.-H.; Rauscher, S. LAWS: Local alignment for water sites—Tracking ordered water in simulations. *Biophysical Journal* **2022**, in press, DOI: 10.1016/j.bpj.2022.09.012.

(44) Cerutti, D. S.; Trong, I. L.; Stenkamp, R. E.; Lybrand, T. P. Simulations of a Protein Crystal: Explicit Treatment of Crystallization Conditions Links Theory and Experiment in the Streptavidin-Biotin Complex. *Biochemistry* **2008**, *47*, 12065–12077, DOI: 10.1021/bi800894u.

(45) Schoenborn, B. P. Solvent effect in protein crystals. *Journal of Molecular Biology* **1988**, *201*, 741–749, DOI: 10.1016/0022-2836(88)90470-6.

(46) Arakawa, T.; Bhat, R.; Timasheff, S. N. Preferential interactions determine protein solubility in three-component solutions: the magnesium chloride system. *Biochemistry* **1990**, *29*, 1914–1923, DOI: 10.1021/bi00459a036.

(47) Fenwick, R. B.; van den Bedem, H.; Fraser, J. S.; Wright, P. E. Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proceedings of the National Academy of Sciences* **2014**, *111*, E445–E454, DOI: 10.1073/pnas.1323440111.

(48) Huang, Q.; Gruner, S. M.; Kim, C. U.; Mao, Y.; Wu, X.; Szebenyi, D. M. E. Reduction of lattice disorder in protein crystals by high-pressure cryocooling. *Journal of Applied Crystallography* **2016**, *49*, 149–157, DOI: 10.1107/S1600576715023195.

(49) Cerutti, D. S.; Case, D. A. Molecular dynamics simulations of macromolecular crystals. *WIREs Computational Molecular Science* **2018**, *9*, e1402, DOI: 10.1002/wcms.1402.

(50) Wall, M. E.; Calabró, G.; Bayly, C. I.; Mobley, D. L.; Warren, G. L. Biomolecular Solvation Structure Revealed by Molecular Dynamics Simulations. *Journal of American Chemical Society* **2019**, *141*, 4711–4720, DOI: 10.1021/jacs.8b13613.

(51) Caldararu, O.; Ignjatović, M. M.; Oksanen, E.; Ryde, U. Water structure in solution and crystal molecular dynamics simulations compared to protein crystal structures. *RSC Advances* **2020**, *10*, 8435–8443, DOI: 10.1039/c9ra09601a.

(52) Williams, C. J.; Headd, J. J.; Moriarty, N. W.; Prisant, M. G.; Videau, L. L.; Deis, L. N.; Verma, V.; Keedy, D. A.; Hintze, B. J.; Chen, V. B.; Jain, S.; Lewis, S. M.; Arendall, W. B.; Snoeyink, J.; Adams, P. D.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science* **2018**, *27*, 293–315, DOI: 10.1002/pro.3330.

(53) Kopec, W.; Thomson, A. S.; de Groot, B. L.; Rothberg, B. S. Interactions between selectivity filter and pore helix control filter gating in the MthK channel. *Journal of General Physiology* **2023**, *155*, e202213166, DOI: 10.1085/jgp.202213166.

(54) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *Journal of the American Chemical Society* **2018**, *140*, 2386–2396, DOI: 10.1021/jacs.7b12191.

(55) Oide, M.; Sugita, Y. Protein folding intermediates on the dimensionality reduced landscape with UMAP and native contact likelihood. *Journal of Chemical Physics* **2022**, *157*, 075101, DOI: 10.1063/5.0099094.

(56) DePristo, M. A.; de Bakker, P. I.; Blundell, T. L. Heterogeneity and Inaccuracy in Protein Structures Solved by X-Ray Crystallography. *Structure* **2004**, *12*, 831–838, DOI: 10.1016/j.str.2004.02.031.

(57) Kuzmanic, A.; Pannu, N. S.; Zagrovic, B. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nature Communications* **2014**, *5*, 3220, DOI: 10.1038/ncomms4220.

(58) Fuentes, E. J.; Der, C. J.; Lee, A. L. Ligand-dependent Dynamics and Intramolecular Signaling in a PDZ Domain. *Journal of Molecular Biology* **2004**, *335*, 1105–1115, DOI: 10.1016/j.jmb.2003.11.010.

(59) Doyle, D. A.; Lee, A.; Lewis, J.; Kim, E.; Sheng, M.; MacKinnon, R. Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell* **1996**, *85*, 1067–1076, DOI: 10.1016/s0092-8674(00)81307-0.

(60) Kozlov, G.; Banville, D.; Gehring, K.; Ekiel, I. Solution Structure of the PDZ2 Domain from Cytosolic Human Phosphatase hPTP1E Complexed with a Peptide Reveals Contribution of the $\beta_2-\beta_3$ Loop to PDZ Domain-Ligand Interactions. *Journal of Molecular Biology* **2002**, *320*, 813–820, DOI: 10.1016/s0022-2836(02)00544-2.

(61) Ivanova, M. E.; Fletcher, G. C.; O'Reilly, N.; Purkiss, A. G.; Thompson, B. J.; McDonald, N. Q. Structures of the human Pals1 PDZ domain with and without ligand suggest gated access of Crb to the PDZ peptide-binding groove. *Acta Crystallographica Section D Biological Crystallography* **2015**, *71*, 555–564, DOI: 10.1107/s139900471402776x.

(62) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **2008**, *29*, 1859–1865, DOI: 10.1002/jcc.20945.

(63) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *Journal of Chemical Theory and Computation* **2011**, *7*, 2284–2295, DOI: 10.1021/ct200133y.

(64) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* **1983**, *79*, 926–935, DOI: 10.1063/1.445869.

(65) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **1995**, *117*, 5179–5197, DOI: 10.1021/ja00124a002.

(66) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19–25, DOI: 10.1016/j.softx.2015.06.001.

(67) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *Jour-*

nal of Computational Chemistry **1997**, *18*, 1463–1472, DOI: 10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.0.co;2-h.

(68) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *Journal of Chemical Physics* **1995**, *103*, 8577–8593, DOI: 10.1063/1.470117.

(69) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *Journal of Chemical Physics* **2007**, *126*, 014101, DOI: 10.1063/1.2408420.

(70) Taulier, N.; Chalikian, T. V. Compressibility of protein transitions. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* **2002**, *1595*, 48–70, DOI: 10.1016/s0167-4838(01)00334-x.

(71) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics* **1984**, *81*, 3684–3690, DOI: 10.1063/1.448118.

(72) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **1981**, *52*, 7182–7190, DOI: 10.1063/1.328693.

(73) Gowers, R.; Linke, M.; Barnoud, J.; Reddy, T.; Melo, M.; Seyler, S.; Domański, J.; Dotson, D.; Buchoux, S.; Kenney, I.; Beckstein, O. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. Proceedings of the 15th Python in Science Conference. 2016; pp 98–105, DOI: 10.25080/Majora-629e541a-00e.

(74) Humphrey, W.; Dalke, A.; Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **1996**, *14*, 33–38, DOI: 10.1016/0263-7855(96)00018-5.

(75) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Meng, E. C.; Couch, G. S.; Croll, T. I.;

Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science* **2021**, *30*, 70–82, DOI: 10.1002/pro.3943.

(76) Kitao, A.; Hirata, F.; Gō, N. The effects of solvent on the conformation and the collective motions of protein: Normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chemical Physics* **1991**, *158*, 447–472, DOI: 10.1016/0301-0104(91)87082-7.

(77) García, A. E. Large-amplitude nonlinear motions in proteins. *Physical Review Letters* **1992**, *68*, 2696–2699, DOI: 10.1103/physrevlett.68.2696.

(78) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**, DOI: 10.48550/ARXIV.1802.03426.

(79) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(80) Harrigan, M. P.; Sultan, M. M.; Hernández, C. X.; Husic, B. E.; Eastman, P.; Schwantes, C. R.; Beauchamp, K. A.; McGibbon, R. T.; Pande, V. S. MSMBuilder: Statistical Models for Biomolecular Dynamics. *Biophysical Journal* **2017**, *112*, 10–15, DOI: 10.1016/j.bpj.2016.10.042.

(81) Sculley, D. Web-scale k-means clustering. WWW '10: Proceedings of the 19th international conference on World wide web. 2010; pp 1177–1178, DOI: 10.1145/1772690.1772862.

(82) Weber, M.; Kube, S. Robust Perron Cluster Analysis for Various Applications in Computational Life Science. In *Lecture Notes in Computer Science*; Springer Berlin Heidelberg, 2005; pp 57–66, DOI: 10.1007/11560500_6.

(83) Kells, A.; Annibale, A.; Rosta, E. Limiting relaxation times from Markov state models. *Journal of Chemical Physics* **2018**, *149*, 072324, DOI: 10.1063/1.5027203.

(84) Elkins, J. M.; Papagrigoriou, E.; Berridge, G.; Yang, X.; Phillips, C.; Gileadi, C.; Savitsky, P.; Doyle, D. A. Structure of PICK1 and other PDZ domains obtained with the help of self-binding C-terminal extensions. *Protein Science* **2007**, *16*, 683–694, DOI: 10.1110/ps.062657507.

(85) The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.