bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Folding-upon-binding pathways of an intrinsically disordered protein from a deep Markov state model

Thomas Sisk & Paul Robustelli†

Dartmouth College, Department of Chemistry, Hanover, NH, 03755

† To whom correspondence should be addressed.

Paul Robustelli:

E-mail: Paul.J.Robustelli@Dartmouth.edu

Address: 6128 Burke Laboratory

Department of Chemistry

Hanover, NH, 03755

Keywords: Intrinsically Disordered Proteins, Molecular Recognition, Markov State Models, Deep Learning, Molecular Dynamics

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

A central challenge in the study of intrinsically disordered proteins is the characterization of the mechanisms by which they bind their physiological interaction partners. Here, we utilize a deep learning based Markov state modeling approach to characterize the folding-upon-binding pathways observed in a long-time scale molecular dynamics simulation of a disordered region of the measles virus nucleoprotein N_{TAIL} reversibly binding the X domain of the measles virus phosphoprotein complex. We find that folding-upon-binding predominantly occurs via two distinct encounter complexes that are differentiated by the binding orientation, helical content, and conformational heterogeneity of N_{TAIL}. We do not, however, find evidence for the existence of canonical conformational selection or induced fit binding pathways. We observe four kinetically separated native-like bound states that interconvert on time scales of eighty to five hundred nanoseconds. These bound states share a core set of native intermolecular contacts and stable N_{TAIL} helices and are differentiated by a sequential formation of native and non-native contacts and additional helical turns. Our analyses provide an atomic resolution structural description of intermediate states in a folding-upon-binding pathway and elucidate the nature of the kinetic barriers between metastable states in a dynamic and heterogenous, or “fuzzy”, protein complex.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Intrinsically disordered proteins (IDPs) are proteins that do not adopt stable tertiary structures in isolation under physiological conditions. IDPs are ubiquitous in eukaryotic proteomes and viruses; and play crucial functional roles in many cellular processes.[1-3] The biological functions of IDPs are often mediated by short sequence segments, referred to as linear motifs or molecular recognition elements, that interact with structured partner proteins.[4-6] The molecular recognition elements of IDPs populate a structurally diverse set of conformations in their unbound states and can adopt a similarly diverse set of conformations when bound to different physiological interaction partners.[7-10] This conformational plasticity enables IDPs to function as hubs in cellular signaling pathways, where they can form specific interactions with multiple binding partners.[11-13] The relative affinities of these interactions can be tuned by post-translational modifications or changes in the cellular environment allowing for sensitive spatial and temporal regulation of cellular processes mediated by IDP interactions.[11, 14-18]

The thermodynamics of IDP interactions are complex, and the relationships between their free and bound state structures are not straightforward.[19.] In some instances, IDPs undergo disorder-to-order transitions and adopt stable tertiary structures when bound to physiological binding partners; a process referred to as "folding-upon-binding".[5, 9, 20-22] In other instances, IDPs retain a substantial amount of conformational disorder in their bound states.[23-26] Such dynamic and heterogenous complexes are sometimes referred to as "fuzzy" complexes.[27, 28] Substantial effort has been made to characterize the kinetics and thermodynamics of IDP binding events[6, 9, 29-31], as elucidating the relationship between the free and bound states of IDPs will enable a more predictive understanding of their roles in biological pathways and human disease.[11, 32]

Stopped-flow and temperature-jump kinetics measurements[31, 33, 34], NMR spectroscopy[35-39], single molecule FRET[40-43] and protein engineering techniques[44-46] have emerged as powerful tools for characterizing the binding processes of IDPs. While these experimental techniques provide detailed mechanistic insight into IDP binding pathways, the data generated by these approaches are generally insufficient to obtain atomic resolution descriptions of the conformational states populated in IDP binding pathways. Atomistic descriptions of IDP binding intermediates and the

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

conformational states populated by IDPs in complexes with their physiological interaction partners are highly desirable as they may facilitate the development of rational drug design strategies for modulating the activity of IDPs implicated in the pathogenesis of diseases.[17,47, 48]

All-atom molecular dynamics (MD) computer simulations provide a powerful complement to biophysical experiments for characterizing conformational ensembles,[49-53] binding pathways[44, 46, 54-56] and bound states of IDPs.[48-53, 56-59] Long timescale MD simulations run with an accurate physical model, or *force field*, can provide atomically detailed structural descriptions of conformational substates involved in IDP binding. MD simulations with sufficient statistical sampling of binding events also provide the equilibrium populations of these states and the rates of transitions between them.[54, 55] Recent improvements to molecular mechanics force fields have dramatically enhanced the accuracy of MD simulations of disordered proteins and have shown promise for describing molecular recognition mechanisms of IDPs.[48, 52, 56, 58, 60, 61] As IDP binding pathways occur on rugged and high-dimensional free energy surfaces, identifying mechanistically meaningful metastable states in MD simulations of IDP remains a substantial challenge.

Markov State Models (MSMs) describe the dynamics of stochastic systems as a transition network of memoryless, probabilistic jumps between sets of states. MSMs are a powerful approach for obtaining mechanistic insight from MD simulations[62, 63] and have provided insights into protein conformational transitions[51, 64, 65], protein folding[66], protein-ligand binding[47, 55, 67] and protein-protein complex formation.[47, 54, 55, 66-69] The accuracy, interpretability, and relevance of information extracted from MSMs are, however, highly dependent on the input features used to describe a simulated system, the methods used to reduce the dimensionality of the input feature space and the partitioning of simulation frames into Markov states.[62, 70, 71] These tasks are particularly challenging when building MSMs to describe the high-dimensional conformational space of disordered proteins.[47, 51, 72]

In recent years, theoretical advancements and applications of machine learning techniques have facilitated the construction of MSMs from MD simulation data.[73] Automated feature selection, dimensionality reduction, and feature scoring methods can be applied to guide and validate the selection of molecular features to construct MSMs.[74-78] These methods identify subsets of slowly

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

evolving structural features, or *collective variables,* that can be used to partition MD trajectories into metastable Markov states that accurately model the kinetics of simulated conformational transitions.[76, 79, 80] The variational approach to Markov processes (VAMP) has emerged as a powerful framework to identify molecular features that describe the slowest evolving degrees of freedom in a simulated system.[80-83] In this approach a scoring function is used to quantify how effectively a set of features describes the kinetics of slow conformational transitions observed in MD simulations, and this score is maximized to identify optimal collective variables for MSM construction. The VAMP method has been extended to a deep learning framework where neural networks (referred to as "VAMPnets") are optimized to identify metastable conformational states directly from molecular features.[84] VAMPnet approaches have been further extended to include physical constraints in the training of neural networks that enable MSMs to be learned directly from simulation data.[85] These models, referred to as "deep reversible MSMs", "deep MSMs", or "Koopman Models", allow for the construction of kinetic models comprised of probabilistic states that may be differentiated by only subtle conformational features.[51, 85]

In this investigation, we have built a conventional MSM and a deep learning based MSM (or "deep MSM") to characterize the folding-upon-binding pathways observed in a 200μs unbiased MD simulation of the α-helical molecular recognition element of the measles virus nucleoprotein N_{TAIL} reversibly binding the X domain (XD) of the measles virus phosphoprotein complex.[56] The conformational dynamics of measles virus N_{TAIL} in solution and the folding-upon-binding of N_{TAIL} to XD have been extensively characterized by a variety of experimental[33, 36, 86-91], and computational methods.[56, 92-94] Here, we construct a hidden Markov state model[95] using time-lagged independent component analysis (tICA)[79, 80, 96, 97], a linear dimensionality reduction technique, and a deep MSM by applying the VAMPnet approach with physical constraints.[85] Our deep MSM employs a multi-input neural network architecture that utilizes a combination of convolutional and fully connected neural network layers to merge structural descriptors with different inherent dimensionalities.

We find that the deep MSM identifies several states that were not identified by a conventional hidden Markov state model. The hidden Markov state model identifies a single heterogenous encounter complex state between N_{TAIL} and XD and a single heterogenous non-native complex

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where N_{TAIL} binds on the opposite face of XD relative native binding site. The deep MSM resolves two structurally and kinetically distinct encounter complex states that are differentiated by the binding orientation and helical content of N_{TAIL} as well as a kinetic trap on the native folding upon binding pathway. The deep MSM also identifies a network of several distinct non-native bound complexes. The hidden Markov state model and deep MSM both resolve 4 kinetically separated bound native-like states that interconvert on time scales of eighty to five hundred nanoseconds. These bound states share a core set of native intermolecular contacts and stable helices and are differentiated by a sequential formation of non-native contacts that facilitate the folding of additional helical turns. Interestingly, the detailed molecular mechanisms of folding-upon-binding revealed by our MSMs are not consistent with canonical conformational selection or induced-fit folding-upon-binding mechanisms. We find that encounter complexes that contain highly helical N_{TAIL} conformations proceed to the fully folded N_{TAIL}:XD complex through a similar network of states as encounter complexes where N_{TAIL} has little helical structure.

Our analyses provide an atomic resolution structural and kinetic description of intermediate states in a folding-upon-binding pathway and elucidate the nature of the kinetic barriers between metastable states in a dynamic and heterogenous, or "fuzzy", protein complex[10, 26-28, 98] formed by an IDP and a structured binding partner. The neural network architecture designed here to train a deep MSM merges convolutional neural network layers that reduce the dimensionality of intermolecular contact matrices with fully connected network layers to describe global structural features. This neural network identifies several conformational states that were not resolved utilizing a reaction coordinate approach, time-lagged independent component analysis (tICA), or a conventional neural network architecture employing only fully connected neural network layers. These states enhance the resolution of the folding-upon-binding mechanism and suggest that folding-upon-binding proceeds through binding pathways that are inconsistent with canonical conformational selection or induced-fit binding mechanisms. This multi-input neural network approach may provide a general strategy for building deep MSMs to model the highly dynamic conformational states of IDPs and protein complexes with substantial conformational disorder.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Molecular dynamics simulation of the measles virus nucleoprotein N_{TAIL} and the X domain of the measles virus phosphoprotein complex. A 200μs explicit solvent unbiased MD simulation of a 21-residue partially helical molecular recognition element of the measles virus nucleoprotein N_{TAIL} (residues 484-504, henceforth referred to as "N_{TAIL}") and the X domain (XD) of the measles virus phosphoprotein complex was previously performed by Robustelli et. al[56] using the Anton[99] supercomputer. This simulation was performed at 400 K using the a99SB-disp protein force field and a99SB-disp water model.[52] A temperature of 400 K was selected for long time scale folding-upon-binding simulations as it was found to be near the simulated melting temperature of the N_{TAIL}:XD complex and enabled an efficient sampling of binding and unbinding transitions in an equilibrium simulation. This simulation was initiated from an unbound conformation of N_{TAIL} and contains 36 binding and 36 unbinding events, where binding and unbinding events are defined using the fraction of native intermolecular contacts (Q)[56, 100] as a reaction coordinate (See Methods). Here, we observed that XD unfolds at the beginning of this trajectory and refolds to its native state after 3 μs of simulation time and that XD unfolds and refolds multiple times in the final 30 μs of the trajectory. As we are only interested in modeling the binding pathways of N_{TAIL} to the native state of XD, we restricted our analysis to a continuous 167 μs subset of the original MD trajectory (from t=3 μs to t=170 μs) where XD remained in its native conformation. This 167 μs segment of the original trajectory contains 831701 frames, spaced with an interval of 200 ps per frame. We refer to this 167 μs segment as the "full trajectory".

Markov state model input features. We considered a set of input features containing 1029 intermolecular distances (one distance between each of the 21x49 intermolecular pairs of residues in N_{TAIL} and XD), 21 binary features based on the DSSP secondary structure assignment[101] of each residue of N_{TAIL}, and 15 features consisting of the value of the helical order parameter Sα[102] for each consecutive seven residue fragment of N_{TAIL} (See Methods). We refer to sum of Sα values for all 15 seven residue fragments of N_{TAIL} as "N_{TAIL} Sα". We consider a total of 1065 features for each MD simulation frame to build an 831701 x 1065 input feature matrix.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Constructing a hidden Markov state model (HMSM) from time-lagged independent component analysis (tICA). We utilized time-lagged independent component analysis (tICA)[79, 80, 96, 97] to reduce the dimensionality of the N_{TAIL}:XD input feature matrix and build an initial MSM. tICA was performed on the input feature matrix using a lag time of 6 ns and the resulting data were projected onto the first ten tICA eigenvectors. Initial analyses revealed that the binary DSSP assignment features had no impact on tICA projections and subsequent analyses, and they were subsequently excluded from the input features for building MSMs from tICA (See Methods). We visualize the free energy surface of the N_{TAIL}:XD folding-upon-binding MD trajectory as a function of the two dominant time-lagged independent components (TICs) in Supplementary Figure 1. We observe that this projection resolves 4 distinct bound-state free energy basins that resemble the native N_{TAIL}:XD complex observed by x-ray crystallography (PDB ID 1T6O)[86] . We determined an initial estimate of the optimal number of states for an MSM derived from the first ten tICA eigenvectors by iteratively applying the *k*-means algorithm with an increasing number of clusters until the resultant states no longer had statistically distinguishable properties in terms of the fraction of native intermolecular contacts (Q), Sα, radius of gyration (R_g) and root mean squared deviation (RMSD) from the native complex. Using this approach, we found seven clusters to be optimal. We estimated a traditional MSM using these clusters as state definitions and a lag time of 24 ns. The implied timescales (ITS) of this model, however, were not converged or fully resolved. This MSM also failed to satisfy the generalized Chapman- Kolmogorov (CK) test[62] (eq. 5), failing to reproduce the fastest processes observed in this system (data not shown).

To produce a valid model, we constructed an MSM with a larger numbers of initial states and coarse grained them to a smaller number states via the HMSM formulism introduced by Noe et al.[95] We found that coarsening an initial twelve state MSM with seven resolved implied timescales (including the stationary process) to a seven state HMSM with a lag time of 6 ns yielded resolved and converged implied timescales and a valid CK-test (Supplementary Figure 2). We refer to this model as the "tICA HMSM". We number these states HMSM state 1-7 in ascending order based on their similarity to the native complex, as assessed by the average values of the native intermolecular contact fraction ($\langle Q \rangle$), N_{TAIL} Sα ($\langle N_{TAIL}$ Sα $\rangle$), R_g ($\langle R_g \rangle$) and RMSD from the crystal structure of the native complex calculated from all structures in each state (Supplementary Figure 3 and Supplementary Table 1). A network representation of the tICA HMSM with structural

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

depictions of each state with the calculated mean first passage times (MFPTs) between them is displayed in Figure 1.

The HMSM state assignments are projected onto the two dominant tICs in Supplementary Figure 4. We visualize the free energy surface of each HMSM state as a function of the fraction of native intermolecular contacts (Q) and N_{TAIL} Sα in Supplementary Figure 5. The average values and standard deviations of Q and N_{TAIL} Sα for each HMSM state are compared in Supplementary Table 1 and Supplementary Figure 6. The populations of native and non-native N_{TAIL}:XD intermolecular contacts and the N_{TAIL} helical propensities for each tICA HMSM state are compared in Supplementary Figure 7. The transition matrix of the HMSM is shown in Supplementary Figure 8 and the calculated MFPTs are shown in Supplementary Figure 9.

In HMSM state 1 N_{TAIL} adopts highly helical conformations ($\langle N_{TAIL}$ Sα$\rangle$ = 10.9). These conformations have comparable helicity to the N_{TAIL} conformation observed in the native N_{TAIL}:XD complex (N_{TAIL} Sα = 12.8 in PDB 1T6O) with the exception of helical fraying observed in the N-terminal N_{TAIL} residues G484-D487 and the C-terminal N_{TAIL} residues A502-I504. The average values of native intermolecular contacts $\langle Q \rangle$ are 0.93, 0.91, 0.79. and 0.78 and the average values of $\langle N_{TAIL}$ Sα$\rangle$ are 10.9, 7.7, 5.6 and 4.8 for HMSM states 1-4, respectively. These 4 states contain stable helical conformations from N_{TAIL} A502 to A494 and are differentiated by the extension of stable N_{TAIL} helical conformations from N_{TAIL} A502 to D493, N_{TAIL} A502 to S491, and N_{TAIL} A502 to D487 in HMSM states 3, 2 and 1 respectively (Supplementary Figure 7). The R_g of the bound states increases from HMSM state 1 to HMSM state 4 as an increasing number of N-terminal residues of less helical conformations of N_{TAIL} extend outward from XD into solution (Supplementary Table 1). Our tICA HMSM also identifies a weakly bound state (HMSM state 5) with a small fraction of native intermolecular contacts and little helical content ($\langle Q \rangle$ = 0.16, $\langle N_{TAIL}$ Sα$\rangle$ = 3.1), a state where N_{TAIL} and XD are largely unbound (HMSM state 6) with a substantially elevated R_g ($\langle Q \rangle$ = 0.01, $\langle N_{TAIL}$ Sα$\rangle$ = 1.4, $\langle R_g \rangle$ = 1.8 nm) and a more compact non-native complex (HMSM state 7) with very few native contacts ($\langle Q \rangle$ = 0.02, $\langle N_{TAIL}$ Sα$\rangle$ = 3.6, $\langle R_g \rangle$ = 1.3 nm) but more N_{TAIL} helical content than unbound N_{TAIL} conformations.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We observe that HMSM state 5 functions as a kinetic hub between unbound conformations in HMSM state 6 and the 4 native-like bound states (Figure 2, Supplementary Figure 8). HMSM State 5 can therefore be interpreted as an on-pathway encounter complex in the folding-upon-binding of pathway N_{TAIL}. The most probable transitions from HMSM state 5 to the native-like bound states are to states 3 and 4, where N_{TAIL} is partially folded, with transition probabilities of $2.53 \pm 0.4\%$ and $1.48 \pm 0.3\%$, respectively (error estimates computed with a Bayesian HMSM and Gibbs sampling approach[80,112], See Methods). From HMSM state 3, transitions to state 2 ($7.33 \pm 0.53\%$) are significantly more probable than to the less helical state 4 ($4.85 \pm 0.4\%$). HMSM state 4 has a relatively large probability of transitioning to state 3 ($12.7 \pm 1.03\%$) and very low probabilities of transitioning to states 1 ($0.4 \pm 0.1\%$) and 2 ($1.1 \pm 0.2\%$). Using Transition path theory (TPT)[103-105], we find that folding-upon-binding pathways from HMSM state 6 (unbound) to states 1 and 2 (most native-like bound states) that exclude visits to state 4 comprise 74.8% of the total probability flux and that the pathway with the maximum flux (46.1%) proceeds through states 5, 3, and 2. We conclude that HMSM state 4 is largely off pathway to the more folded, bound states. We observe that HMSM state 7 consists of a non-native N_{TAIL}:XD complex where N_{TAIL} is bound on the opposite face of XD relative to the native binding groove. Conformations in HMSM state 7 predominantly transition back to unbound conformations in HMSM state 6 (transition probability of $4.17 \pm 0.78\%$) and very rarely transition directly to HMSM state 5 (transition probability of $0.1 \pm 0.1\%$).

Constructing a deep Markov state model with a multi-input neural network architecture. We sought to improve the resolution of our kinetic model and obtain greater mechanistic insight into N_{TAIL}:XD folding-upon-binding by employing the deep learning "VAMPnet" approach with physical constraints to build a deep MSM.[85] In this approach, the variational approach to Markov processes (VAMP) is integrated into a deep learning framework that combines feature selection, dimensionality reduction, state discretization, and kinetic modeling into a continuous pipeline for constructing MSMs. The VAMP provides a "VAMP score" that estimates how well a set of features describes the kinetics of the slowest evolving transitions observed in an MD simulation.[76, 81-83] In a VAMPnet, a neural network is trained to learn a non-linear function that transforms input features into probabilistic state assignments that maximize the VAMP score. A VAMPnet outputs a probabilistic (or "fuzzy") Markov state assignment for each frame of an MD simulation

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

trajectory. Probabilistic state assignments describe the probability that each trajectory frame is a member of each Markov state. Higher VAMP scores result from probabilistic MSM state assignments that maximize the autocorrelation of each state assignment. Training neural networks to maximize VAMP scores therefore identifies slowly evolving state definitions describing metastable intermediates in long timescale processes.

Mardt et. al. extended the VAMPnet approach to learn a stochastic and reversible transition matrix defining the transition probabilities between fuzzy states obtained from an unconstrained VAMPnet.[84, 85] A reversible and stochastic transition matrix adheres to detailed balance and has all positive elements so each element can therefore be interpreted as a transition probability. The learned deep MSM state assignments and reversible, stochastic transition matrix define a kinetic model from which the stationary distribution of states and their interconversion rates can be computed. These models have been referred to as "deep MSMs", "VAMPnets with physical constraints" and "Koopman models" in previous studies due to their relationship with Koopman operator theory.[106] Deep MSMs pose a great advantage over traditional MSMs as the utilization of neural networks in these models allow for the optimization of non-linear state membership functions.

We used the full set of 1065 input features to learn a deep MSM with a VAMPnet with physical constraints. We refer to this MSM as the "deep MSM". To optimally integrate features that describe the helical content of N_{TAIL} (Sα and binary DSSP) and features that describe the position and orientation of N_{TAIL} relative to XD (the N_{TAIL}:XD intermolecular distance matrix) in our VAMPnet, we designed a multi-input neural network architecture. A schematic illustration of this multi-input neural network architecture is presented in Figure 2. This neural network architecture employs a combination of convolutional network layers and fully connected network layers to merge structural descriptors with different dimensionalities. Convolutional neural networks provide dramatic performance advantages for deep learning tasks involving image data.[107] Recognizing that the intermolecular distances matrix (or intermolecular "contact map") between N_{TAIL} and XD obtained in each frame of the simulation can be interpreted as an image, we sought to leverage the local spatial coherence in these contact maps by transforming them with convolutional neural network layers in our VAMPnet. We then combine the information obtained

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from convolutional neural network layer transformations of intermolecular contact maps with information obtained from fully connected dense neural network layer transformations of the Sα and binary DSSP helical assignment features.

The three neural network inputs (intermolecular distance matrices, N_{TAIL} Sα values and binary DSSP N_{TAIL} helical assignments) are transformed separately in three branches, applying convolutional neural network layers to transform intermolecular contact maps and fully connected neural network layers to transform the vector quantities of Sα and binary DSSP helix assignments (See Methods). The resulting outputs from each branch of the network are combined and transformed by a final set of fully connected neural network layers. The details of the final architecture of this neural network are described and illustrated in Supplementary Figure 10. The initial fully connected neural network layers used to transform Sα values and binary helical DSSP assignments increase the dimensionality of these data to better capture relationships between different sequence regions in N_{TAIL} and the initial convolutional network layers reduce the dimensionality of intermolecular contact maps to better capture essential relationships between intermolecular contacts in different regions of the N_{TAIL}:XD complex with a coarser representation of intermolecular distances.

We determined the final architecture of our neural network implementation and VAMPnet hyperparameters (batch size, learning rate, epsilon parameter, model lag time, and number of states) by iteratively optimizing the VAMP2 score (eq. 8) of an unconstrained neural network (See Methods). We found that using 12 output states and a lag time of 2 ns to train unconstrained VAMPnets maximized the VAMP2 score and consistently produced the same set of 12 distinguishable states. We characterize the latent space and state assignments of the initial unconstrained VAMPnet in Figure 3.

We constructed our final deep MSM by retraining the initial unconstrained VAMPnet with physical constraints to learn a reversible and stochastic transition matrix defining the transition probabilities between the 12 states identified by the unconstrained VAMPnet (See Methods).[84, 85] The Chapman-Kolmogrov (CK) test[62], implied timescales, and steady state distributions for the deep MSM estimated at a lag time of 6 ns are shown in Supplementary Figure 11. We refer to the

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

12 states obtained from the deep MSM as deep MSM states 1-12. We number the states of the deep MSM in ascending order based on their similarity to the native N_{TAIL}:XD complex in terms of the fraction of native intermolecular contacts (Q), N_{TAIL} Sα, radius of gyration and RMSD from the native complex (Supplementary Figure 12). We visualize the free energy surface of each deep MSM state as a function of Q and N_{TAIL} Sα in Supplementary Figure 13. We compare the average values and standard deviations of Q, N_{TAIL} Sα and the radius of gyration for each deep MSM state in Supplementary Table 2 and Supplementary Figure 14. We compare the populations of native and non-native N_{TAIL}:XD intermolecular contacts and the N_{TAIL} helical propensities for each deep MSM state in Supplementary Figure 15. The transition matrix and the mean first passage times for the deep MSM are shown in Supplementary Figures 16 and 17, respectively.

A transition network representation of the deep MSM with structural depictions of each state and the mean first passage times between states is displayed in Figure 4. We observe that 5 of the deep MSM states closely resemble 5 of the tICA HMSM states. Deep MSM states 1-4 closely resemble the 4 native-like HMSM bound states (HMSM states 1-4). Deep MSM state 8, where N_{TAIL} is unbound, closely resembles HMSM state 6. In the tICA HMSM, we resolve a single heterogenous encounter complex state (HMSM state 5). The deep MSM increases the resolution of our model and effectively fine grains this heterogenous encounter complex into 3 distinct states: deep MSM states 5, 6 and 7. These states are substantially more homogenous than HMSM state 5 and are differentiated by the helical content of N_{TAIL}, the orientation of N_{TAIL} relative to XD, the conformational heterogeneity of N_{TAIL} and the populations of native and non-native intermolecular contacts (Figures 4-5, Supplementary Figures 12-15).

The deep MSM similarly fine grains HMSM state 7, the heterogenous non-native complex where N_{TAIL} is bound to the opposite face of XD relative to the native binding site, into 3 distinct states (deep MSM states 10-12, Figure 5, Supplementary Figures 12-15). In these more homogenous states N_{TAIL} is bound in different locations on XD and contains distinct populations of helical content. In addition, the VAMPnet also identifies a rare conformational state (deep MSM state 9, steady-state population $p = 0.1 \pm 0.02\%$) in which N_{TAIL} is inserted between the three helical bundles of XD.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A deep Markov state model resolves two structurally and kinetically distinct encounter complex states and a kinetic trap. In the tICA HMSM, most of the probability flux from unbound N_{TAIL} states to native-like bound states flows through a single Markov state (tICA HMSM state 5) which functions as an encounter complex and kinetic hub for transitions between bound and unbound conformations (Figure 2, Supplementary Figures 8-9). tICA HMSM state 5 has a steady-state population (p) of $p = 8.7 \pm 1.1\%$ and contains a small fraction of native intermolecular contacts ($\langle Q \rangle = 0.22$) and relatively little helical content ($\langle N_{TAIL}\ S\alpha \rangle = 3.1$). In the deep MSM this state has effectively been split into three states: deep MSM states 5, 6, and 7 (Figures 4-5). Deep MSM states 5, 6 and 7 have steady state populations of $p = 1.1 \pm 0.2\%$, $p = 5.7 \pm 0.5\%$ and $p = 3.0 \pm 0.3\%$, respectively. We observe that the populations of helical N_{TAIL} conformations are substantially smaller in deep MSM state 5 ($\langle N_{TAIL}\ S\alpha \rangle = 1.4$) and deep MSM state 6 ($\langle N_{TAIL}\ S\alpha \rangle = 2.0$) compared to deep MSM state 7 ($\langle N_{TAIL}\ S\alpha \rangle = 5.1$). We find that deep MSM states 5, 6 and 7 have similar fractions of native intermolecular contacts ($\langle Q \rangle = 0.19$, $\langle Q \rangle = 0.19$, and $\langle Q \rangle = 0.18$, respectively) but observe that there is a large difference in the subsets of the intermolecular residue pairs that form native and non-native intermolecular contacts in each state (Figure 5).

N_{TAIL} residues L495 and L498 insert into the hydrophobic binding groove of XD in the native complex. In deep MSM state 6 these leucine residues form similar populations of native and non-native intermolecular contacts and N_{TAIL} is not restricted to native-like binding orientations, and instead samples a relatively isotropic distribution of rotational orientations. In deep MSM state 7, native intermolecular contacts formed by N_{TAIL} L498 have substantially higher populations than native intermolecular contacts formed by N_{TAIL} L495, and N_{TAIL} L495 forms highly populated non-native intermolecular contacts. Visual inspection of deep MSM state 6 and state 7 reveals that N_{TAIL} L498 binds at similar positions in the native XD hydrophobic binding groove in both states (Figure 5). In deep MSM state 7, however, N_{TAIL} L495 is inserted into a non-native binding site in the hydrophobic binding groove of XD that orients N_{TAIL} in the opposite (or "upside-down") orientation relative to the N_{TAIL} orientation observed in the native N_{TAIL}:XD bound complex. We define a rotational order parameter in the form of an angle to quantify the orientation of N_{TAIL} relative to the native binding face of XD in each deep MSM state in Supplementary Appendix 1 and present the distribution of this order parameter for each deep MSM state in Supplementary Figure 18.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

N_{TAIL} conformations in deep MSM state 6 have a similar helical propensity to unbound states of N_{TAIL}, except for a slightly elevated helical propensity observed in residues A492-L495 (Figure 5, Supplementary Figure 15). In deep MSM state 7, N_{TAIL} has a higher helical propensity that more closely resembles the less helical native-like bound states (deep MSM states 3 and 4). One might therefore hypothesize that deep MSM state 6 functions as an encounter complex for a binding pathway resembling an "induced fit" mechanism, where the formation of native intermolecular contacts proceeds the subsequent folding of secondary structure elements formed the bound state, while deep MSM state 7 functions as an encounter complex for a parallel binding pathway resembling a "conformational selection" mechanism, where preformed native-like secondary structure elements bind XD before the subsequent formation of native intermolecular contacts. A detailed inspection of the transition probabilities and transition rates among deep MSM states, however, reveals that N_{TAIL} binding pathways do not fall into such a dichotomy (Figure 4, Supplementary Figures 16-17).

While N_{TAIL} conformations in deep MSM state 7 are substantially more helical than N_{TAIL} conformations in state 6, we do not observe greater transition probabilities from state 7 to the more helical native-like bound states 1 and 2 (Figure 5, Supplementary Figure 16). The transition probabilities from deep MSM state 7 to states 1 and 2 are $0.1 \pm 0.01\%$ and $0.5 \pm 0.1\%$, respectively. These values are smaller than the transition probabilities observed from the less helical encounter complex (deep MSM state 6) to states 1 and 2 ($0.6 \pm 0.1\%$ and $1.7 \pm 0.2\%$, respectively). The highest transition probabilities from deep MSM state 7 are to state 6 ($15.3 \pm 0.8\%$) and state 4 ($4.9 \pm 0.7\%$), states where N_{TAIL} is substantially less helical.

These observations contrast with the classical paradigm of conformational selection, where a stable, preformed helix binds and remains helical for the duration of a binding event. We observe that the transition rates from the two deep MSM encounter complex states (states 6 and 7) to the deep MSM native-like bound states (states 1-4) are within statistical error (Supplementary Figure 17) and that deep MSM states 6 and 7 are most clearly kinetically distinguished based on incoming transitions from unbound and non-native conformations (Supplementary Figure 16). These results indicate that while we identify distinct encounter complex states with different N_{TAIL} helical

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

propensities and conformational pathways leading to their formation, these states ultimately transition to native-like bound states with similar rates and ultimately form the same network of partially bound and folded fuzzy complexes that subsequently transition to the most native-like state. Consequently, we conclude that folding-upon-binding pathways originating from these encounter complex states are not well described by an induced fit / conformational selection dichotomy.

Deep MSM state 5 transitions almost exclusively to state 6 which is the only state that has an appreciable probability of transitioning to state 5 (Supplementary Figure 16). Consequently, we identify deep MSM state 5 as an off-pathway kinetic trap on folding-upon-binding pathways that proceed through state 6. Deep MSM state 5 is similar to state 6 but N_{TAIL} has an elevated helical propensity in residues A492-L495 (Figure 5). Deep MSM state 5 contains more highly populated non-native contacts between N_{TAIL} residues L495 and L496 and XD residues Y480, L481, L484, F497, and I504 (average population of 55.7 $\pm$ 7.33%) than state 6 and state 7 (average populations of 19.7 + 2.4% and 12.3 $\pm$ 4.7%, respectively). We thus identify the stabilization of helical conformations of N_{TAIL} by the formation of non-native contacts as the basis for the substantial kinetic barrier observed between deep MSM state 5 and the native-like bound states.

Kinetic barriers between native-like bound states originate from non-native contacts. N_{TAIL} folding-upon-binding pathways from encounter complex states (deep MSM states 6 and 7) to the nost native-like bound state (deep MSM state 1) are largely mediated by the sequential formation and subsequent breakage of two distinct sets of non-native intermolecular contacts. The majority of the probability flux from the deep MSM encounter complex states to the native state states proceeds through deep MSM states 3 and 4. These states contain similar N_{TAIL} helical propensites and populations of native intermolecular contacts, but are differentiated by an elevated population of a cluster of non-native intermolecular contacts between N_{TAIL} residues A492 and D493 and XD residues D487, I488, and D493 in deep MSM state 3 (Figure 6). This cluster of non-native intermolecular contacts is highlighted by a dotted rectangle in Figure 6A, and representative depictions of these contacts are shown in Figure 6B. The average population of the non-native contacts between these groups of residues is 32.9 $\pm$ 0.7% in deep MSM state 3 compared to 3.9 $\pm$ 3.8% in state 4.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Deep MSM state 3 contains a substantially populated intramolecular salt bridge between residues N_{TAIL} R497 and N_{TAIL} D493. We define this salt bridge as being formed in trajectory frames where one of the carbonyl oxygens of N_{TAIL} D493 is within 3.5 Å of a guanidinium nitrogen of N_{TAIL} R497. By this definition, the N_{TAIL} R497:D493 salt bridge has a population of 4.1 $\pm$ 0.7% in state 4 and 26.6 $\pm$ 0.1% in state 3. These results suggests that the kinetic barrier between deep MSM state 4 and state 3 partially results from the process of forming and breaking the intramolecular N_{TAIL} R497:D493 salt bridge and non-native intermolecular contacts between N_{TAIL} A492 and D493 and XD residues D487, I488, and D493. We observe that the process of forming these contacts is substantially faster than the process of breaking them (MFPT = 80.0 $\pm$ 3.4 ns for transitions from deep MSM state 4 to state 3 and MFPT = 274.1 $\pm$ 27.4 ns for transitions from state 3 to state 4). Interestingly, it has been observed that the N_{TAIL} mutation R497G substantially diminishes the affinity of N_{TAIL} to XD.[108] K_D values of 3.0 $\pm$ 0.2 μM and 44.4 $\pm$ 2.2 μM were measured for wild type and R497G N_{TAIL}, respectively. N_{TAIL} R497 forms stable native intermolecular contacts with XD in all the deep MSM native-like bound states. The absence of these native intermolecular interactions should destabilize the native complex between the N_{TAIL} R497G mutant and XD. The absence of an intramolecular salt bridge between N_{TAIL} R497 and D493 may further destabilize deep MSM state 3. As most of the total probability flux (70.7 $\pm$ 6.0%) from the unbound state (deep MSM state 8) to most native-like bound (state 1) proceeds through state 3, this additional destabilization of state 3 may contribute to the dramatic affinity loss observed for N_{TAIL} R497G observed in previous studies.

The formation of non-native intermolecular contacts in deep MSM state 3 coincides with the transient formation of several weakly populated native intermolecular contacts between N_{TAIL} residues R490 and S491 with XD residues D487, I488, and D493 (average population of 14.0 $\pm$ 4.4%, dark rectangle, Figure 6B). These native contacts subsequently become “locked in” after transitions to deep MSM state 2, where they have an average population of 86.7 $\pm$ 5.4%. The formation of these stable native intermolecular contacts is accompanied by a substantial increase in the population of intermolecular hydrogen bonds between the sidechain hydroxyl hydrogen of N_{TAIL} S491 and the carboxylic acid oxygens of XD D493 and the hydroxyl oxygen of N_{TAIL} S491 and the backbone amide hydrogen of XD K489. These hydrogen bonds are observed in the x-ray

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

structure of the N_{TAIL}:XD complex[86] and the N_{TAIL} mutation S491L was previously demonstrated to reduce the affinity of N_{TAIL} to XD beneath the detection limits of ITC[108], underscoring the importance of these intermolecular hydrogen bonds in stabilizing the N_{TAIL}:XD complex. These hydrogen bonds have a population of 53.0 $\pm$ 0.4% in deep MSM state 2 compared to in 6.6 $\pm$ 0.1% and 0.2 $\pm$ 0.1% of frames in states 3 and 4, respectively. The formation of this cluster of native contacts in deep MSM state 2 is accompanied by an increase in the helical propensities of N_{TAIL} residues S491-D493, and the formation of several non-native intermolecular contacts between N_{TAIL} residue R489 and XD residues T483, D486 and D487 (average population = 49.3 $\pm$ 24.3%). The strongest non-native intermolecular contacts in this cluster occur between N_{TAIL} R489 and XD D487 (p = 82.5 $\pm$ 0.76%) and N_{TAIL} R489 and XD D486 (p = 40.2 $\pm$ 0.4%), demonstrating the importance of non-native intermolecular salt bridge interactions in stabilizing this state.

The stability of non-native contacts formed by N_{TAIL} R489 and XD residues T483, D486 and D487 appear to substantially contribute to the kinetic barrier between deep MSM state 2 and state 1. These contacts have an average population of 49.3 $\pm$ 24.3% in deep MSM state 2 but are nearly absent in state 1 (average population = 2.0 $\pm$ 1.7%). Transitions from deep MSM state 2 to state 1 are also accompanied by the formation of stable helical conformations from N_{TAIL} S491 to D487 and the formation of a final set of native intermolecular contacts between N_{TAIL} D487 and XD D487 and N_{TAIL} D467 and XD K489 (p = 37.7 $\pm$ 0.3% and p = 43.2 $\pm$ 0.3% respectively in deep MSM state 1). These native intermolecular contacts are indicated by a solid block box in Figure 4A. Transitions between deep MSM state 2 and state 1 are relatively fast (MFPT = 109.1 ± 7.2 ns for transitions from state 2 to state 1 and MFPT = 99.8 ± 10.7 ns for transitions from state 1 to state 2) and are among the fastest of the transitions observed between native-like bound states. This transition involves the cooperative extension of the N_{TAIL} helix by 4 residues, whereas the helix of N_{TAIL} is extended by only a single residue in transitions from deep MSM state 4 to state 3. The transition from deep MSM state 2 to state 1 involves the formation of a favorable salt bridge between N_{TAIL} D487 and XD K489 in a conformation where the aliphatic residues of N_{TAIL} D487 and XD D487 sidechains are in contact, but the negatively charged carboxylic acid moieties are orientated to minimize unfavorable charge interactions. We speculate that the strong electrostatic attractions and repulsions between this set of charged sidechains may facilitate the relatively fast transitions observed between deep MSM state 2 and state 1.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Comparison of Markov state models with a 1D reaction coordinate for folding-upon-binding. In a previous investigation by Robustelli et. al[56] a 1D reaction coordinate was optimized to characterize the folding-upon-binding mechanism observed in the MD simulation analyzed here. This reaction coordinate was derived using the fraction of native intermolecular contacts (Q) between N_{TAIL} and XD as an initial reaction coordinate and employing the variational optimization approach of Best and Hummer[109] to reweight the contribution of each native intermolecular contact to produce a new reaction coordinate (R). This optimization was carried out to increase the maximum value of the conditional probability distribution p(TP|R), where p(TP|R) is the probability that a frame of the MD trajectory is on transition path at a given value of the optimized reaction coordinate R.

A projection of the MD trajectory onto the previously calculated 1D reaction coordinate R was found to contain three apparent free-energy minima separating unbound and native-like bound conformations (Supplementary Figure 19). It is, however, unclear if the apparent free-energy barriers observed in this projection are kinetically meaningful. We have calculated the probability distribution of the value of the reaction coordinate R for each kinetically distinct deep MSM state in Supplementary Figure 19. We observe that the two primary encounter complex states identified in this investigation (deep MSM states 6 and 7) are largely indistinguishable based on this reaction coordinate. We also observe that native-like bound states of the deep MSM (deep MSM states 1-4) are similarly indistinguishable based on this reaction coordinate. This result is unsurprising given the importance of non-native contacts in differentiating the Markov states of our deep MSM and underscores the complementary insights that MSMs can provide to low dimensional reaction coordinate approaches for describing protein folding and disordered protein folding-upon-binding.

Discussion

We report the construction of Markov state models (MSMs) to structurally and kinetically characterize folding-upon-pathways observed in an unbiased long time scale MD simulation of a disordered molecular recognition element of the measles virus nucleoprotein N_{TAIL} reversibly binding the X domain of the measles virus phosphoprotein complex. We constructed a hidden Markov state model (HMSM) using time-lagged independent component analysis (tICA), a linear

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

dimensionality reduction technique, and a deep learning based MSM (or "deep MSM") using the VAMPnet approach with physical constraints with a multi-input neural network architecture. The MSMs constructed with these two approaches both resolve an unbound state and 4 kinetically separated native-like bound states that interconvert on time scales of eighty to five hundred nanoseconds. In the HMSM built using tICA, we observe that transitions between unbound N_{TAIL} conformations and native-like bound states of N_{TAIL}:XD complexes predominantly occur through a single conformationally heterogenous Markov state, which we refer to as an "encounter complex" state. In contrast, the deep MSM built using the reversible VAMPnet approach resolves several additional structurally and kinetically distinct states including two encounter complexes and an off-pathway kinetic trap.

In both encounter complex states identified in the deep MSM N_{TAIL} residue L498 is inserted into the hydrophobic binding groove of XD in its native binding site. These encounter complex states are differentiated by the binding orientation, helical content, and conformational heterogeneity of N_{TAIL}. In one encounter complex state N_{TAIL} adopts relatively disordered conformations with similar helical content to unbound N_{TAIL} conformations and samples a relatively isotropic distribution of rotational orientations relative the binding face of XD. In the second encounter complex state N_{TAIL} adopts a more ordered set of conformations with substantially more helical content than is observed in its unbound state and predominantly binds XD in a single orientation that is "upside-down" relative to its orientation in the native complex. This upside-down binding pose is stabilized by the insertion of N_{TAIL} residue L495 into a non-native binding site in the hydrophobic binding groove of XD.

We highlight that while N_{TAIL} conformations in the more disordered N_{TAIL}:XD encounter complex state have similar helical propensities to unbound conformations of N_{TAIL} and N_{TAIL} conformations in the more ordered encounter complex state have similar helical propensities to those observed in the native N_{TAIL}:XD complex, the deep MSM does not suggest the presence of parallel "induced-fit" and "conformational selection"-type pathways. Transitions from both encounter complex states to the most native-like bound states proceed through similar pathways, illustrating that helical content formed early in folding-upon-binding transitions paths is not necessarily indicative of a conformational selection mechanism. This result is consistent with a previous 1D reaction

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

coordinate transition path analyses of N_{TAIL}:XD folding-upon-binding where it was observed that helical content formed early in transition paths frequently breaks to enable the formation of additional native intermolecular contacts before refolding.[56]

There is substantial experimental and computational evidence demonstrating that many IDPs maintain significant conformational disordered when bound to their physiological interaction partners.[23-25, 57] This phenomenon is frequently referred to as the formation of a "fuzzy" protein complex, and is often explained using the energy-landscape theory inspired concept of conformational frustration.[26, 57, 110-115] Conformational frustration describes the existence of multiple competing favorable interactions that cannot be simultaneously satisfied and therefore result in a dynamic equilibrium between distinct conformational states. While the existence of fuzzy complexes and the role of conformational frustration in these complexes is well appreciated, few studies have provided atomic resolution molecular mechanisms that rationalize the kinetics of the conformational transitions among the conformational states of IDPs in fuzzy complexes.[55, 57, 67, 98] The MSMs reported here identify a network of conformationally frustrated bound states of the N_{TAIL}:XD complex that share a core set of native intermolecular contacts and are differentiated by the sequential formation of non-native intermolecular and intramolecular contacts that facilitate the folding of additional helical turns. Our analyses provide atomic resolution descriptions of conformationally frustrated states of an IDP in a fuzzy protein complex and quantitative estimates of the time scales of transitions between these states. Our results underscore that an interplay between native intermolecular contacts, non-native intermolecular contacts, and non-native intramolecular contacts produce kinetic barriers between conformationally frustrated states of an IDP in a fuzzy protein complex.[116, 117] The insights generated from this study and future atomistic studies of fuzzy IDP complexes may ultimately facilitate the design of conformationally frustrated protein complexes with rationally tunable binding affinities.

It was previously noted[56] that the folding-upon-binding pathways observed in the MD trajectory analyzed here are broadly consistent with previously reported NMR experiments[87], stopped-flow kinetics measurements[33] and ϕ-value analyses of measles virus N_{TAIL}:XD binding.[89] Stopped-flow kinetics measurements clearly resolve separate rates for the formation of an initial encounter complex between N_{TAIL} and XD and the subsequent folding of N_{TAIL}[33], and protein engineering

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ϕ-values indicate that encounter complex formation is mediated by hydrophobic residues (A494,L495,L498, and A502) in the central helix of N_{TAIL}.[89] While the simulation analyzed in here was run at higher temperature (400 K) than previous experimental investigations, the MSMs derived in this investigation are broadly consistent with these previously published experimental data.

A recent study of measles virus nucleoprotein and phosphoprotein interactions underlying liquid-like phase separation reported a small set of ^{15}N NMR relaxation dispersion data to characterize the binding equilibrium of the measles virus N_{TAIL}:XD complex[91]. These data were well fit by a 2-state binding model, suggesting that only one dominant kinetic barrier is resolved in these NMR experiments. As MSMs reported here were derived from MD simulations performed at 400 K and NMR measurements in this experimental investigation were performed at 298K, it is not possible to directly compare the simulated and experimentally measured rates and state populations in these two studies. Building MSMs of N_{TAIL}:XD binding at physiological temperatures by combining the VAMPnet approach developed in this work with adaptive sampling strategies could, however, enable a direct comparison between simulated and experimental rates in this system. The recently developed augmented Markov model formalism, where MSM state populations and transition rates are refit using maximum-entropy methods to match agreement with experimental data, provides an eloquent approach to assess the agreement between MSMs and NMR relaxation data.[118] Such studies may illuminate deficiencies in current molecular mechanics force fields used to study IDP folding-upon-binding, and ultimately facilitate the design of fuzzy protein complexes between IDPs and structured binding partners.

It is interesting to consider the conformational properties of the native-like bound states of the measles virus N_{TAIL}:XD complex resolved in this study in the context of previously reported NMR relaxation dispersion measurements used to characterize the binding mechanism of the homologous sendai virus N_{TAIL} molecular recognition element to the homologous sendai phosphoprotein X domain (sendai XD) in unprecedented detail.[22] In this study, unbound sendai N_{TAIL} was found to be in equilibrium with two bound states, with a population ratio of ~3:1, that were characterized by chemical shift differences with the unbound state. The more populated bound state was found to contain an elevated population of helical elements relative to apo sendai

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

N_{TAIL} (as assessed by large changes in backbone carbon chemical shifts) but to remain relatively nonspecifically bound (as assessed by relatively small changes in nitrogen and proton backbone chemical shifts in residues at the sendai N_{TAIL}:XD binding interface). The less populated bound state has NMR chemical shifts consistent with the fully folded and ordered sendai native N_{TAIL}:XD complex. The authors of this study note that the NMR relaxation dispersion data reported are insufficient to provide atomic resolution descriptions of these states, and do not contain information on the relative position of N_{TAIL} on the surface of XD in the more populated bound conformation. This lack of information makes it challenging to understand the microscopic nature of the kinetic barriers between these states.

It is important to caveat that there are substantial differences in the sequences of N_{TAIL} and XD in the sendai and measles viruses. The α-helical molecular recognition element sendai N_{TAIL} has more charged residues (9) than the α-helical molecular recognition element of measles virus N_{TAIL} (5) and the measles virus N_{TAIL}:XD binding interface is substantially more hydrophobic than the sendai N_{TAIL}:XD binding interface, suggesting that electrostatics and polar interactions are likely to play a larger role in the sendai N_{TAIL}:XD binding mechanism.[22,87] While one expects there will be appreciable differences in the binding mechanism and bound ensembles of sendai N_{TAIL}:XD and measles virus N_{TAIL}:XD complexes it is interesting to speculate that the experimentally observed kinetic barriers observed in the bound states of the sendai N_{TAIL}:XD complex may share some features with the kinetic barriers identified here. The network of measles virus N_{TAIL}:XD bound states reported here contains kinetic barriers that result from the formation of non-native intermolecular and intramolecular contacts that must be broken to facilitate the formation of the fully folded native complex. An analogous set of interactions, perhaps with greater electrostatic contributions resulting from native and non-native salt bridges that confer greater conformational frustration, may underlie the experimentally observed kinetic barriers between bond states of the sendai N_{TAIL}:XD complex. Investigating differences in the binding mechanisms of measles virus N_{TAIL} and Sendai N_{TAIL} will be of interest in future investigations. Accurately describing differences in these binding mechanisms will present a stringent test of the quality of MD force fields used to study IDP folding-upon-binding.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Lastly, we have demonstrated the utility of a multi-input neural network framework for describing the conformational dynamics of a highly dynamic intrinsically disordered protein. The approach presented here, where convolutional neural network layers are utilized to reduce the dimensionality of interatomic distance matrices while fully connected dense neural network layers are used to process lower dimensional order parameters describing the helical content of an IDP before combining all features in a fully connected dense neural network, provides a high degree of flexibility for identifying optimal combinations of molecular feature sets with different inherent dimensionalities and embeddings. We demonstrated that this approach distinguishes several structurally and kinetically distinct Markov states that were not resolved using the traditional linear dimensionality reduction tICA approach. We speculate that the deep learning strategy employed here may provide a generalizable approach for learning low dimensional representations of high dimensional IDP simulation data that are best described by multiple distinct degrees of freedom. We plan to investigate the utility of this approach for building MSMs of monomeric IDPs and for identifying collective variables for enhanced sampling methods and diffusion models in future studies.

Methods

Markov State Models. Markov state models (MSMs) are stochastic dynamical models that approximate the kinetics of molecules as memoryless, probabilistic jump processes between sets of states.[62] MSMs utilize a time reversible transition matrix[119] containing conditional probabilities of transitioning between states. The transition matrix of a MSM is reversible and functions as a transfer operator that propagates a distribution of states, $p(t)$, forward (and backward) in time by $k\tau$ discrete steps where k is a positive integer and τ is the lag time of the model.

$$p^T(t + k\tau) \approx p(t)^T T^k(\tau) \quad (1)$$

The optimal lag time of a MSM can be determined by ploting the implied time scales (ITS) as a function of the lag time and choosing the lag time at which the implied time scales[120] are approximatly constant[62]. Additionally, the time resolution of the model can be determined by checking that ITS are above the lag time at which the model is estimated (Supplementary Figures 2 and 11). Implied time scales are determined from the eigen values, λ_i, of the transition matrix .

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$t_i(\tau) = -\frac{\tau}{ln|\lambda_i|} \quad (2)$$

By definition, MSM transition matrices have a maximum eigen value of 1 whose eigen vector corresponds to the steady state or stationary population, $\vec{\pi}$, of states as time approaches infinity.[121]

$$T(\tau)\vec{\pi} = \vec{\pi} \quad (3)$$

Using the the stationary distribution and the transition matrix of a MSM, the mean first passage times between pairs of states ($MFPT_{ij}$) can be determined from an N_{states} by N_{states} system of equations (Supplemenary Figures S9 and S17).[122]

.

$$MFPT_{ij} = 1 + \sum_{k \neq j} T_{ik} MFPT_{kj} \quad (4)$$

$$MFPT_{ii} = {}^{1}/_{\pi_i}$$

In addition to ITS, validation of MSMs and their transtion matrices is determined by the Chapman-Kolmogrov equation[62, 121],

$$T(\tau k) = T(\tau)^k \quad (5)$$

in which the ability of a transition matrix to reproduce transistion probabilities at longer timesales is evaluated (Supplementary Figures 2 and 11).

Input Data for Markov State Models. We utilized the 200 μs unbiased MD trajectory from Robustelli et. al[56] which contains N_{TAIL} residues 484-504, XD residues 458-506 and 20mM of NaCl in a 72 Å per side cubic box. This trajectory was parametrized using the a99SB-disp force field, a99SB-disp water model[52] and contained 1,000,000 frames with a spacing of 200ps. For the construction of our MSMs[62, 70], we only considered a continuous 167μs subset (from 3μs to 170μs)

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of the original trajectory in which XD predominantly remains in its folded state. We generated the molecular features for MSM construction and neural network training by calculating intermolecular distances between all residues of N_{TAIL} and XD using the minimum distance between heavy atoms. Additionally, we computed the α-helical order parameter Sα [102] and identified helical conformations using the DSSP[101] algorithm. The order parameter Sα quantifies the helical content of each 7-residue segment of a peptide chain and is computed by the following,

$$S_\alpha = \sum_i^N \frac{1-\left(\frac{RMSD\alpha_i}{r_0}\right)^8}{1-\left(\frac{RMSD\alpha_i}{r_0}\right)^{12}} \quad (6)$$

.

where $RMSD\alpha_i$ denotes the root mean squared deviation between each 7-residue segment of N_{TAIL} and a geometrically perfect alpha helix comprised of the same residues. The exponential terms in the equation act as a switching function to output values between 0 (not helical) and 1 (perfectly helical) for each segment. The threshold of the switching function is tuned by the parameter r_0, which was chosen to be 0.8 Å. Setting the parameter r_0 to 0.8 Å has the effect of reducing $RMSD\alpha_i$ values > 2.5 Å to nearly zero and $RMSD\alpha_i$ values < 0.5 Å to nearly 1. For constructing MSMs, we chose to omit the summation in (eq. 6) to retain a more localized description of the helical content of N_{TAIL}. As a result of the 7-residue sliding window used in the computation of Sα and N_{TAIL} being 21 residues long, we compute a length 15 vector for each time step of the simulation describing the helical content of every possible contiguous 7 residue segment of N_{TAIL}. We note that for broad statistical characterizations (such as in Figure 1), the summation in equation 1 is retained to provide an estimate of the total helicity of N_{TAIL} per simulation frame ("N_{TAIL} Sα").

We constructed the second α-helical descriptor for N_{TAIL} using the DSSP Algorithm. The DSSP algorithm uses dihedral angles and hydrogen bonding analysis to classify the secondary structure of each residue in a peptide chain. The secondary structure predictions given by DSSP were then numericized by equating helical classifications to 1 and all others to zero. As a result, the processed binary DSSP assignments produce a vector of length 21 for each time step of the simulation with values indicating if each residue of N_{TAIL} is in a helical conformation (value of 1) or not (value of 0). Both Sα and binary DSSP features were considered in quantifying the helical content of N_{TAIL}

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

as they evaluate helical content using distinct metrics and as a result, produce differing degrees of locality in the descriptions they provide.

We tested several feature sets for state discretization including combinations of interatomic distances, dihedral angles, fraction native intermolecular contacts (Q), binary DSSP assignments and Sα values. We assessed the quality of feature sets by comparing VAMP2 scores[76, 83], the spectral gap observed among the eigenvalues of the dominant tICA eigenmodes,[79, 96, 97] and the ability of each feature set to resolve conformationally distinct free energy basins in low dimensional tICA projections. For tICA, we found the combination of intermolecular residue distances and Sα best satisfied these metrics and that the addition of DSSP features had negligible effect. We subsequently omitted the DSSP features from our tICA analysis and used only intermolecular distances and Sα order parameters. In contrast, we found that including DSSP features in our VAMPnet increased the model's ability to differentiate N_{TAIL} conformations differing only in the helical content of residues near the termini; thus, we used a feature set containing intermolecular residue-residue distances, Sα, and binary DSSP helical assignments as input data in our VAMPnet implementation.

Construction of a hidden Markov state model (HMSM). To construct an initial MSM, we performed tICA on a feature set comprised of the nearest-heavy-atom intermolecular distances between all residues of N_{TAIL} and XD and Sα values. The tICA lag time, number of tICA components (tICs) used for clustering, and the number of *k-means* clusters were optimized based on the interpretability and distinctness of the structural properties of the resulting clusters. We iteratively computed tICA with varying lag times and clustered the resulting tICs using a varying number of components and k-means clusters. We characterized the structural properties of clusters by computing their distributions of the fraction of native intermolecular contacts (Q), N_{TAIL} $S\alpha$, Radius of gyration (R_g), intermolecular contact probabilities and helical assignments from the DSSP algorithm. We found that using a lag time of 6 ns for tICA, clustering conformations using the ten time independent components (tICs) with the largest eigenvalues and implementing *the k-means* algorithm with seven cluster centers produced the most interpretable and conformationally distinct clusters. However, upon estimating MSMs from these clusters over a range of lag times, we found that for lag times up to 24 ns, these models produced resolved, but non-converged

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

implied timescales (data not shown). These MSMs also failed to reproduce transition probabilities for non-native bound states at longer timescales.

To produce MSMs with both converged time scales and robust CK-tests, we employed hidden Markov state models (HMSMs). HMSMs are an effective tool for building robust and reproducible MSMs for high dimensional systems where finding a set of Markov states that pass validation tests is challenging.[95] Projected HMSMs are estimated from transitional MSMs; the slowest relaxing timescales of the original MSM are used to coarse grain its states to a smaller number of metastable sets. The number of metastable sets used to build an HMSM should be equal to or less then the number of resolved timescales in the conventional MSM they're estimated from. We built our HMSM by estimating a series of HMSMs from MSMs with varying numbers of states and lag times. We increased the number of states in the initial MSMs by employing the k-means clustering algorithm with larger numbers of centroids to cluster the same ten tICs we previously found to be optimal to prevent the HMSM coarse graining from reducing our model to too few states. We found that using a lag time of 6 ns, twelve initial clusters and coarsening to seven states produced robust HMSMs (in terms of timescales and CK-tests) with the fewest number of states (Supplementary Figure 2).

Unconstrained VAMPnet and neural network architectures. The feature set used to train the deep MSM was comprised of the intermolecular distances between all residue pairs of N_{TAIL} and XD, Sα order parameters and binary DSSP assignments. We employed a multi-input deep learning approach where each feature type was processed separately before being aggregated with the other features to make state predictions. This approach allows for the input feature set to be optimized internally and each feature type to be processed using neural network layers that best suit its inherent data structure. This approach enabled us to treat the matrix of intermolecular distances (or "contact map") calculated in each frame of the simulation as an image and utilize convolutional neural network layers to leverage the local spatial coherence in this representation. We utilized separate sets of fully connected neural network layers to process the Sα and binary DSSP feature sets. Each instantaneous set of intermolecular residue distances were arranged into a 49 by 21 matrix where each index represents the intermolecular distance between each residue in XD (49 residues) and N_{TAIL} (21 residues). Each set of Sα and binary DSSP values were placed into length

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

15 and 21 vectors, respectively. In aggregate, the VAMPnet dataset is comprised of 3 distinct feature sets, each processed separately by distinct sets of neural network layers (or lobes), before being aggregated and transformed through a final lobe, containing fully connected neural network layers (Figure 2). The output of the final lobe is capped with a SoftMax activation function to produce a normalized distribution that describes the probability of a frame being assigned to each Markov state.

We determined the architecture of our neural network by varying the number of layers and their widths in each lobe of the neural network. To reduce computational overhead, we constrained our optimization of the neural network architecture by requiring that each lobe contain the same number of layers and that the lobes used to transform the N_{TAIL} Sα and DSSP helical order parameters be identical apart from their input layers. In addition, the possible configurations of the convolutional layers used to transform intermolecular distance matrices were constrained based on the shape the input (49 XD residues by 21 N_{TAIL} residues). We determined our architecture by first performing a grid search over a range of configurations and then performed a Bayesian optimization around the optimal parameters identified in the initial grid search. For the Bayesian optimization, we used the tree-structured Parzen estimator algorithm[123, 124] implemented in the *optuna*[125] software. A detailed diagram of the final neural network architecture determined from the Bayesian optimization procedure is displayed in Supplementary Figure 10. After determining the neural network architecture, we employed this procedure to determine the optimal batch size, optimizer learning rate and epsilon parameter. We found that using learning rate of 5e-6, a batch size of 16384 and an epsilon parameter of 1e-7 produced optimal results.

Additional hyperparameters of VAMPnets include the lag time of the model and the number of output states. To determine these hyperparameters, we conducted optimization runs incrementally increasing the values of each hyperparameter while holding the other hyperparameters constant. We judged the success of these trials based on the maximization of the VAMP score relative to its highest possible value and the interpretability of the learned state assignments in terms of the fraction of native contacts (Q), Sα, radius of gyration and RMSD from the native complex. We found that using 12 output states and a lag time of 2 ns to train the unconstrained VAMPnet best satisfied these conditions and consistently produced similar sets of states. The final architecture of

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

multi-input neural network used in our VAMPnet implementation is shown in Supplementary Figure 10.

We trained our initial unconstrained VAMPnet using the VAMP2 score. The VAMP2 score evaluates the so-called kinetic variance between each neural network transformed sample, $\chi_0(x_t)$, of the dataset and it's time-lagged analogue, $\chi_\tau(x_{t+\tau})$, where χ_0 and χ_τ are neural network transformations that convert molecular features into probabilistic Markov state assignments and x_t and $x_{t+\tau}$ are instantaneous sets of molecular features at times t and t+τ.[84] Optimizing the VAMP2 score of transformations $\chi_0(x_t)$ and $\chi_\tau(x_{t+\tau})$ is analogous to solving the problem of finding orthonormal transformations of x_t and $x_{t+\tau}$ with maximal time-correlations and corresponds to finding the best linear approximation[84] to the following,[83]

$$\mathbb{E}[\chi_\tau(x_{t+\tau})] \approx \widetilde{K}^T \mathbb{E}[\chi_0(x_t)] \tag{7}$$

where $\widetilde{K}^T$ is the finitely estimated Koopman matrix that transforms a potentially non-linear dynamical system or dataset into a latent space which, on average, evolves linearly in time. The VAMP2 score is defined as the Frobenius norm or sum of the squared singular values (σ_i) of the half-weighted Koopman matrix, ${C_{00}}^{-\frac{1}{2}} C_{0\tau} {C_{\tau\tau}}^{-\frac{1}{2}}$.

$$VAMP2 = ||{C_{00}}^{-\frac{1}{2}} C_{0\tau} {C_{\tau\tau}}^{-\frac{1}{2}}||_F^2 + 1 = \sum \sigma_i^2 \tag{8}$$

Where the covariance matrices, C_{00}, $C_{0\tau}$ and $C_{\tau\tau}$ are defined by mean free neural network transformed instantaneous and time lagged data as follows.

$$\begin{aligned} C_{00} &= \mathbb{E}_t[\chi_0(x_t)\chi_0(x_t)^T], \\ C_{0\tau} &= \mathbb{E}_t[\chi_0(x_t)\chi_\tau(x_{t+\tau})^T], \\ C_{\tau\tau} &= \mathbb{E}_{t+\tau}[\chi_\tau(x_{t+\tau})\chi_\tau(x_{t+\tau})^T] \end{aligned} \tag{9}$$

We note that in general, neural network transformations, χ_0 and χ_τ can be distinct neural network architectures with independently trained weights, however, in our implementation $\chi_0 \equiv \chi_\tau$.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Training the constrained VAMPnet to construct a deep MSM. After determining the optimal architecture and hyperparameters for the unconstrained VAMPnet, we proceeded to build a constrained VAMPnet using the same architecture with the addition of two constraint layers. In the constrained VAMPnet[85], the constraint layers (u and S) are implemented to ensure the learned transition matrix is both stochastic (all positive elements) and reversible (obeys detailed balance). Constraint u is a vector of length equal to the number of states used to weight data towards equilibrium and constraint S is matrix of shape N_{states} by N_{states} used to estimate a reversible transition matrix. The constrained VAMPnet was trained with a modified version of VAMP-E score that incorporates the constraints u and S.

$$VAMP - E = tr\left[S^T C_{00} S C_{\gamma\gamma} - 2S^T C_{0\gamma}\right], \tag{10}$$

where

$$C_{00} = \mathbb{E}[\chi(x_t)\chi(x_t)^T],$$
$$C_{0\gamma} = \mathbb{E}[\chi(x_t)\gamma(x_{t+\tau})^T],$$
$$C_{\gamma\gamma} = \mathbb{E}[\gamma(x_{t+\tau})\gamma(x_{t+\tau})^T],$$
$$\gamma(x) = \chi(x)\chi(x)^T u.$$

Here, gamma is a weighted state representation used to compensate for non-equilibrium state assignment probabilities. We trained our constrained VAMPnet 30 separate times starting from the same initial unconstrained VAMPnet.

In the constrained VAMPnet procedure, both the weights of the unconstrained VAMPnet and constraint layers are optimized, thus, retraining only the constrained VAMPnet also modifies the weights of the initial, unconstrained VAMPnet. We note that using the same unconstrained VAMPnet in each optimization of the constrained VAMPnet produces small error estimates that may be underestimated compared error estimates obtained from retraining the unconstrained VAMPnet multiple times. Given the large number of parameters in our neural network architecture (~4e6 parameters), we used this approach to circumvent considerable computational costs and consider these error estimates as lower bounds of the trial errors. As outlined in its original

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

implementation[85], it is recommended to include an initial step in which only the constraints of the constrained VAMPnet are trained using batches containing all training data. When training the unconstrained VAMPnet and the constraints together (a separate step), we attempted to stay consistent with this strategy and used the largest batch size possible given our computational resources which was 56,000 time-lagged pairs of data. To estimate the implied timescales and CK-tests, we retrained only the constraints of the constrained VAMPnet at integer multiples of the initial lag time (6 ns) which was done for all 30 optimization runs. We chose to use a lag time of 6 ns for the constrained VAMPnet based on the results of these validation measures which we found to produce the most reproducible and robust models in a series of initial estimations of the constrained VAMPnet at varying lag times (Supplementary Figure 11).

Neural network training. In both the unconstrained and constrained VAMPnets, we used a 9:1 train-validation split, randomly shuffled time lagged pairs of data and implemented early stopping to prevent overfitting where we saved network weights each time the VAMP score reach a new maximum. We implemented all neural networks in using the deep learning library *PyTorch*[126].

Estimation of trajectory observables and error analysis. For the HMSM, all MSM observables and error estimates were computed using the *pyemma*[70] and *deeptime*[127] software packages via Bayesian hidden markov models which use a gibbs sampling scheme to resample the transition matrix. Here, we estimated errors by resampling the HMSM transition matrix using 100 trials. All HMSM trajectory observables are the bootstrap mean and its associated 95% confidence intervals computed from the results of the resampling procedure. For the deep MSM, we trained the final model using 30 independent trials and computed both MSM and trajectory observables from the trained models. All statistical analysis of the trajectory observables of the deep MSM states and MSM observables are computed by bootstrapping / aggregating the results of these 30 trials, e.g. average values, 95% confidence intervals of averages, standard deviations, weighted historgrams and discrete probability distributions. Trajectory observables from the deep MSM states were computed from the probabilistic state assignments produced from each optimization run by the following,

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\langle \hat{O}_{state_i} \rangle = \frac{\int \hat{O}\,(t)\chi(t)_{state_i} dt}{\int \chi(t)_{state_i} dt} \tag{12}$$

where $\widehat{\boldsymbol{O}}(\boldsymbol{t})$ represents an arbitrary trajectory observable computed for every frame (t) of the trajectory and $\boldsymbol{\chi}(\boldsymbol{t})_{\boldsymbol{state_i}}$ is a probabalistic state assignment for every frame (t) of the trajectory. Using this definition, we can also compute the standard deviation of trajectory observables by the following equation.

$$SD_{\hat{O}_{state_i}} = \sqrt{\langle (\hat{O}_{state_i})^2 \rangle - \langle \hat{O}_{state_i} \rangle^2} \tag{13}$$

We combine uncertaines computed from separate trials and contact popualtions for different residue pairs by combining variances,

$$SD_C^2 = \frac{\sum_i n_i (SD_i^2 - (\bar{X}_i - \bar{X}_C)^2)}{\sum_i n_i}$$
$$\bar{X}_C = \frac{\sum_i n_i \bar{X}_i}{\sum_i n_i} \tag{14}$$

Where SD_C^2 is the combined variance, n_i are the number of trials used to compute the mean and standard deviation of each statistic to be combined, $\bar{X}_i$ are the means of each statistic to be combined and $\bar{X}_C$ is the combined mean.

Fraction of Native Intermolecular Contacts. The fraction of native intermolecular contacts (Q), as defined in Robustelli et al[56, 100], was used to characterize the formation of the N_{TAIL}:XD complex. The fraction of native contacts at each simulation time step, (t), was calculated by the following,

$$Q(t) = \frac{\sum_{i=1}^{N} \frac{1}{1 + e^{10(d_i(t) - x_0)}}}{N} \tag{15}$$

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where d_i represents the nearest neighborh heavy atom distance between each pair of native contacts, N is the total number of native contact pairs and x_0 is a cutoff distance of 5Å. Native intermolecular contacts were previosuly defined as those contacts which remained stable (populated > 80%) in an MD simulation of the native N_{TAIL}:XD complex run at 400 K, to match the temperature of the equilibrium folding-upon-binding simulation analyzed here.[56]

Color gradients of structural snapshots. We computed the color gradients of the structural snapshots of N_{TAIL}:XD using a modified version of the fraction of native intermolecular contacts based only on the crystal structure of the native complex (PDB 1T6O).[86] For establishing color gradients, we defined native contacts as any intermolecular residue pair between N_{TAIL} and XD with a minimum heavy atom distances less than 5 Å in PDB 1T6O. Correspondingly, we define non-native contacts for each residue as all other possible intermolecular contacts that have not been identified as native. In each simulation frame two residues are considered to be in contact if their nearest heavy atom distance is less than 5 Å. We compute the average population of the native and non-native contacts of every residue in each Markov state. For coloring structures, we normalize native and non-native fractions by dividing each by the largest fraction observed in any Markov state (~ 0.99 and ~0.14 for native and non-native fractions, respectivly) which assigns a value between 0 and 1 for each residue in each Markov state. We then set a color gradient ranging from 0 to 1 in the molecular visualization software *pymol*[128]and set the beta value of each residue (alpha carbon) to the normalized fraction of native and non-native contacts. The normalization step allows the scale of the color gradients to be the same across all structures, thus allowing for quantitative comparision of the contact profiles of each Markov state via their structural snapshots.

Data & Code Availability

All code used for trajectory analyses and the construction and validation of the hidden Markov state model and deep Markov state model are freely available from GitHub (https://github.com/paulrobustelli/Sisk_NTAIL_DeepMSM_2023). The 200 μs N_{TAIL}:XD MD trajectory analyzed here is available for non-commericial use by request from D.E. Shaw Research (Trajectories@DEShawResearch.com).

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements

This work was supported by the National Institutes of Health under award R35GM142750.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures

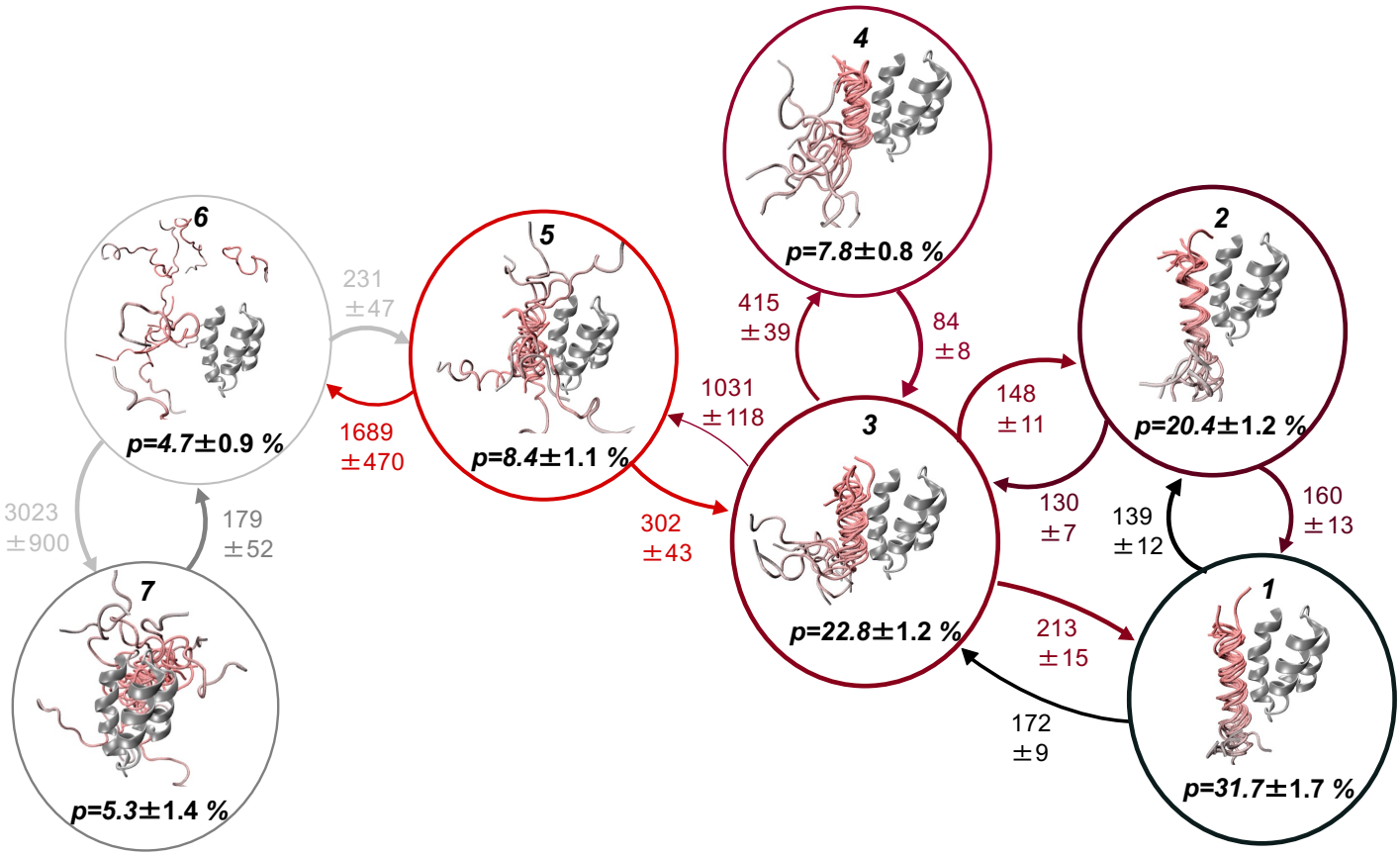

Figure 1. Transition network representation of a conventional hidden Markov state model of N_{TAIL}:XD folding-upon-binding derived from a long-time scale equilibrium molecular dynamics simulation. Network representation of the transition matrix obtained from a hidden Markov state model (HMSM) derived from time-lagged independent component analysis (tICA) of a long-time scale MD simulation. Representative structures of each Markov state are displayed in circles along with their stationary probabilities (*p*). The thickness of circles is proportional to the stationary probability of each state. In representative structures of each state N_{TAIL} is colored with a gray-to-red gradient from the N-terminus to the C-terminus and XD is colored gray. Transition probabilities between states are indicated with directional arrows, and the thickness of the arrows is proportional the magnitude of the transition probability between states. Mean first passage times between states are reported in nanoseconds. All errors indicate the mean of the upper and lower deviations of the 95% confidence interval calculated from bootstrapping using 100 samples.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

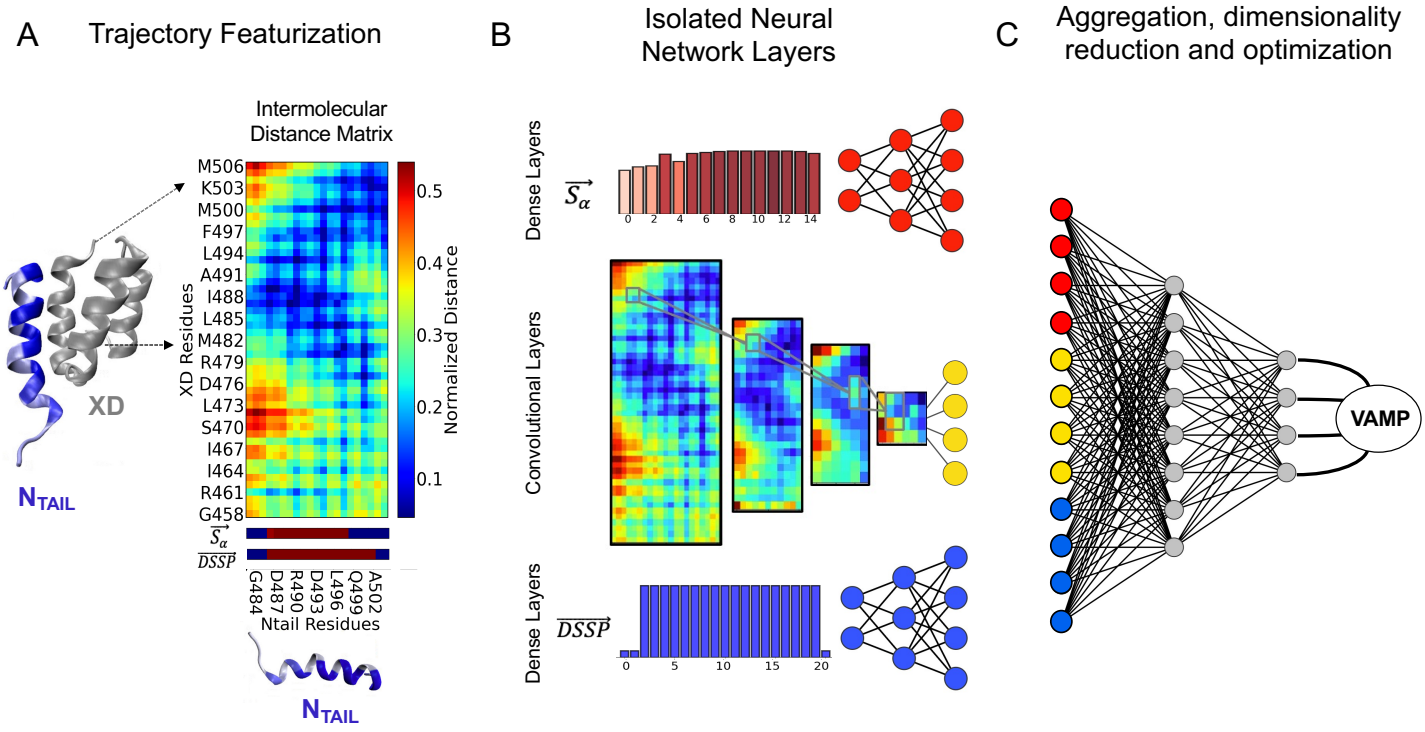

Figure 2. **Multi-input neural network architecture used for building a deep Markov state model of N_{TAIL}:XD folding-upon-binding. (A)** Structural representation of the native N_{TAIL}: XD complex . XD is colored gray and N_{TAIL} is colored in a gray-to-blue gradient proportional to each residues fraction of native contacts in deep MSM state 1. The set of deep MSM input features (intermolecular distances between N_{TAIL} and XD, N_{TAIL} Sα order parameters, binary DSSP helical assignments) are shown for the structure in A (right). **(B)** Schematic representation of the isolated neural network layers used to process each feature type based on its inherent dimensionality. Sα (red) and binary DSSP (blue) features are treated as 1D vectors and are processed with dense neural network layers. The intermolecular distance matrix between N_{TAIL} and XD is processed with convolutional neural network layers to take advantage of the spatial coherence of data points in its matrix form. **(C)** A qualitative schematic showing the aggregation and further processing of output features from the 3 isolated sets of layers. Upon aggregation, the processed output features from each isolated layer are combined by a final set of dense layers to reduce the dimensionality of the output to a normalized probability distribution over Markov states. The output probability distributions are used to compute a VAMP score for batches of time-lagged data pairs.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

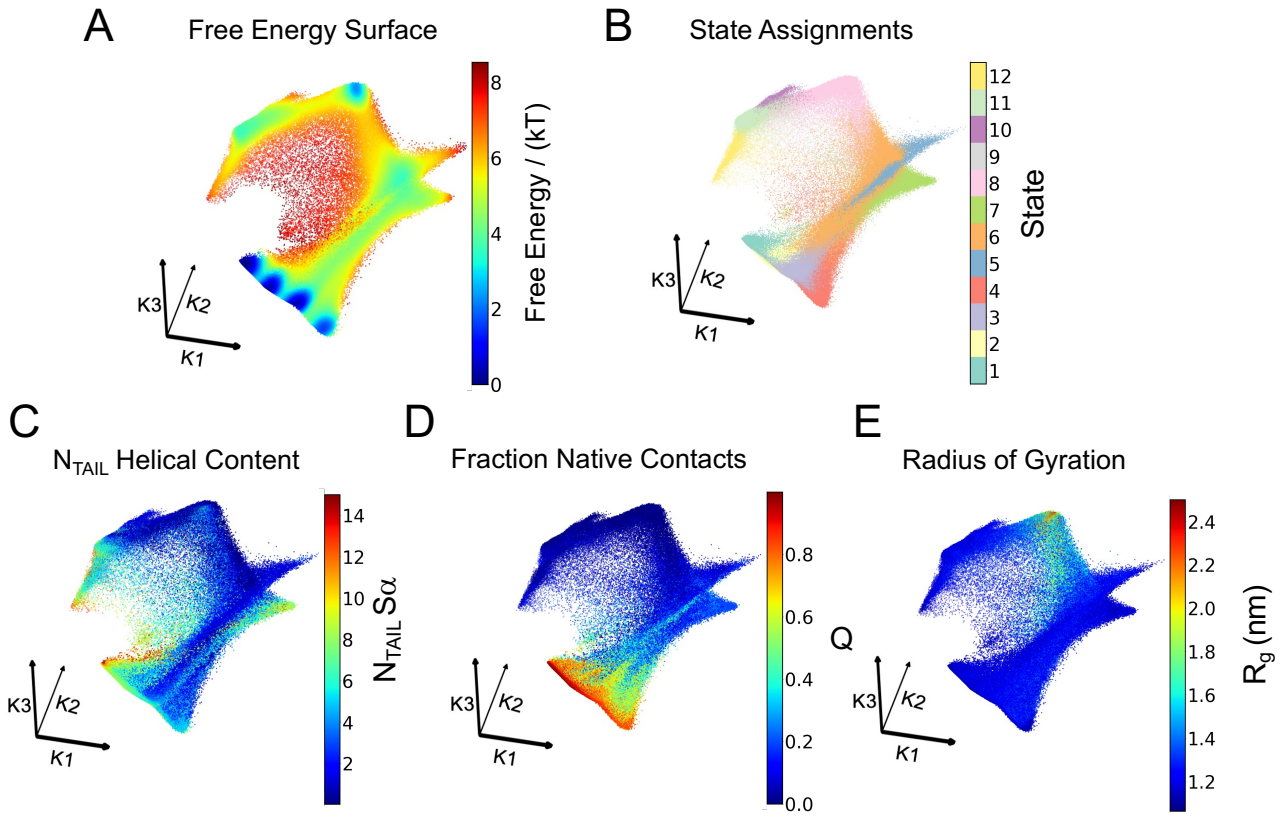

Figure 3. VAMPnet latent space and state assignments used to construct a deep Markov state model of N_{TAIL}:XD folding-upon-binding. We characterize the latent space of our N_{TAIL}:XD VAMPnet by projecting MD observables onto the left singular functions (or "Koopman modes") K1, K2, and K3 of the half-weighted Koopman matrix estimated from an initial unconstrained VAMPnet. Truncating the singular value decomposition to 3 singular vectors gives a 3-dimensional latent space or set of singular functions where points are embedded in a kinetically meaningful way. We characterize the latent space representation of each MD simulation frame by coloring each data point by **(A)** the apparent free energy obtained by taking the negative natural log of a gaussian kernel density estimate over the 3-dimensional latent space-projected data, **(B)** its crisp Markov state assignment **(C)** the fraction of native intermolecular contacts (Q) **(D)** the sum of the N_{TAIL} α-helical folding order parameter Sα for each 7 residue segment of N_{TAIL} and **(E)** the radius of gyration (R_g) of all Cα carbons of N_{TAIL} and XD.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

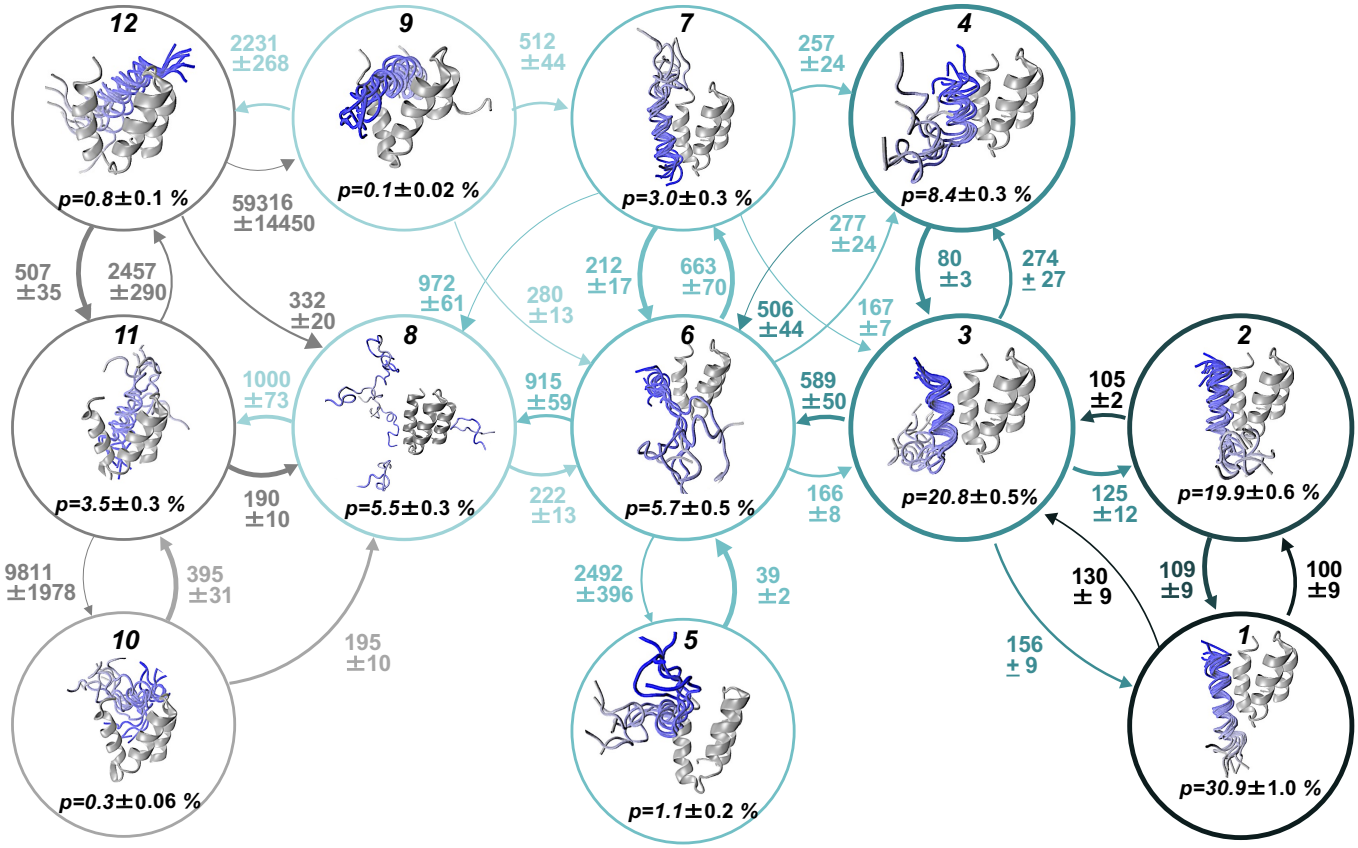

Figure 4. Transition network representation of a deep Markov state model of N_{TAIL}:XD folding-upon-binding derived from a long-time scale molecular dynamics simulation using a multi-input neural network architecture. Network representation of the transition matrix of a deep Markov state model (MSM) obtained from a multi-input neural network architecture. Representative structures of each Markov state are displayed in circles along with their stationary probabilities (*p*). The thickness of circles is proportional to the stationary probability of each state. In the representative structures of each state, N_{TAIL} is colored by a gray-to-blue gradient from the N-terminus to the C-terminus and XD is colored gray. The transition probability between states is indicated with directional arrows, and the thickness of the arrows is proportional the magnitude of the transition probability between states. Mean first passage times between states are reported in nanoseconds. The values and errors reported here are the bootstrap means and their 95% confidence intervals, obtained from 30 independent optimization runs of the constrained VAMPnet.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

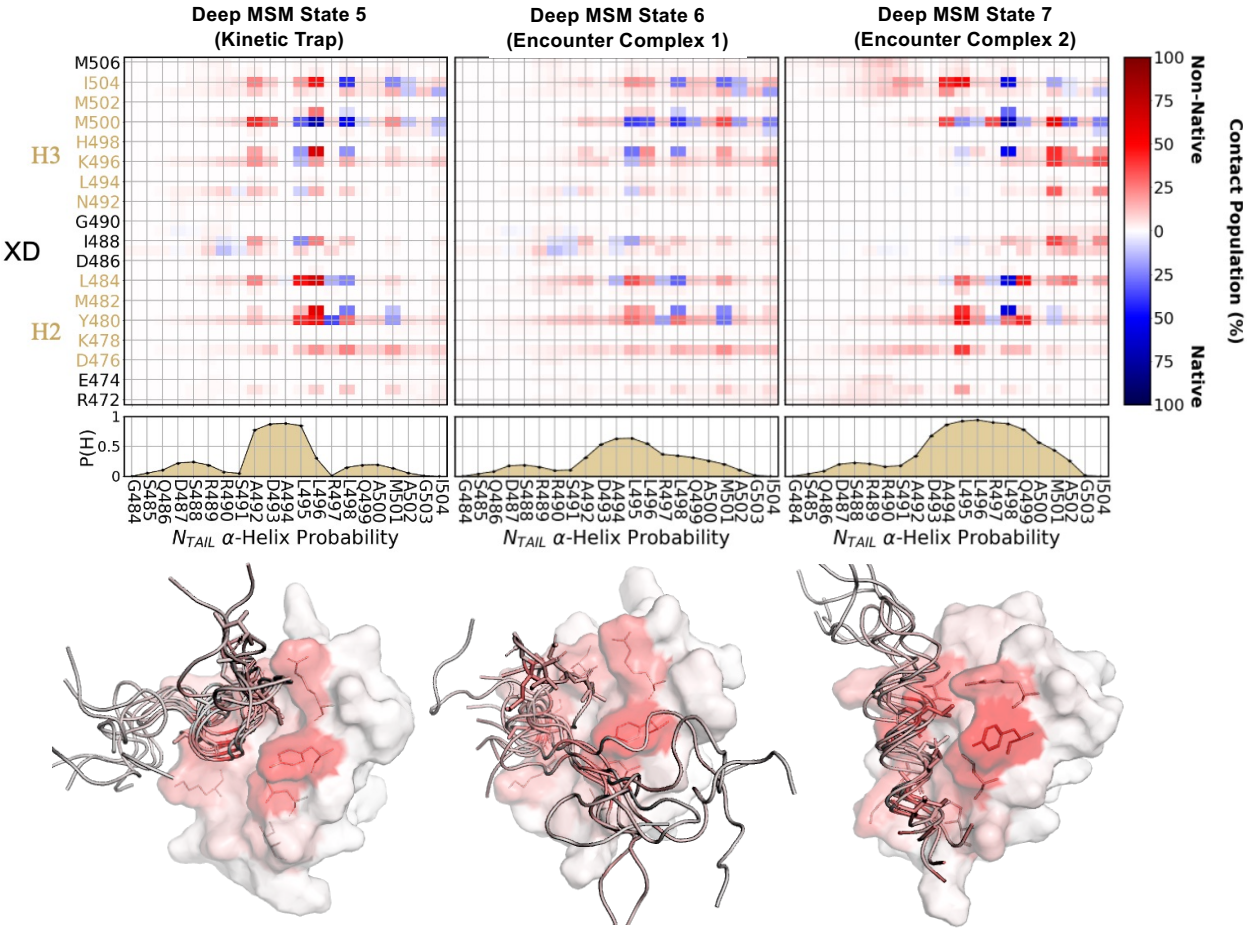

Figure 5. A deep Markov state model of N_{TAIL}:XD folding-upon-binding pathways resolves two distinct encounter complex states and a kinetic trap state. State averaged intermolecular contact populations and N_{TAIL} helical propensities for deep MSM states 5, 6 and 7. Intermolecular contacts are defined as occurring in all frames where the minimum distance between heavy atoms of two residues is less than 5.0 Å. Native intermolecular contacts are colored blue and non-native contacts are colored red**.** Native contacts are defined as those present in the crystal structure (PDB 1T6O) using the same criteria. Helical propensities (P(H)) were calculated using DSSP. Structural representations contain an overlay of multiple representative N_{TAIL} structures with one surface representation of XD. The residues of N_{TAIL} and XD are colored by a gray-to-red gradient that represents the fraction of frames where non-native intermolecular contacts are formed by each residue in each state.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

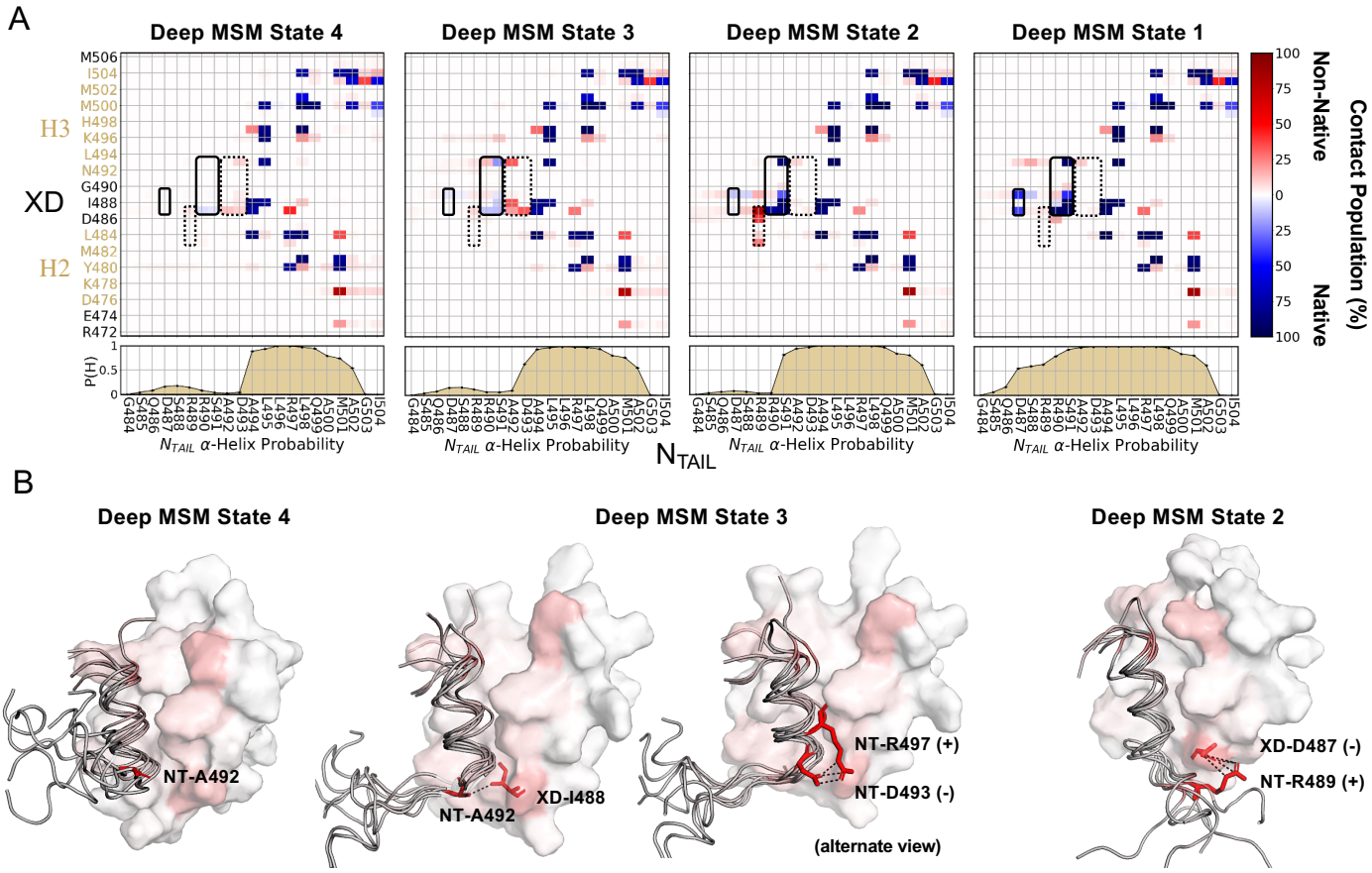

Figure 6. Kinetic barriers between native-like N_{TAIL}:XD bound states originate from non-native intermolecular and intramolecular contacts. **(A)** State averaged intermolecular N_{TAIL}:XD contact populations and N_{TAIL} helical propensities for native-like N_{TAIL}:XD bound states. Intermolecular contacts are defined as occurring in all frames where the minimum distance between heavy atoms of two residues is less than 5.0 Å. Native intermolecular contact pairs are colored blue and non-native intermolecular contact pairs are colored red**.** Native contacts are defined as those present in the crystal structure (PDB 1T6O) using the same criteria. **(B)** Structural representations of native-like N_{TAIL}:XD bound states. Each state representation is an overlay of multiple representative N_{TAIL} structures with one surface representation of XD. The residues of N_{TAIL} and XD are colored by a gray-to-red gradient that represents the fraction of frames where non-native intermolecular contacts are formed by each residue in each state. Selected sidechains of N_{TAIL} (NT) and XD are shown as sticks to illustrate important non-native contacts in different states.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

(1) Babu, M. M.; Kriwacki, R. W.; Pappu, R. V. Versatility from Protein Disorder. *Science* **2012**, *337* (6101), 1460-1461.
(2) Davey, N. E.; Travé, G.; Gibson, T. J. How viruses hijack cell regulation. *Trends in Biochemical Sciences* **2011**, *36* (3), 159-169.
(3) Habchi, J.; Tompa, P.; Longhi, S.; Uversky, V. N. Introducing protein intrinsic disorder. *Chemical Reviews* **2014**, *114* (13), 6561-6588.
(4) Cheng, Y.; Oldfield, C. J.; Meng, J.; Romero, P.; Uversky, V. N.; Dunker, A. K. Mining α-Helix-Forming Molecular Recognition Features with Cross Species Sequence Alignments. *Biochemistry* **2007**, *46* (47), 13468-13477.
(5) Oldfield, C. J.; Cheng, Y.; Cortese, M. S.; Romero, P.; Uversky, V. N.; Dunker, A. K. Coupled Folding and Binding with α-Helix-Forming Molecular Recognition Elements. *Biochemistry* **2005**, *44* (37), 12454-12470.
(6) Vacic, V.; Oldfield, C. J.; Mohan, A.; Radivojac, P.; Cortese, M. S.; Uversky, V. N.; Dunker, A. K. Characterization of Molecular Recognition Features, MoRFs, and Their Binding Partners. *Journal of Proteome Research* **2007**, *6* (6), 2351-2366.
(7) Camacho-Zarco, A. R.; Schnapka, V.; Guseva, S.; Abyzov, A.; Adamski, W.; Milles, S.; Jensen, M. R.; Zidek, L.; Salvi, N.; Blackledge, M. NMR Provides Unique Insight into the Functional Dynamics and Interactions of Intrinsically Disordered Proteins. *Chemical Reviews* **2022**, *122* (10), 9331-9356.
(8) Jensen, M. R.; Ruigrok, R. W. H.; Blackledge, M. Describing intrinsically disordered proteins at atomic resolution by NMR. *Current Opinion in Structural Biology* **2013**, *23* (3), 426-435.
(9) Receveur-Bréchot, V.; Bourhis, J.-M.; Uversky, V. N.; Canard, B.; Longhi, S. Assessing protein disorder and induced folding. *Proteins: Structure, Function, and Bioinformatics* **2005**, *62* (1), 24-45.
(10) Tompa, P.; Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends in Biochemical Sciences* **2008**, *33* (1), 2-8.
(11) Babu, M. M.; van der Lee, R.; de Groot, N. S.; Gsponer, J. Intrinsically disordered proteins: regulation and disease. *Current opinion in structural biology* **2011**, *21* (3), 432-440.
(12) Dunker, A. K.; Cortese, M. S.; Romero, P.; Iakoucheva, L. M.; Uversky, V. N. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS Journal* **2005**, *272* (20), 5129-5148.
(13) van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.; et al. Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews* **2014**, *114* (13), 6589-6631.
(14) Gsponer, J.; Futschik, M. E.; Teichmann, S. A.; Babu, M. M. Tight Regulation of Unstructured Proteins: From Transcript Synthesis to Protein Degradation. *Science* **2008**, *322* (5906), 1365-1368.
(15) van der Lee, R.; Lang, B.; Kruse, K.; Gsponer, J.; Sánchez de Groot, N.; Huynen, Martijn A.; Matouschek, A.; Fuxreiter, M.; Babu, M. M. Intrinsically Disordered Segments Affect Protein Half-Life in the Cell and during Evolution. *Cell Reports* **2014**, *8* (6), 1832-1844.
(16) Bah, A.; Forman-Kay, J. D. Modulation of Intrinsically Disordered Protein Function by Post-translational Modifications. *Journal of Biological Chemistry* **2016**, *291* (13), 6696-6705.
(17) Csizmok, V.; Forman-Kay, J. D. Complex regulatory mechanisms mediated by the interplay of multiple post-translational modifications. *Current Opinion in Structural Biology* **2018**, *48* (48), 58-67.
(18) Zhou, J.; Zhao, S.; Dunker, A. K. Intrinsically Disordered Proteins Link Alternative Splicing and Post-Translational Modifications to Complex Cell Signaling and Reulation. *Biophysical Journal* **2018**, *114* (3), 79a.
(19) Oldfield, C. J.; Meng, J.; Yang, J. Y.; Yang, M. Q.; Uversky, V. N.; Dunker, A. K. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* **2008**, *9* (S1).
(20) Bourhis, J.-M.; Johansson, K.; Receveur-Bréchot, V.; Oldfield, C. J.; Dunker, K. A.; Canard, B.; Longhi, S. The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. *Virus Research* **2004**, *99* (2), 157-167.
(21) Dogan, J.; Gianni, S.; Jemth, P. The binding mechanisms of intrinsically disordered proteins. *Phys. Chem. Chem. Phys.* **2014**, *16* (14), 6323-6331.
(22) Schneider, R.; Maurin, D.; Communie, G.; Kragelj, J.; Hansen, D. F.; Ruigrok, R. W. H.; Jensen, M. R.; Blackledge, M. Visualizing the Molecular Recognition Trajectory of an Intrinsically Disordered Protein Using Multinuclear Relaxation Dispersion NMR. *Journal of the American Chemical Society* **2015**, *137* (3), 1220-1229.
(23) Sigalov, A. B.; Zhuravleva, A. V.; Orekhov, V. Y. Binding of intrinsically disordered proteins is not necessarily accompanied by a structural transition to a folded form. *Biochimie* **2007**, *89* (3), 419-421.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(24) Mittag, T.; Orlicky, S.; Choy, W.-Y.; Tang, X.; Lin, H.; Sicheri, F.; Kay, L. E.; Tyers, M.; Forman-Kay, J. D. Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proceedings of the National Academy of Sciences* **2008**, *105* (46), 17772-17777.
(25) Mittag, T.; Kay, L. E.; Forman-Kay, J. D. Protein dynamics and conformational disorder in molecular recognition. *Journal of Molecular Recognition* **2009**, *23* (2), 105-116.
(26) Freiberger, M. I.; Wolynes, P. G.; Ferreiro, D. U.; Fuxreiter, M. Frustration in Fuzzy Protein Complexes Leads to Interaction Versatility. *The Journal of Physical Chemistry B* **2021**, *125* (10), 2513-2520.
(27) Sharma, R.; Raduly, Z.; Miskei, M.; Fuxreiter, M. Fuzzy complexes: Specific binding without complete folding. *FEBS Letters* **2015**, *589* (19PartA), 2533-2542.
(28) Miskei, M.; Antal, C.; Fuxreiter, M. FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies. *Nucleic Acids Research* **2016**, *45* (D1), D228-D235.
(29) Gianni, S.; Dogan, J.; Jemth, P. Coupled binding and folding of intrinsically disordered proteins: what can we learn from kinetics? *Current Opinion in Structural Biology* **2016**, *36*, 18-24.
(30) Rogers, J. M.; Oleinikovas, V.; Shammas, S. L.; Wong, C. T.; De Sancho, D.; Baker, C. M.; Clarke, J. Interplay between partner and ligand facilitates the folding and binding of an intrinsically disordered protein. *Proceedings of the National Academy of Sciences* **2014**, *111* (43), 15420-15425.
(31) Shammas, S. L.; Crabtree, M. D.; Dahal, L.; Wicky, B. I. M.; Clarke, J. Insights into Coupled Folding and Binding Mechanisms from Kinetic Studies. *Journal of Biological Chemistry* **2016**, *291* (13), 6689-6695.
(32) Iakoucheva, L. M.; Brown, C. J.; Lawson, J. D.; Obradović, Z.; Dunker, A. K. Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. *Journal of Molecular Biology* **2002**, *323* (3), 573-584.
(33) Dosnon, M.; Bonetti, D.; Morrone, A.; Erales, J.; di Silvio, E.; Longhi, S.; Gianni, S. Demonstration of a Folding after Binding Mechanism in the Recognition between the Measles Virus N_{TAIL} and X Domains. *ACS Chemical Biology* **2014**, *10* (3), 795-802.
(34) Rogers, J. M.; Wong, C. T.; Clarke, J. Coupled Folding and Binding of the Disordered Protein PUMA Does Not Require Particular Residual Structure. *Journal of the American Chemical Society* **2014**, *136* (14), 5197-5200.
(35) Arai, M.; Sugase, K.; Dyson, H. J.; Wright, P. E. Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding. *Proceedings of the National Academy of Sciences* **2015**, *112* (31), 9614-9619.
(36) Bourhis, J.-M.; Receveur-Bréchot, V.; Oglesbee, M.; Zhang, X.; Buccellato, M.; Darbon, H.; Canard, B.; Finet, S.; Longhi, S. The intrinsically disordered C-terminal domain of the measles virus nucleoprotein interacts with the C-terminal domain of the phosphoprotein via two distinct sites and remains predominantly unfolded. *Protein Science* **2005**, *14* (8), 1975-1992.
(37) Longhi, S.; Receveur-Bréchot, V.; Karlin, D.; Johansson, K.; Darbon, H.; Bhella, D.; Yeo, R.; Finet, S.; Canard, B. The C-terminal Domain of the Measles Virus Nucleoprotein Is Intrinsically Disordered and Folds upon Binding to the C-terminal Moiety of the Phosphoprotein. *Journal of Biological Chemistry* **2003**, *278* (20), 18638-18648.
(38) Schneider, R.; Blackledge, M.; Jensen, M. R. Elucidating binding mechanisms and dynamics of intrinsically disordered protein complexes using NMR spectroscopy. *Current Opinion in Structural Biology* **2019**, *54*, 10-18.
(39) Sugase, K.; Dyson, H. J.; Wright, P. E. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* **2007**, *447* (7147), 1021-1025.
(40) Ferreon, A. C. M.; Ferreon, J. C.; Wright, P. E.; Deniz, A. A. Modulation of allostery by protein intrinsic disorder. *Nature* **2013**, *498* (7454), 390-394.
(41) Gambin, Y.; VanDelinder, V.; Ferreon, A. C. M.; Lemke, E. A.; Groisman, A.; Deniz, A. A. Visualizing a one-way protein encounter complex by ultrafast single-molecule mixing. *Nature Methods* **2011**, *8* (3), 239-241.
(42) Marino, J.; Buholzer, K. J.; Zosel, F.; Nettels, D.; Schuler, B. Charge Interactions Can Dominate Coupled Folding and Binding on the Ribosome. *Biophysical Journal* **2018**, *115* (6), 996-1006.
(43) Sturzenegger, F.; Zosel, F.; Holmstrom, E. D.; Buholzer, K. J.; Makarov, D. E.; Nettels, D.; Schuler, B. Transition path times of coupled folding and binding reveal the formation of an encounter complex. *Nature Communications* **2018**, *9* (1).
(44) Karlsson, E.; Andersson, E.; Dogan, J.; Gianni, S.; Jemth, P.; Camilloni, C. A structurally heterogeneous transition state underlies coupled binding and folding of disordered proteins. *Journal of Biological Chemistry* **2019**, *294* (4), 1230-1239.
(45) Malagrinò, F.; Diop, A.; Pagano, L.; Nardella, C.; Toto, A.; Gianni, S. Unveiling induced folding of intrinsically disordered proteins – Protein engineering, frustration and emerging themes. *Current Opinion in Structural Biology* **2022**, *72* (72), 153-160.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(46) Toto, A.; Camilloni, C.; Giri, R.; Brunori, M.; Vendruscolo, M.; Gianni, S. Molecular Recognition by Templated Folding of an Intrinsically Disordered Protein. *Scientific Reports* **2016**, *6* (1).
(47) Löhr, T.; Kohlhoff, K.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. A Small Molecule Stabilizes the Disordered Native State of the Alzheimer's Aβ Peptide. *ACS Chemical Neuroscience* **2022**, *13* (12), 1738-1745.
(48) Zhu, J.; Salvatella, X.; Robustelli, P. Small molecules targeting the disordered transactivation domain of the androgen receptor induce the formation of collapsed helical states. *Nature Communications* **2022**, *13* (1).
(49) Camilloni, C.; Vendruscolo, M. Statistical Mechanics of the Denatured State of a Protein Using Replica-Averaged Metadynamics. *Journal of the American Chemical Society* **2014**, *136* (25), 8982-8991.
(50) Lindorff-Larsen, K.; Trbovic, N.; Maragakis, P.; Piana, S.; Shaw, D. E. Structure and Dynamics of an Unfolded Protein Examined by Molecular Dynamics Simulation. *Journal of the American Chemical Society* **2012**, *134* (8), 3787-3791.
(51) Löhr, T.; Kohlhoff, K.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. A kinetic ensemble of the Alzheimer's Aβ peptide. *Nature Computational Science* **2021**, *1* (1), 71-78.
(52) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences* **2018**, *115* (21), E4758-E4766.
(53) Robustelli, P.; Trbovic, N.; Friesner, R. A.; Palmer, A. G. Conformational Dynamics of the Partially Disordered Yeast Transcription Factor GCN4. *Journal of Chemical Theory and Computation* **2013**, *9* (11), 5190-5200.
(54) Paul, F.; Noé, F.; Weikl, T. R. Identifying Conformational-Selection and Induced-Fit Aspects in the Binding-Induced Folding of PMI from Markov State Modeling of Atomistic Simulations. *The Journal of Physical Chemistry B* **2018**, *122* (21), 5649-5656.
(55) Paul, F.; Wehmeyer, C.; Abualrous, E. T.; Wu, H.; Crabtree, M. D.; Schöneberg, J.; Clarke, J.; Freund, C.; Weikl, T. R.; Noé, F. Protein-peptide association kinetics beyond the seconds timescale from atomistic simulations. *Nature Communications* **2017**, *8* (1).
(56) Robustelli, P.; Piana, S.; Shaw, D. E. Mechanism of Coupled Folding-upon-Binding of an Intrinsically Disordered Protein. *Journal of the American Chemical Society* **2020**, *142* (25), 11092-11101.
(57) Peiffer, A. L.; Garlick, J. M.; Joy, S. T.; Mapp, A. K.; Brooks, C. L. Allostery in the dynamic coactivator domain KIX occurs through minor conformational micro-states. *PLOS Computational Biology* **2022**, *18* (4), e1009977.
(58) Robustelli, P.; Ibanez-de-Opakua, A.; Campbell-Bezat, C.; Giordanetto, F.; Becker, S.; Zweckstetter, M.; Pan, A. C.; Shaw, D. E. Molecular Basis of Small-Molecule Binding to α-Synuclein. *Journal of the American Chemical Society* **2022**, *144* (6), 2501-2510.
(59) Best, R. B.; Zheng, W.; Mittal, J. Correction to Balanced Protein–Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *Journal of Chemical Theory and Computation* **2015**, *11* (4), 1978-1978.
(60) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *The Journal of Physical Chemistry B* **2015**, *119* (16), 5113-5123.
(61) Piana, S.; Robustelli, P.; Tan, D.; Chen, S.; Shaw, D. E. Development of a Force Field for the Simulation of Single-Chain Proteins and Protein–Protein Complexes. *Journal of Chemical Theory and Computation* **2020**, *16* (4), 2494-2507.
(62) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov models of molecular kinetics: Generation and validation. *The Journal of Chemical Physics* **2011**, *134* (17), 174105.
(63) Noé, F.; Rosta, E. Markov Models of Molecular Kinetics. *The Journal of Chemical Physics* **2019**, *151* (19), 190401.
(64) Kohlhoff, K. J.; Shukla, D.; Lawrenz, M.; Bowman, G. R.; Konerding, D. E.; Belov, D.; Altman, R. B.; Pande, V. S. Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nature Chemistry* **2013**, *6* (1), 15-21.
(65) Raich, L.; Meier, K.; Günther, J.; Christ, C. D.; Noé, F.; Olsson, S. Discovery of a hidden transient state in all bromodomain families. *Proceedings of the National Academy of Sciences* **2021**, *118* (4).
(66) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. *Journal of the American Chemical Society* **2011**, *133* (45), 18413-18419.
(67) Plattner, N.; Noé, F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nature Communications* **2015**, *6* (1).
(68) Chakrabarti, K. S.; Olsson, S.; Pratihar, S.; Giller, K.; Overkamp, K.; Lee, K. O.; Gapsys, V.; Ryu, K.-S.; de Groot, B. L.; Noé, F.; et al. A litmus test for classifying recognition mechanisms of transiently binding proteins. *Nature Communications* **2022**, *13* (1).

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(69) Plattner, N.; Doerr, S.; De Fabritiis, G.; Noé, F. Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nature Chemistry* **2017**, *9* (10), 1005-1011.
(70) Scherer, M. K.; 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. *Journal of Chemical Theory and Computation* **2015**, *11* (11), 5525-5542.
(71) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *Journal of the American Chemical Society* **2018**, *140* (7), 2386-2396.
(72) Herrera-Nieto, P.; Pérez, A.; De Fabritiis, G. Characterization of partially ordered states in the intrinsically disordered N-terminal domain of p53 using millisecond molecular dynamics simulations. *Scientific Reports* **2020**, *10* (1).
(73) Konovalov, K. A.; Unarta, I. C.; Cao, S.; Goonetilleke, E. C.; Huang, X. Markov State Models to Study the Functional Dynamics of Proteins in the Wake of Machine Learning. *JACS Au* **2021**, *1* (9), 1330-1341.
(74) McGibbon, R. T.; Pande, V. S. Variational cross-validation of slow dynamical modes in molecular kinetics. *The Journal of Chemical Physics* **2015**, *142* (12), 124105.
(75) Litzinger, F.; Boninsegna, L.; Wu, H.; Nüske, F.; Patel, R.; Baraniuk, R.; Noé, F.; Clementi, C. Rapid Calculation of Molecular Kinetics Using Compressed Sensing. *Journal of Chemical Theory and Computation* **2018**, *14* (5), 2771-2783.
(76) Scherer, M. K.; Husic, B. E.; Hoffmann, M.; Paul, F.; Wu, H.; Noé, F. Variational selection of features for molecular kinetics. *The Journal of Chemical Physics* **2019**, *150* (19), 194108.
(77) Sengupta, U.; Carballo-Pacheco, M.; Strodel, B. Automated Markov state models for molecular dynamics simulations of aggregation and self-assembly. *The Journal of Chemical Physics* **2019**, *150* (11), 115101.
(78) Ravindra, P.; Smith, Z.; Tiwary, P. Automatic mutual information noise omission (AMINO): generating order parameters for molecular systems. *Molecular Systems Design & Engineering* **2020**, *5* (1), 339-348.
(79) Naritomi, Y.; Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *The Journal of Chemical Physics* **2011**, *134* (6), 065101.
(80) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *The Journal of Chemical Physics* **2013**, *139* (1), 015102.
(81) Noé, F.; Nüske, F. A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems. *Multiscale Modeling & Simulation* **2013**, *11* (2), 635-655.
(82) Nüske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S. J. S.; Noé, F. Variational Approach to Molecular Kinetics. *Journal of Chemical Theory and Computation* **2014**, *10* (4), 1739-1752.
(83) Wu, H.; Noé, F. Variational Approach for Learning Markov Processes from Time Series Data. *Journal of Nonlinear Science* **2019**, *30* (1), 23-66.
(84) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nature Communications* **2018**, *9* (1), 1-11.
(85) Mardt, A.; Pasquali, L.; Noé, F.; Wu, H. Deep learning Markov and Koopman models with physical constraints. *Proceedings of Machine Learning Research* **2020**, *107*, 451-475.
(86) Kingston, R. L.; Hamel, D. J.; Gay, L. S.; Dahlquist, F. W.; Matthews, B. W. Structural basis for the attachment of a paramyxoviral polymerase to its template. *Proceedings of the National Academy of Sciences of the United States of America* **2004**, *101* (22), 8301-8306.
(87) Gely, S.; Lowry, D. F.; Bernard, C.; Jensen, M. R.; Blackledge, M.; Costanzo, S.; Bourhis, J.-M.; Darbon, H.; Daughdrill, G.; Longhi, S. Solution structure of the C-terminal X domain of the measles virus phosphoprotein and interaction with the intrinsically disordered C-terminal domain of the nucleoprotein. *Journal of Molecular Recognition* **2010**, *23* (5), 435-447.
(88) Jensen, M. R.; Communie, G.; Ribeiro, E. A.; Martinez, N.; Desfosses, A.; Salmon, L.; Mollica, L.; Gabel, F.; Jamin, M.; Longhi, S.; et al. Intrinsic disorder in measles virus nucleocapsids. *Proceedings of the National Academy of Sciences* **2011**, *108* (24), 9839-9844.
(89) Bonetti, D.; Troilo, F.; Toto, A.; Brunori, M.; Longhi, S.; Gianni, S. Analyzing the Folding and Binding Steps of an Intrinsically Disordered Protein by Protein Engineering. *Biochemistry* **2017**, *56* (29), 3780-3786.
(90) Bonetti, D.; Troilo, F.; Brunori, M.; Longhi, S.; Gianni, S. How Robust Is the Mechanism of Folding-Upon-Binding for an Intrinsically Disordered Protein? *Biophysical Journal* **2018**, *114* (8), 1889-1894.
(91) Guseva, S.; Milles, S.; Jensen, M. R.; Salvi, N.; Kleman, J.-P.; Maurin, D.; Ruigrok, R. W. H.; Blackledge, M. Measles virus nucleo- and phosphoproteins form liquid-like phase-separated compartments that promote nucleocapsid assembly. *Science Advances* **2020**, *6* (14).

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(92) Ozenne, V.; Schneider, R.; Yao, M.; Huang, J.-r.; Salmon, L.; Zweckstetter, M.; Jensen, M. R.; Blackledge, M. Mapping the Potential Energy Landscape of Intrinsically Disordered Proteins at Amino Acid Resolution. *Journal of the American Chemical Society* **2012**, *134* (36), 15138-15148.
(93) Wang, Y.; Chu, X.; Longhi, S.; Roche, P.; Han, W.; Wang, E.; Wang, J. Multiscaled exploration of coupled folding and binding of an intrinsically disordered molecular recognition element in measles virus nucleoprotein. *Proceedings of the National Academy of Sciences* **2013**, *110* (40), E3743-E3752.
(94) Han, M.; Xu, J.; Ren, Y.; Li, J. Simulation of coupled folding and binding of an intrinsically disordered protein in explicit solvent with metadynamics. *Journal of Molecular Graphics and Modelling* **2016**, *68*, 114-127.
(95) Noe, F.; Wu, H.; Prinz, J. H.; Plattner, N. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. *J Chem Phys* **2013**, *139* (18), 184114.
(96) Molgedey, L.; Schuster, H. G. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters* **1994**, *72* (23), 3634-3637.
(97) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *Journal of Chemical Theory and Computation* **2013**, *9* (4), 2000-2009.
(98) Hadži, S.; Loris, R.; Lah, J. The sequence–ensemble relationship in fuzzy protein complexes. *Proceedings of the National Academy of Sciences* **2021**, *118* (37).
(99) Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; et al. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM* **2008**, *51* (7), 91-97.
(100) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334* (6055), 517-520.
(101) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22* (12), 2577-2637.
(102) Pietrucci, F.; Laio, A. A Collective Variable for the Efficient Exploration of Protein Beta-Sheet Structures: Application to SH3 and GB1. *Journal of Chemical Theory and Computation* **2009**, *5* (9), 2197-2201.
(103) E, W.; Vanden-Eijnden, E. Towards a Theory of Transition Paths. *Journal of Statistical Physics* **2006**, *123* (3), 503-523.
(104) Metzner, P.; Christof, S.; Vanden-Eijnden, E. Transition Path Theory for Markov Jump Processes. *Multiscale Modeling and Simulation* **2009**, *7* (3), 1192-1219.
(105) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences* **2009**, *106* (45), 19011-19016.
(106) Koopman, B. O. Hamiltonian Systems and Transformation in Hilbert Space. *Proceedings of the National Academy of Sciences* **1931**, *17* (5), 315-318.
(107) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* **2012**, *60* (6), 84-90.
(108) Bloyet, L.-M.; Brunel, J.; Dosnon, M.; Hamon, V.; Erales, J.; Gruet, A.; Lazert, C.; Bignon, C.; Roche, P.; Longhi, S.; et al. Modulation of Re-initiation of Measles Virus Transcription at Intergenic Regions by PXD to NTAIL Binding Strength. *PLOS Pathogens* **2016**, *12* (12), e1006058.
(109) Best, R. B.; Hummer, G. Reaction coordinates and rates from transition paths. *Proceedings of the National Academy of Sciences* **2005**, *102* (19), 6732-6737.
(110) Jenik, M.; Parra, R. G.; Radusky, L. G.; Turjanski, A.; Wolynes, P. G.; Ferreiro, D. U. Protein frustratometer: a tool to localize energetic frustration in protein molecules. *Nucleic Acids Research* **2012**, *40* (W1), W348-W351.
(111) Jemth, P.; Mu, X.; Engström, Å.; Dogan, J. A Frustrated Binding Interface for Intrinsically Disordered Proteins. *Journal of Biological Chemistry* **2014**, *289* (9), 5528-5533.
(112) Lawrence, C. W.; Kumar, S.; Noid, W. G.; Showalter, S. A. Role of Ordered Proteins in the Folding-Upon-Binding of Intrinsically Disordered Proteins. *The Journal of Physical Chemistry Letters* **2014**, *5* (5), 833-838.
(113) Parra, R. G.; Schafer, N. P.; Radusky, L. G.; Tsai, M.-Y.; Guzovsky, A. B.; Wolynes, P. G.; Ferreiro, D. U. Protein Frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrostatics. *Nucleic Acids Research* **2016**, *44* (W1), W356-W360.
(114) Stelzl, L. S.; Mavridou, D. A. I.; Saridakis, E.; Gonzalez, D.; Baldwin, A. J.; Ferguson, S. J.; Sansom, M. S. P.; Redfield, C. Local frustration determines loop opening during the catalytic cycle of an oxidoreductase. *eLife* **2020**, *9*.
(115) Gianni, S.; Freiberger, M. I.; Jemth, P.; Ferreiro, D. U.; Wolynes, P. G.; Fuxreiter, M. Fuzziness and Frustration in the Energy Landscape of Protein Folding, Function, and Assembly. *Accounts of Chemical Research* **2021**, *54* (5), 1251-1259.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.21.550103; this version posted July 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(116) Papoian, G. A.; Wolynes, P. G. The physics and bioinformatics of binding and folding-an energy landscape perspective. *Biopolymers* **2003**, *68* (3), 333-349.
(117) Chen, T.; Song, J.; Chan, H. S. Theoretical perspectives on nonnative interactions and intrinsic disorder in protein folding and binding. *Current Opinion in Structural Biology* **2015**, *30*, 32-42.
(118) Olsson, S.; Wu, H.; Paul, F.; Clementi, C.; Noé, F. Combining experimental and simulation data of molecular processes via augmented Markov models. *Proceedings of the National Academy of Sciences* **2017**, *114* (31), 8265-8270.
(119) Trendelkamp-Schroer, B.; Wu, H.; Paul, F.; Noé, F. Estimation and uncertainty of reversible Markov models. *The Journal of Chemical Physics* **2015**, *143* (17), 174101.
(120) Buchete, N.-V.; Hummer, G. Coarse Master Equations for Peptide Folding Dynamics. *The Journal of Physical Chemistry B* **2008**, *112* (19), 6057-6069.
(121) Pavliotis, G. A. *Stochastic Processes and Applications*; Springer, 2014.
(122) Hoel, P. G.; Port, S. C.; Stone, C. J. *Introduction to stochastic processes*; Waveland Press, 1987.
(123) Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for Hyper-Parameter Optimization. *Neural Information Processing Systems* **2011**, *24*.
(124) Bergstra, J.; Yamins, D.; Cox, D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *proceedings.mlr.press* **2013**, 115-123.
(125) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* **2019**.
(126) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury Google, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*; 2019. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.
(127) Hoffmann, M.; Scherer, M.; Hempel, T.; Mardt, A.; de Silva, B.; Husic, B. E.; Klus, S.; Wu, H.; Kutz, N.; Brunton, S. L.; et al. Deeptime: a Python library for machine learning dynamical models from time series data. *Machine Learning: Science and Technology* **2021**, *3* (1), 015009.
(128) Rigsby, R. E.; Parker, A. B. Using the PyMOL application to reinforce visual understanding of protein structure. *Biochemistry and Molecular Biology Education* **2016**, *44* (5), 433-437.