Domino Effect in Allosteric Signaling of Peptide Binding

Pablo Andrés Vargas-Rosales and Amedeo Caflisch *

Department of Biochemistry, *University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland*

Correspondence to Amedeo Caflisch: *caflisch@bioc.uzh.ch (A. Caflisch), @caflischgroup (A. Caflisch), @pavaro2906 (P. A. Vargas-Rosales)*
https://doi.org/10.1016/j.jmb.2022.167661
Edited by Igor Berezovsky

Abstract

While being a thoroughly studied model of dynamic allostery in a small protein, the pathway of signal transduction in the PDZ3 domain has not been fully determined. Here, we investigate peptide binding to the PDZ3 domain by conventional and fully data-driven analyses of molecular dynamics simulations. First, we identify isoleucine 37 as a key residue by widely used computational procedures such as cross-correlation and community network analysis. Simulations of the Ile37Ala mutant show disruption of the coordinated movements of spatially close regular elements of secondary structure. Then, we employ a recently developed unsupervised, data-driven procedure to determine an optimized reaction coordinate (slowest-relaxation eigenvector) of peptide binding. We use this reaction coordinate to improve sampling by restarting additional simulations from the transition state region. Significant differences in the distributions of some of the pairwise residue distances in the bound and unbound states emerge from the projection onto the optimized reaction coordinate. The unsupervised analysis shows that allosteric signaling is transduced from the β2 strand, which forms part of the peptide binding site, to the spatially adjacent β3 and β4 strands, and from there to the α3 helix. The domino-like transmission of a (peptide binding) signal along β strands and α helices that are close in three-dimensional space is likely to be a general mechanism of allostery in single-domain proteins.
© **2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).**

Introduction

Allostery is a term coined in the 1960s to describe the phenomenon by which the binding of a ligand to a protein is affected by the interaction of another ligand at a different binding site.[1–2] The initial models of allostery were proposed to describe cooperativity on multimeric complexes, hallmarked by a conformational change along the subunits.[3] The realization that proteins are dynamic instead of rigid structures has called for new methods to describe allosteric effects which have been observed also in single-domain proteins. The new methods do not necessarily rely on a specific conformational change of the backbone or a two-state model, but rather consider the whole ensemble of states of the protein and the motion of its side chains.[4] Computational methods can be employed to explore the atomistic mechanisms of allosteric signal transmission, the influence of thermodynamic and kinetic components of the allosteric effects, and rationally design systems that show allosteric behavior.[5] Advanced simulation algorithms and protocols are needed to sample events that occur on timescales not accessible by conventional molecular dynamics, e.g., the transition path between the unliganded (T) and tetraoxygenated (R) structures of hemoglobin.[6] Concerning the analysis of the simulations, dynamic cross-

0022-2836/© 2022 The Authors. Published by Elsevier Ltd.This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

correlation and correlation networks are frequently employed to discern the pathways through which the signal is transmitted.[7–9]

PDZ domains are α/β domains of about 100 residues that mediate protein-peptide interactions in several hundred proteins, usually recognizing C-terminal segments of their target proteins.[10] The peptide ligand binds in an extended conformation into a groove between a two-strand antiparallel β sheet (β2-β3) and an α helix (α2), and non-covalently extends the β2-β3 meander into a three-strand β sheet (Figure 1(c)). The PDZ3 domain from *Rattus norvegicus* PSD-95 has an additional α-helix at its C-terminal end (α3), the removal of which has been shown to reduce peptide affinity by about 20-fold, without an impact on the global structure.[11] The C-terminal α3 helix has been defined as residues Pro93-Arg98,[11] while we consider Pro93-Ala101 as the α3 helix based on the results of the simulations (see Results and Discussion section) and to compare with a time-resolved spectroscopy study of α3 helix (un)folding.[12] Helix α3 does not directly interact with the peptide ligand, and its residue Tyr96 maintains a distance of at least 6 Å to the glutamine residue of the C-terminal segment of the CRIPT protein. Nevertheless, α3 has been reported to form stabilizing interactions with the β2-β3 loop (residues Gly29-Gly34) which favor binding, explaining the behavior upon truncation.[13] It was also shown that phosphorylation

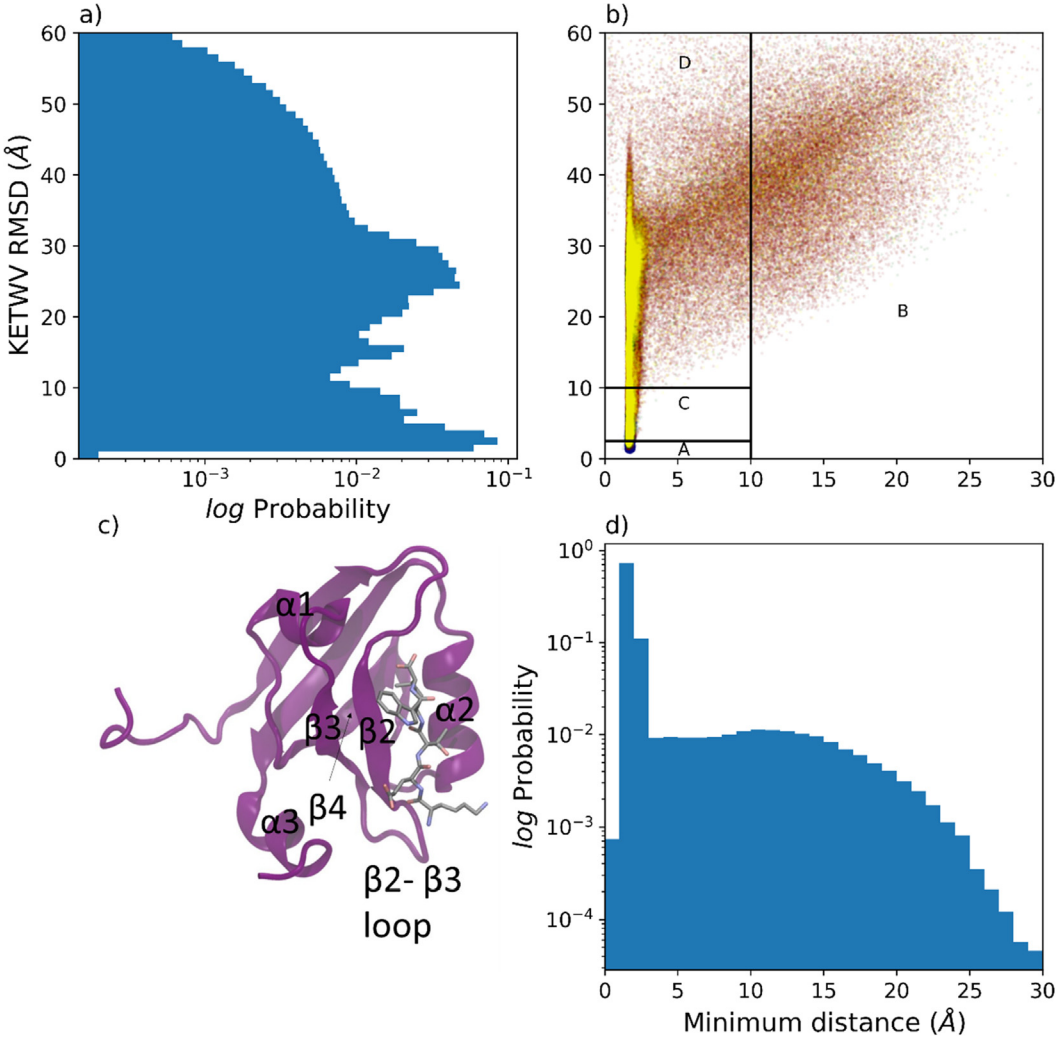

Figure 1. Geometric analysis of PDZ3/peptide simulation. (a) Distribution of RMSD values of the Cα atoms of the KETWV peptide with respect to the crystal structure, centered using Cα atoms of PDZ3. (b) Scatterplot of peptide RMSD vs minimum distance to PDZ3 colored by the trajectory of origin (blue, MD runs from the bound state; red, from unbound runs; green, from SAPPHIRE/PI-based reseeding, yellow, from eigenvector reseeding). Four states are assigned: A: crystal-like bound state, B: fully unbound from PDZ3, C: Encounter complex and incomplete binding, D: noncanonical binding to other regions of PDZ3. (c) Crystal structure of PDZ3 with relevant secondary structure elements labeled (PDB ID: 1TP5, residues Leu302 to Asn403 which are re-numbered in the text as Leu1-Asn102). (d) Distribution of minimum distances between the PDZ3 domain and KETWV peptide.

of Tyr96 in the first turn of the α3 helix perturbs its secondary structure, and thus modulates peptide binding by altering the β2-β3 loop.[13–15]

The group of Peter Hamm has shed light (both literally and metaphorically) on allosteric signaling in PDZ domains by a series of elegant time-resolved spectroscopy studies.[12] They have covalently linked a photoswitchable azobenzene to the α3 helix of PSD95-PDZ3, in the positions of Glu94 and Ala101, both mutated to Cys. These residues are separated by two turns of the α helix, or approximately 11 Å, which is the expected separation between the Cys94 and Cys101 anchors of the azobenzene in the *cis* state.[16] The *cis* to *trans* transition of the photoswitchable linker was used to study the influence of the α3 helix unfolding on the binding affinity of a pentapeptide ligand, which does not interact directly with either the α3 helix or the linker.[17] The stabilizing effect of the α3 helix (azobenzene linker in the *cis* state) results in a temperature-dependent increase in affinity of up to 120-fold. Conversely, peptide binding influences the rate of *cis*-to-*trans* isomerization of the azobenzene. The strength of the allosteric force exerted by the azobenzene conformation switching on peptide binding has been determined to be of 1 nN, as measured by the *cis*-to-*trans* enthalpy difference and the 3 Å change in the distance of the two anchoring residues of the photoswitchable linker.[17] Time-resolved spectroscopy has revealed a 4 ns timescale for the helix unfolding, and a 200 ns timescale for the allosteric signal to reach the binding pocket.[12]

Previous simulation studies in our group have revealed the mechanism of conformational selection in the PDZ3 domain, as peptide binding favors a reduced aperture of the groove between the α2 helix and β2 strand.[18] Similar results have been obtained in a more recent simulation study which used experimental data as input.[19] In particular, the closing of the binding site emerged as the principal component with most influence in a dimensionality reduction analysis. Furthermore, the stabilization of the β2-β3 loop by the α3 helix was observed.[19] Multiple μs simulations of spontaneous binding to the PDZ2 domain of PTP1E have revealed electrostatic steering by the formation of non-native salt bridges between carboxy groups of the peptide and basic side chains in the β2-β3 loop and β3 strand.[20] These findings are consistent with simulation studies of the PSD95-PDZ3 domain,[13] which show the importance of the ionic interactions of the peptide with the loop for binding stability. Furthermore, it has been proposed that the allosteric modulation depends on the shifting of hydrogen bonding networks, also implicating an interaction between the β2-β3 loop and the C-terminal helix.[21] Some simulation studies compare the bound and ligand-free PDZ3 domain, and find the effects of peptide binding by comparing different descriptors for each residue. These descriptors include the position of the residue to the center of mass, the root mean square deviation (RMSD), non-bonding interactions, etc.[22], correlated motions on THz scale,[23] sectors of residues with related covariance.[24] Other studies rely on the effects of single-residue mutations on residue communication[25–26] and energy transduction[27–28] to investigate the flow of information in PDZ3. A recent analysis based on dynamic communities show the importance of the α3 helix in mediating the communication between different regions of the PDZ3 domain.[26] The study focuses on the impact of single point mutations, though, not on the process of peptide (un)binding.

Molecular dynamics simulations of biomolecules yield a particularly high-dimensional data, consisting of three-dimensional coordinates of thousands of atoms, making it almost impossible to analyze visually. A common strategy to better understand the simulations is therefore to project the multidimensional trajectories onto a one-dimensional reaction coordinate (RC), which can be used to describe a complex process (e.g., protein folding) in an intuitive way. Then, the dynamics of the process can be described as diffusion on the free energy profile (FEP) along the RC, with basins describing states and the height of the barriers determining the rates of interconversion. This RC can either be a single geometric variable, such as an interatomic distance, or a combination of distances. It can also be generated by a dimensionality reduction strategy. Traditional dimensionality reduction procedures are prone to fail due to how they treat the redundancy of the degrees of freedom and the loss of dynamic information, so a more tailored approach is required to analyze the folding process or allosteric effects. If the RC is poorly chosen, for example by using only simple geometric measures, such as RMSD or interatomic distances, the information lost during dimensionality reduction masks the true FEP barriers, yielding sub-diffusive dynamics along the FEP.[29] For an optimal RC the kinetics of the system are preserved, and therefore, dynamics are diffusive on the projected FEP. Furthermore, the optimal RC yields the highest possible cut profile, and a partition function which is independent of the timestep used for its calculation.[30–32] Therefore, the choice of an optimal RC is essential to correctly describe such processes like protein folding or signal transduction.

One example of an optimal RC is the committor, which given two boundary states A and B, describes the probability of reaching state B before reaching state A from any point in the phase space.[33–34] Sergei Krivov has developed a non-parametric method for calculating the committor, and a metric based on the cut function $Z_{C,1}$ to evaluate the optimality of the obtained RC.[35] A drawback of the committor is the necessity of defining boundary states. More recently, the same author has proposed a fully unsupervised method

for optimization of the RC that approximates the slowest-relaxing eigenvectors of the transfer operator.[36] No boundary states need to be defined. The RC is iteratively optimized with features such as interatomic distances or angles, transformed by a polynomial function.[32] Additionally, a criterion based on cut profiles can be employed for further validation.[36]

Here we combine the eigenvector optimization procedure[36] and the SAPPHIRE (States And Pathways Projected with HIgh REsolution) plot-based analysis[37] to investigate allosteric signaling upon peptide binding to PDZ3. We reasoned that a fully data-driven optimization of the RC(s) is most adequate for investigating *a priori* unknown allosteric transitions. Multiple unbiased molecular dynamics (MD) simulations were started from the crystal structure of the PDZ3 domain in complex with the Lys-Glu-Thr-Trp-Val ligand (*holo* state of PDZ3), and from the fully dissociated pentapeptide (*apo* PDZ3). The MD trajectories were first analyzed with traditional procedures based on intuitive geometric variables. Then the slowest-relaxation eigenvector was used as progress index for the SAPPHIRE plot to identify hidden states in the free energy surface. The combined analyses reveal a domino-like effect of allosterism in which the residues and secondary structure elements that mediate the transduction of the allosteric signal are contiguous in the native structure of the PDZ3 domain.

Results and discussion

Multiple independent MD simulations were performed to analyze peptide (un)binding (Table 1). Initially, 20 binding runs were started with the KETWV peptide randomly positioned far away from the protein domain. In addition, the bound state was investigated by six runs started from the crystal structure (PDB ID: 1TP5) of the complex with the peptide inside the binding groove. The obtained sampling was analyzed using the progress index[38] and SAPPHIRE[37] methods. From the transition region (highest barrier) of the SAPPHIRE plot, 20 frames were selected and simulations were restarted from them. After an initial blind (i.e., unsupervised) analysis with the optimal reaction coordinate framework, 16 additional runs were launched. The cumulative sampling amounts to 34 μs.

We first present the analysis of the MD simulations by conventional protocols which make use of geometric variables, e.g., RMSD of the peptide backbone. For this analysis, the bound and unbound states were defined based on geometric criteria alone, namely RMSD and minimum distance between peptide and protein. We also monitor the distance between Glu94 and Ala101, both part of the α3 C-terminal helix and located two turns apart. These are the same residues mutated in the experimental study to cysteines and used as anchor for the azobenzene photoswitch.[17] We then present the unsupervised analysis based on the optimization of the slowest-relaxation eigenvectors and their use for projecting the free energy of the binding process. We calculate several statistical measures between inter-residue distances along the FEP to discern residues relevant to the transduction of the allosteric signal. In the following text, the three-letter notation is used for the residues of the PDZ3 domain while the one-letter abbreviation is employed for the residues of the peptide ligand.

Conventional analysis of the MD simulations

Seven complete binding events of the KETWV peptide were sampled in the runs started from the apo structure of PDZ3 (two of them are shown in the Supplementary Movies S1 and S2) and the reseeded trajectories (see subsection Reseeded trajectories in the Materials and Methods). In agreement with the ms time scale of KETWV peptide unbinding from PDZ3,[12] there was no full unbinding during the 6 μs of cumulative sampling starting from the peptide/PDZ3 complex crystal structure. Furthermore, only two dissociation events took place from the trajectories that were restarted from the transition region. In contrast, partial unbinding with the C-terminal valine of the KETWV peptide still buried in the binding site was observed frequently. The projection of the free energy along an intermolecular distance that reflects the burial of the side chain of the C-terminal V shows that the highest barrier on both

Table 1 Summary of performed simulations.

Simulation Type	Number of runs	Length per run (ns)	Starting structures	Number of binding events
Unbinding	6	1200	PDB: 1TP5	-
Binding	20	1000	1TP5 with peptide randomly placed	2
SAPPHIRE-guided reseeding	20	100	Barrier region on geometric variable-based SAPPHIRE plot	1
Optimal-RC reseeding	16	300	Barrier region on optimal RC-based SAPPHIRE plot	4
Ile37Ala	24	200	1TP5 with Ile37Ala mutation	1

the cut- and histogram-based Free Energy Profiles corresponds to a distance of about 8 Å (Figure S 12). The aforementioned partial unbinding and the free energy profile suggest that the rate limiting step for peptide (un)binding is the insertion of the V side chain. These results are in agreement with a previous simulation study of a PDZ2 domain,[20] where the rate-limiting step of binding was found to be the burial of the C-terminal valine of the EQVSAV peptide.

The two-dimensional projection of the phase space onto the peptide RMSD and the minimum distance between peptide and protein defines four regions of peptide-protein interaction (Figure 1(b)). The fully bound structure, with a peptide RMSD smaller than 2.5 Å, is populated during 22% of the simulation. The fully dissociated state, with a minimum intermolecular distance of more than 10 Å, is present in 9% of the frames. The encounter complex, defined as the state with RMSD between 2.5 Å and 10 Å, is observed in 14% of the frames. In the remaining 55% of the sampling, the peptide is non-natively bound. In the subsection Unsupervised analysis, the free energy states will be defined by a data-driven procedure and analyzed in detail.

We first investigate whether a direct comparison of *holo* and *apo* PDZ3 sections of the sampling would give any insight into allosteric effects, by analyzing 10 segments of 20 ns in which the peptide was either fully bound or unbound. The root mean square fluctuation (RMSF) analysis shows that peptide binding stabilizes the protein structure compared to the *apo* state (Figure 2(a), (b)). The lower fluctuations of the bound state of PDZ3 (on a 5-ns time scale) are observed not only for the secondary structure elements of the binding groove, i.e., the α2 helix and β2 strand, but also for segments that do not interact directly with the peptide ligand, e.g., the α1 helix and the C-terminal segment. While this observation is congruent with an allosteric effect upon ligand binding it does not explain the mechanism of signal transduction. Note that a higher stability of *holo* PDZ3 vs. *apo* was reported also in previous simulation studies.[18,21,39]

In the bound state, the KETWV peptide is involved in specific interactions with α2 and β2, as well as stabilizing contacts of the W side chain and the PDZ3 residues Phe39 and Leu41 (Figure 2(c)). The contact map can also be used to identify differences in intra-protein contacts for the sampling with bound or unbound peptide (Figure 2(c)). It emerges that the contacts between the β2-β3 loop and α2 helix are present only in bound segments of the MD trajectories (solid circle in Figure 2(c)). Another set of exclusive contacts in *holo* stretches of the MD sampling is the interaction of the α1 helix and β1 strand (dashed circle in Figure 2(c)). In the unbound state there are a few additional contacts in the antiparallel β sheet formed by strands β2 and β3 (red dots in Figure 2(c)). The similar contact maps of PDZ3 in the presence and absence of the peptide ligand provide further evidence that simple geometric variables (e.g., the inter-residue distances) are not adequate to capture allosteric effects due to peptide binding. Moreover, the contact map does not take into account dynamic information, an important component of MD simulations.

The analysis of the dynamic cross correlation (DCC) of PDZ3 and the peptide ligand KETWV gives a dynamic picture of the interactions between the different residues by considering how residues displace with respect to others. This is in contrast to contact maps which only provide static information. The DCC matrix shows a strong correlation of the bound peptide with the secondary structure elements of the PDZ3 binding sites, namely α2 and β2, but also other regions such as β3 and α1 (Figure 3(a)). A negative correlation is observed between the N-terminal segment and the peptide. Several regions of the DCC differ in the bound and unbound trajectory segments. A positive correlation between the α1 helix and both the β1 strand and the β1-β2 loop is observed in the bound state and not in the unbound state. Another difference is present in the α2 interaction with β2. These two regions which bind to KETWV are positively correlated in the bound state, while in the unbound there is a low negative correlation (Figure 3(b)). In general, the α2 helix shows some negative cross-correlation to the rest of the PDZ3 (β2, β3, and α1) in the unbound segments, while there is less negative cross correlation and even some positive correlation in the bound state. This is to be expected from the coupling effect of the peptide, which bridges the relatively flexible α2 to the core of the domain. These DCC results are consistent with the hinge-like motion of α2, and its locking due to peptide binding.[18]

As for the C-terminal region, in the bound state there is a negative cross correlation to α2, which is only partially present in the unbound state. Furthermore, Ser97-Arg98 show a positive correlation to the β2-β3 loop and the region around α1, which is broken upon peptide binding. This behavior has been shown in previous studies describing an interaction between both regions altered by peptide binding, after which the loop interacts more with the peptide.[14] The information contained in the DCC matrices is detailed but hard to interpret. Thus, a simpler way to illustrate the correlated displacement is to build dynamic communities from residues with similar correlated motions.

From the cross-correlation matrices of the bound and unbound states, allosteric communities were calculated using the bio3d R library (see Methods).[41] Residues with closely-correlated motions are grouped together into communities,

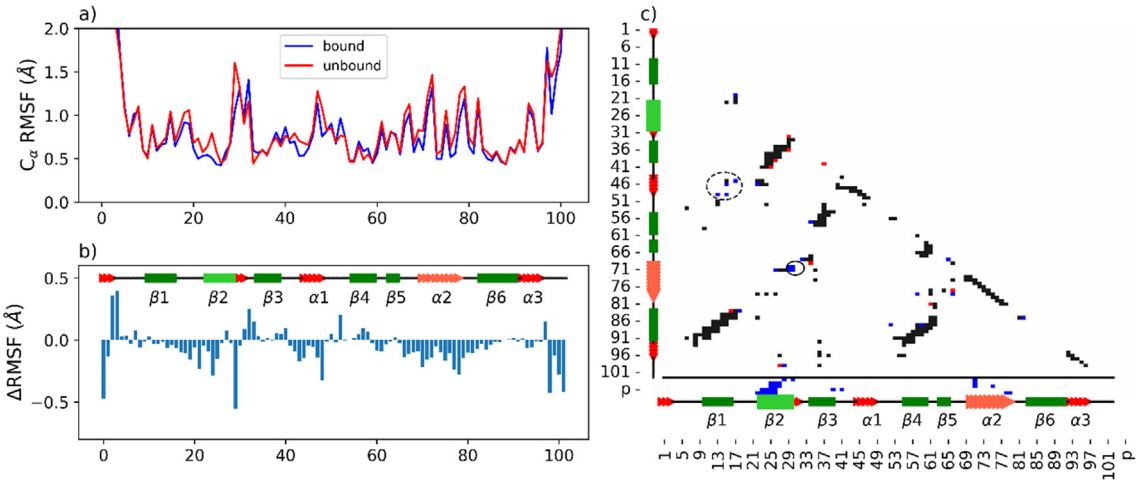

Figure 2. Differences in PDZ3 backbone flexibility in the presence and absence of the KETWV peptide. (a) Sequence profile of Cα RMSF averaged every 5 ns for 10 segments of *holo* (blue) and 10 of *apo* trajectories (red), of 20 ns each. (b) Secondary structure, with the naming of elements from,[40] and difference in RMSF (bound-unbound). (c) Contact map calculated with a distance threshold of 5 Å and 80% occurrence. Contacts in black are conserved on bound and unbound trajectories, while contacts in blue and red are exclusive for bound and unbound trajectories, respectively. The horizontal line at the bottom of the contact map separates the KETWV peptide (abbreviated as p) from the PDZ3 domain. The interaction between the β2-β3 loop and α2 (solid circle), and between β1 and α1 (dashed circle) are present only in the bound state.

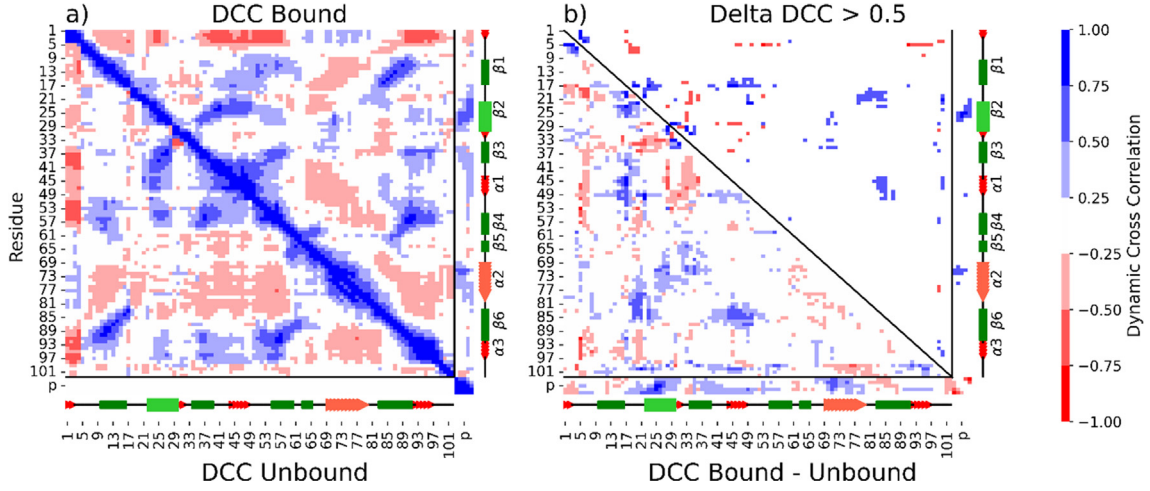

Figure 3. Analysis of correlated displacement of residue pairs. (a) Dynamic Cross-Correlation for bound (upper half) and unbound (lower half) segments of the trajectories. (b) Difference in cross-correlation (DCC_{bound} - $DCC_{unbound}$) (lower) and residue pairs with high difference in cross-correlation (upper). In both panels, the axes are labeled with the sequence numbering of PDZ3. The pairs that involve the peptide (abbreviated as p) are in the bottom and on the right of each panel (separated by a horizontal and vertical line, respectively).

and these nodes are connected to each other according to the closeness of correlated motions between them. There are substantial differences between the dynamic communities of *holo* and *apo* states (Figure 4). First of all, the communities (spheres in Figure 4) are larger in the bound state. In both *apo* and *holo* states, the α2 helix (which is part of the binding groove) forms its own community, underlining its conformational independence. In the *holo* state, the peptide is in the same community with the β2 strand. When bound, the peptide is extended and augments the β2-β3 meander into a three-strand β sheet. Importantly, the community of the α3 helix in the bound state (large gray sphere in the bottom left of Figure 4(a)) includes also Ile37 on the β3 strand and is closely tied to the β2-β3 loop (yellow sphere in Figure 4 (a)). Thus, the β3 strand is likely to mediate the

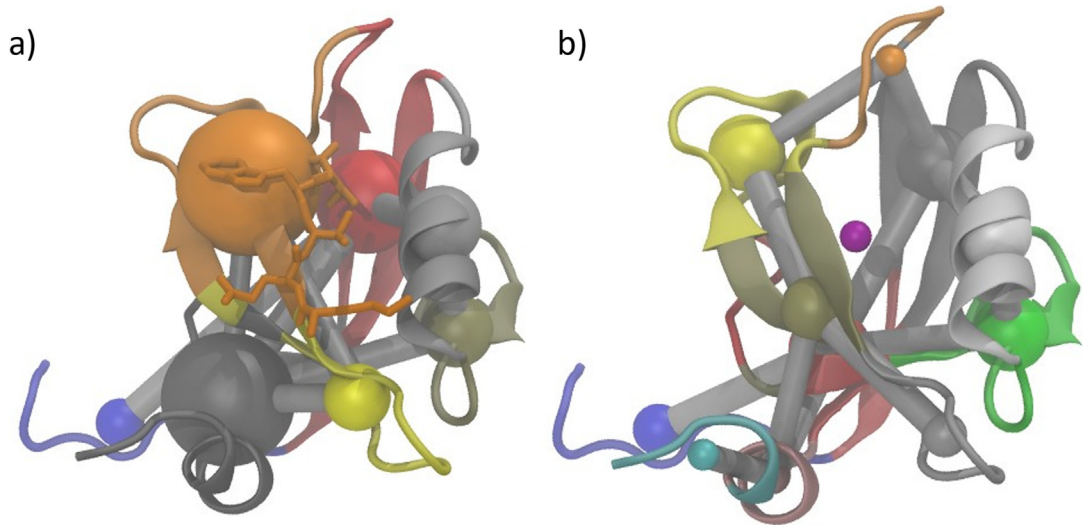

Figure 4. Dynamic networks for (a) bound protein and (b) unbound. Nodes (colored spheres) represent dynamic communities, formed by tightly cross-correlated residues, while the edges (cylinders) connecting them have a diameter proportional to the dynamic cross-correlation between communities. Both in the bound and the unbound system, the α3 helix (dark gray and pink/cyan respectively) is connected to the binding site through the β2-β3 meander (orange bound, and olive unbound). The β2-β3 loop (yellow in bound) is connected to the α3 helix only in the bound state. As expected, the KETWV peptide is part of the β2 community (orange). The α2 helix (dark gray in bound, light grey in unbound) is disconnected from the β2 strand in both states.

interactions between the β2 strand and the C-terminal α3 helix (see also below). In the unbound state, there is no interaction between the α3 helix and the β2-β3 loop. Finally, it is worth noting that in the unbound state the α3 helix is divided in two small dynamic communities. This analysis suggests that peptide binding helps maintaining the structure of α3 together, which would otherwise move more freely when the peptide is unbound.

A simple geometric variable like the RMSD can be used for defining the bound state for the calculation of binding times directly from the trajectories. Using as threshold a peptide RMSD < 5 Å (Figure S 1) we can estimate the association rate constant k_{on} even if only seven binding events were observed (at 223, 330, 348, 387, 603, 1019, and 1047 ns, respectively) on the 1.05-μs time scale of the binding runs. For this purpose, the cumulative distribution of the binding times is fitted by a single exponential function $f(t) = \exp(-t/\tau)$ which yields a characteristic time τ of 3.6 μs (Figure S 2). Taking into account the concentration of the peptide in the simulation box (5 mM), the binding rate constant k_{on} is 55.5 $ms^{-1}mM^{-1}$. Considering a k_{off} of (1/200) ms^{-1},[12] the equilibrium dissociation constant is $K_d = 0.09$ μM (at the simulation temperature of 26.5 °C) which is a factor of about 15 more favorable than the value of 1.4 μM measured at 30 °C for the photoswitchable PDZ3.[17] This discrepancy is due, at least in part, to the slightly faster self-diffusion coefficient of bulk water in the TIP3P water model which is nearly three times larger than the experimental value.[42] The values of k_{on} extracted from the MD trajectories of binding are robust to a range of RMSD thresholds between 3 and 7.5 Å, and also comparable to those obtained using the optimized RC as threshold for binding (Figure S 3).

Simulation analysis of the Ile37Ala point mutant

Both the Correlation Network Analysis (Figures 3 and 4) and the unsupervised optimal reaction coordinate framework (see below) suggest that Ile37, which is in the β3 strand, plays a significant role in the allosteric signal transmission upon peptide binding. We decided to provide further evidence of the role of Ile37 by additional simulations of the Ile37Ala point mutant, starting from the unbound peptide. We focused on the apo structure of the mutant and did not start simulations from the bound state because of the long timescale of peptide unbinding reported experimentally, and the lack of unbinding events in our simulations of the wild type PDZ3. The secondary structure analysis showed a preference for a one-turn α-helix in the C-terminal region, present between Pro93 and Tyr96 (Figure S 13). Consistently, the distribution of the Glu94-Ala101 distance, which monitors two turns of the α3 helix, is shifted in the mutant towards a longer distance in comparison to the wild type (Figure 5(a)). This observation is consistent with the role of Ile37 as allosteric mediator, since the mutation in the β3 strand reduces the stability of the α3 helix. Additional evidence on the importance of Ile37 on peptide binding emerges from the Community Network Analysis which shows a significantly

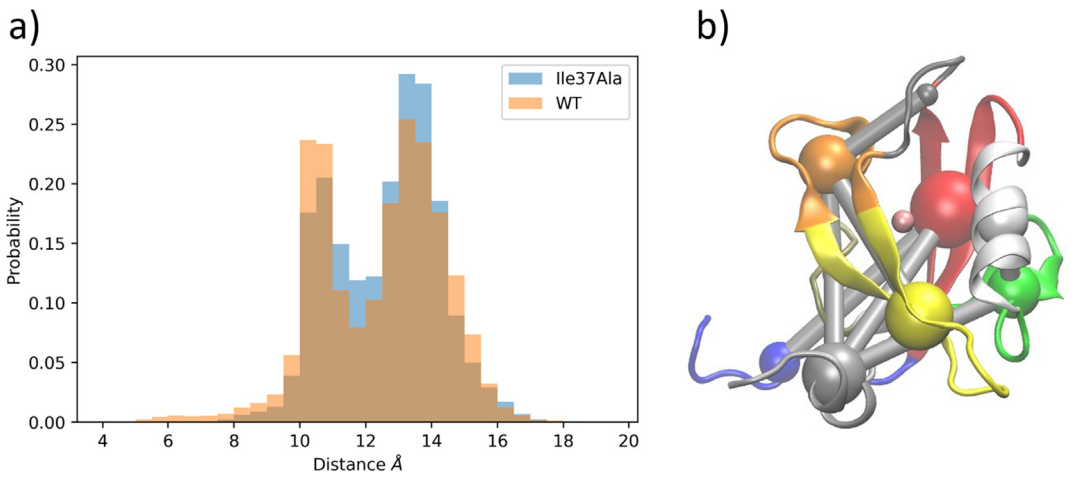

Figure 5. Results of Ile37Ala mutation. a) Distribution of Glu94-Ala101 distance for the mutant PDZ3 (blue) and wild type (orange). The first peak at 10.5 Å reflects two full turns of an α helix. The distribution in the mutant is shifted towards a longer distance which reflects the reduced stability of the second turn of the helix. b) Correlation Network of Ile37Ala-PDZ3. In contrast to the wild type, the α3 helix (gray sphere) in the Ile37Ala mutant is not linked to the β2-β3 meander (yellow sphere), and thus the domino-like allosteric signaling is compromised.

different cross-correlation profile of the Ile37Ala mutant (Figure 5(b)) with respect to the wild-type (Figure 4(b)). In the mutant, the motion of the α3 helix is correlated with the β1 and β6 strands while it is not linked to the β2 and α2 secondary structure elements which form the peptide binding site. In contrast, for the wild type the communities of α3 and β2 are connected in both the *apo* and *holo* trajectory segments (Figure 4). This provides evidence of the reduced collective motions between the β2, β3, and α3 elements upon mutation of Ile37 into Ala.

Unsupervised analysis

As a point of departure with respect to previous studies, we employ here an unsupervised, fully data-driven framework based on the slowest-relaxation eigenvector(s) as optimal RC(s).[36,43] The slowest-relaxation eigenvector is calculated for each snapshot of the MD trajectory (also called *frame* hereafter) by iteratively and recursively updating a RC functional $r(k\Delta t_0)$ (where k is the snapshot index and Δt_0 is the trajectory saving interval) using at each iteration a single variable chosen from a set of interatomic distances or dihedral angles.[32] The functional form of the RC approximates the slowest-relaxation eigenvector if it minimizes the total square displacement $\Delta r^2(\Delta t) = \Sigma[r(k\Delta t_0 + \Delta t) - r(k\Delta t_0)]^2$ (where Δt is the lag time which is equal to Δt_0 or its multiple), under the constraint that the sum over all snapshots of the squared values of the RC is equal to one, i.e., $\Sigma_k^N r^2(k\Delta t_0) = 1$, and that the calculated RC is othogonal to all previously optimized eigenvectors. This optimization strategy is equivalent to a maximization of the autocorrelation function.[36] The iterative optimization of the RC $r(k\Delta t_0)$ is based on the nonlinear combination of a large set of variables, e.g., pairwise residue distances. Although the result of each optimization step depends on the residue-pair distance chosen, the overall RC is independent of the distances used and their order. Furthermore, since no definition of edge states or knowledge of the system is needed the method is fully unsupervised, i.e., "blind".[36] This is in contrast to the committor optimization method, where two edge states must be defined initially. In the present study we take into account the different rates of optimization previously described,[35] making it adaptive. This is achieved by an evaluation of the optimality along the calculated reaction coordinate. MD frames with a suboptimal value of the RC are allowed to vary, while frames better optimized are kept fixed and don't contribute to the optimization, which prevents overfitting.

The optimal RC methodology was originally validated on simulations of protein folding[36,43] while it is adopted here to analyze peptide binding and allosteric signaling. Random intra-protein distances (80%) and protein-peptide intermolecular distances (20%) are used to optimize a RC that approximates the first eigenvector of an implicit Markov model describing the system. The RC approximates the optimal RC more accurately along the transition region than at the free energy minima, as shown by the optimality criteria (Figure 6). The first metric, the optimality criterion θ, is used to assess the quality of a RC approximating the slowest-relaxing eigenvector.[36] In the case of optimality, it is constant along the values of the RC and close to zero. For the obtained RC, θ is constant and close to zero for most of the transition region, with nonzero values denoting suboptimality in the free energy basins (Figure 6(a)). The second metric is the committor

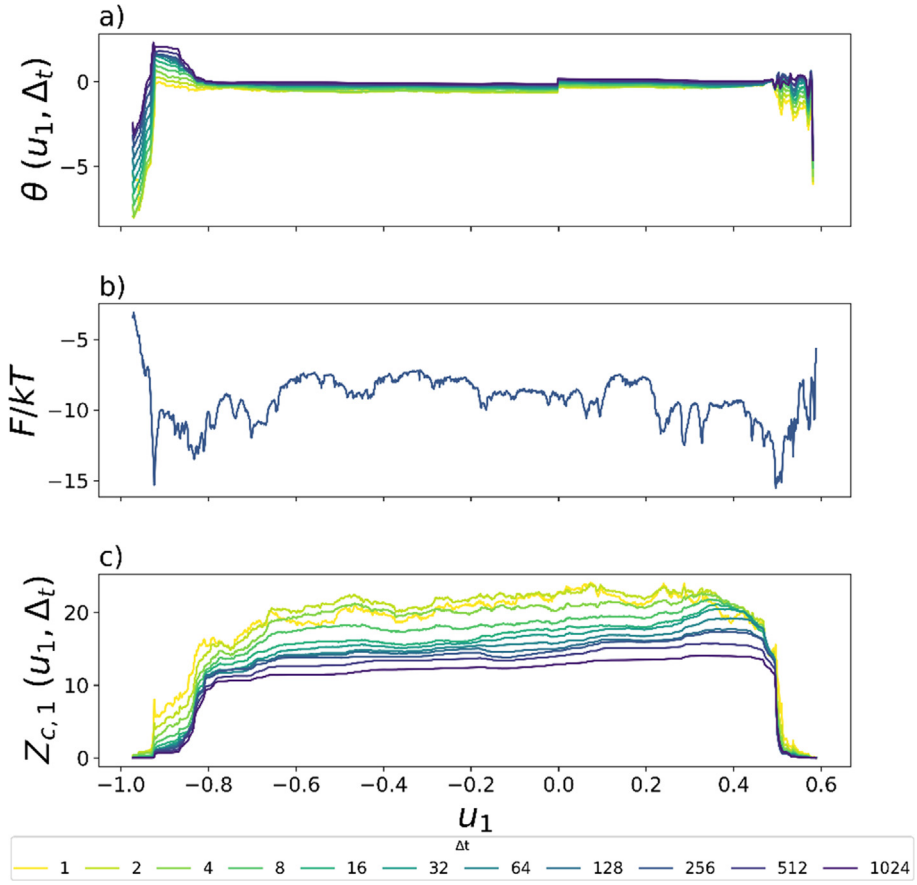

Figure 6. Reaction coordinate (RC) quality metrics applied to the optimized slowest-relaxation eigenvector u_1. (a) Eigenvector optimality criterion $\theta(u_1, \Delta t)$. Optimal RCs have a $\theta(x, \Delta t)$ which is constant and close to zero, and independent of Δt. (b) Free energy profile projected on the optimized RC u_1. (c) Committor optimality criterion $Z_{C,1}$. The committor shows a $Z_{C,1}(x, \Delta t)$ which is almost constant at the barrier, and converges for large values of Δt. The convergence to a value of about 13–14 is consistent with the seven binding events observed in the total sampling. The profiles of $\theta(u_1, \Delta t)$ and $Z_{C,1}(u_1, \Delta t)$ show that the transition region is very well optimized, while the two main basins are not. In both (a) and (c) different colors correspond to different lag times $\Delta t = [1, 2, \ldots, 1024]$ used for the calculation of both $\theta(u_1, \Delta t)$ and $Z_{C,1}(x, \Delta t)$.

optimality criterion $Z_{C,1}$, which shows a similar behavior to the θ criterion. In an equilibrated simulation, $Z_{C,1}$ should converge to twice the number of transitions between the edge states defined for committor calculation.[35] The profiles converge at $Z_{C,1} \approx 13$, which is nearly twice the number of observed binding events (seven). The projection of the MD sampling into the two RCs representing the slowest-relaxing eigenvectors shows that there is one main (un)binding pathway (Figure 7(c)). There are two pronounced minima in the projection onto the first (i.e., slowest-relaxation) eigenvector u_1 at $u_1 \approx -0.9$ and 0.5. These minima are the basins of the bound state ($u_1 \approx -0.9$) and a state that includes non-natively bound and fully unbound frames (Figure 7(a)). These two states are separated by a broad transition state region. The second eigenvector u_2 discriminates the transition region between the two main basins of u_1 (Figure 7(d) and Figure S 7). In addition, the minor basins around $u_2 = 0.9$ and $u_2 = 0.5$ include snapshots with non-native associations of the peptide to the N- and C-terminal regions of the protein, respectively (Figure S 11). The optimized reaction coordinates have also been transformed to their "natural" counterparts, in which the diffusion coefficient is constant and equal to one.[44] The overall shape of basins and large transition region are similar as for the non-transformed eigenvectors (Supplementary Information Figure S 4, Figure S 5, and Figure S 6). The projection of the free energy along the optimized RC shows the height of the individual barriers but does not illustrate directly the correspondence between basins (minima) and protein conformations or peptide binding modes.

The SAPPHIRE plot is a one-dimensional projection of the free energy which is useful to illustrate and characterize all free energy basins.[37–38,45] As an extension to the original method, we employ here the optimized eigenvector

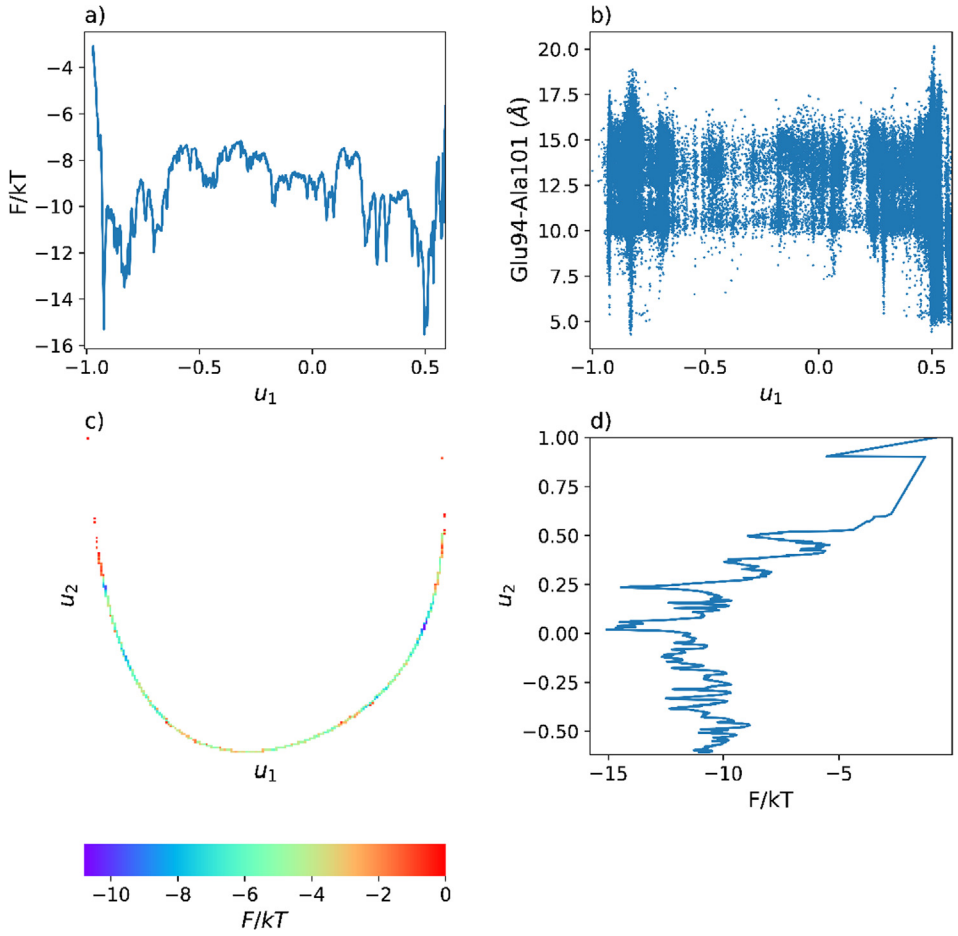

Figure 7. Free energy surfaces of the process of KETWV-peptide binding process to the PDZ3 domain. (a) Projection of the free energy on the slowest-relaxation eigenvector u_1. The Free Energy Profile shows the different basins from the bound state at around −0.9 to the unbound states at 0.5. (b) Distribution of distances between the Cα of residues Glu94 and Ala101 in the α3 helix projected on the u_1 eigenvector. (c) Two-dimensional, histogram-based, Free Energy Profile of trajectory frames binned according to their u_1 and u_2 values. The use of the two first eigenvectors, describing slowest relaxing processes, allows ruling out the presence of secondary pathways. (d) Projection of the free energy on the second eigenvector u_2.

RC u_1 as progress index to order the individual snapshots (Figure 8). Note that we use the term progress *index* instead of progress *variable* because the former is defined only for the snapshots of the MD sampling while the latter is a function that can be evaluated for any coordinate set. For the structural annotation, 17 inter-residue distances were selected, 12 of which correspond to peptide-protein distances with the highest mutual information calculated with respect to the different basins (relative entropy, see Unsupervised analysis subsection of Materials and Methods). The other 5 distances are intra-protein distances that show a mutual information above the 99.9 quantile (Table S 1). These distances are useful to describe the binding process by monitoring the evolution of contacts along the RC.

The $u_1 \approx -0.9$ basin corresponds to the native-bound state, and includes two subbasins defined mainly by the orientation of the T side chain of the peptide towards the α2 helix or towards β2 (His71-T and Val27-T distances, Figure 8) and the contacts of W to the β3 strand, which are stronger in the first subbasin. The distance between the secondary structure elements that make up the peptide binding groove, i.e., α2 helix and β2 strand (Phe24-Lys75), is in general closer in the bound basin, being more frequently less than 10 Å compared to the rest of the RC. The second basin ($u_1 \approx 0.5$), corresponds to the non-natively bound frames and comprises both fully unbound and peptide-protein interactions outside the binding site. The basin shows no strict clustering according to minimum distance between protein and peptide of the fully bound states. There is, though, a separation between the regions where the KETWV peptide is dissociated from the protein, closer to the transition states, and where the peptide is bound far away from the recognition pocket, at the end of the RC (see minimum

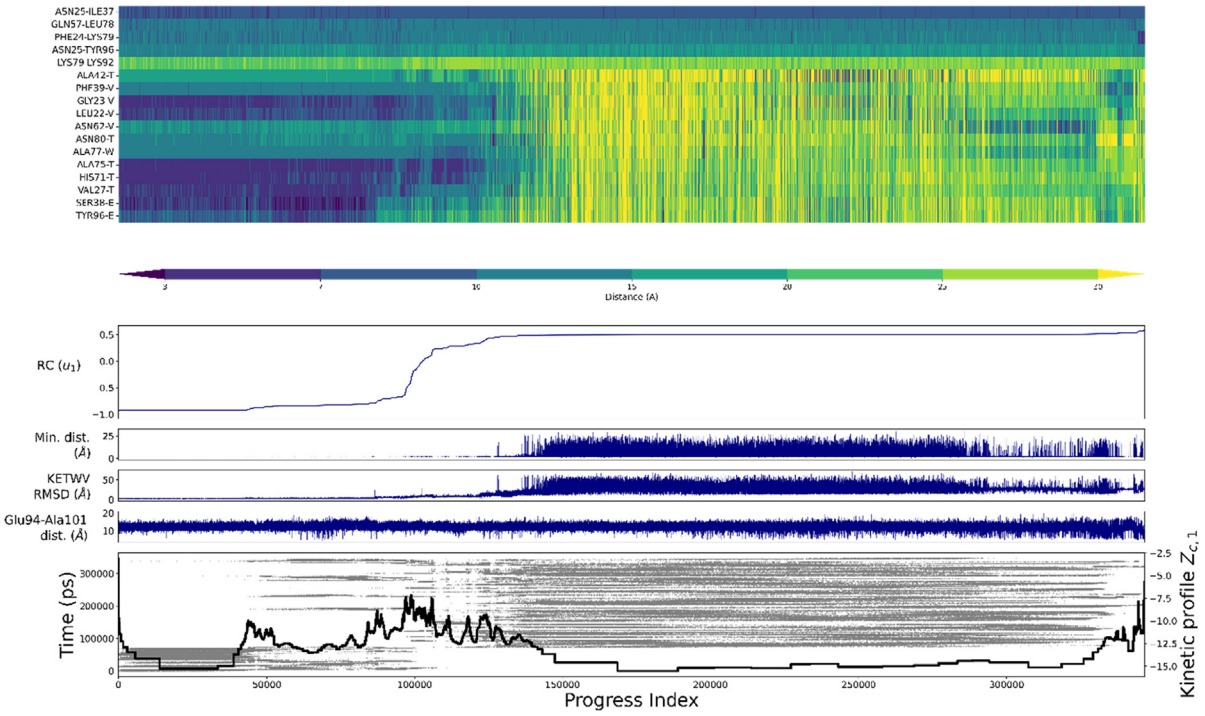

Figure 8. SAPPHIRE plot of total sampling using the optimized first eigenvector u_1 as progress index. From top to bottom: structural annotation with interatomic distances; value of the slowest eigenvector RC, u_1; minimum distance between peptide and protein; RMSD of peptide Cα with respect to crystal structure; distance between residues Glu94-Ala101 (which are the two residues connected by the azobenzene linker in[17]); kinetic and temporal annotation. The kinetic annotation corresponds to the cut-based free energy profile of the RC as shown in Figure 7(*a*). The temporal annotation shows the ordering of trajectories in real-time.

distance panel or Asn62-V distance in Figure 8, also visible in u_2 as seen in Figure S 7). The intercalation of fully unbound frames and non-natively bound ones reflects the necessary unbinding of the peptide after unstable contacts with non-canonical sites of the protein. Furthermore, apart from the basins at $u_1 \approx -0.9$ and the transition regions, no other protein-peptide interaction shows its own basin along the projection of the FEP. These results provide evidence that non-native binding events are unstable and are not populated for a significant amount of time, as there are many transitions between the different non-canonical bindings. This is also clear when looking at the interaction map of the peptide for the basin at $u_1 \approx -0.9$ (Figure S 9) where the peptide remains tightly bound to the canonical binding site for a majority of the frames. In contrast, for the basin at $u_1 \approx 0.5$ (Figure S 10) there is no stable interaction and the peptide comes into contact with most residues with similar probability. The movies Movie S1 and Movie S2 illustrate the structural stability of the native bound state and the transient character of the non-native peptide/protein interactions.

The geometric annotation at the transition state region can be used to describe the process of (un) binding from a structural point of view. In general, the different basins show the importance of V for the binding process, as they present a native or near-native binding of V along the transition state region. The importance of V has been previously described for PDZ2.[20] We found that the correct burial of the V side chain on the binding site by contact to the side chain of Leu 78 is a main barrier in the binding process of PDZ3 as well (Figure S 12). The encounter complex, between $u_1 \approx -0.5$ to 0.4, shows the initial interaction of V with α2 (Leu22-V, Gly23-V). Meanwhile, T has a nonnative contact to Ala42, and afterwards to Ala75 which is in contact with T in the crystal structure. Overall, the geometric annotation suggests a binding process by which the peptide first interacts via its V with the β2 strand (Leu22-V, Gly23-V), and via T to the α1 (Ala42-T) or α2 helix (His71-T, Ala75-T). The SAPPHIRE plot also shows that W interacts with Ala77 on the α2 helix (Ala77-W). The nonnative contacts then need to be broken in the high free energy barrier. The E residue remains largely unbound in the transition state region (Ser38-E), and K does not appear at all in the important residue pairs, due to its flexibility. Some intra protein residue pairs also show closer distances on these regions, such as Asn25-Ile37, and Asn25-Tyr96. Finally, the side chain of E positions itself correctly, shown by the distances to Ser38 and Tyr96, and so the

bound state is reached. The E-Tyr96 distance of between 5 Å and 10 Å has been previously reported, and implies that the α3 helix remains mostly docked to its site near the binding site while the peptide is in the binding site.[11]

We now focus the analysis on the α3 helix for which there is experimental evidence of allosteric signaling upon peptide binding.[11,17] The distance between residues Glu94 and Ala101 is the one chosen to report on the structure of the C-terminal region. These residue positions are the ones used in the time-resolved spectroscopy experiments with the photoswitch.[17] The selected residues are seven residues apart and thus they span two turns of an α helix, which correspond to a separation of 10.5 Å. A *dssp* analysis of these residues on the bound and unbound stretches shows that α3 can comprise either one or two turns (Figure S 14 and Figure S 15). The two peaks of the distribution of the Glu94-Ala101 distance at 10.5 Å and 13.5 Å correspond to a two-turn and one-turn α3 helix, respectively.

The distance between the Cα atoms of Glu94 and Ala101 fluctuates frequently during the sampling (Figure 7(b)) in accordance to an unfolding timescale of 4 ns as measured by time-resolved spectroscopy.[12] The larger distance at around 15 Å corresponds to a partial unfolding of the helix and stretching of the Glu94-Ala101 segment, which can be seen when visualizing the trajectories (Movies S1 and S2). In the photocontrollable PDZ3, the azobenzene isomerizes from the *cis* conformation, which favors the α-helix conformation, to the loop-favoring *trans* conformer. This imposes a change of roughly 4 Å between the anchoring points,[46] which is only part of the fluctuation range observed in the simulations. Conversely, a Glu94-Ala101 distance substantially shorter than 10 Å corresponds to a coil conformation that cannot be reached in the presence of the photoswitchable linker. As can also be seen in the SAPPHIRE plot (Figure 8), this distance fluctuates greatly, from less than 5 Å to more than 20 Å, in both the first and last basin, while it remains rather constrained between 10 Å and 16 Å in the transition region. The highest fluctuations (down to values shorter than 10 Å) occur when the C-terminal helix is dislodged away from the β2 strand, meaning a Val27 CB to Phe99 CZ distance greater than 12.5 Å. In contrast, fluctuations of the Glu94-Ala101 distance between 10 and 16 Å are observed when the distance between α3 and Val27 is below 10 Å, although at a separation of 5 Å the distance between Glu94 and Ala101 can fluctuate down to 8 Å. This behavior shows the importance of the packing of the α3 helix towards the protein and how it impacts its own secondary structure.

The optimal RC framework and SAPPHIRE plot analysis allow for a fully data-driven definition of the main free energy basins which is useful for further geometric analysis. We used the trajectory reordered according to the newly obtained RC to calculate a series of metrics between distributions of pairwise distances along the trajectory. First, we compare the distributions of distances in the bound and non-natively-bound and unbound basin (Figure 9, top panel). The first metric used is the Kolmogorov-Smirnov (KS) test value, which takes the maximum distance between two empirical distributions, while the second is a measure of distance between distributions, namely the Kullback-Leibler (KL) divergence, which is based on their relative entropy. Afterwards, (Figure 9, bottom panel), we chose metrics that capture the information content of the distances projected on the RC. One such measure is the mutual information, for which the relative entropies of each basin were calculated, cutting at points u_1 = [−0.887, −0.74, −0.552, −0.325, 0.094, 0.227, 0.269, 0.375, 0.561]. The other was the correlation of the distances to the RC itself.

For the C-terminal region, the interaction presenting the greatest change in distance distribution is the one between α3 and the α2 binding helix. Another important variation is observed between α3 and the β2-β3 loop (Figure 9). We have also checked which distance pairs, excluding peptide-protein interactions, belong to the 99.9 percentile of the different values tested (Table S 1). Many of these interactions involve either Leu78 and Lys79 in helix α2 or Asn25 in the β2 strand (both forming the binding groove) and other regions of the protein. Ile37, which was discovered by the conventional analyses, also shows up, with the Ile37-Asn25 and Ile37-Lys79 distances shown as highly variable ones. The distance between Lys79 and Asn25 reports on the degree of opening of the binding site, and has previously been used as RC[18] and shown to have a big variance in molecular simulations.[19] In this trajectory, it fluctuates on a lower range in the bound basin compared to the unbound one, which is consistent with what was found previously. Another interesting interaction is the one between Lys79 and Glu57, located in the β4 strand, which is also in the vicinity of α3. Further interaction is seen between Lys79 and residues Gln90-Glu94 of α3, which have both a high KS test value and mutual information. This means their distance distribution varies significantly between bound and unbound states, and it is a distance with a high relative entropy when partitioned in the basins studied.

Meanwhile, correlation and KL divergence inform particularly on the interactions of Asn25 and Leu78 with other residues. Leu78, like its neighboring Lys79 on α2, has a significative KL divergence to the residues Gln57-Leu59 of β4, which are in close contact with the α3 helix as mentioned above. Leu78 also reports on the interactions with

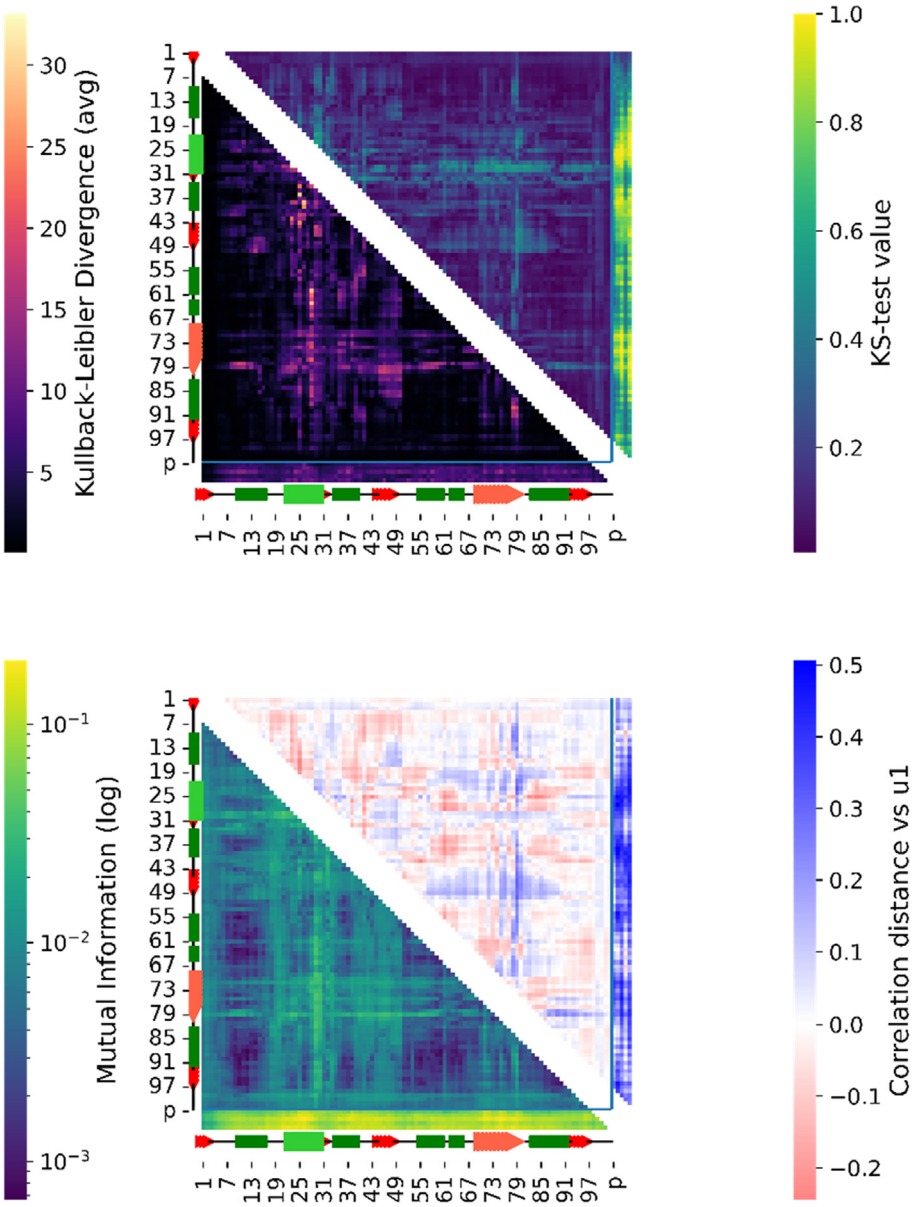

Figure 9. Analysis of pairwise distances in the bound and unbound states. (Top) Differences in pairwise distances between bound ($u_1 \approx -0.9$) and unbound ($u_1 \approx 0.5$) basins. (Upper diagonal) Kolmogorov-Smirnov (KS) test value for pairwise distance distributions. (Lower diagonal) Kullback-Leibler (KL) divergence for pairwise distance distributions. These metrics and distance measures between empirical distributions show residue pairs with significant variation between the bound and unbound basins of the FEP projected on u_1. (Bottom) Information content of pairwise distances. (Lower diagonal) Mutual Information (log) of pairwise distances clustered by basin. (Upper diagonal) Pearson correlation of pairwise distances and RC. Mutual Information captures the relative entropy of each distance distribution along the RC, and the correlation shows the similarity of the distance between the residues to the RC itself. In all, the pairs of residues with more influence on the optimized reaction coordinate are elucidated.

the β2 strand (mainly through Asn25), and other regions of the protein, such as Phe36-Ser38 located between β2 and α1, and Lys92 at the beginning of the C-terminal α3 helix.

Overall, these findings provide evidence that the signal is transmitted from the β2 strand, which forms antiparallel β-sheet hydrogen bonds with the peptide ligand, to the Phe36-Ser38 and Arg53-Ile58 segments of the strands β3 and β4, respectively. From the latter strands, the signal is transduced to the spatially close C-terminal α3 helix, in a domino-like process.[47] The importance of Ile37 has already been highlighted by the community network analysis in the Conventional analysis of the MD simulations section. Some of these residues, especially Asn25, Ile37, and Leu78 have

been identified in previous MD and energy perturbation studies as possible carriers of allosteric signals, although those simulations did not investigate the unbound state.[25,27–28,24] In contrast to our simulation results, a study of electrostatic interactions highlighted the importance of the β2-β3 loop in allosteric regulation,[21] which does not seem to emerge from our unsupervised analysis.

Finally, the transmission of the peptide binding signal by a domino effect does not necessarily exclude other mechanisms of signal transduction. Although the binding signal is transmitted to the α3 helix mainly via spatially adjacent elements of secondary structure, some remote regions of the protein like the β1 strand and α1 helix seem also to react to the binding of the peptide according to the comparison of inter-residue distance distributions (Figure 9). A more diffuse propagation of the signal is congruent with the observation that allosteric transitions are "felt" by all residues in the protein, though in different ways depending on their environment and properties.[22]

Conclusions

We have investigated allosteric signaling in the PDZ3 domain by multiple molecular dynamics simulations started from the KETWV-peptide bound state or the fully dissociated state amounting to 34 μs of sampling. The analysis with conventional methods, such as the contact maps and cross-correlation analyses, has provided information on pairs of residues potentially involved in the transmission of the signal upon peptide binding. The community network analysis has identified Ile37 (on the β3 strand) as a key residue in the allosteric communication. It also suggested a potential role of the β2 and β3 strands in propagating the signal from the peptide to the C-terminal helix α3. Thus, we decided to investigate in depth the role of Ile37 by additional simulations of the Ile37Ala-PDZ3 mutant. These simulations show that the Ile to Ala mutation at position 37 reduces the correlated motions, and therefore the allosteric signal transduction.

An important difference with respect to previous simulation studies of PDZ domains is the unsupervised analysis of the MD trajectories based on the optimal RC framework. We have projected the free energy along the slowest-relaxation eigenvector. This is an optimal RC that captures the process of peptide (un)binding while resolving intermediate transition state regions. Importantly, the optimized RC was used to guide further sampling of the transition regions which has resulted in four additional binding events. This RC is able to elucidate important structural changes between the basins on the projected FEP which are comparable to, but more detailed than, those obtained by more traditional methods such as cross-correlation. Furthermore, the RC clearly distinguishes the bound and unbound states, and shows some partition in the bound basin. In contrast to geometric approaches, the use of the slowest-relaxation eigenvector as the progress index in the SAPPHIRE plot results in a description of the whole binding process. The combination of slowest-relaxation eigenvector as optimal RC, SAPPHIRE analysis, and methods based on information theory provides evidence that the signal of peptide binding to PDZ3 is transmitted from the binding groove to the α3 helix in a domino-like cascade through regular elements of secondary structure that are spatially adjacent.

The overall agreement between our results and previous experimental and computational studies on PDZ3 suggests that the optimal RC approach could prove valuable for other complex systems for which the allosteric pathways have not been described. Further analyses such as the pairwise comparison of additional free energy basins, for example using KL divergence of residue distances, could provide even a richer description of the binding process, showing the key interactions at each barrier of the FEP. We plan to further investigate the transduction of the allosteric signal by MD simulations of forced (un)folding of the helix α3, which will directly emulate the photoswitching of the azobenzene linker.[48] Furthermore, the residue pairs and interactions highlighted in the present simulation study could be analyzed by mutational studies *in vitro*. More directly, their role in the wild type PDZ3 domain could be verified by means of NMR spectroscopy (^{15}N backbone relaxation and side chain ^{2}H-methyl relaxation experiments) and other biophysical techniques (e.g., isothermal titration calorimetry) which have already proved useful in studying allosteric effects in PDZ domains.

Materials and methods

Molecular dynamics simulations

Molecular dynamics simulations were prepared using the PDZ3 domain of *Rattus norvegicus* PSD-95 in complex with the KETWV pentapeptide, which has the same sequence as in the experimental study.[17] The PDZ3 residues from Leu302 to Asn403 are here re-numbered from 1 to 102, and are shown with the three-letter code, while the peptide is unnumbered and designated with its one-letter code. All simulations were run by GROMACS 2020.5[49] using the CHARMM36 force field[50] and the TIP3P water model.[51] Six independent runs were started from the crystal structure of the complex (PDB ID: 1TP5), immersed in a 6.9-nm cubic box, and neutralized with Na^+ and Cl^- ions at a concentration of 150 mM. The system was first minimized for 500,000 steps, and then underwent a 2 ns isothermal-isobaric (NPT) ensemble equilibration to 300 K and 1 bar, using the Berendsen barostat with a coupling time of 2 ps.[52] Afterwards, production runs were carried out in the canonical (NVT) ensemble, utilizing the velocity rescaling thermostat with a coupling time of 1 ps.[53] Furthermore, 20 binding runs were started from random positions and orientations of the peptide in the simulation box which were generated using the GROMACS *insert-molecules* command. The equilibration and production phases were similar

as for the runs started from the complex. All simulations were carried out with periodic boundary conditions, and long-range interactions were treated by the particle mesh Ewald method with a cutoff of 12 Å.[54] The same 12 Å cutoff applied to van der Waals interactions. The time step of integration was 2 fs. Production sampling was collected to 1200 ns, and energies and coordinates were saved every 25 ps for analysis. Considering simulations restarted from transition regions, as described afterwards, the total sampling collected amounts to 34 μs.

Structural analyses

Native contacts were determined from a crystal structure, while other descriptive contacts between peptide and protein were extracted from one of the binding trajectories. The trajectory was split into four states according to the RMSD of the Cα atoms of the peptide (upon structural overlap of the PDZ3) and minimum distance criteria. Ten intervals of 20 ns each were selected from the unbinding trajectories in regions where the Cα RMSD of the peptide was below 2.5 Å. Ten sections of 20 ns were sampled from the unbinding trajectories in which the minimum distance between peptide and protein is above 10 Å (Figure S 8). The root mean square fluctuation (RMSF) was calculated independently for bound and unbound sections of the trajectory, using intervals of 5 ns for structure averaging as previously done for PDZ3.[18] The trajectory was reordered geometrically by a *progress index*,[45] and visualized using the States and Pathways Projected on High Resolution (SAPPHIRE) visualization method to find structural insights about the states found.[37] This visualization was used to reseed 20 new trajectories (see Reseeded trajectories subsection).

A strategy commonly used for the analysis of allostery is the use of Dynamic Cross-Correlation of atomic displacements to find correlated motions.[9,55] Such an analysis was performed using the Bio3d R library.[41] The cross-correlations were calculated for all heavy atoms, and averaged per residue. Cross-correlations were calculated for trajectory segments where peptide was bound and where it was completely unbound, as previously defined. Furthermore, the cross-correlation difference was calculated to find regions where a change in the collective movements is caused by the presence of the peptide. From the cross-correlation matrix, a community network can be constructed by clustering residues with similar cross-correlation patterns into macro-nodes of highly dynamically connected residues.[7] Such an analysis was performed with Bio3d and visualized on VMD.[56]

A kinetic analysis of the binding events was performed by fitting an exponential curve $f(t) = \exp(-t/\tau)$ to the probability of the peptide being unbound at each time. This was calculated for the 20 binding trajectories. In addition, binding events during reseeded trajectories were also selected. These are one of the SAPPHIRE reseeded trajectories, and four of the eigenvector projection reseeded trajectories, yielding a total of 25 binding trajectories (Table 1). For the reseeded trajectories, the time of binding was calculated as the time of the trajectory plus the time of the sampled frame in the original trajectory. Thus, the data are not fully independent as each of the five reseeded trajectories shares a segment with one of the original binding trajectories. This approximation introduces an error that is smaller than a complete neglect of the original trajectory which would result in too rapid binding times. The probability of the peptide being unbound was calculated by setting 25 binary vectors of length 1000, and setting the value of each frame to 1 if unbound. For each binding event, the vector was set to zero for all frames after the RMSD threshold was crossed. The binding rate constant k_{on} is the reciprocal of the product of the exponential decay factor τ and the peptide concentration (5 mM according to the size of the simulation box). The K_D dissociation constant was calculated as k_{off}/k_{on}, using an unbinding rate constant k_{off} of (1/200) ms^{-1} as previously reported[12]. The unbinding rate cannot be calculated as only one unbinding event was observed. This is expected from the experimentally measured ms time scale of the dissociation process. To assess the robustness of the threshold used to define a binding event, the time for binding was calculated for three values of the peptide RMSD (3, 5, and 7.5 Å). Furthermore, the kinetic constants were also calculated using the value of the optimized reaction coordinate, described in the following section, and setting 0.1 and −0.75 as threshold values (Figure S 3).

Mutation studies

The geometric analysis revealed the role of Ile37 as essential to signal transmission. Thus we decided to perform MD simulations of the Ile37Ala mutant of PDZ3. A set of 24 independent trajectories, of 200 ns each, was seeded. The initial topology was obtained by mutating Ile37 to Ala utilizing an in-house developed graphical user interphase (ACGUI) which accesses the function from the software CAMPARI.[57] Due to the long timescale of peptide unbinding, only binding simulations were carried out, in which the peptide was randomly positioned around the Ile37Ala-PDZ3. The simulations were performed as previously described in the section Molecular dynamics simulations. RMSD of the peptide with respect to the (mutated) crystal structure was monitored to determine binding events.

To explore the effect of the mutation on the structure of the C-terminal α3 helix, the secondary structure of Lys92-Asn102 was determined using the MDTraj python library.[58] Furthermore, the distance between Glu94 and Ala101 was monitored and compared with that of the wild type PDZ3. The dynamic cross correlation was calculated for the mutated trajectories, excluding three of them where the peptide bound to the binding groove. The cross correlation matrix was used to calculate the correlation networks as previously described.

Unsupervised analysis

An approximation to the first eigenvector of the transfer operator of the underlying dynamics of the system was calculated as previously published,[36] adapting the code described in[43,59]. The eigenvector time-series is approximated by a linear combination of basis functions. In short, a seed reaction coordinate is iteratively and recursively optimized using randomly sampled collective variables from the trajectory coordinates. In this case, interatomic distances between pairs of residues were chosen. For each step of the optimization, the RC is updated as a function of the RC itself and a new, randomly chosen, distance. Although the result of each optimization step might depend on the distance chosen, overall the optimization converges independently of the order and distances used. Furthermore, the high number of steps ensures all distances are considered. A basis function combines the information from the chosen collective variable and the existing RC. The coefficients of the basis functions are chosen such that they provide a solution to the generalized eigenvalue problem as described in the original publication,[36] and result in a minimum of the total displacement along the RC. In this way, a value of the RC is assigned to each frame in the trajectory. This RC approximates the slowest-relaxing eigenvectors of the system.

The MD trajectories were sub-sampled with a timestep Δt_0 = 100 ps and then concatenated. A Δt of 256*Δt_0 (25.6 ns) was chosen as lag time to calculate the eigenvectors. A slightly larger lag time helps mask the suboptimality of the eigenvector. A lag time Δt_∞ of 1024*Δt_0 (102.4 ns) is used as reference to test the convergence of the optimization. This value should be smaller than the characteristic lag time of the process to be described. In this case, the signal transduction occurs on a timescale of 200 ns, while peptide unbinding is expected to happen after 200 ms.[12] For the iterative optimization of the RC, interatomic distances were calculated between (non-replacement) combinations of peptide Cα, protein Cα atoms, and protein side chain atoms. In each iteration step, the distances

were chosen randomly. Furthermore, to avoid dominance of intra-protein distances, peptide-protein intermolecular distances were chosen every five iteration steps. This can be understood as assigning a 20% weight to the peptide protein distances (510 distances), against an 80% weight to intra-protein (Cα and side chain) distances (18537 distances). Distances were transformed by a sigmoidal function $f(x) = 1 - (1 + exp(-(x - \chi)/\tau))^{-1}$ centered at $\chi = 7$ Å and with a sharpness of $\tau = 2$ Å. With these values of χ and τ the sigmoidal function captures the formation of van der Waals contacts and hydrogen bonds and decays to zero for large separations at which the energy contribution is essentially zero. In other words, the sigmoidal transformation is equivalent to a contact map transform.

The committor calculation framework introduces adaptivity as a way to avoid overfitting of the transition regions and sub-optimality in the basin areas.[35] Adaptivity is based on scanning of the profile of an optimality criterion along different values of the RC. Frames for which the value of the RC shows optimality are kept constant on the next iteration of the RC optimization. In this case, we use the eigenvector optimality criterion θ(x, Δt), which for an optimal RC is constant and close to zero along x (reaction coordinate) and Δt. The difference in the optimality criterion θ is calculated for different timesteps, to find the timestep which shows the highest difference along the θ profile. This is described as $\Delta\theta(x, \Delta t_i, \Delta t) = \theta(x, \Delta t_i) - \theta(x, \Delta t)$, with Δt_i equal the timestep for which $\Delta\theta(x, \Delta t_i, \Delta t)$ has the largest range. Then, the suboptimality at each point x of the RC is defined as $s(x) = \exp(\Delta\theta(x, \Delta t_i, \Delta t) - \max_x(\Delta\theta(x, \Delta t_i, \Delta t)))$. Finally, s(x) is normalized to the 0 to 1 range. This suboptimality is used as probability for a Bernoulli random variable which fixes each x with probability s(x). The optimization was carried out for 80,000 steps. A cut-based free energy profile $F_{C,1} = -kTln(Z_{C,1}(x))$ was calculated from the partition function $Z_{C,1}(x)$. At a point x of a reaction coordinate, the partition function considers one half of the sum of the distances in the coordinate of all the steps going through this point.[30] Based on the FEPs projected on the two slowest-relaxation eigenvectors, a two-dimensional histogram was obtained by binning the values of each trajectory frame based on the value of the frame on each of the two coordinates. This projection elucidates parallel pathways along the RC. The slowest-eigenvector projection was used to determine the poorly sampled transition regions between the bound and unbound states, from which 16 new simulations were started as mentioned above (subsection Reseeded trajectories).

For each interatomic distance, a series of statistics were calculated based on the basins obtained from the FEP. Distance value distributions p_i were calculated as a histogram, with the binning being the maximum between the Sturges and Freedman-Diaconis estimators using NumPy 1.20.1 functions.[60] The RC was discretized at the points $u_1 = [-0.887, -0.74, -0.552, -0.325, 0.094, 0.227, 0.269, 0.375, 0.561]$, and the mutual information for each distance x was calculated as $I = H - \sum_{k=1}^{Nbasins} q_k H(x|a_k)$. $H = -\sum_{i=1} p_i log(p_i)$ describes the total entropy of the distance values, calculated with SciPy 1.5.3. Meanwhile, $H(x|a_k)$ is the entropy of basin k, with q_k being the size of basin k, measured as the number of frames in it.[61–62]

The difference in the distribution of distances at the natively bound and non-natively bound basins was calculated using the two-sample Kolmogorov-Smirnov (KS) test and the Kullback-Leibler (KL) divergence. The two-sample KS test was calculated with the SciPy python library on empirical distributions of pairwise distances. The assumed null hypothesis is that $p_{native} = p_{nonnat}$, The test value D is calculated as $D = sup_x|X_{native}(x) - X_{nonnat}(x)|$. The test finds the least upper bound of the difference in the empirical distribution $\hat{X}$ of distance X between the native and nonnative basins. Distance pairs with a two-tailed p-value higher than 5% were omitted for the analysis, being considered non-significant. The KL divergence, was calculated as $\sum p_{native}(i) log(\frac{p_{native}(i)}{p_{nonnat}(i)})$, or $\sum p_{nonnat}(i) log(\frac{p_{nonnat}(i)}{p_{native}(i)})$ if the first expression is undefined, for each bin i of the distance distribution.[63] Additionally, the correlation of each pairwise distance to the RC itself was also calculated. Once calculated, the RC was used as Progress Index for a SAPPHIRE-style analysis. The structural annotation was determined by the previous analysis of all pairwise distances, choosing distances with high mutual information. Twelve peptide-protein and five intra-protein pairwise distances were chosen in the end to describe the process.

Reseeded trajectories

An initial analysis of the six unbinding and twenty binding trajectories was performed using the SAPPHIRE visualization method,[37] in which frames are ordered according to their pairwise geometric similarity.[45] For the calculation of the progress index, a set of 105 peptide-protein distances was used as geometric variable and the Euclidean distance was employed for pairwise comparison of snapshots. From the kinetic annotation, the transition region was defined and 20 frames were selected as coordinate sets for restarting 20 new trajectories, respectively, using newly generated velocities (Table 1). Production MD was collected for 100 ns for each of the 20 reseeded runs. Two runs, one with a binding and one with an unbinding event were then extended to 300 ns. This reseeding was inspired by a previously published adaptive sampling procedure called progress index guided sampling.[38]

A further reseeding was undertaken using the projection onto the slowest-relaxation eigenvector calculated previously (see section Unsupervised analysis). From the initial (un)binding runs, and the reseeded trajectories mentioned above, the free energy profile of peptide (un)binding was calculated. From the transition state region, 16 frames were sampled and their coordinates used to restart trajectories with new velocities. Production runs were collected for 300 ns (Table 1).

CRediT authorship contribution statement

Pablo Andrés Vargas-Rosales: Methodology, Investigation, Writing – original draft, Writing – review & editing. **Amedeo Caflisch:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing, Funding acquisition.

DATA AVAILABILITY

Data will be made available on request.

Acknowledgements

We thank Sergei Krivov for sharing helpful insights about his methods as well as giving us early access to his code. We also thank Francesco Cocina, Ioana Ilie, Julian Widmer, and Peter Hamm for interesting discussions. We also wish to thank the anonymous reviewers for their valuable comments. This work was supported financially by an Excellence grant of the Swiss National Science Foundation to A.C. We acknowledge access to Eiger@Alps at the Swiss National Supercomputing Centre, Switzerland.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2022.167661.

Received 21 January 2022;
Accepted 24 May 2022;
Available online 28 May 2022

Keywords:
slowest-relaxation eigenvector;
molecular dynamics;
free energy profile;
optimal reaction coordinate;
PDZ3 domain

Abbreviations:
RC, reaction coordinate; MD, molecular dynamics; RMSF, root mean square fluctuation; RMSD, root mean square deviation; DCC, Dynamic Cross-Correlation; FEP, free energy profile; KL, Kullback-Leibler divergence; KS, Kolmogorov-Smirnov test

References

1. Monod, J., Wyman, J., Changeux, J.P., (1965). On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12**, 88–118. https://doi.org/10.1016/S0022-2836(65)80285-6.
2. Koshland, D.E.J., Némethy, G., Filmer, D., (1966). Comparison of experimental binding data and theoretical models in proteins containing subunits*. *Biochemistry* **5**, 365–385. https://doi.org/10.1021/BI00865A047.
3. Cui, Q., Karplus, M., (2008). Allostery and cooperativity revisited. *Protein Sci.* **17**, 1295–1307. https://doi.org/10.1110/PS.03259908.
4. Liu, J., Nussinov, R., (2016). Allostery: An overview of its history, concepts, methods, and applications. *PLOS Comput. Biol.* **12**, https://doi.org/10.1371/JOURNAL.PCBI.1004966 e1004966.
5. Wodak, S.J., Paci, E., Dokholyan, N.V., Berezovsky, I.N., Horovitz, A., Li, J., Hilser, V.J., Bahar, I., Karanicolas, J., Stock, G., Hamm, P., Stote, R.H., Eberhardt, J., Chebaro, Y., Dejaegere, A., Cecchini, M., Changeux, J.P., Bolhuis, P.G., Vreede, J., Faccioli, P., Orioli, S., Ravasio, R., Yan, L., Brito, C., Wyart, M., Gkeka, P., Rivalta, I., Palermo, G., McCammon, J.A., Panecka-Hofman, J., Wade, R.C., Di Pizio, A., Niv, M.Y., Nussinov, R., Tsai, C.J., Jang, H., Padhorny, D., Kozakov, D., McLeish, T., (2019). Allostery in its many disguises: from theory to applications. *Structure.* **27**, 566–578. https://doi.org/10.1016/J.STR.2019.01.003.
6. Fischer, S., Olsen, K.W., Nam, K., Karplus, M., (2011). Unsuspected pathway of the allosteric transition in hemoglobin. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 5608–5613. https://doi.org/10.1073/PNAS.1011995108/-/DCSUPPLEMENTAL.
7. S. Bowerman, J. Wereszczynski, Detecting Allosteric Networks Using Molecular Dynamics Simulation, in: Methods Enzymol., Academic Press Inc., 2016: pp. 429–447. https://doi.org/10.1016/bs.mie.2016.05.027.
8. Hünenberger, P.H., Mark, A.E., van Gusteren, W.F., (1995). Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. *J. Mol. Biol.* **252**, 492–503. https://doi.org/10.1006/JMBI.1995.0514.
9. Ichiye, T., Karplus, M., (1991). Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins Struct. Funct. Bioinforma.* **11**, 205–217. https://doi.org/10.1002/prot.340110305.
10. Lee, H.J., Zheng, J.J., (2010). PDZ domains and their binding partners: Structure, specificity, and modification. *Cell Commun. Signal.* **8**, 1–18. https://doi.org/10.1186/1478-811X-8-8.
11. Petit, C.M., Zhang, J., Sapienza, P.J., Fuentes, E.J., Lee, A.L., (2009). Hidden dynamic allostery in a PDZ domain. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 18249–18254. https://doi.org/10.1073/pnas.0904492106.
12. Bozovic, O., Ruf, J., Zanobini, C., Jankovic, B., Buhrke, D., Johnson, P.J.M., Hamm, P., (2021). The speed of allosteric signaling within a single-domain protein. *J. Phys. Chem. Lett.* **12**, 4262–4267. https://doi.org/10.1021/acs.jpclett.1c00915.
13. Mostarda, S., Gfeller, D., Rao, F., (2012). Beyond the binding site: The role of the β2 - β3 loop and extra-domain structures in PDZ domains. *PLoS Comput. Biol.* **8**, https://doi.org/10.1371/journal.pcbi.1002429 e1002429.
14. Murciano-Calles, J., Corbi-Verge, C., Candel, A.M., Luque, I., Martinez, J.C., (2014). Post-translational modifications modulate ligand recognition by the third PDZ domain of the MAGUK protein PSD-95. *PLoS ONE* **9**, https://doi.org/10.1371/journal.pone.0090030 e90030.
15. Zhang, J., Petit, C.M., King, D.S., Lee, A.L., (2011). Phosphorylation of a PDZ domain extension modulates binding affinity and interdomain interactions in postsynaptic density-95 (PSD-95) protein, a membrane-associated guanylate kinase (MAGUK). *J. Biol. Chem.* **286**, 41776–41785. https://doi.org/10.1074/jbc.M111.272583.
16. Woolley, G.A., (2005). Photocontrolling peptide α helices. *Acc. Chem. Res.* **38**, 486–493. https://doi.org/10.1021/AR040091V.
17. Bozovic, O., Jankovic, B., Hamm, P., (2020). Sensing the allosteric force. *Nat. Commun.* **11**, 1–7. https://doi.org/10.1038/s41467-020-19689-7.
18. Steiner, S., Caflisch, A., (2012). Peptide binding to the PDZ3 domain by conformational selection. *Proteins Struct. Funct. Bioinforma.* **80**, 2562–2572. https://doi.org/10.1002/PROT.24137.
19. Dudola, D., Hinsenkamp, A., Gáspári, Z., (2020). Ensemble-based analysis of the dynamic allostery in the PSD-95 PDZ3 domain in relation to the general variability of PDZ structures Page 8348. 21 (2020) 8348 *Int. J. Mol. Sci.* **21** https://doi.org/10.3390/IJMS21218348.
20. Blöchliger, N., Xu, M., Caflisch, A., (2015). Peptide binding to a PDZ domain by electrostatic steering via nonnative salt

bridges. *Biophys. J.* **108**, 2362–2370. https://doi.org/10.1016/J.BPJ.2015.03.038.
21. Kumawat, A., Chakrabarty, S., (2017). Hidden electrostatic basis of dynamic allostery in a PDZ domain. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E5825–E5834. https://doi.org/10.1073/PNAS.1705311114/-/DCSUPPLEMENTAL.
22. Hayatshahi, H.S., Ahuactzin, E., Tao, P., Wang, S., Liu, J., (2019). Probing protein allostery as a residue-specific concept via residue response maps. *J. Chem. Inf. Model.* **59**, 4691–4705. https://doi.org/10.1021/ACS.JCIM.9B00447/SUPPL_FILE/CI9B00447_SI_001.PDF.
23. Conti Nibali, V., Morra Bc, G., Havenith, M., D'angelo D, G., Colombo, G., (2018). Concerted motions in allosteric model proteins at terahertz frequencies. *Atti Della Accad. Peloritana Dei Pericolanti - Cl. Di Sci. Fis. Mat. e Nat.* **96**, 6. https://doi.org/10.1478/AAPP.961A6.
24. B. Lakhani, K.M. Thayer, E. Black, D.L. Beveridge, Spectral analysis of molecular dynamics simulations on PDZ: MD sectors, Https://Doi.Org/10.1080/07391102.2019.1588169. 38 (2019) 781–790. https://doi.org/10.1080/07391102.2019.1588169.
25. Karami, Y., Bitard-Feildel, T., Laine, E., Carbone, A., (2018). "Infostery" analysis of short molecular dynamics simulations identifies highly sensitive residues and predicts deleterious mutations. *Sci. Rep.* **2018 81.8**, 1–18. https://doi.org/10.1038/s41598-018-34508-2.
26. Atilgan, C., Guclu, T.F., Atilgan, A.R., (2021). Dynamic community composition unravels allosteric communication in pdz3. *J. Phys. Chem. B.* **125**, 2266–2276. https://doi.org/10.1021/ACS.JPCB.0C11604/SUPPL_FILE/JP0C11604_SI_001.PDF.
27. Wang, W.B., Liang, Y., Zhang, J., Wu, Y.D., Du, J.J., Li, Q. M., Zhu, J.Z., Su, J.G., (2018). Energy transport pathway in proteins: Insights from non-equilibrium molecular dynamics with elastic network model. *Sci. Rep.* **2018 81.8**, 1–13. https://doi.org/10.1038/s41598-018-27745-y.
28. Gulzar, A., Valiño Borau, L., Buchenberg, S., Wolf, S., Stock, G., (2019). Energy transport pathways in proteins: a non-equilibrium molecular dynamics simulation study. *J. Chem. Theory Comput.* **15**, 5750–5757. https://doi.org/10.1021/ACS.JCTC.9B00598/SUPPL_FILE/CT9B00598_SI_001.PDF.
29. Krivov, S.V., (2010). Is protein folding sub-diffusive? *PLoS Comput. Biol.* **6**, 1000921. https://doi.org/10.1371/journal.pcbi.1000921.
30. Krivov, S.V., (2012). On reaction coordinate optimality. *J. Chem. Theory Comput.* **9**, 135–146. https://doi.org/10.1021/CT3008292.
31. Banushkina, P.V., Krivov, S.V., (2016). Optimal reaction coordinates. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **6**, 748–763. https://doi.org/10.1002/WCMS.1276.
32. Banushkina, P.V., Krivov, S.V., (2015). Nonparametric variational optimization of reaction coordinates. *J. Chem. Phys.* **143**, https://doi.org/10.1063/1.4935180 184108.
33. Berezhkovskii, A.M., Szabo, A., (2019). Committors, first-passage times, fluxes, Markov states, milestones, and all that. *J. Chem. Phys.* **150**, https://doi.org/10.1063/1.5079742 054106.
34. Roux, B., (2021). String method with swarms-of-trajectories, mean drifts, lag time, and committor. *J. Phys. Chem. A.* **125**, 7558–7571. https://doi.org/10.1021/ACS.JPCA.1C04110.
35. Krivov, S.V., (2018). Protein folding free energy landscape along the committor - the optimal folding coordinate. *J. Chem. Theory Comput.* **14**, 3418–3427. https://doi.org/10.1021/acs.jctc.8b00101.
36. Krivov, S.V., (2021). Blind analysis of molecular dynamics. *J. Chem. Theory Comput.* **17**, 2725–2736. https://doi.org/10.1021/ACS.JCTC.0C01277.
37. Blöchliger, N., Vitalis, A., Caflisch, A., (2014). High-Resolution Visualisation of the States and Pathways Sampled in Molecular Dynamics Simulations. *Sci. Rep.* **2014 41.4**, 1–5. https://doi.org/10.1038/srep06264.
38. Bacci, M., Vitalis, A., Caflisch, A., (1850). A molecular simulation protocol to avoid sampling redundancy and discover new states. *Biochim. Biophys. Acta - Gen. Subj.* **2015**, 889–902. https://doi.org/10.1016/j.bbagen.2014.08.013.
39. Morra, G., Genoni, A., Colombo, G., (2014). Mechanisms of differential allosteric modulation in homologous proteins: Insights from the analysis of internal dynamics and energetics of PDZ domains. *J. Chem. Theory Comput.* **10**, 5677–5689. https://doi.org/10.1021/CT500326G/SUPPL_FILE/CT500326G_SI_001.PDF.
40. Doyle, D.A., Lee, A., Lewis, J., Kim, E., Sheng, M., MacKinnon, R., (1996). Crystal Structures of a Complexed and Peptide-Free Membrane Protein-Binding Domain: Molecular Basis of Peptide Recognition by PDZ. *Cell* **85**, 1067–1076. https://doi.org/10.1016/S0092-8674(00)81307-0.
41. Grant, B.J., Rodrigues, A.P.C., ElSawy, K.M., McCammon, J.A., Caves, L.S.D., (2006). Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **22**, 2695–2696. https://doi.org/10.1093/BIOINFORMATICS/BTL461.
42. P. Mark, L. Nilsson, Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K, (2001). https://doi.org/10.1021/jp003020w.
43. Krivov, S.V., (2021). Nonparametric Analysis of Nonequilibrium Simulations. *J. Chem. Theory Comput.* **17**, 5481. https://doi.org/10.1021/ACS.JCTC.1C00218.
44. Krivov, S.V., Karplus, M., (2008). Diffusive reaction dynamics on invariant free energy profiles. *Proc. Natl. Acad. Sci.* **105**, 13841–13846. https://doi.org/10.1073/PNAS.0800228105.
45. Blöchliger, N., Vitalis, A., Caflisch, A., (2013). A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems. *Comput. Phys. Commun.* **184**, 2446–2453. https://doi.org/10.1016/j.cpc.2013.06.009.
46. Zhang, F., Zarrine-Afsar, A., Al-Abdul-Wahid, M.S., Prosser, R.S., Davidson, A.R., Woolley, G.A., (2009). Structure-based approach to the photocontrol of protein folding. *J. Am. Chem. Soc.* **131**, 2283–2289. https://doi.org/10.1021/JA807938V/SUPPL_FILE/JA807938V_SI_001.PDF.
47. Kornev, A.P., Taylor, S.S., (2015). Dynamics-Driven Allostery in Protein Kinases. *Trends Biochem. Sci.* **40**, 628–647. https://doi.org/10.1016/J.TIBS.2015.09.002.
48. Jankovic, B., Gulzar, A., Zanobini, C., Bozovic, O., Wolf, S., Stock, G., Hamm, P., (2019). Photocontrolling Protein-Peptide Interactions: From Minimal Perturbation to Complete Unbinding. *J. Am. Chem. Soc.* **141**, 10702–10710. https://doi.org/10.1021/JACS.9B03222/SUPPL_FILE/JA9B03222_SI_001.PDF.
49. Abraham, M.J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., Lindah, E., (2015). Gromacs: High performance molecular simulations through multi-level

parallelism from laptops to supercomputers. *SoftwareX.* **1–2**, 19–25. https://doi.org/10.1016/j.softx.2015.06.001.
50. Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., De Groot, B.L., Grubmüller, H., MacKerell, A.D., (2017). CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods.* **14**, 71. https://doi.org/10.1038/NMETH.4067.
51. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., Klein, M.L., (1998). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926. https://doi.org/10.1063/1.445869.
52. Berendsen, H.J.C., Postma, J.P.M., Van Gunsteren, W.F., Dinola, A., Haak, J.R., (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690. https://doi.org/10.1063/1.448118.
53. Bussi, G., Donadio, D., Parrinello, M., (2007). Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, https://doi.org/10.1063/1.2408420 014101.
54. Darden, T., Perera, L., Li, L., Pedersen, L., (1999). New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure.* **7**, R55–R60. https://doi.org/10.1016/S0969-2126(99)80033-1.
55. Lange, O.F., Grubmüller, H., (2006). Generalized correlation for biomolecular dynamics. *Proteins Struct. Funct. Genet.* **62**, 1053–1061. https://doi.org/10.1002/prot.20784.
56. Humphrey, W., Dalke, A., Schulten, K., (1996). VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.
57. Vitalis, A., Pappu, R.V., (2009). Chapter 3 Methods for Monte Carlo Simulations of Biomacromolecules. *Annu. Rep. Comput. Chem.* **5**, 49–76. https://doi.org/10.1016/S1574-1400(09)00503-9.
58. McGibbon, R.T., Beauchamp, K.A., Harrigan, M.P., Klein, C., Swails, J.M., Hernández, C.X., Schwantes, C.R., Wang, L.-P., Lane, T.J., Pande, V.S., (2015). MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109**, 1528–1532. https://doi.org/10.1016/J.BPJ.2015.08.015.
59. Banushkina, P.V., Krivov, S.V., (2015). Fep1d: A script for the analysis of reaction coordinates. *J. Comput. Chem.* **36**, 878–882. https://doi.org/10.1002/JCC.23868.
60. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., (2020). Array programming with NumPy. *Nat.* **2020 5857825.585**, 357–362. https://doi.org/10.1038/s41586-020-2649-2.
61. Cover, T.M., Thomas, J.A., (2006). 2.3 Relative Entropy and Mutual Information. *In: Elem. Inf. Theory , 2nd ed.*. John Wiley & Sons, Inc., Hoboken, NJ, pp. 16–20.
62. Victor, J.D., (2002). Binless strategies for estimation of information from neural data. *Phys. Rev. E.* **66**, https://doi.org/10.1103/PhysRevE.66.051903 051903.
63. S. Kullback, R.A. Leibler, On Information and Sufficiency, Https://Doi.Org/10.1214/Aoms/1177729694. 22 (1951) 79–86. https://doi.org/10.1214/AOMS/1177729694.