Title

Kinetics-Based State Definitions for Discrete Binding Conformations of T4 L99A in MD via Markov State Modeling.

Journal

Journal of chemical information and computer sciences, 64(23)

Authors

Zhang, Chris

Osato, Meghan

Mobley, David

Publication Date

2024-12-09

DOI

10.1021/acs.jcim.4c01364

Peer reviewed

HHS Public Access

Author manuscript

J Chem Inf Model. Author manuscript; available in PMC 2025 February 11.

Published in final edited form as:

J Chem Inf Model. 2024 December 09; 64(23): 8870–8879. doi:10.1021/acs.jcim.4c01364.

Kinetics-Based State Definitions for Discrete Binding Conformations of T4 L99A in MD via Markov State Modeling

Chris Zhang[†], **Meghan Osato**[‡], **David L. Mobley**[*,†,‡]

[†]Department of Chemistry, University of California, Irvine, 1120 Natural Sciences II, Irvine, California 92697, United States

[‡]Department of Pharmaceutical Sciences, University of California, Irvine, 856 Health Sciences Road, Irvine, California 92697, United States

Abstract

As a model system, the binding pocket of the L99A mutant of T4 lysozyme has been the subject of numerous computational free energy studies. However, previous studies have failed to fully sample and account for the observed changes in the binding pocket of T4 L99A upon binding of a congeneric ligand series, limiting the accuracy of results. In this work, we resolve the closed, intermediate and open states for T4 L99A previously reported in experiment in MD and establish definitions for these states based on the dynamics of the system. From this analysis, we arrive at two primary conclusions. Firstly, assignment of simulation trajectories into discrete states should not be done simply based on RMSD to crystal structures as this can result in misassignment of states. Secondly, the different metastable conformations studied here need to be carefully treated, as we estimate the timescales for conformational interconversion to be on the order of 10^2 to 10^3 ns – far longer than timescales for typical binding calculations. We conclude with a discussion on the need to develop enhanced sampling methods to generally account for significant changes in protein conformation due to relatively small ligand perturbations.

Graphical Abstract

We map the discrete states of the T4 L99A binding pocket onto a 2D projection using the slowest motions of the system.

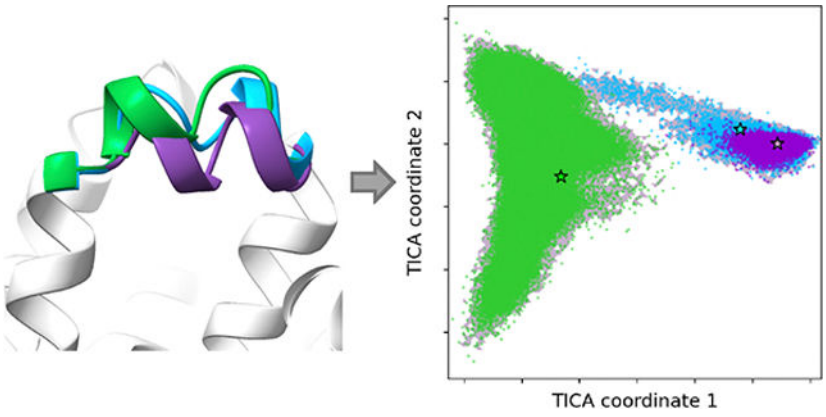
[*] dmobley@uci.edu .

Disclosures

The authors declare the following competing financial interest(s): DLM serves on the scientific advisory boards of Anagenex and OpenEye Scientific Software, Cadence Molecular Sciences. He also is an Open Science Fellow with Psivant.

Supporting Information Available

All Supporting Information is made freely available. Included are plots of the timescales and slow transitions for each system and snapshots of trajectory frames along various points of the TICA surface.



Introduction

Early stage drug discovery is an important and painstaking process. Identification and validation of compounds in pre-clinical trials span many years with thousands of potential candidates whittled down to a single approved drug.[1,2] Modern developments to *in silico* methods seek to remedy the time and cost drain of early stage discovery, providing information on ligand properties without the need for time- and resource-intensive experiments.[3] In particular, relative binding free energy (RBFE) calculations have been widely adopted due to their ability to provide thermodynamic information on potential drug candidates while only requiring simulation data generated from molecular dynamics (MD).[4–7]

The lysozyme of the T4 bacteriophage serves as an ideal test system for binding free energy calculations. A unique leucine to alanine mutation (L99A) creates a solvent-inaccessible and virtually apolar binding pocket with high binding rates for small, organic molecules. The properties of the binding pocket prevent motion of bulk water in and out of the binding pocket, decreasing computational cost and increasing the ease of performing RBFE calculations.[7–9] However, when bound to a congeneric series of alkyl benzenes, the T4 L99A binding pocket is experimentally observed to adopt three discrete conformations (Figure 1), each with increasingly more area open to bulk solvent.[10] The emergence of several discrete binding pocket states upon binding of congeneric ligands has been observed in other systems as well,[10] suggesting that many protein binding pockets may undergo significant changes for even relatively small ligand perturbations.

Prior investigations into simulating congeneric series of alkyl benzene ligands bound to T4 L99A have shown that choice of protein starting conformation impacts the protein conformations sampled in MD.[11] In particular, Bradford et al. noted that high-energy barriers separate the distinct conformational states of the T4 L99A F-helix, requiring additional simulation time or enhanced sampling techniques to overcome.[12] While RBFE calculations in principle capture protein conformational changes if the underlying simulations are run long enough, these conformational changes have often not been thoroughly assessed in prior studies on this system.[11] Interconversion between binding pocket states is slow, requiring significant simulation time to capture even a single event.

Improving RBFE calculations may require knowledge of the conformational changes of a system and their relative timescales, which is often difficult to define and sufficiently sample in the course of MD timescales typically used for these calculations.[5]

Building Markov state models (MSMs) can lead to better understanding of the conformational states of a system.[13–15] MSMs allow for the construction of a transition probability matrix for a system at equilibrium, which can be used to obtain the relative populations of conformational states and the transition timescales between them.[16,17] This can be accomplished by running short MD simulations of a system in replicate, decreasing overall wall clock time compared to running a single long trajectory. To build an MSM, relevant features of the system are selected and transformed using time-lagged independent component analysis (TICA) to resolve coordinates corresponding to the slowest motions of the system.[18] Features can be selected using known information of a biological system, using common elements such as torsional angles or pairwise distances among backbone atoms, or using methods which measure the quantity of kinetic variance for features.[13,19,20] From there, a set of metastable states can be identified by clustering along the TICA space. A transition probability matrix can be calculated by estimating the number of transitions between metastable states at some fixed lag time, τ.

For a transition matrix constructed under a specific set of conditions, the eigenvector corresponding to the largest eigenvalue of the transition matrix is equal to one and describes the stationary distribution of the system. Sorting the remaining eigenvalues of the transition matrix in descending order, the corresponding eigenvectors describe the subsequent slowest motions of the system.[16] The timescale for each slow motion, also known as the **implied timescale**, can be calculated as:

$$t_n = \frac{-\tau}{\ln \lambda_n} \tag{1}$$

where is t_n is the n^{th} implied timescale of the system, τ is the MSM lag time and λ_n is the n^{th} largest eigenvalue of the transition matrix.[21]

Slow motions for a given system are found by evaluating plots of the implied timescales of a system as a function of selected lag times. Processes with timescales shorter than the lag time are not considered slow motions and cannot be resolved by the MSM. The number of metastable states for a system can be selected based on the number of slow motions in the system, and validated using the Chapman-Kolmogorow (CK) test to ensure the MSM is Markovian.[19]

In this study, we seek to capture transitions among the experimentally observed conformational states of the T4 L99A F-helix and make order of magnitude estimates for the timescales of these transitions, to help guide future binding studies and other research. We begin by defining the discrete conformational states of the T4 L99A F-helix based on the slow dynamics of the system. Using Markov state models (MSMs), we estimate the timescales for transitions between discrete states in MD. Furthermore, we show how

defining metastable states of a system from the slowest motions of the system results in more interpretable and clearly separated states compared to using RMSD to experimental crystal structures. Our approach demonstrates how changes in protein conformation for the T4 L99A system are not observed on MD timescales typical for binding studies. Thus, there is a need to apply Markov state modeling to identify potential changes in protein conformation, and potentially the need for future binding studies to employ enhanced sampling techniques or alternate approaches to carefully treat these conformational changes.

Methods

Different simulated systems.

In this study, we sought to better understand the closed, intermediate and open states originally defined by Merski et al. and use crystal structures provided in that manuscript as our starting structures for MD.[10] The authors reported the structures for a number of congeneric ligands bound to T4 L99A and the proportion of observed electron density occupied by the closed, intermediate and open states for each system. We took the ligand-bound system with the highest relative population for a given discrete state as the representative starting structure of that state in this study. Using this methodology, we arrived at the structure of benzene bound to T4 L99A (PDB ID: 4w52) as our starting structure for the closed state, butylbenzene bound to T4 L99A (PDB ID: 4w57 using alternate location A for the sidechains and conformer A for the ligand) for the intermediate state and hexylbenzene bound to T4 L99A bound (PDB ID: 4w59 using alternate location A for the sidechains and conformer A for the ligand). We initialized systems for all combinations of these three protein structures (4w52, 4w57 and 4w59) and their three bound ligands (benzene, butylbenzene and hexylbenzene) to generate a total of nine different systems (Table 1). Additional structures extending this series of congeneric ligands bound to T4 L99A are available in the literature, notably from Bradford et al.[12] which contains structures at both cryo and room temperatures. However, we found that for the purposes of this study – where we focus on the F-helix of the T4 L99A structure – these additional structures did not provide any additional insights (Figure S13).

We refer to the three systems where the bound ligand is the ligand from the crystal structure as **native systems**. The remaining six systems where the bound ligand is different from the ligand in the crystal structure are referred to as **mixed systems**. We set up simulations for all nine systems based on the procedure outlined in the following subsection. For the sake of brevity, we reference various systems throughout this work in the following manner: [discrete state]–[ligand name] (Table 1). This should be interpreted as “the system of [ligand] bound to T4 L99A started from the [discrete state] structure”. As an example, “closed–benzene system” should be read as “the system of benzene bound to T4 L99A started from the closed structure”.

Preparation and parametrization of proteins and ligands.

We prepared the T4 lysozyme protein and ligand structures as input structures for MD simulations. The topology and coordinate input files for GROMACS simulations[22,23] can be found in on GitHub at https://github.com/MobleyLab/T4_MSM.

We prepared the structure for each protein of interest (4w52 conformation A, 4w57 conformation A, and 4w59 conformation A) using Open-Eye Spruce[24] to add hydrogens and any missing loops. We solvated each protein with TIP3P waters and ions to achieve a concentration of 150 mM. We then parameterized the solvated systems using the Amber ff14SB force field.[25]

We assigned partial charges for each ligand of interest – benzene, butylbenzene and hexylbenzene – using OpenEye's AM1-BCC charge engine[26] and parameterized them using Open Force Field version 2.0.0.[27]

We combined the solvated protein structures and prepared ligands into native and mixed system complexes as described in the previous subsection to produce a total of nine starting structures (Table 1). We used the crystallographic ligand pose from the native structure for each protein-ligand complex.

Running molecular dynamics in GROMACS.

We simulated all protein-ligand complex systems using GROMACS (v.2021.2). Prior to production, we performed energy minimization for up to 1500 steps using steepest descents. We then equilibrated the systems in two phases, using a 20 ps NVT simulation followed by a 5 ns NPT simulation. We ran production simulations for 100 ns per replicate, with a total of 10 replicates per system. Although we initialized some ligands in their non-native crystal structures for mixed system simulations, we found that any potential clashes were resolved throughout minimization and equilibration steps. We note that there may be instances where clashes do not resolve, although this was not observed in our study. We provide plots for where structures fall in TICA space after each phase of equilibration for each system in the Supporting Information (Figure S1).

Additionally, we evaluated both protein and ligand convergence for each system (Figures S15–S32). To assess protein convergence, we looked at the time series of an atom on a residue outside of the F-helix (Tyr88). For convergence in ligand orientation, we evaluated the distance between the same atom and a selected atom on the ligand. We observed little motion in the protein and ligand time series in all benzene-/butylbenzene-bound systems and the hexylbenzene-bound native system (S15–S26; S31, S32). In the some repeats of the closed–hexylbenzene (S27, S28) and int–hexylbenzene (S29, S30) mixed systems, the time series were less stable. In these repeats, we saw correlation between the protein and ligand motions. As these are systems where we expect the protein may shift to accommodate a larger nonnative ligand, we conclude our simulations are largely converged. The MDP files for GROMACS simulations can be found on GitHub at https://github.com/MobleyLab/T4_MSM.

Markov state model construction.

We built all MSMs in this study using the PyEMMA (v2.5.12)[19] and deeptime (v.0.4.4)[28] Python packages. We started by choosing features along the F-helix region of the protein (residues 107–115) to characterize the changes in the T4 L99A binding pocket. Using MDTraj (v.1.9.9),[29] we calculated the pairwise distance between all $C_\alpha - C_\alpha$ and $C_\beta - C_\beta$ pairs for all residues in the F-helix for a total of 51 distances. We used an implementation of

TICA from deeptime (v0.4.4), selecting a TICA lag time of 0.02 ns (10 frames) to transform the set of features into coordinates that described the slowest motions of the system. We then applied *k*-means clustering on the projection of the trajectory frames onto the two largest TICA components in order to resolve metastable states for each system. We used 1000 cluster centers in order to adequately sample the surface, specifically targeting sampling regions of lower density. We selected a lag time of $\tau = 0.5$ ns (250 frames) to construct each MSM; this was the shortest lag time beyond which the implied timescales for each system plateaued (Figure S2).

Calculating RMSD to reference crystal structures.

We calculated the root-mean-square deviation (RMSD) between each frame of our trajectories and the closed (PDB ID: 4w52), intermediate (PDB ID: 4w57) and open (PDB ID: 4w59) crystal structures using the rmsd function in MDTraj (v.1.9.9). We used this function to center each trajectory and calculate the distance between atoms in our trajectories and each reference crystal structure. We calculated RMSD for atoms in both the F-helix (residues 107 – 115) and for atoms in a region we defined as the protein binding pocket. We defined the binding pocket as all atoms 0.5nm away from the benzene ligand in the 4w52 structure of T4 L99A.[8] For RMSD calculations in both the F-helix and binding pocket, we evaluated an all heavy atom variation as well as a variation only using C_α and C_β atoms of the selection (Figure S3, S4).

Estimating MSM mean first passage time.

We estimated timescales among our discrete states using the `mfpt` function from the deeptime package. The mean first passage time (MFPT) is defined as the expected time (reported in units of number of simulation frames) to reach one state when starting in another.[30] To calculate the mean first passage time, we provided the function with a transition probability matrix for our metastable states as well as the lag time used to construct the MSM. We obtained these coarse-grained transition probability matrices using the `pcca` function from deeptime.[31]

Results

Slowest motion of congeneric ligand systems corresponds to opening of the F-helix.

We combine simulation data from all three native systems together to construct a trajectory where the slowest motion of our system corresponds to the opening of the F-helix and/or any important differences between these native systems.[13,14] For clarity, these three **native systems** are benzene bound to T4 L99A started from the closed structure (closed–benzene), butylbenzene bound to T4 L99A started from the intermediate structure (int–butylbenzene) and hexylbenzene bound to T4 L99A started from the open structure (open–hexylbenzene) (see Methods for more explanation on terminology). Specifically, we expect these native systems to occupy three distinct states with no or few transitions between them. Thus, a concatenated trajectory of these native systems would show distinct transitions between states at the end of each of the individual component trajectories, turning these state transitions into easily recognizable “slow motions”. This does not mean that the opening of the F-helix is a fundamental transition seen in all systems, but rather allows us to resolve

a consistent set of time-lagged independent components (tICs) that can be used to categorize the frames of each individual system into discrete states.

We find that our 2D TICA surface, constructed from $C_\alpha - C_\alpha$ and $C_\beta - C_\beta$ pairwise distances within the protein F-helix (residues 107–115), consists of three primary regions of density. We refer to these regions as the half-moon region on the left, an oval-shaped well on the right and a bridge region connecting them (Figure 2). To better understand how regions of the TICA space correspond to different F-helix conformations, we project the three native crystal structures onto our 2D TICA surface. We find that the open crystal structure is roughly centered in the half-moon region, the intermediate crystal structure is right of the bridge region and the closed crystal structure is centered in the oval-shaped well (Figure 2). Based on the position of each crystal structure in the 2D TICA surface, we determine that the differences among discrete states is predominately captured by variation in the first independent component (Figure 2).

Time-lagged independent component analysis provides an alternative to RMSD for defining discrete states.

Using the tICs we previously define, we map each native system trajectory onto the 2D TICA surface. Frames of the closed–benzene system map around the oval-shaped well and are centered around the closed crystal structure, suggesting this area corresponds to the closed conformation of the F-helix (Figure 3A). Simulations of the int–butylbenzene system sample all regions of the oval-shaped well in addition to the bridge region linking the two wells together (Figure 3B). The overlap between the regions covered by the closed–benzene and int–butylbenzene trajectories is expected as the int–butylbenzene system is reported to sample both closed and intermediate states.[10] However, the bridge region is *not* sampled by the closed–benzene system, which suggests that it is the region of TICA space corresponding to the intermediate conformation. Lastly, simulations of the open–hexylbenzene system exclusively sample the half-moon region on the left side of the TICA plot, indicating this region corresponds to the open conformation (Figure 3C). In the absence of tIC1, we find that separation among the closed, intermediate and open crystal structures is much less distinct. We observe much greater overlap between the native system simulations (Figure S5).

We find that using the RMSD to reference crystal structures to define states shows some agreement with our TICA-based approach. In the majority of cases, native system frames close to a reference crystal structure in TICA space also tend to also have low RMSD to that crystal (Figure 3D–F). This does not have to be the case as distance in TICA space does not necessarily correspond to structural similarity,[18] since TICA is based on kinetics. However, here, RMSD seems to be a poor way to define the state boundaries. For example, frames in both the half-moon and oval shaped wells show low RMSD to the closed crystal, although the latter simulation is initialized from the open state (Figure 3D).

Choice of protein starting structure affects sampling of protein conformational states.

We proceed to construct Markov state models (MSMs) for each native system to resolve metastable states based on slow dynamics. By analyzing the eigenvectors of the estimated

transition probability matrix from an MSM – corresponding to the timescales of the slowest motions of the system – we can identify the number of macrostates for each system, and validate that the MSM is still Markovian. We use plots of the implied scales for each system to determine whether processes occur on timescales which can resolved by our MSMs. Processes faster than the MSM lag time are not considered to be kinetically meaningful transitions.[19]

We observe that the closed–benzene system undergoes no kinetically meaningful transitions in the elapsed simulation time frame. The slowest motions of the system decay faster than the MSM lag time and correspond to random fluctuations (Figure S6A). Based on these findings, we determine there is only a single state sampled in the simulations of the closed–benzene system (Figure 4A), which is consistent with the region of TICA space we identify as the closed region (Figure 3A).

In simulations of the int–butylbenzene system, we find that the slowest motion of the system is an exchange within the oval-shaped well of the 2D TICA surface, splitting the region into two kinetically distinct macrostates (Figure S7B). We characterize this split macrostate later in the Results. Interestingly, the *second* slowest motion of the int–butylbenzene system corresponds to changes in density between the bridge and oval-shaped regions of the TICA space (Figure S7B), which suggests those regions correspond to the closed and intermediate states respectively (Figure 3B). However, this process decays faster than the MSM lag time and thus we do not consider the intermediate state to be a discrete state based on analysis of this system (Figure S2).

Lastly, we determine there is only a single state sampled in simulations of the open–hexylbenzene system (Figure 4C). While there are motions in the system occurring on slow timescales (Figure S2), they correspond to fluctuations within the half-moon shaped well, and are primarily described by differences in the second tIC (Figure S8C). Given we primarily attribute the F-helix opening motion to the first tIC (Figure 2), we do not consider these to be states relevant to our analysis.

Simulation of mixed protein-ligand systems allows for estimation of discrete state transition timescales.

In attempts to observe transitions within our simulation timescales, we also simulate and build MSMs for a series of **mixed** systems. These are simulations begun from a crystallographic state belonging with a different ligand than the one they are simulated with, so we expect these simulations to transition to a different state while they are run, even if only once (see Methods).

For the benzene-bound mixed systems, we find that simulations of both the int–benzene (Figure 5A) and open–benzene (Figure 5B) systems primarily sample the closed state. In the int–benzene system, we observe a single slow motion (Figure S6B) corresponding to an exchange between two regions of the oval-shaped well (Figure 5A).

For the butylbenzene-bound mixed systems, we observe a single state for the closed–butylbenzene system (Figure 5C, Figure S7A) and find that this system does not sample

the bridge region of the TICA space that is present in the native butylbenzene system (Figure 4B). However in the open–butylbenzene system, not only is the bridge region sampled, but the slowest motion of the system corresponds to an exchange between the bridge and oval-shaped regions of the TICA space (Figure S7C). We subsequently identify two states for the open–butylbenzene system (Figure 5D).

Lastly, we find multiple states for both mixed systems containing the hexylbenzene ligand. The closed–hexylbenzene system interconverts between the closed and open states (Figure 5E). Interestingly, the slowest motion of this closed–hexylbenzene system is an exchange between those two discrete states, but we do not observe any sampling of the intermediate state (Figure S8A). The int–hexylbenzene system samples five distinct states (Figure 5E). Four of the five states are observed in other systems, corresponding to the closed, intermediate and open states as well as the state observed in the int–benzene (Figure 5A) system. The final discrete state falls between the closed and open states, similar to the intermediate state. However, it falls lower along the second tIC and is distinct from the previously established intermediate state.

For each system with multiple MSM-identified states, we estimated the timescales of the transitions between them by discretizing our trajectories to calculate a mean-first passage time (see Methods). A summary of which systems contain which transitions is in Table 2. This data allows us to estimate the transition rates shown in Tables 3–5. We note that these rates are highly approximate, given that we have a small number of transitions between states, transitions may be unidirectional, not all transitions are observed in each system and ligands are modeled in their non-native crystal structures. However, these estimates may be useful in providing a rough idea of the order of magnitude for the timescales of these transitions in MD.

Primarily, our timescale estimates suggest that discrete state transitions can be observed in MD but not on relatively short timescales. The majority of transition estimates are in the order of 10^2–10^3 ns (Table 3–5). Our estimates of timescales also further the idea that the intermediate state is a relatively short-lived state. In both mixed systems with transitions between the closed and intermediate states (Figure 5D, F), we estimate that going from intermediate to closed is an order of magnitude faster than going from closed to intermediate (Table 3). However, for the int–hexylbenzene system where we observe transitions between the intermediate and open states (Figure 5E), we instead find that going from intermediate to open is about an order of magnitude *slower* than going from open to intermediate (Table 4).

Physical Features of Different Macrostates.

In order to gain insight into the macrostates identified in this study, we evaluate distributions for three different features of the F-helix: H-bond distances, helix end-to-end distances and distances between residues that roughly approximate helix coil diameter (Figures S10–S12). For each macrostate, we use the difference in feature distributions to identify unique physical feature(s) characterizing the state. We find that the closed state is characterized by larger values for the distance between the Gly110-C_α and Ala112-C_α atoms (Figure S12) and relatively shorter end-to-end helix distance (Gly107-C_α and Thr115-C_α) (Figure S11). For the intermediate state, we observe a much larger distance between the Gly110-O

and Phe114-H atoms (Figure S10), as well as larger distance between the Val111-C_α and Gly113-C_α atoms (Figure S12). The open state is characterized by shorter distance between the Val111-O and Thr115-H atoms (Figure S10), as well as slightly longer distance between the Gly107-O and Val111-H atoms (Figure S10).

We also observed two macrostates distinct from the closed, intermediate and open states surfaced in the MSMs for various systems. One of these macrostates splits the oval-shaped well on the right side of the TICA surface and is observed in the int–butylbenzene, int–benzene and open–benzene systems (Figure 4B; 5A, F). It is characterized by a larger distance between the Gly113-C_α and Thr115-C_α atoms (Figure S12). The other macrostate is only observed in the int–hexylbenzene system (Figure 5F) and forms a bridge between the left and right wells in TICA space, further below the intermediate state. This state is characterized by larger end-to-end helix distance (Gly107-C_α and Thr115-C_α) (Figure S11) and distance between the Glu108-C_α and Gly110-C_α atoms (Figure S12).

Discussion

In this study, we use MD to resolve several discrete states of the T4 L99A binding pocket also observed in experiment. To aid in selection of features to guide this analysis, we concatenate parallel trajectories of benzene, butylbenzene and hexylbenzene bound to their native crystal structures so that the slowest motion of the resulting combined trajectory corresponds to the transition from the closed to intermediate to open states of the T4 L99A binding pocket.[10] In doing so, we are able map simulations of both native and mixed ligand systems onto a two-dimensional TICA surface. We find based on the relative positions of the crystal structures in TICA space that the closed and intermediate structures are more kinetically similar states, whereas the open conformation is more kinetically distinct from the two.

Evidence of the intermediate observed in MD.

Firstly, we find evidence that the intermediate state is a kinetically distinct state. The trajectory of the native int–butylbenzene system samples the bridge region of the TICA surface (Figure 3B), but this bridge region is not identified as a discrete state in the MSM for the system (Figure 4B). However, the MSM built from simulations of the open–butylbenzene system (Figure 5D) does identify the bridge region of the TICA space as a kinetically distinct macrostate. This suggests that while the intermediate state may be observed in our MD simulations of the int–butylbenzene system, transitions to and from the state are too transient for it to be considered a kinetically distinct state. The intermediate state is identified as a kinetically distinct state when we attempt to force the transition in simulations of the open–butylbenzene mixed system.

TICA-based state definitions may be preferable to RMSD-based definitions.

Secondly, we observe that using TICA and slow motions yield clearer state definitions than RMSD. This is in line with previous reports in the literature of using transition rates to define macrostates.[32] RMSD-based state definitions lead to fuzzy boundaries between states. For example, we observe that trajectory frames close to the open crystal structure in TICA

space generally have low RMSD to that structure (Figure 3F). However, some of these same frames also have low RMSD to the int crystal (Figure 3E), which would complicate any analysis which attempted to assign states based on RMSD cutoffs alone. Given the int and open structures are considered distinct in experiment, considering a structure to be similar to both states is confusing. We additionally test several variations of our RMSD analysis, calculating the RMSD by looking at all heavy atom and C_α atoms, as well as C_β atom-only variations both in the F-helix and a region we defined as the binding pocket (see Methods). By including atoms outside of the F-helix for our binding pocket RMSD calculations, we account for any potential biases due to limited atom selection. We find that our conclusions hold true regardless of which atoms are selected for the RMSD calculation (Figure S3, S4).

Additionally, structures with similar RMSD to the same reference structure are not necessarily similar to each other. As an example, we observe that frames in the top left corner of the half-moon shaped well and frames in the right side of the oval-shaped well are similarly distant to the open structure based on RMSD (Figure 3F). However, we instead find with our kinetics-based definition that the entirety of the half-moon shaped well describes the open state (Figure 4C). While the top and bottom portions of the half-moon shaped well can be further separated into kinetically distinct states (Figure S8C), this separation is entirely along the second tIC, which does not primarily correspond to opening of the F-helix.

One implication of this finding is that previous work looking at the transitions between states for this system may have used incorrect state definitions and potentially miscounted transitions[11] and misdiagnosed the extent to which protein conformational transitions are adequately sampled on simulation timescales. It may be worth revisiting this work and applying these TICA-based state definitions to evaluate how much enhanced sampling methods are able to accelerate transitions. Furthermore, using slow system dynamics to define individual states could be extended to other systems where there are known conformational changes in response to a congeneric series of ligands. Examples of known systems with such behavior include the heat shock protein 90 (HSP90)[33,34] and the human estrogen receptor alpha (ERα).[10] This approach could especially be useful in systems where there may not necessarily be reference crystal structures for each discrete state, since in such cases an RMSD-based analysis is not possible.

Other Structures.

As outlined in the Methods section, the structures we use for the closed, intermediate and open states in this study are taken from the conformation with the highest electron density with each selected ligand from Merski et al.[10] However, these structures are by no means a comprehensive set of all available structures or all possible protein conformations. Particularly, Bradford et al.[12] pointed out the existence of temperature effects for cryo structures for this same protein mutant, providing a number of different protein-ligand structures obtained at different temperatures which differed in various ways. We were interested in assessing how the structures from Bradford et al.'s work interface with the present study, so we projected both room temperature and cryo crystal structures for an extended series of alkyl benzene ligands from this work onto our TICA surface (Figure

S13). We observe two distinct groupings for all structures, indicating that these structures are similar to the structures from Merski et al. along the primary tICs we use for this study. Thus, we exclude these additional structures from our analysis because for the purposes of this study, which seeks to identify the larger kinetic differences between the closed, intermediate and open states, more granular differences among structures within the same discrete state are not as important.

Implications for future binding studies.

One caveat of this study is that the procedure presented may not be suitable for some biological systems. For the T4 L99A system, we have access to crystal structures for all discrete states, allowing us to effectively reverse engineer a trajectory in which the slowest motion of the system corresponds to the opening of the F-helix observed in experiment. In many cases, it may not be known what prompts conformational changes or whether the system undergoes any conformational change at all on binding different ligands. We caution readers that TICA may not always be a suitable approach, as particularly for larger proteins, TICA coordinates have been found to contain little information about actual protein dynamics.[35]

Furthermore, we find in this study that even when running MD for timescales much longer than those run for typical binding studies, our native system trajectories typically remain in their starting conformations. This point is further reinforced by a test we run using only the first versus last 50 ns of each simulation replicate. Generally, conclusions from analysis of the two halves of our data are similar. While the TICA components do not change depending on which portion of the data we analyze, we find that with less simulation time, we fail to sample the bridge-like region corresponding to the intermediate state (Figure S14). This is not surprising, as conformational changes here have been previously reported to have large energetic barriers so simply running longer simulations (unless they are orders of magnitude longer) may not resolve issues with sampling.[12,36,37]

Binding studies, including in L99A, are typically done on much shorter timescales than those studied here. Thus, it is vital for researchers running such studies to have an idea of the relevant timescales for protein conformational sampling, and ensure that (a) they are using the appropriate protein structure(s) for their ligand, or (b) they are somehow enhancing sampling of protein conformational transitions, or (c) they diagnose the quality of protein sampling. Given that it is relatively common for structure-based design studies to uncover multiple distinct protein conformations upon binding of different ligands, it may be possible to anticipate some of these conformational changes.[7,10,38,39] However, timescales for transitions between relevant protein conformations might be slow and uncharacterized, in which case researchers need to be cautious. We argue that it may be necessary to further develop and employ enhanced sampling methods for dealing with protein conformational changes.[36,37,40–42]

Lastly, we make all files necessary to reproduce our trajectories freely available. We also include the code used for analysis so that researchers can access our state analyses. To aid in further investigation of the macrostates presented in this work, we include snapshots of

structures along various points in the TICA surface in the Supporting Information, should researchers wish to analyze this data (Figure S10).

Conclusion

Using Markov state modeling, we verify that the discrete states of the T4 L99A binding pocket previously established in experiment[10] can also be observed in MD, although we estimate that the conformational changes occur on timescales between 10^2 and 10^3 ns. Our timescale estimates are consistent with previous findings, where it was shown that for the T4 L99A system, the choice of starting protein conformational state heavily influences which conformations are sampled through the course of the simulation.[11] For known systems undergoing discrete conformational changes upon binding of a congeneric ligand series, we suggest combining short, replicate MD simulations of different bound ligands with MSMs to resolve biologically-relevant slow motions of the system. The coordinates resolved from kinetic information offer an alternative to RMSD-based definitions and can be used to provide more clarity on states of a biological system by allowing us to define macrostates by a collection of structures rather than a single static crystal structure.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Léa El Khoury and Yunhui Ge for helpful discussions and feedback. We thank the National Institutes of Health (R35GM148236) for funding, and OpenEye for an academic software license used in this work.

Data and Software Availability

All code, data and results from this study can be found at on GitHub at https://github.com/MobleyLab/T4_MSM.

References

(1). Reymond J-L; Awale M Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database. ACS Chem. Neurosci. 2012, 3, 649–657. [PubMed: 23019491]

(2). Rennane S; Baker L; Mulcahy A Estimating the Cost of Industry Investment in Drug Research and Development: A Review of Methods and Results. Inquiry. 2021, 58.

(3). Leelananda SP; Lindert S Computational Methods in Drug Discovery. Beilstein J. Org. Chem. 2016, 12, 2694–2718. [PubMed: 28144341]

(4). Cournia Z; Allen B; Sherman W Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. J. Chem. Inf. Model. 2017, 57, 2911–2937. [PubMed: 29243483]

(5). Mobley DL; Klimovich PV Perspective: Alchemical Free Energy Calculations for Drug Discovery. J. Chem. Phys. 2012, 137, 230901 1–12. [PubMed: 23267463]

(6). Wang L; Wu Y; Deng Y; Kim B; Pierce L; Krilov G; Lupyan D; Robinson S; Dahlgren MK; Greenwood J; Romero DL; Masse C; Knight JL; Steinbrecher T; Beuming T; Damm W; Harder E; Sherman W; Brewer M; Wester R; Murcko M; Frye L; Farid R; Lin T; Mobley DL; Jorgensen WL; Berne BJ; Friesner RA; Abel R Accurate and Reliable Prediction of Relative Ligand

Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. J. Am. Chem. Soc. 2015, 137, 2695–2703. [PubMed: 25625324]

(7). Mobley DL; Graves AP; Chodera JD; McReynolds AC; Shoichet BK; Dill KA Predicting Absolute Ligand Binding Free Energies to a Simple Model Site. J. Mol. Biol. 2007, 371, 1118–1134. [PubMed: 17599350]

(8). Mondal J; Ahalawat N; Pandit S; Kay LE; Vallurupalli P Atomic Resolution Mechanism of Ligand Binding to a Solvent Inaccessible Cavity in T4 Lysozyme. PLoS Comput. Biol. 2018, 14(5), e1006180. [PubMed: 29775455]

(9). Wang L; Berne BJ; Friesner RA On Achieving High Accuracy and Reliability in the Calculation of Relative Protein–Ligand Binding Affinities. Proc. Natl. Acad. Sci. U.S.A. 2012, 109, 1937–1942. [PubMed: 22308365]

(10). Merski M; Fischer M; Balius TE; Eidam O; Shoichet BK Homologous Ligands Accommodated by Discrete Conformations of a Buried Cavity. Proc. Natl. Acad. Sci. U.S.A. 2015, 112, 5039–5044. [PubMed: 25847998]

(11). Lim NM; Wang L; Abel R; Mobley DL Sensitivity in Binding Free Energies Due to Protein Reorganization. J. Chem. Theory Comput. 2016, 12, 4620–4631. [PubMed: 27462935]

(12). C. Bradford SY; Khoury LE; Ge Y; Osato M; L. Mobley D; Fischer M Temperature Artifacts in Protein Structures Bias LigandBinding Predictions. Chem. Sci. 2021, 12, 11275–11293. [PubMed: 34667539]

(13). Ge Y; Kier BL; Andersen NH; Voelz VA Computational and Experimental Evaluation of Designed -Cap Hairpins Using Molecular Simulations and Kinetic Network Models. J. Chem. Inf. Model. 2017, 57, 1609–1620. [PubMed: 28614661]

(14). Ge Y; Borne E; Stewart S; Hansen MR; Arturo EC; Jaffe EK; Voelz VA Simulations of the Regulatory ACT Domain of Human Phenylalanine Hydroxylase (PAH) Unveil Its Mechanism of Phenylalanine Binding. J. Biol. Chem. 2018, 293, 19532–19543. [PubMed: 30287685]

(15). Buch I; Giorgino T; De Fabritiis G Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. Proc. Natl. Acad. Sci. U.S.A. 2011, 108, 10184–10189. [PubMed: 21646537]

(16). Husic BE; Pande VS Markov State Models: From an Art to a Science. J. Am. Chem. Soc. 2018, 140, 2386–2396. [PubMed: 29323881]

(17). Chodera JD; Noé F Markov State Models of Biomolecular Conformational Dynamics. Curr. Opin. Struct. Biol. 2014, 25, 135–144. [PubMed: 24836551]

(18). Pérez-Hernández G; Paul F; Giorgino T; De Fabritiis G; Noé F Identification of Slow Molecular Order Parameters for Markov Model Construction. J. Chem. Phys. 2013, 139, 015102. [PubMed: 23822324]

(19). Scherer MK; Trendelkamp-Schroer B; Paul F; Pérez-Hernández G; Hoffmann M; Plattner N; Wehmeyer C; Prinz J-H; Noé F PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. J. Chem. Theory Comput. 2015, 11, 5525–5542. [PubMed: 26574340]

(20). Wu H; Noé F Variational Approach for Learning Markov Processes from Time Series Data. J. Nonlinear Sci. 2020, 30, 23–66.

(21). Bowman GR In An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation; Bowman GR, Pande VS, Noé F, Eds.; Advances in Experimental Medicine and Biology; Springer Netherlands, 2014; pp 7–22.

(22). Berendsen H; van der Spoel D; van Drunen R GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation. Comput. Phys. Commun. 1995, 91, 43–56.

(23). Abraham MJ; Murtola T; Schulz R; Páll S; Smith JC; Hess B; Lindahl E GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. SoftwareX 2015, 1–2, 19–25.

(24). OpenEye Spruce 1.5.3.3: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com. 1/1/23.

(25). Maier JA; Martinez C; Kasavajhala K; Wickstrom L; Hauser KE; Simmerling C ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J. Chem. Theory Comput. 2015, 11, 3696–3713. [PubMed: 26574453]

(26). OpenEye Quacpac 2.2.3.3: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com. 1/1/23.

(27). Wagner J; Thompson M; Dotson D; hyejang; Boothroyd S; Rodríguez-Guerra J openforcefield/openff-forcefields: Version 2.0.0 “Sage”. 2021; DOI:10.5281/ZENODO.5214478.

(28). Hoffmann M; Scherer M; Hempel T; Mardt A; de Silva B; Husic BE; Klus S; Wu H; Kutz N; Brunton SL; Noé F Deeptime: A Python Library for Machine Learning Dynamical Models from Time Series Data. Mach. Learn. Sci. Technol. 2021, 3, 015009.

(29). McGibbon RT; Beauchamp KA; Harrigan MP; Klein C; Swails JM; Hernández CX; Schwantes CR; Wang L-P; Lane TJ; Pande VS MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophys. J. 2015, 109, 1528–1532. [PubMed: 26488642]

(30). Hoel PG; Port SC; Stone CJ Introduction to Stochastic Processes; Waveland Press, 1986.

(31). Röblitz S; Weber M Fuzzy Spectral Clustering by PCCA+: Application to Markov State Models and Data Classification. Adv. Data Anal. Classif. 2013, 7, 147–179.

(32). Swope WC; Pitera JW; Suits F; Pitman M; Eleftheriou M; Fitch BG; Germain RS; Rayshubski A; Ward TJC; Zhestkov Y; Zhou R Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 2. Example Applications to Alanine Dipeptide and a -Hairpin Peptide. J. Phys. Chem. B 2004, 108, 6582–6594.

(33). Street TO; Lavery LA; Agard DA Substrate Binding Drives Large-Scale Conformational Changes in the Hsp90 Molecular Chaperone. Mol. Cell 2011, 42, 96–105. [PubMed: 21474071]

(34). Amaral M; Kokh DB; Bomke J; Wegener A; Buchstaller HP; Eggenweiler HM; Matias P; Sirrenberg C; Wade RC; Frech M Protein Conformational Flexibility Modulates Kinetics and Thermodynamics of Drug Binding. Nat. Commun. 2017, 8, 2276. [PubMed: 29273709]

(35). Schultze S; Grubmüller H Time-Lagged Independent Component Analysis of Random Walks and Protein Dynamics. J. Chem. Theory Comput. 2021, 17, 5766–5776. [PubMed: 34449229]

(36). Burley KH; Gill SC; Lim NM; Mobley DL Enhancing Side Chain Rotamer Sampling Using Nonequilibrium Candidate Monte Carlo. J. Chem. Theory Comput. 2019, 15, 1848–1862. [PubMed: 30677291]

(37). Gill SC; Lim NM; Grinaway PB; Rustenburg AS; Fass J; Ross GA; Chodera JD; Mobley DL Binding Modes of Ligands Using Enhanced Sampling (BLUES): Rapid Decorrelation of Ligand Binding Modes via Nonequilibrium Candidate Monte Carlo. J. Phys. Chem. B 2018, 122, 5579–5598. [PubMed: 29486559]

(38). Meirovitch H; Cheluvaraja S; White RP Methods for Calculating the Entropy and Free Energy and Their Application to Problems Involving Protein Flexibility and Ligand Binding. Curr Protein Pept Sci. 2009, 10, 229–243. [PubMed: 19519453]

(39). Stank A; Kokh DB; Fuller JC; Wade RC Protein Binding Pocket Dynamics. Acc. Chem. Res. 2016, 49, 809–815. [PubMed: 27110726]

(40). Ahmad K; Rizzi A; Capelli R; Mandelli D; Lyu W; Carloni P Enhanced-Sampling Simulations for the Estimation of Ligand Binding Kinetics: Current Status and Perspective. Front. Mol. Biosci. 2022, 9.

(41). Gallicchio E; Lapelosa M; Levy RM The Binding Energy Distribution Analysis Method (BEDAM) for the Estimation of Protein-Ligand Binding Affinities. J. Chem. Theory Comput. 2010, 6, 2961–2977. [PubMed: 21116484]

(42). Smith L; Novak B; Osato M; Mobley DL; Bowman GR PopShift: A Thermodynamically Sound Approach to Estimate Binding Free Energies by Accounting for Ligand-Induced Population Shifts from a Ligand-Free Markov State Model. J. Chem. Theory Comput. 2024, 20, 1036–1050. [PubMed: 38291966]

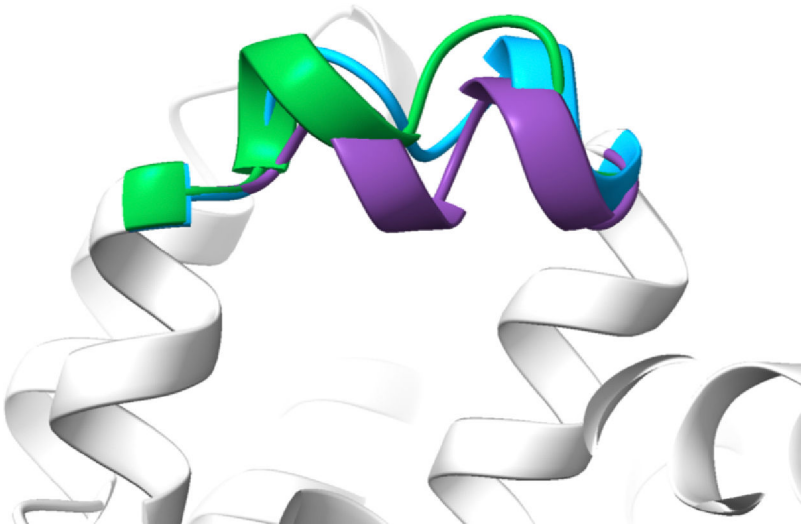


Figure 1.

Discrete conformations of the T4 L99A F-helix. The F-helix region (residues 107–115) of T4 L99A is reported[10] to adopt three distinct conformations as observed crystallographically: closed, intermediate and open. Here we show three overlays of the F-helix region from the crystal structures we use to represent each state in this work. Shown is the benzene-bound crystal structure (4w52) to represent the experimentally defined closed state (purple), the butylbenzene-bound crystal structure (4w57) to represent the experimentally defined intermediate state (cyan), and the hexylbenzene-bound crystal structure (4w59) to represent the experimentally defined open state (green).

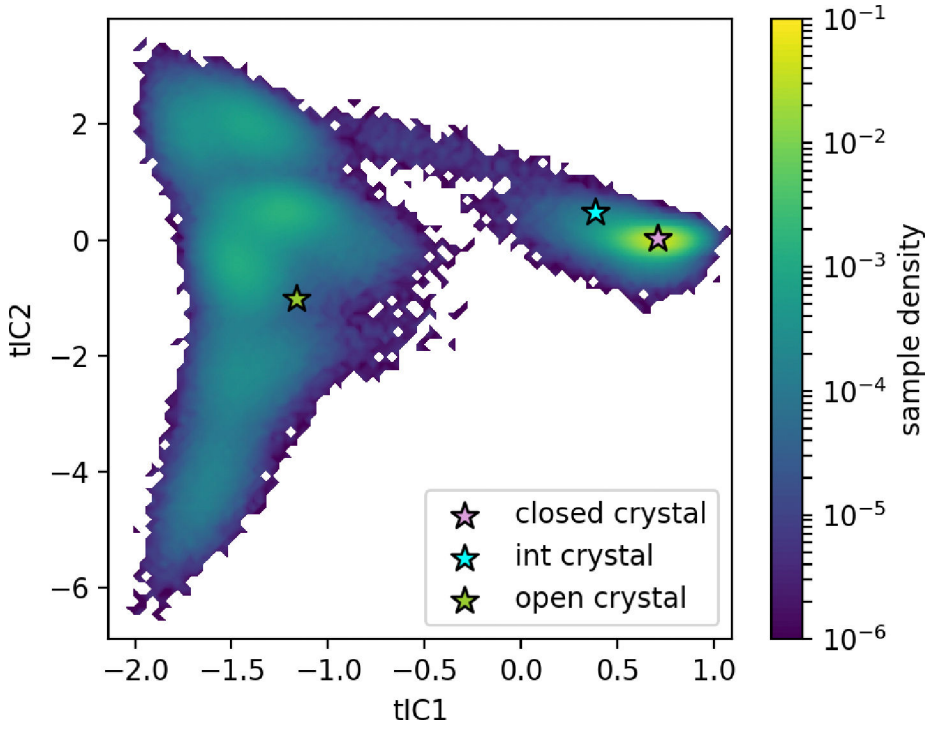


Figure 2.

Concatenated trajectory data projected onto a 2D TICA landscape. We concatenate six 100 ns parallel trajectories each of benzene, butylbenzene and hexylbenzene bound to their native crystal structures (4w52, 4w57, 4w59, respectively) to form a concatenated trajectory (18 trajectories, total 1.8 μs). In this concatenated trajectory, we select the pairwise distances between all C_α–C_α and C_β–C_β atoms in the F-helix as features. Using TICA to resolve coordinates along the slowest motions of the system, we show the trajectory projected onto the top two tICs. We map where the crystal structures for closed (purple star), intermediate (cyan star) and open (green star) fall on this landscape. The arrangement of the crystal structures suggest that traversing along the first TICA component captures the discrete changes in the F-helix.

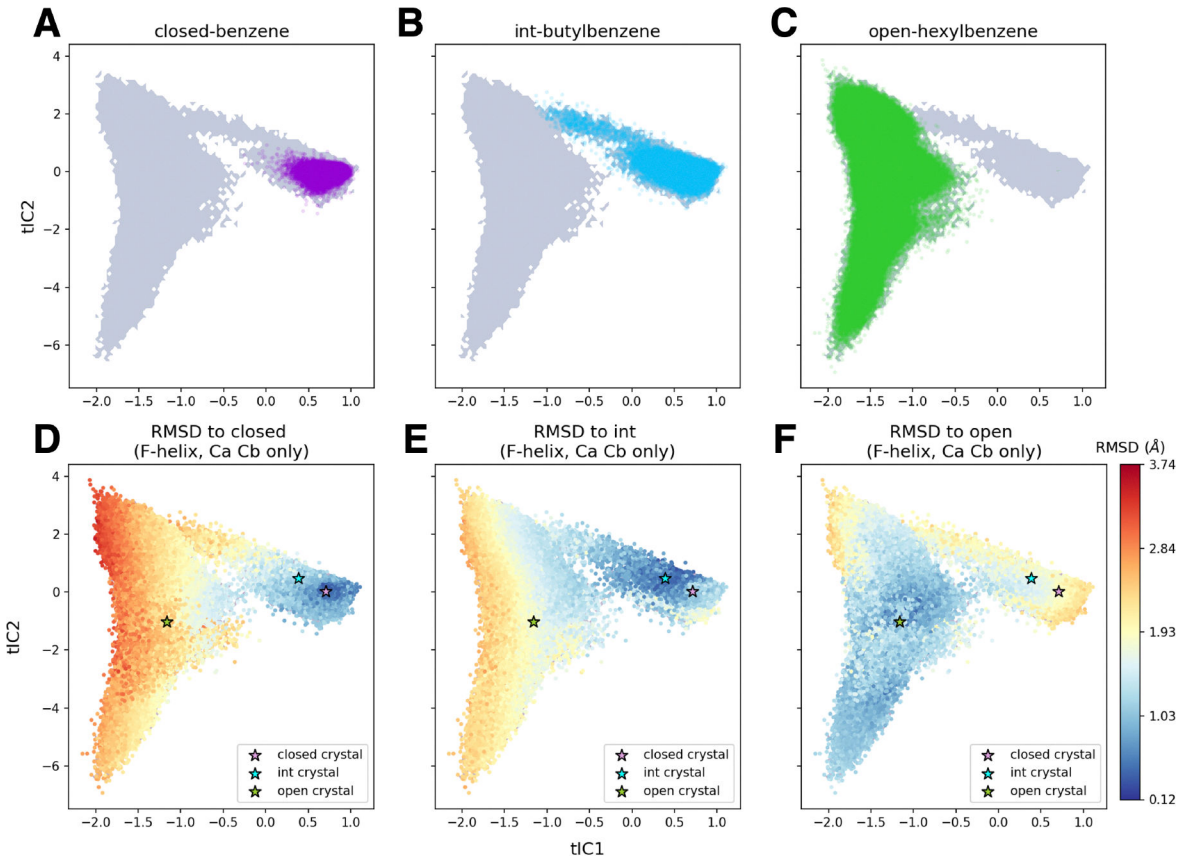


Figure 3.

RMSD of trajectory frames in 2D TICA space relative to experimental crystal structures. (A–C) On top of the density of the concatenated trajectory (gray), we project the frames from simulations of the (A) closed–benzene (purple dots), (B) int–butylbenzene (cyan dots) and (C) open–hexylbenzene (green dots) systems in 2D TICA space. (D–F) For each frame of the concatenated trajectory, we calculate the RMSD of the F-helix to different crystal structures. Shown is RMSD to (D) the benzene-bound crystal structure (closed), (E) the butylbenzene-bound crystal structure (intermediate) and (F) the hexylbenzene-bound crystal structure (open). The heatmap signifies the RMSD of each frame of the concatenated trajectory to the specified reference crystal structure. We additionally map where the closed (purple star), intermediate (cyan star) and open (green star) crystal structures fall on this landscape for reference. Frames closer to each respective crystal structure in TICA space tend to have lower RMSD to that structure.

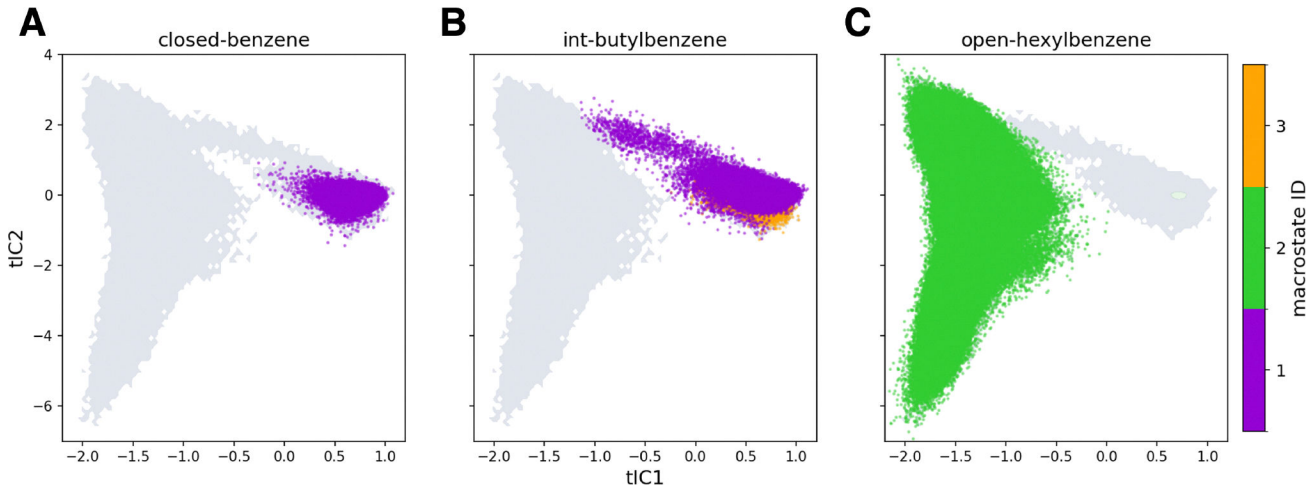


Figure 4.

MSM-resolved states for each native system. We construct MSMs to resolve discrete states based on the slowest processes in each of the three native systems. On top of the density of the concatenated trajectory (gray), we project the frames of simulations of the (A) closed–benzene, (B) int–butylbenzene and (C) open–hexylbenzene systems. We color frames based on their MSM-assigned macrostates, keeping macrostate definitions consistent across the different native systems. In the elapsed simulation time (1 μs), we find that native systems are mostly confined to their starting protein conformational states.

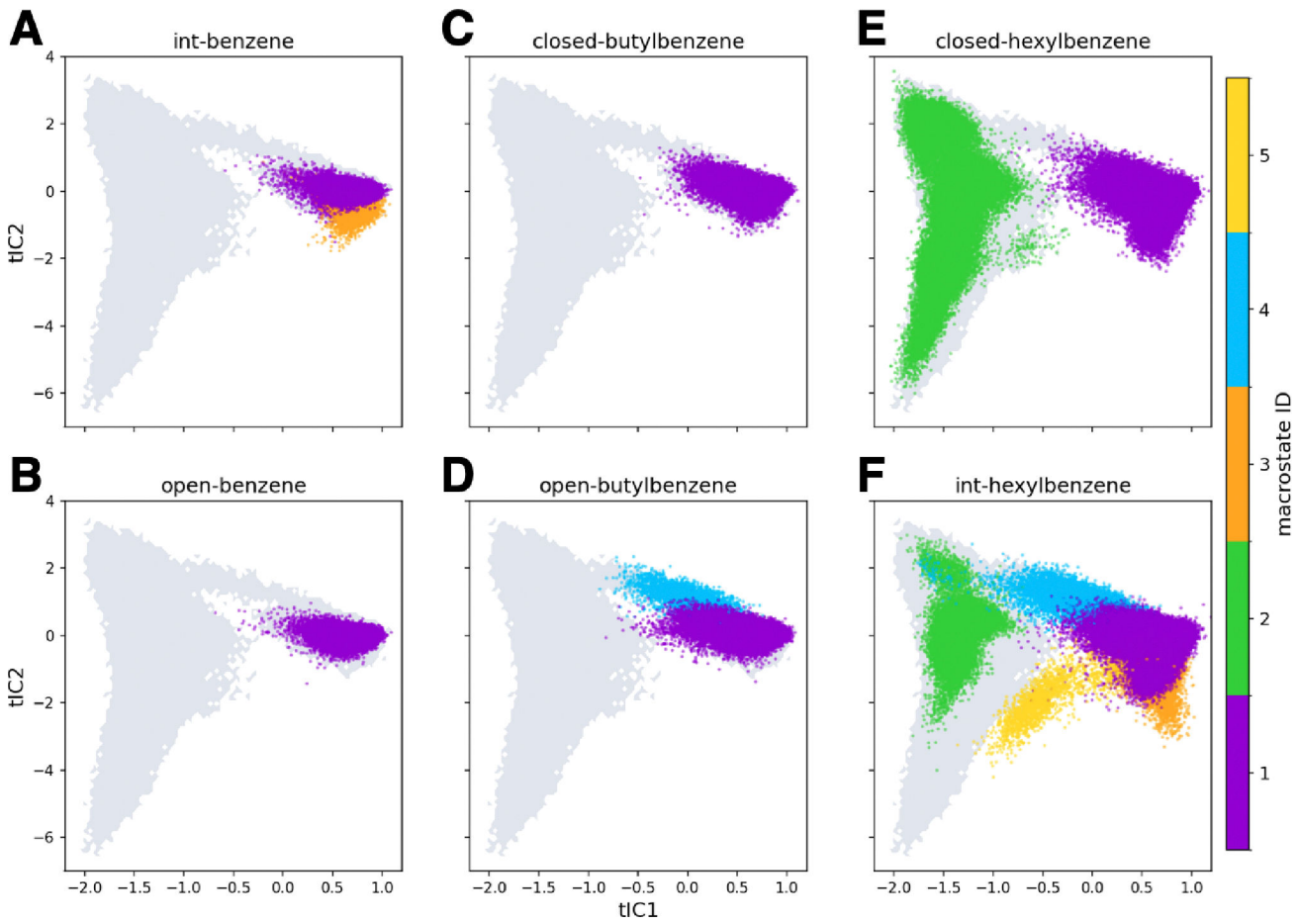


Figure 5. MSM-resolved states for each mixed system.

We initialize benzene, butylbenzene and hexylbenzene ligands each in their respective non-native crystal structures to observe potential changes in the F-helix. On top of the density of the concatenated trajectory (gray), we project the frames corresponding to simulations of the (A) int–benzene, (B) open–benzene, (C) closed–butylbenzene, (D) open–butylbenzene, (E) closed–hexylbenzene and (F) int–hexylbenzene systems. We color frames based on their MSM-assigned macrostates, keeping macrostate definitions consistent across the different native systems. On these timescales, simulations of benzene in T4 L99A primarily stay in the closed conformation when simulated beginning from either the (A) int or (B) open structures. Simulations of butylbenzene only sample the closed state when begun from the (C) closed structure but undergo interconversion with the intermediate state when begun from the (D) open structure. Simulations of hexylbenzene interconvert between the closed and open states without sampling the intermediate state when begun from the (E) closed structure. When begun from the (F) int structure, simulations of hexylbenzene sample the closed, intermediate and open states. In the int-hexylbenzene system, we also observe a state (yellow) not found in any other systems.

Table 1.

Matrix of different simulated systems. Entries in bold denote systems in which the bound ligand is the native ligand of the corresponding starting crystal structure.

Bound ligand	**Starting protein structure**		
	closed (4w52)	**int (4w57)**	**open (4w59)**
benzene	**closed–benzene**	int–benzene	open–benzene
butylbenzene	closed–butylbenzene	**int–butylbenzene**	open–butylbenzene
hexylbenzene	closed–hexylbenzene	int–hexylbenzene	**open–hexylbenzene**

J Chem Inf Model. Author manuscript; available in PMC 2025 February 11.

Table 2.

Observed transitions for all closed and mixed systems.

system	closed ↔ int	int ↔ open	open ↔ closed
closed–benzene	X	X	X
int–benzene	X	X	X
open–benzene	X	X	X
closed–butylbenzene	X	X	X
int–butylbenzene	X	X	X
open–butylbenzene	YES	X	X
closed–hexylbenzene	X	X	YES
int–hexylbenzene	YES	YES	YES
open–hexylbenzene	X	X	YES

Table 3.

Order of magnitude estimates for transitions between closed and intermediate states.

system	closed to int	int to closed
int–hexyl	2×10^2 ns	1×10^1 ns
open–butyl	5×10^2 ns	4 ns

Table 4.

Order of magnitude estimates for transitions between intermediate and open states.

system	int to open	open to int
int–hexyl	9×10^2 ns	1×10^2 ns

Table 5.

Order of magnitude estimates for transitions between closed and open states.

system	closed to open	open to closed
closed–hexyl	1×10^3 ns	1×10^2 ns
int–hexyl	1×10^3 ns	4×10^1 ns