The Journal of Chemical Physics

RESEARCH ARTICLE | OCTOBER 03 2024

Self-organizing maps of unbiased ligand–target binding pathways and kinetics

Special Collection: Machine Learning for Biomolecular Modeling

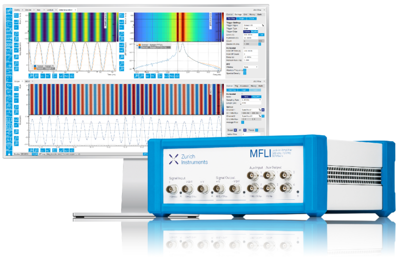
Lara Callea ; Camilla Caprai ; Laura Bonati ; Toni Giorgino ; Stefano Motta

Check for updates

J. Chem. Phys. 161, 135102 (2024)
https://doi.org/10.1063/5.0225183

View Online

Export Citation



AIP Publishing

Self-organizing maps of unbiased ligand–target binding pathways and kinetics

Cite as: J. Chem. Phys. 161, 135102 (2024); doi: 10.1063/5.0225183
Submitted: 23 June 2024 • Accepted: 18 September 2024 •
Published Online: 3 October 2024

View Online Export Citation CrossMark

Lara Callea,[1] Camilla Caprai,[2,3] Laura Bonati,[1] Toni Giorgino,[3,a)] and Stefano Motta[1,a)]

AFFILIATIONS

[1] Department of Earth and Environmental Sciences, University of Milano-Bicocca, Piazza della Scienza 1, Milan 20126, Italy
[2] Department of Biosciences, University of Milan, via Celoria 26, Milan 20133, Italy
[3] National Research Council of Italy, Biophysics Institute (CNR-IBF), Via Celoria 26, Milan 20133, Italy

Note: This paper is part of the JCP Special Topic on Machine Learning for Biomolecular Modeling.
[a)] **Authors to whom correspondence should be addressed:** toni.giorgino@cnr.it and stefano.motta@unimib.it

ABSTRACT

The interpretation of ligand–target interactions at atomistic resolution is central to most efforts in computational drug discovery and optimization. However, the highly dynamic nature of protein targets, as well as possible induced fit effects, makes difficult to sample many interactions effectively with docking studies or even with large-scale molecular dynamics (MD) simulations. We propose a novel application of Self-Organizing Maps (SOMs) to address the sampling and dynamic mapping tasks, particularly in cases involving ligand flexibility and induced fit. The SOM approach offers a data-driven strategy to create a map of the interaction process and pathways based on unbiased MD. Furthermore, we show how the preliminary SOM mapping is complementary to kinetic analysis, with the employment of both network-based approaches and Markov state models. We demonstrate the method by comprehensively mapping a large dataset of 640 μs of unbiased trajectories sampling the recognition process between the phosphorylated YEEI peptide and its high-specificity target lck-SH2. The integration of SOM into unbiased simulation protocols significantly advances our understanding of the ligand binding mechanism. This approach serves as a potent tool for mapping intricate ligand–target interactions with unprecedented detail, thereby enhancing the characterization of kinetic properties crucial to drug design.

INTRODUCTION

The elucidation of ligand–target interactions at the atomistic level is central to pharmaceutical research and drug development.[1] Despite the substantial progress achieved through conventional computational approaches such as docking studies, the challenges posed by the dynamic nature of protein targets and potential induced fit effects continue to impede a reliable quantitative description of protein–ligand interactions, essential for drug discovery and optimization.[2] Molecular dynamics (MD) simulations, offering a dynamic and realistic depiction of biomolecular interactions by considering the inherent flexibility and adaptability of protein targets, have emerged as versatile tools to unravel the complexities associated with ligand binding.[3,4] Classical all-atom MD approaches, for example, model all the atoms in a system as particles interacting via bonded and non-bonded potentials described by empirical force fields.[5] MD-based approaches overcome the limitations encountered in rigid docking models, allowing for a more accurate representation of induced fit effects and capturing subtle changes in protein conformation during ligand binding events. However, despite holding great promise, their effective utilization requires addressing challenges related to computational expense and the need for extensive sampling to comprehensively explore the conformational landscape.[6,7] This is particularly crucial for achieving atomistic resolution in ligand–target interactions, a prerequisite for robust drug discovery efforts.

Broadly speaking, MD-based methods can be divided into *biased* (or enhanced) and *unbiased* sampling approaches. Enhanced sampling methods address the issue of computational demands by applying perturbations to the Hamiltonian of the system, enabling the reconstruction of free-energy landscapes and other properties.[6,7] Techniques such as alchemical free-energy perturbation,[8] steered MD,[9,10] metadynamics,[11–13] Gaussian-accelerated MD,[14] supervised MD,[15] random acceleration MD,[16] and many others[17–21] have enjoyed remarkable success, with the drawback that they usually rely on a choice of collective variables to be performed *a priori* and/or they introduce a bias that can influence the natural dynamics of the system. Recently, with the increase in computational power, long-timescale processes have become computationally tractable with *unbiased* simulations.[4,22,23] In unbiased MD, the system evolves only subject to force-field based potentials and thermostats, without any external bias, providing in principle a faithful representation of the system's dynamics. This characteristic enables the extraction of kinetic information, such as transition rates and timescales, directly from the trajectories. However, unbiased MD generates a large amount of data that are not straightforward to analyze. The calculation of kinetic properties in unbiased MD is facilitated by employing advanced analysis techniques, with Markov State Models (MSMs) being a prominent example.[24–26] The construction of MSM, while a powerful tool for capturing the kinetic aspects of molecular processes, often encounters challenges. One notable difficulty arises from the sheer complexity and dimensionality of the data, making convergence a non-trivial task. The identification of suitable metrics for validation purposes further adds to the intricacy.

In response to these challenges, this paper proposes an application of Self-Organizing Maps (SOMs)[27,28] as a novel strategy to account for the complexities associated with ligand flexibility and induced fit phenomena. We demonstrate the approach in a particularly challenging scenario, namely a flexible tetrapeptide ligand (pYEEI), in a high-specificity target (p56lck Src homology 2 domain). Src homology 2 (SH2) domains provide phospho rylation-dependent signaling receptor domains that recognize short peptides with very high sequence specificity and affinity.[29] The signaling function is of high physiological interest: at least one hundred human proteins contain an SH2 domain (InterPro), and phosphopeptide mimetics targeting SH2 have been explored to reduce proliferation in *in vitro* breast cancer models, e.g., inhibiting the STAT3 pathway.[30]

The SH2:pYEEI association occurs within a dynamic landscape on both interacting sides, making it very challenging to map out systematically. SOM offers unique clarity in understanding these complexities, providing a natural map of the interactive process and a robust means of verifying the outcomes. We demonstrate the method on a large dataset of 640 μs of unbiased trajectories, manifold extending the ones used in previous studies.[23] Furthermore, we show how SOM mapping is highly suitable for kinetic analysis, facilitating an in-depth investigation of the temporal dynamics of the process. The method provides opportunities for several systematic kinetic analysis strategies, including both network-based communities and MSM. Of particular interest is the integration of SOM with MSM as an alternative to the initial clustering stage, thus offering the potential for a seamless integration of these two analytical paradigms. This integrative methodology enhances our understanding of ligand binding mechanisms and holds significant promise for optimizing the drug discovery process.

METHODS

Self-organizing maps

A Self-Organizing Map (SOM) is an unsupervised learning approach that facilitates the projection of high-dimensional data into a lower-dimensional space.[27,31] This technique has been widely used in biomolecular simulation analyses, with applications that include clustering of conformations[32–35] for the analysis of pathways in enhanced sampling MD simulations.[36–38] For these purposes, we previously released PathDetect-SOM,[39,40] a tool based on SOMs, that was here applied to a set of unbiased simulations of the SH2:pYEEI complex.[23] The SOM algorithm starts with the choice of a set of features that describes each data point (here a set of distances for each frame). Then, the map is initialized and trained with the input vectors containing the values of the selected features for all the simulation frames. Each frame is considered as a data point and assigned to the neuron with most similar feature values. During the training process, the feature values of a neuron and its neighbors are adjusted toward the values of the input vector assigned to that neuron. This process continues over multiple cycles to achieve an accurate low-dimensional representation of the data. The final prototype vector of each output neuron summarizes the conformations associated with the neuron, and groups of similar conformations are mapped to neighboring neurons. In this work, the training was performed over 5000 cycles using a 20 × 20 sheet-shaped SOM with a hexagonal lattice shape and without periodicity across the boundaries. After training, each frame of the simulation set is assigned to a neuron on the map (hexagons), and each neuron represents a geometric microstate of the complex. In a second step, the neurons are further grouped in an optimal number of clusters by agglomerative hierarchical clustering, using Euclidean distances and complete linkage. The optimal number of clusters was selected according to the first maximum in the silhouette profile ranging in a reasonable interval (5–20 clusters). Properties such as the ligand RMSD from the x-ray structure[41] were displayed assigning to the neuron a color code proportional to the average value of the property for the frames belonging to that neuron.

Network and connectivity analysis

An approximate transition matrix was estimated by counting the transitions between each pair of neurons in all the simulations. A graph was then built with nodes represented by neurons, and edges were set to the negative logarithm of the transition probability between the corresponding neurons. For the sake of representation, transitions with fewer than ten counts were not represented in the graph. For the analysis that makes use of the transition probability matrix, the whole data were instead used. The distance between two nodes in the graph was calculated along the shortest path connecting them as the negative logarithm of the product of the pairwise transition probabilities between neurons along the path. A committor analysis was also performed computing the probability of hitting a set of states A before set B, starting from different initial states. In this case, the two extremes were the bound and unbound states, as detailed in the Results section. To validate the obtained results, we employed a progressive bootstrap analysis to estimate the average standard deviation on the computed committor, as a function of the number of replicas used. We randomly selected subsets of replicates of increasing size. For each subset, we performed 100 bootstrap resamples re-computing the transition matrix and estimating the average standard deviation of the committor computed on each neuron. Moreover, a bootstrap analysis was performed using 2/3 of the replicas, and the committor probability values were then obtained.

This process was repeated 250 times, allowing the calculation of the average and the standard deviation on the obtained values. This was used to show the average value of the committor probability for each neuron and the standard deviation obtained from the bootstrap. All the analyses were performed in the R statistical environment using the kohonen,[42,43] igraph,[44] and markovchain[45] packages.

Estimation of binding kinetics on-rates from trajectories

The kinetics of the SH2:pYEEI association can be obtained from the distribution of times taken to reach the bound state. As a first approximation, one can model the process as a two-state irreversible transition between the unbound and bound states; this model enables the estimation of the binding rate, assumed constant in time, from the ratio of binding events over the time sampled in the unbound state, after normalizing by the effective concentration.[23] Confidence intervals for the incidence rate (also known as Poisson rate exact confidence intervals) are then provided by Ulm's formula[46] (see the supplementary material).

MSM models of molecular systems rely on partitioning the conformational space into distinct states and estimate the transition probabilities between these states from simulated trajectories.[47–49] Unbiased MD generates the necessary data for constructing a MSM, enabling a detailed understanding of the underlying kinetics of ligand–target interactions. The transition probability matrix can then be used to estimate thermodynamic (e.g., asymptotic state probabilities) and kinetic (e.g., rates) properties. The success of a Markovian description is sensitive to the precise partitioning of the state space: different partitions will be closer or further from Markovianity,[50] thus providing better or worse extrapolations on the long timescales. We therefore built a set of Markov models based on the states and communities identified by the SOM mapping procedure using the transition count estimators provided by the Deeptime Python library.[51]

The SH2–phosphopeptide system as a high-specificity recognition model

The crystallographic structure of SH2 has been first reported by Waksman,[52] who showed (albeit in a static structure) a complex two-pronged binding mechanism, likely occurring as a consequence of electrostatic steering at the N end of the peptide (namely pTyr), as well as induced fit in a hydrophobic region holding the C end. In a previous study, some of us[23] analyzed the dynamics of the two-pronged binding mechanism of the pYEEI peptide to the p56lck SH2 domain through multiple parallel unbiased MD simulations, from which only five spontaneous binding events could be recovered.[23] The system has been modeled based on the crystal structure of human p56lck tyrosine kinase SH2 in complex with the pYEEI phosphopeptide at 1 Å resolution by Tong *et al.* (PDB: 1LKK[41]), where ligand was displaced by 40 Å to obtain an unbound starting conformation. The system was parameterized with the CHARMM27 force field. Water molecules present in the crystal structure were retained, and the system was solvated in TIP3P explicit water and 150 mM NaCl, leaving a buffer of 52 Å of water around the protein in the direction of the ligand, and at least 12 Å in the other directions. Equilibration included 10 ns of simulation in the constant-pressure ensemble (Berendsen thermostat) followed by 20 ns in the constant-volume ensemble. The ligand was prevented from diffusing during equilibration by 1 kcal/mol/Å^2 harmonic restraints applied to its Cα atoms.[53] The resulting simulation box was 60 × 66 × 98 Å^3, with a smaller 40 × 40 × 60 Å^3 flat-bottom box restraining the center of mass of the ligand to a 20 mM effective concentration. The equilibrated systems were finally simulated with the ACEMD software on the GPUGRID.net distributed simulation network[53] in multiple replicas at 295 K, with the particle-mesh Ewald treatment of long-range electrostatics. Further details of the simulation protocol are as previously reported.[23] In this work, to build a statistically relevant structural decomposition of the process, we produced a dataset widely extending the previous analysis; the extended dataset built for this paper consists of 772 unbiased SH2:pYEEI MD trajectories, almost all 800 ns long (distribution in the supplementary material), with snapshots taken every 1 ns. The new dataset contains a total of 640 μs, extending the previous sampling over fourfold and enabling the use of statistical and graph-theoretical methods over the SOM map. The extended dataset is publicly available in full in Zenodo (see "Data Availability").

RESULTS

SOM clustering of unbiased trajectories

The trained SOM is represented in Fig. 1, where each neuron (hexagon) corresponds to a configurational microstate, i.e., a peptide binding mode defined by specific values of the intermolecular distances used as input features. Neurons close to each other represent similar configurations. The map depicts a distribution of states ranging from the unbound state (bottom left of the map) to the crystallographic-like bound state (top center) or alternative bound states (top right). This is due to the sampling of the binding process, providing sufficient representation to the macroscopically meaningful states, which are then captured during the training process. Hierarchical clustering grouped the 400 neurons into 11 clusters (represented with different colors on the map and labeled as A–K) that coarsely represent the binding geometries explored by the system during the simulations (Fig. 1); one representative conformation from each cluster is shown to provide an overview of the corresponding macrostate. It is possible to identify a cluster representative of the unbound state, namely cluster A (blue in Fig. 1); a series of clusters (B, C, D, F, G, H, J, K) describing the possible pre-bound states in which the peptide begins the first contacts with the protein; and two different bound states contained in clusters E and I (purple and yellow in Fig. 1, respectively). Cluster E is characterized by conformations in which the ligand binds like in the crystallographic geometry, while cluster I represents an alternative bound state in which the ligand is rotated by 180° with respect to the x-ray geometry.[41]

Binding pathway analysis

Tracing the pathway followed by each trajectory on the SOM reveals a remarkable heterogeneity as each of them evolves following a different sequence of clusters. This is expected because, unlike simulations performed with MD methods in which a bias guides the

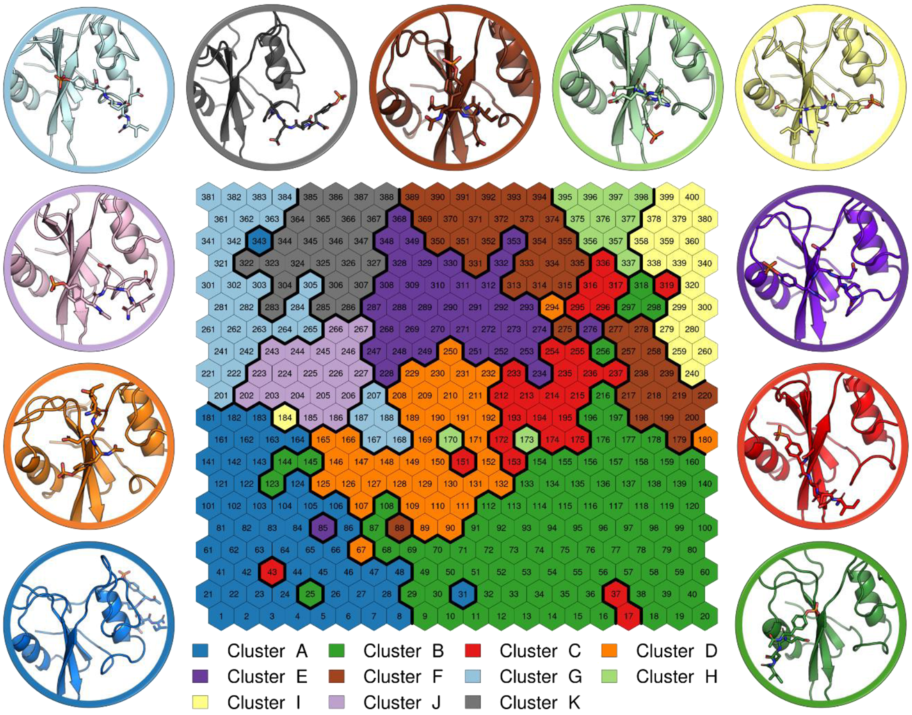


FIG. 1. SOM clustering of unbiased simulations of pYEEI binding to the SH2 domain. The representative conformation of each cluster is depicted in cartoons with ligand in sticks.

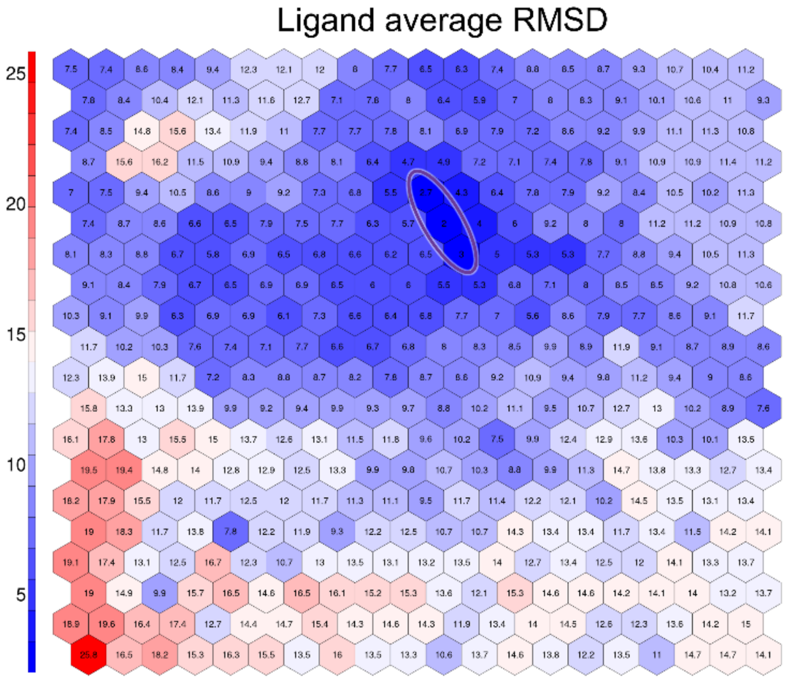


FIG. 2. SOM colored according to ligand average RMSD with respect to the bound state (values increasing from blue to red). States 272, 291, and 311 (circled) have RMSD ≤ 3 Å.

system along a selected collective variable (CV), unbiased simulations evolve without following a specific direction and, in general, explore transient pockets and kinetic traps.[4,47]

All simulations start from the unbound state (cluster A), but only a minority reaches (or ends in) the crystallographic-like bound state (cluster E) within the sampled time. Plotting the average ligand RMSD values from the x-ray structure of the frames belonging to each neuron on the SOM (Fig. 2), the lowest values are found for neurons 311, 291, and 272 (purple circle in Fig. 2), within cluster E, which we hence assume as the proper bound state. Twenty-two out of the 772 trajectories (2.8%) end in one of the three states. Using the three states as the definition of the bound state yields a k_{on} of 3.5×10^6 s^{-1} M^{-1} (95% CI: 2.5–4.8 $\times 10^6$ s^{-1} M^{-1}).

The binding process as a transition network

The transition network analysis allowed us to summarize all the sampled pathways in a graph (Fig. 3). All the simulations start from the unbound state (nodes in blue). Reading the transition network from the top, in yellow one finds nodes that lead to the alternative bound state in which the ligand is rotated 180° with respect to the x-ray geometry. Moving downward from the unbound state toward the right side of the graph, it is possible to observe the pathways that lead to the crystallographic-like bound state (purple nodes) and, in particular, toward the three neurons (311, 291, and 272) with the lowest ligand RMSD values from the x-ray structure.

Community detection identifies kinetically connected and transition states

From the transition network, by analyzing the number of transitions between nodes, it is possible to detect "highly connected communities."[54] Community analysis can be interpreted as a form of kinetic clustering, independent of any information about the bound and unbound states.

While geometric clustering yields information about conformational similarities by grouping together neurons having resembling features, community detection provides an overview of the kinetic relationships of the system. Community-based clustering

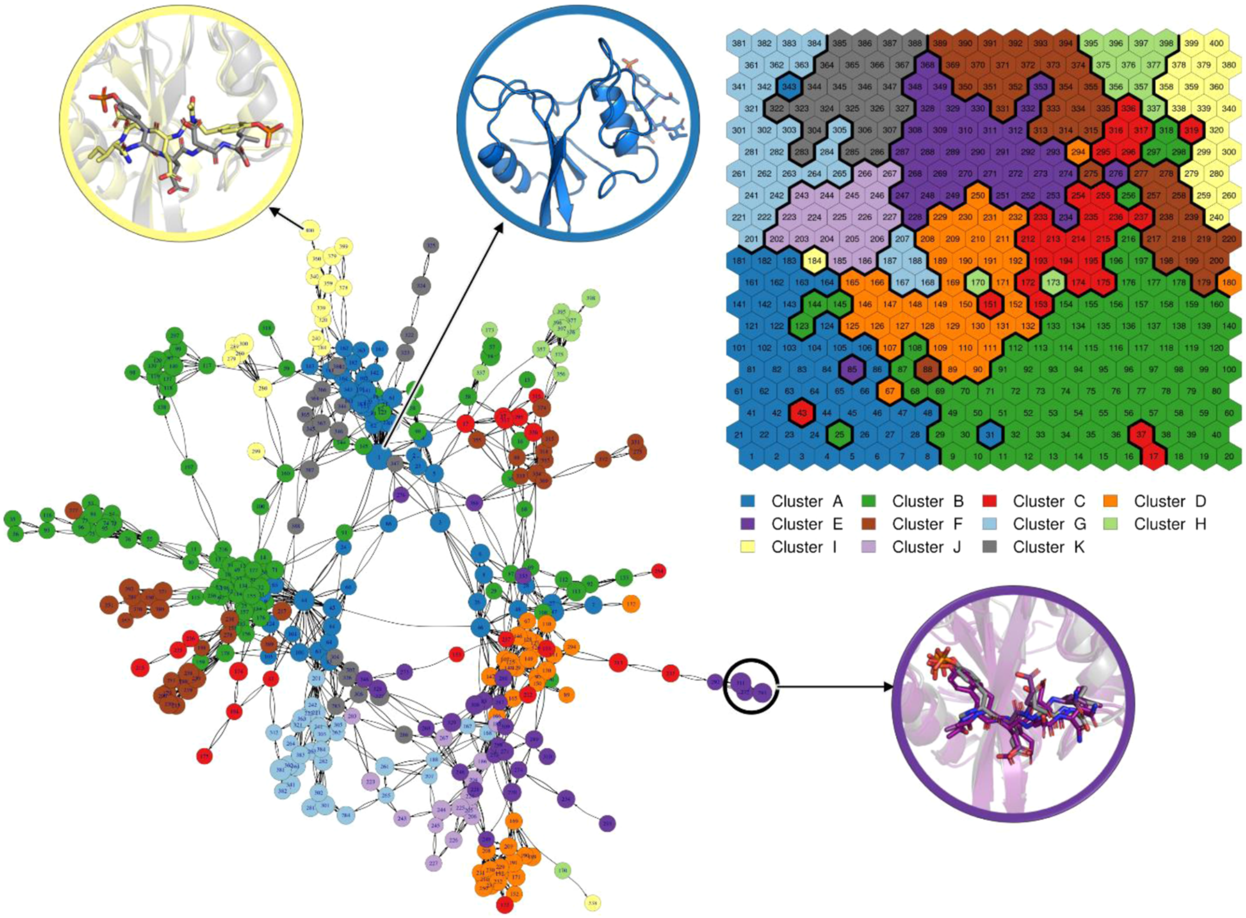


FIG. 3. Transition network analysis: transition network with nodes colored according to the SOM clusters. Representative conformations of neurons that characterize the unbound, crystallographic-like bound, and alternative bound states are represented within circles using blue, magenta, and yellow, respectively, for the cartoons and the ligand carbon atoms; the cartoons and the ligand carbon atoms of the experimental structure (PDB ID 1LKK[41]) are shown in gray.

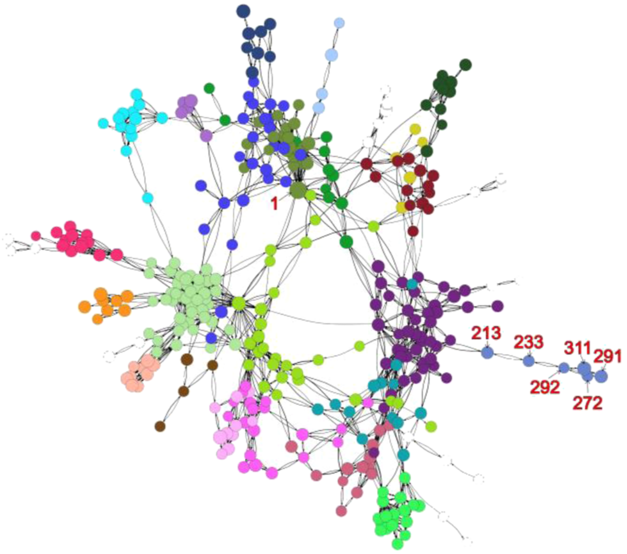


FIG. 4. Transition network colored according to the detected communities. Community detection identifies nodes (neurons) corresponding to the bound (272, 291, and 311) and transition (233) states, highlighted in steel blue. Nodes with insufficient transitions to be correctly assigned to a community are grayed out.

based on the Walktrap algorithm was able to separate the nodes (neurons) belonging to the bound state from the surrounding ones, which the geometric clustering instead clumps together. This is evident by comparing the purple region (cluster E) in Fig. 1, with the one highlighted in Fig. 4, both corresponding to the crystallographic-like bound state in the two different clustering. Cluster E is significantly larger, including many more neurons than the community indicated in steel blue in Fig. 4, which encompasses a group of only six nodes. In addition, community analysis identifies the transition state (neuron 233) and places it in the same community as the bound state. Notably, a committor analysis (discussed in more detail later in the text) confirms that neuron 233 is the closest to the iso-committor surface (committor value of 0.43) and, therefore, a likely representative of the transition state. The presence of a single node in the transition state may lead to the interpretation that this binding process can be indeed modeled as a two-state process with a downhill pathway after a single transition saddle point.

Shortest-path analysis shows the prototypical binding pathway

To better understand the steps involved in the pYEEI peptide binding to the SH2 domain, we computed the sub-optimal (yellow in Fig. 5) and the shortest (orange in Fig. 5) pathways linking the unbound state (neuron 1) with the crystallographic-like bound state (neuron 291). The shortest path is defined as the one that minimizes the cumulative sum of edge weights between the starting node and the ending node.

Furthermore, we compared the previously proposed mechanism[23] for peptide binding with the present proposal based on the analysis of the shortest path. As shown in Fig. 6, initially the peptide is in the bulk [panel (a)]. The first step is characterized by the initial contact of the pY group of the peptide with the protein, which involves the formation of two salt bridges with residues R154 and K182 [panel (b)]. This interaction is further stabilized in the subsequent step by the formation of a hydrogen-bond network with residues S156, S158, and A160, which traps pY in the bound state [panel (c)]. The following step involves the contact of the second E residue through a salt bridge with R184 [panel (d)]. Next, the

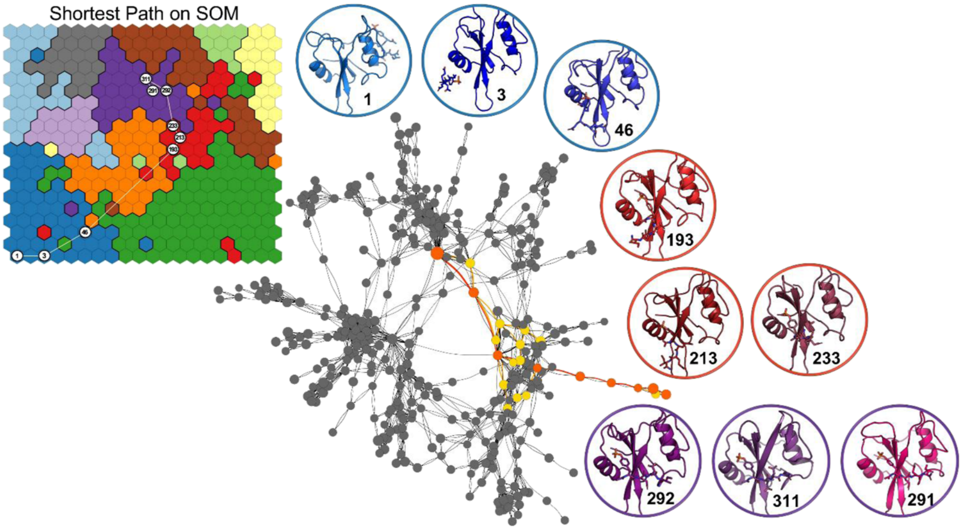


FIG. 5. Pathway analysis on the SOM and in the transition network. All sub-optimal pathways are plotted in the transition network with nodes colored in yellow. The shortest path is mapped on the graph with orange nodes and marked on the SOM with white circles. The representative conformations of neurons that characterize the shortest path are depicted in cartoons with the ligand in sticks.

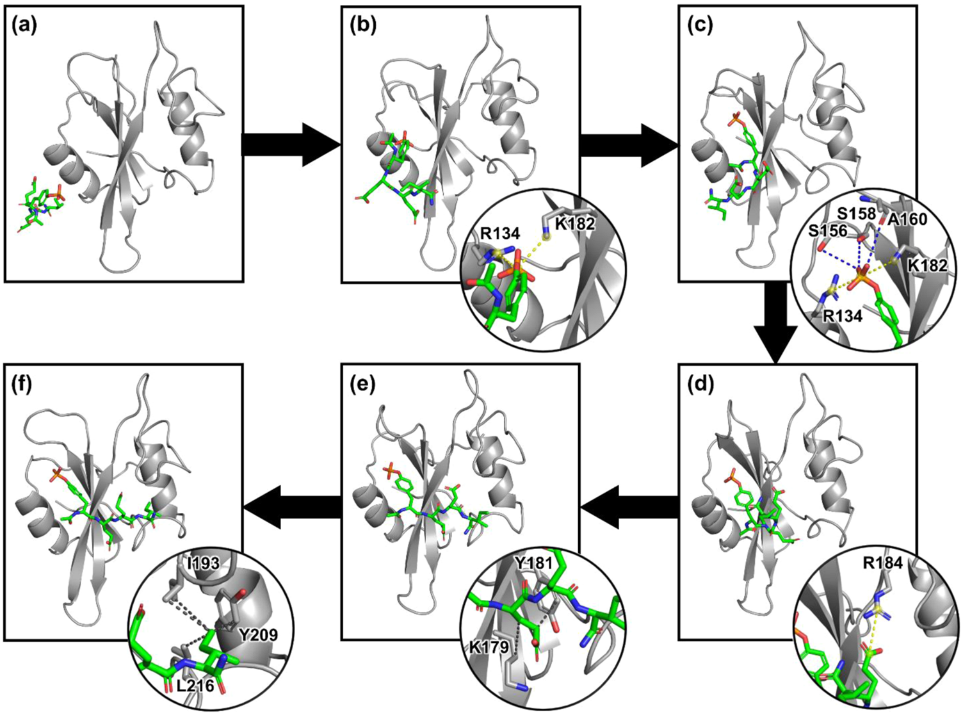


FIG. 6. Insight into the peptide binding mechanism: schematic representation of the different steps: (a) unbound, (b) initial contact, (c) anchoring of pTyr, (d) salt bridge of the second Glu residue with R184, (e) positioning of the second Glu residue, and (f) positioning of the C-terminal residue of the peptide between the EF and BG loops. Cartoon and gray sticks represent the protein and the residues involved in the interaction with the pYEEI peptide, represented in green stick. Yellow dashes represent the salt bridges, blue dashes represent the hydrogen bonds, and dark gray dashes represent the hydrophobic interactions.

first E residue of the peptide stabilizes in the bound position, forming a hydrophobic interaction with both K179 and Y181 [panel (e)]. The final step for peptide binding involves positioning of the I residue between the EF and BG loops, favored by a network of hydrophobic interactions with residues I193, Y209, and L216 [panel (f)]. Based on the above discussion, the two mechanisms appear to be consistent.

Bound-state heterogeneity

All simulations reaching the bound state showed that the bound geometry is characterized by different conformations well represented by neurons 291 and 311, which exhibit similar and very low ligand RMSD values to the x-ray structure. Interaction characteristics of the two microstates are slightly different: neuron 311 shows a salt bridge with K179, that is lacking in neuron 291, and neuron 291 presents a more extended network of hydrophobic interactions involving the EF and BG loops (as shown in Fig. S2). The fact that different simulations reach bound states with slightly different conformations is expected, as individual replicas may explore distinct, yet closely related, conformational states that, however, share the same key interactions. Upon plotting the number of simulations that ended in each neuron on the SOM (left-hand panel of Fig. 7), we observed that, after neuron 1 (representing the unbound state), neuron 311 exhibited the highest occurrence, capturing the largest number of replicas. This analysis also revealed that neurons 381 and 400 (at the top left and right vertices of the SOM), neuron 351 (just above neuron 311), and neuron 55 (at the bottom right corner) have 8 to 10 simulations ending in them, suggesting that they could be kinetic traps. The section on "Markovian modeling of the binding process" will provide estimates of the asymptotic (equilibrium) probabilities of each state computed through Markov state models.

Committor analysis identifies the transition state

We conducted a committor analysis (right-hand panel of Fig. 7) to calculate the probability of the system to access the bound conformation (neuron 311) before reaching the unbound conformation (neuron 1), starting from each neuron. The convergence of the calculation was assessed by evaluating the average standard deviation of the computed committors through a progressive bootstrap analysis (see the Methods section for further details and Fig. S3). The analysis revealed that once the number of replicas reaches ~500, the average standard deviation of the committor probabilities stabilizes, confirming the convergence of the calculation. As the transition matrix was constructed from unbiased simulations, conformations with a committor of around 0.50 can be considered in proximity of the transition states. In this case, the energy barrier appears to be located around neuron 233 (committor value 0.43), in the conformations of which the pY group of the peptide is fixed to the protein and

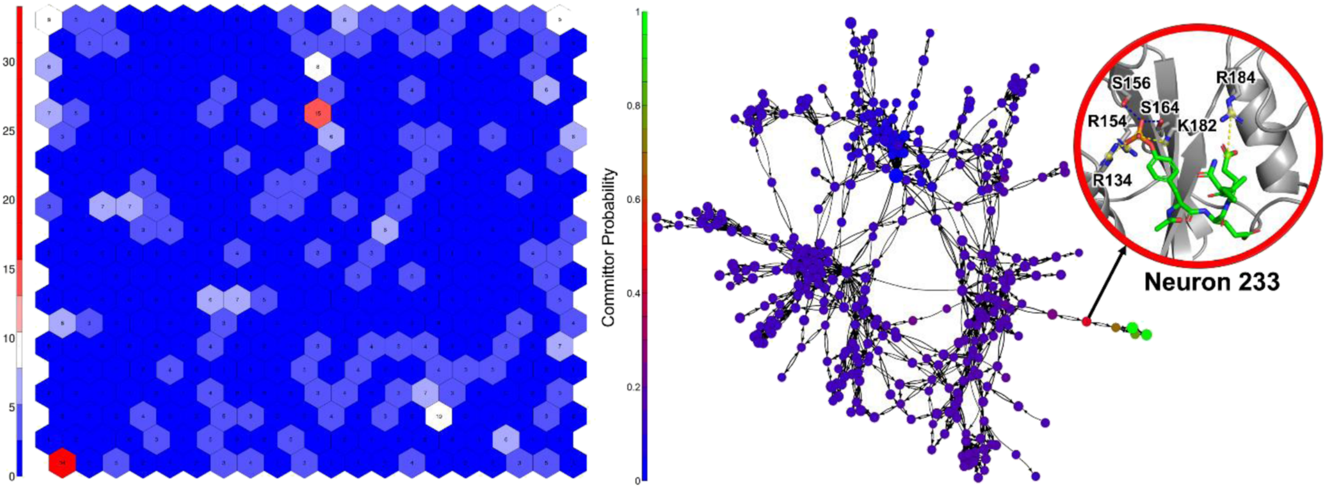


FIG. 7. On the left, SOM is annotated with the number of simulations ending in each neuron (values increasing from blue to red); on the right, graph network with nodes colored according to the committor probability analysis (increasing values from blue to green). The representative conformation of neuron 233 is depicted in cartoons with the ligand and the residues involved in the interactions in green and gray sticks, respectively. To verify the robustness of our model and validate the previously discussed results, we conducted a bootstrap analysis (see the Methods section and Fig. S4).

a salt bridge is formed between the second E residue of the peptide and R184.

The barrier appears to be due to the probability that the collision between the two molecules occurs with the correct orientation, allowing the first salt-bridge to form properly and aligning the rest of the molecule correctly to form the second salt-bridge. For the pYEEI peptide binding to the SH2 domain, the initial encounter must position the tyrosine-phosphate (pY) group to form the first salt bridge with residues R154 and K182 on the protein surface. This initial contact is crucial, as it stabilizes the initial complex and orients the peptide for subsequent binding steps. If the peptide collides in an incorrect orientation, the necessary interactions will not form, and the peptide may dissociate or bind in a less favorable mode. This requirement for precise orientation introduces an entropic cost. The system must overcome the disorder associated with the numerous possible orientations during the peptide's diffusion. The correct alignment is less probable, contributing to the observed transition state at neuron 233, identified by a committor analysis. This state marks a critical point where the probability of progressing to the bound state equals that of returning to the unbound state, indicating a significant entropic barrier.

Communities provide a robust basis to compute the binding kinetics

The computation of the binding kinetics based on atomistic data is sensitive to the precise definition of what microstates constitute "the bound state." In this respect, the community analysis graph provides a very robust approach for the computation of kinetics. First, we assume a threshold of ε = 0.5 for the committor value c, and we consider all the states whose committor is $c_i \geq \varepsilon$ as the bound state (otherwise being unbound). This yields 52 binding events and therefore a k_{on} rate of 4.4 (95% CI: 3.3–5.8) $\times 10^6$ s^{-1} M^{-1}. Taking the transition state at $c = 0.5$ is a common assumption; however, it is interesting to do a sensitivity analysis to study how strongly k_{on} depends on the precise value of the threshold. We again computed k_{on} as the event rate, i.e., the ratio of the number of binding events observed to the total unbound sampled time, now as a function of ε (Fig. 8). The rates are insensitive to the precise choice of ε, being essentially constant as for the canonical ε = 0.5 in a wide range of values, thus again confirming that the community analysis projects the complex process into a good reaction coordinate (albeit discretized) with two attraction basins separated by a well-defined transition region.

This contrasts, for example, with an RMSD-based definition computing the similarity of the configuration of the ligand in the final frame of each simulation to the (assumed known) x-ray structure. Such a definition would imply that, for example, 24 simulations (≈3%) end in a frame with RMSD ≤ 2 Å, while 143 simulations (≈18%) end in a state with RMSD ≤ 5 Å. By taking an RMSD threshold between 1.8 and 3.0 Å as a definition of the bound state, the sensitivity analysis of k_{on} identifies binding rates that are relatively constant (Fig. S5 of the supplementary material). Using the larger plateau region between 1.8 and 2.4 Å, one obtains a k_{on} rate of 2.0 (95% CI: 1.3–3.0) $\times 10^6$ s^{-1} M^{-1}. The RMSD-based approach to the definition of the bound state, however, has the drawback that the k_{on} value depends strongly on both (1) the precise knowledge of the bound structure and (2) the RMSD threshold chosen. Other geometric observables could, in principle, be similarly used to define the kinetics, e.g., the fraction of native contacts provides a similar value for k_{on}—albeit with a plateau too narrow to identify reliably (Fig. S6 of the supplementary material). To assess the impact of the simulation size on the estimated k_{on}, we calculated k_{on} as a function of replica length (Fig. S7, top) and the number of replicas (Fig. S7, bottom). The results show that after ~350 ns per replica (total aggregate simulation time of 280 μs), the estimated k_{on} reaches a plateau, indicating the convergence of the calculation.

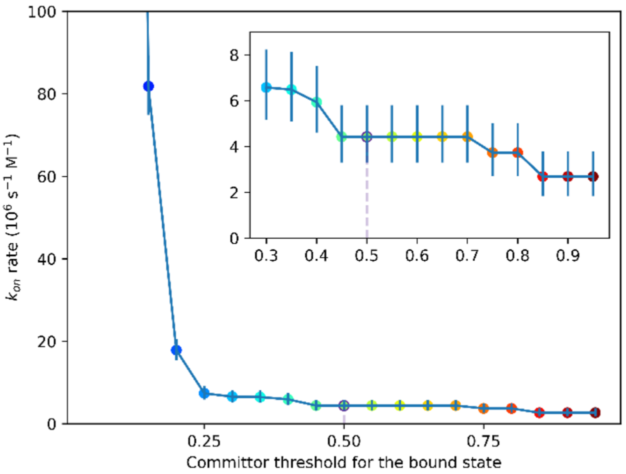


FIG. 8. Sensitivity analysis of the on-rate of the SH2:pYEEI binding process as a function of the committor value used to define the bound state. The community-based projection makes the k_{on} rate remarkably robust to the definition of the bound state (committor threshold, on the horizontal axis; the iso-committor value of 0.5 is highlighted).

Markovian modeling of the binding process

The SOM-based discretization lends itself to be a basis for a Markovian modeling of the association process. An MSM analysis can be conducted either discretizing the state space according to the SOM neurons, or according to the communities previously identified by the community analysis; in both cases, the implied timescales are quite similar (Fig. S8), pointing to relaxation processes on the order of 10 and 2 μs, respectively, around a lag time of 300 ns. MSM defined based on communities displays convergence at slightly lower lag times, despite the coarser states.

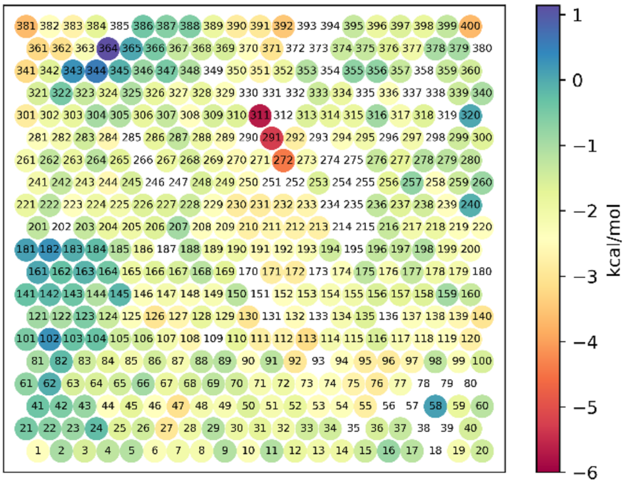


FIG. 9. Standard free energy of each SOM neuron, obtained by Boltzmann inversion of the equilibrium probabilities estimated Markov-modeling the binding process at a lag time of 350 ns. Neurons of the bound state (272, 291, and 311) have the lowest free energy values with a total equilibrium probability of 0.54; neurons not belonging to any community are left blank. The zero of the free energy is assigned to the state unbound at 1M; it was set applying the concentration correction (−2.4 kcal/mol) to state 1, namely unbound at simulation concentration.

We used the MSMs to extrapolate the asymptotic (equilibrium) state probabilities. Equilibrium probabilities have a distinct peak in the 272, 291, and 311 neuron triplet, clearly identifying the bound state. The three states collectively account for 54% of the equilibrium probability distribution. Boltzmann inversion of the equilibrium probabilities provides free energy values. Minor free energy minima are at neuron 400 (reverse-bound, cluster I), neuron 381 ("vertically bound," cluster G), neuron 392 ("L-shaped," cluster F), and few other likely metastable configurations previously identified and discussed in the SOM.

Finally, even though this is not the main objective of the analysis, equilibrium probabilities can be converted into standard free energies of binding by Boltzmann inversion and accounting for the ligand concentration vs the standard state. Figure 9 shows a free energy landscape of the binding process mapped on the SOM neurons, reconstructed by building a MSM on the whole set of unbiased trajectories, which clearly identifies the bound triplet.

CONCLUSIONS

In this work, for the first time, we applied SOM to an extensive dataset of SH2–pYEEI binding trajectories obtained from unbiased MD simulations. This approach provides valuable insights into the evolution of the system by revealing geometric clusters. Modeling the binding process as a transition network enabled the identification of key kinetic stages such as the transition state, the bound state, and potential kinetic traps. A simple two-state treatment of the reaction process projected on the SOM discretization yields association rates in agreement with the experimental values. The combination of Markov models with SOM discretization yielded a robust free energy landscape, clearly pinpointing the bound state as the most energetically favorable configuration. This combined framework provides an accurate description of complex processes characterized by heterogeneous pathways. The results of this method are especially notable as they account for the remarkable flexibility of the tetrapeptide and the proper accounting of induced fit effects on the receptor.

A key advantage of using SOM is its ability to preserve the topological relationships between microstates, which enhances the interpretability of the conformational landscape compared to other traditional clustering methods. Other dimensionality reduction methods, however, exist, such as Principal Component Analysis (PCA). PCA reduces dimensionality by focusing on the variance captured by the first few principal components, but it can overlook subtle yet important features of the data. In contrast, SOMs create a map where, at the same time, data are grouped in microstates and similar data are spatially close, providing a more intuitive and detailed representation. This allows for the detection of geometric patterns and the subsequent construction of a transition network that accurately captures the kinetics of the binding process. The SOM-based approach not only facilitates the identification of key states but also offers a versatile platform for further analysis, such as mapping external properties or performing kinetic clustering, which PCA alone does not inherently provide. Other approaches for pathway detection in MD simulations exist such as the one presented in

Ref. 55. However, these methods are not intended to cluster frames and obtain microstates that form the unique platform offered by SOM for the subsequent analysis.

Despite the significant advantages of SOM, it is not without limitations. In this particularly challenging case, one major restriction is that—due to the inherent flexibility of both the protein and peptide, combined with the limited number of simulations reaching the bound state—we employed a set of "non-blind" distances to enhance the description of the correct bound state. While SOM can operate in a "blind" mode using all intermolecular distances, this approach tends to treat all sampled states equally. However, for this complex scenario, a more selective approach was necessary. In addition, a careful interpretation of the free energy values obtained from the MSM analysis is needed for two key reasons. First, each neuron in the model represents a subset of the phase space of different volume. Second, the equilibrium probabilities identified by MSM are likely underestimated due to insufficient sampling near the bound state, which can be addressed through adaptive sampling schemes.

An additional limitation of the approach presented is the extensive sampling required to obtain reliable statistics on the binding process. While unbiased MD simulations are straightforward, requiring no collective variable definitions and avoiding external biases that could alter the physics of the process, they demand a substantial amount of simulation time, which may be prohibitive for large-scale drug design campaigns. However, the continuous optimization of MD software and the increasing computational power have led to orders-of-magnitude increases in the feasible simulation lengths over the past decade. This trend is expected to continue with advancements in artificial intelligence and quantum computing, making studies like the one presented here more routine in the future. In the context of a drug design campaign, it remains challenging to study a large number of ligands using this approach due to its high computational cost. Nevertheless, analyzing one or a few highly promising ligands could be valuable in identifying the key characteristics that make a ligand attractive or in uncovering potential barriers during the binding process that could be mitigated through proper functionalization, ultimately enhancing the ligand's efficacy.

Finally, while the data-driven approach we presented here has promising implications in drug design and optimization tasks, for pharmaceutical purposes, one may want to investigate the off-rate, which has an important correlation with the biological activity.[56–58] Sampling k_{off} (or residence times) with purely unbiased approaches like the one presented in this paper is much more challenging than k_{on} because of the far longer timescales involved. While feasible for shorter residence times (hundreds of microseconds are becoming accessible),[59,60] biased sampling approaches such as (infrequent) metadynamics,[61] adaptive sampling Markov models,[62] weighted ensemble,[63] and similar may still be approaches of preference for this task. Future work will focus on characterizing and reducing the sensitivity of the method to the *a priori* knowledge of the bound pose, thereby broadening its applicability to even more challenging "blind pathway reconstruction" tasks able to simultaneously recover pose, binding paths, and kinetics. This could be achieved, for example, through the integration of SOM with an adaptive seeding technique to enrich the number of transitions between neurons to improve the robustness and accuracy of the analysis.

SUPPLEMENTARY MATERIAL

The supplementary material contains additional information and analyses supporting the main text. It includes figures detailing the selected atoms for interatomic distance calculation and used for SOM training (Fig. S1), the three-dimensional structures of representative conformations for neurons 291 and 311 with annotated interactions (Fig. S2), the value of the average standard deviation of the computed committors in a progressive bootstrap analysis (Fig. S3), and the average values and standard deviation for committor probabilities obtained from bootstrap analysis (Fig. S4). Sensitivity analyses of the on-rate of the SH2 binding process are provided, examining the influence of the RMSD threshold value (Fig. S5) and the fraction of crystallographic native contacts (Fig. S6). Analysis of the *on* rate of the SH2:pYEEI binding process as a function of the simulation time and the number of replicas (Fig. S7). Implied timescales for a Markov state model built on SOM neurons and communities, showing convergence and relaxation processes, are also included (Fig. S8). The supplementary material details the Poisson rate and confidence interval calculations, providing the formula for estimating incidence rates in a two-state irreversible model, along with the corresponding confidence intervals using Ulm's formula (Poisson rate confidence interval).

ACKNOWLEDGMENTS

T.G. acknowledges financial support from the Spoke 7 of the National Centre for HPC, Big Data and Quantum Computing (Centro Nazionale 01 – CN0000013), funded by the European Union – NextGenerationEU, Mission 4, Component 2, Investment line 1.4, CUP B93C22000620006; from the PRIN 2022 (BioCat4BioPol) from the Ministero dell'Università e Ricerca, funded by the European Union – NextGenerationEU, Mission 4 Component C2, CUP B53D23015140006; and from the project InvAt–Invecchiamento Attivo e in Salute (FOE 2022) CUP B53C22010140001. T.G. thanks the volunteers of GPUGRID.net for donating computing time for the simulations. We acknowledge CINECA awards under the ISCRA initiative and the agreement with the University of Milano-Bicocca, for the availability of high-performance computing resources and support.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Lara Callea: Data curation (lead); Formal analysis (equal); Investigation (lead); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Camilla Caprai**: Visualization (equal); Writing – review & editing (equal). **Laura Bonati**: Conceptualization (equal); Supervision (equal); Writing – review & editing (equal). **Toni Giorgino**: Conceptualization (equal); Formal analysis (equal); Supervision (equal); Writing – review & editing (equal). **Stefano Motta**: Conceptualization (equal); Methodology (equal); Supervision (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The full dataset supporting the findings of this study is openly available in Zenodo at http://doi.org/10.5281/zenodo.12205888.

REFERENCES

[1] R. E. Amaro and A. J. Mulholland, "Multiscale methods in drug design bridge chemical and biological complexity in the search for cures," Nat. Rev. Chem. **2**(4), 0148 (2018).

[2] A. Basciu, L. Callea, S. Motta, A. M. J. J. Bonvin, L. Bonati, and A. V. Vargiu, "No dance, no partner! A tale of receptor flexibility in docking and virtual screening," Annu. Rep. Med. Chem. **59**, 43–97 (2022).

[3] E. Brini, C. Simmerling, and K. Dill, "Protein storytelling through physics," Science **370**(6520), eaaz3041 (2020).

[4] Y. Shan, E. T. Kim, M. P. Eastwood, R. O. Dror, M. A. Seeliger, and D. E. Shaw, "How does a drug molecule find its target binding site?," J. Am. Chem. Soc. **133**(24), 9181–9183 (2011).

[5] S. A. Hollingsworth and R. O. Dror, "Molecular dynamics simulation for all," Neuron **99**(6), 1129–1143 (2018).

[6] M. De Vivo, M. Masetti, G. Bottegoni, and A. Cavalli, "Role of molecular dynamics and related methods in drug discovery," J. Med. Chem. **59**, 4035–4061 (2016).

[7] V. Limongelli, "Ligand binding free energy and kinetics calculation in 2020," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **10**(4), e1455 (2020).

[8] R. W. Zwanzig, "High-temperature equation of state by a perturbation method. I. Nonpolar gases," J. Chem. Phys. **22**(8), 1420–1426 (1954).

[9] B. Isralewitz, M. Gao, and K. Schulten, "Steered molecular dynamics and mechanical functions of proteins," Curr. Opin. Struct. Biol. **11**(2), 224–230 (2001).

[10] T. Giorgino and G. De Fabritiis, "A high-throughput steered molecular dynamics study on the free energy profile of ion permeation through gramicidin A," J. Chem. Theory Comput. **7**(6), 1943–1950 (2011).

[11] A. Laio and M. Parrinello, "Escaping free-energy minima," Proc. Natl. Acad. Sci. U. S. A. **99**, 12562–12566 (2002).

[12] The PLUMED consortium, "Promoting transparency and reproducibility in enhanced molecular simulations," Nat. Methods **16**(8), 670–673 (2019).

[13] L. Callea, L. Bonati, and S. Motta, "Metadynamics-based approaches for modeling the hypoxia-inducible factor 2α ligand binding process," J. Chem. Theory Comput. **17**(7), 3841–3851 (2021).

[14] Y. Miao, A. Bhattarai, and J. Wang, "Ligand Gaussian accelerated molecular dynamics (LiGaMD): Characterization of ligand binding thermodynamics and kinetics," J. Chem. Theory Comput. **16**(9), 5526–5547 (2020).

[15] D. Sabbadin and S. Moro, "Supervised molecular dynamics (SuMD) as a helpful tool to depict GPCR–ligand recognition pathway in a nanosecond time scale," J. Chem. Inf. Model. **54**(2), 372–376 (2014).

[16] D. B. Kokh, M. Amaral, J. Bomke, U. Grädler, D. Musil, H. P. Buchstaller, M. K. Dreyer, M. Frech, M. Lowinski, F. Vallee, M. Bianciotto, A. Rak, and R. C. Wade, "Estimation of drug-target residence times by τ-random acceleration molecular dynamics simulations," J. Chem. Theory Comput. **14**(7), 3859–3869 (2018).

[17] L. Mollica, S. Decherchi, S. R. Zia, R. Gaspari, A. Cavalli, and W. Rocchia, "Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations," Sci. Rep. **5**, 11539 (2015).

[18] A. Spitaleri, S. Decherchi, A. Cavalli, and W. Rocchia, "Fast dynamic docking guided by adaptive electrostatic bias: The MD-binding approach," J. Chem. Theory Comput. **14**(3), 1727–1736 (2018).

[19] S. Motta, L. Callea, S. Giani Tagliabue, and L. Bonati, "Exploring the PXR ligand binding mechanism with advanced Molecular Dynamics methods," Sci. Rep. **8**, 16207 (2018).

[20] J. Rydzewski, "maze: Heterogeneous ligand unbinding along transient protein tunnels," Comput. Phys. Commun. **247**, 106865 (2020).

[21] P. C. T. Souza, S. Thallmair, P. Conflitti, C. Ramírez-Palacios, R. Alessandri, S. Raniolo, V. Limongelli, and S. J. Marrink, "Protein–ligand binding with the coarse-grained Martini model," Nat. Commun. **11**(1), 3714 (2020).

[22] M. Ahmad, W. Gu, and V. Helms, "Mechanism of fast peptide recognition by SH3 domains," Angew. Chem., Int. Ed. **47**(40), 7626–7630 (2008).

[23] T. Giorgino, I. Buch, and G. De Fabritiis, "Visualizing the induced binding of SH2-phosphopeptide," J. Chem. Theory Comput. **8**(4), 1171–1175 (2012).

[24] J. D. Chodera and F. Noé, "Markov state models of biomolecular conformational dynamics," Curr. Opin. Struct. Biol. **25**, 135–144 (2014).

[25] B. E. Husic and V. S. Pande, "Markov state models: From an art to a science," J. Am. Chem. Soc. **140**(7), 2386–2396 (2018).

[26] P. Tiwary, V. Limongelli, M. Salvalaglio, and M. Parrinello, "Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps," Proc. Natl. Acad. Sci. U. S. A. **112**(5), E386–E391 (2015).

[27] T. Kohonen, "Essentials of the self-organizing map," Neural Networks **37**, 52–65 (2013).

[28] G. Bouvier, N. Desdouits, M. Ferber, A. Blondel, and M. Nilges, "An automatic tool to analyze and cluster macromolecular conformations based on self-organizing maps," Bioinformatics **31**(9), 1490–1492 (2015).

[29] S. Zhou, "SH2 domains recognize specific phosphopeptide sequences," Cell **72**(5), 767–778 (1993).

[30] P. Morlacchi, F. M. Robertson, J. Klostergaard, and J. S. McMurray, "Targeting SH2 domains in breast cancer," Future Med. Chem. **6**(17), 1909–1926 (2014).

[31] T. Kohonen, "The self-organizing map," Proc. IEEE **78**, 1464–1480 (1990).

[32] G. Frigerio, E. Donadoni, P. Siani, J. Vertemara, S. Motta, L. Bonati, L. D. Gioia, and C. D. Valentin, "Mechanism of RGD-conjugated nanodevice binding to its target protein integrin $\alpha_V\beta_3$ by atomistic molecular dynamics and machine learning," Nanoscale **16**(8), 4063–4081 (2024).

[33] E. Donadoni, G. Frigerio, P. Siani, S. Motta, J. Vertemara, L. De Gioia, L. Bonati, and C. Di Valentin, "Molecular dynamics for the optimal design of functionalized nanodevices to target folate receptors on tumor cells," ACS Biomater. Sci. Eng. **9**, 6123–6137 (2023).

[34] S. Motta, P. Siani, E. Donadoni, G. Frigerio, L. Bonati, and C. Di Valentin, "Metadynamics simulations for the investigation of drug loading on functionalized inorganic nanoparticles," Nanoscale **15**(17), 7909–7919 (2023).

[35] S. Motta and L. Bonati, "TCDD-induced allosteric perturbation of the AhR:ARNT binding to DNA," Int. J. Mol. Sci. **24**(11), 9339 (2023).

[36] E. Hendrix, S. Motta, R. F. Gahl, and Y. He, "Insight into the initial stages of the folding process in onconase revealed by UNRES," J. Phys. Chem. B **126**(40), 7934–7942 (2022).

[37] T. Li, S. Motta, A. Stevens, S. Song, E. Hendrix, A. Pandini, and Y. He, "Recognizing the binding pattern and dissociation pathways of the p300 Taz2-p53 TAD2 complex," JACS Au **2**(8), 1935–1945 (2022).

[38] Rubina and S. T. Moin, "Attempting well-tempered funnel metadynamics simulations for the evaluation of the binding kinetics of methionine aminopeptidase-II inhibitors," J. Chem. Inf. Model. **63**(24), 7729–7743 (2023).

[39] S. Motta, A. Pandini, A. Fornili, and L. Bonati, "Reconstruction of ARNT PAS-B unfolding pathways by steered molecular dynamics and artificial neural networks," J. Chem. Theory Comput. **17**(4), 2080–2089 (2021).

[40] S. Motta, L. Callea, L. Bonati, and A. Pandini, "PathDetect-SOM: A neural network approach for the identification of pathways in ligand binding simulations," J. Chem. Theory Comput. **18**(3), 1957–1968 (2022).

[41] L. Tong, T. C. Warren, J. King, R. Betageri, J. Rose, and S. Jakes, "Crystal structures of the human p56lck SH2 domain in complex with two short phosphotyrosyl peptides at 1.0 Å and 1.8 Å resolution," J. Mol. Biol. **256**(3), 601–610 (1996).

[42] R. Wehrens and J. Kruisselbrink, "Flexible self-organizing maps in kohonen 3.0," J. Stat. Software **87**, 1–18 (2018).

[43] R. Wehrens and L. M. C. Buydens, "Self- and super-organizing maps in *R*: The kohonen package," J. Stat. Software **21**(5), 1 (2007).

[44] G. Csardi and T. Nepusz, "The igraph software package for complex network research," InterJ., Complex Syst. **1695**, 1–9 (2006).

[45] G. A. Spedicato, "Discrete time Markov chains with R," R J. **9**(2), 84 (2017).

[46] K. Ulm, "Simple method to calculate the confidence interval of a standardized mortality ration (SMR)," Am. J. Epidemiol. **131**(2), 373–375 (1990).

[47] I. Buch, T. Giorgino, and G. De Fabritiis, "Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations," Proc. Natl. Acad. Sci. U. S. A. **108**(25), 10184–10189 (2011).

[48] V. S. Pande, K. Beauchamp, and G. R. Bowman, "Everything you wanted to know about Markov State Models but were afraid to ask," Methods **52**(1), 99–105 (2010).

[49] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, "Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations," Proc. Natl. Acad. Sci. U. S. A. **106**(45), 19011–19016 (2009).

[50] F. Nüske, H. Wu, J.-H. Prinz, C. Wehmeyer, C. Clementi, and F. Noé, "Markov state models from short non-equilibrium simulations—Analysis and correction of estimation bias," J. Chem. Phys. **146**(9), 094104 (2017).

[51] M. Hoffmann, M. Scherer, T. Hempel, A. Mardt, B. de Silva, B. E. Husic, S. Klus, H. Wu, N. Kutz, S. L. Brunton, and F. Noé, "Deeptime: A Python library for machine learning dynamical models from time series data," Mach. Learn.: Sci. Technol. **3**(1), 015009 (2022).

[52] G. Waksman, "Binding of a high affinity phosphotyrosyl peptide to the Src SH2 domain: Crystal structures of the complexed and peptide-free forms," Cell **72**(5), 779–790 (1993).

[53] I. Buch, M. J. Harvey, T. Giorgino, D. P. Anderson, and G. De Fabritiis, "High-throughput all-atom molecular dynamics simulations using distributed computing," J. Chem. Inf. Model. **50**(3), 397–403 (2010).

[54] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," Proc. Natl. Acad. Sci. U. S. A. **101**(9), 2658–2663 (2004).

[55] J. Rydzewski and W. Nowak, "Machine learning based dimensionality reduction facilitates ligand diffusion paths assessment: A case of cytochrome P450cam," J. Chem. Theory Comput. **12**(4), 2110–2120 (2016).

[56] D. C. Swinney, "Biochemical mechanisms of drug action: What does it take for success?," Nat. Rev. Drug Discovery **3**(9), 801–808 (2004).

[57] G. Vauquelin, "Effects of target binding kinetics on *in vivo* drug efficacy: k_{off}, k_{on} and rebinding," Br. J. Pharmacol. **173**(15), 2319–2334 (2016).

[58] S. Decherchi and A. Cavalli, "Thermodynamics and kinetics of drug-target binding by molecular simulation," Chem. Rev. **120**(23), 12788–12833 (2020).

[59] G. Martìnez-Rosell, T. Giorgino, M. J. Harvey, and G. de Fabritiis, "Drug discovery and molecular dynamics: Methods, applications and perspective beyond the second timescale," Curr. Top. Med. Chem. **17**(23), 2617–2625 (2017).

[60] J. B. Greisman, L. Willmore, C. Y. Yeh, F. Giordanetto, S. Shahamadtar, H. Nisonoff, P. Maragakis, and D. E. Shaw, "Discovery and validation of the binding poses of allosteric fragment hits to protein tyrosine phosphatase 1b: From molecular dynamics simulations to x-ray crystallography," J. Chem. Inf. Model. **63**(9), 2644–2650 (2023).

[61] M. Salvalaglio, P. Tiwary, and M. Parrinello, "Assessing the reliability of the dynamics reconstructed from metadynamics," J. Chem. Theory Comput. **10**(4), 1420–1425 (2014).

[62] S. Doerr and G. De Fabritiis, "On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations," J. Chem. Theory Comput. **10**(5), 2064–2069 (2014).

[63] M. C. Zwier, J. L. Adelman, J. W. Kaus, A. J. Pratt, K. F. Wong, N. B. Rego, E. Suárez, S. Lettieri, D. W. Wang, M. Grabe, D. M. Zuckerman, and L. T. Chong, "WESTPA: An interoperable, highly scalable software package for weighted ensemble simulation and analysis," J. Chem. Theory Comput. **11**(2), 800–809 (2015).