High-Resolution Markov State Models for the Dynamics of Trp-Cage Miniprotein Constructed Over Slow Folding Modes Identified by State-Free Reversible VAMPnets

Hythem Sidky,[†] Wei Chen,[‡] and Andrew L. Ferguson[*,†]

†*Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637*

‡ *Department of Physics, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, Illinois 61801*

E-mail: andrewferguson@uchicago.edu

Abstract

State-free reversible VAMPnets (SRVs) are a neural network-based framework capable of learning the leading eigenfunctions of the transfer operator of a dynamical system from trajectory data. In molecular dynamics simulations, these data-driven collective variables (CVs) capture the slowest modes of the dynamics and are useful for enhanced sampling and free energy estimation. In this work, we employ SRV coordinates as a feature set for Markov state model (MSM) construction. Compared to the current state of the art, MSMs constructed from SRV coordinates are more robust to the choice of input features, exhibit faster implied timescale convergence, and permit the use of shorter lagtimes to construct higher kinetic resolution models. We apply this methodology to study the folding kinetics and conformational landscape of the Trp-cage miniprotein. Folding and unfolding mean first passage times are in good agreement with prior literature, and a nine macrostate model is presented. The unfolded ensemble comprises a central kinetic hub with interconversions to several metastable unfolded conformations and which serves as the gateway to the folded ensemble. The folded ensemble comprises the native state, a partially unfolded intermediate "loop" state, and a previously unreported short-lived intermediate that we were able to resolve due to the high time-resolution of the SRV-MSM. We propose SRVs as an excellent candidate for integration into modern MSM construction pipelines.

1 Introduction

Molecular dynamics (MD) simulations are an indispensable tool in the study of the conformational, thermodynamic, and kinetic properties of biomolecular systems. Advances in MD software and hardware have enabled access to millisecond timescales at atomistic resolution, but a major challenge is how to best analyze these large simulated trajectories to extract experimentally-meaningful kinetic and thermodynamic quantities.

Markov State Models (MSMs) have emerged as a powerful framework for analyzing MD simulations and recovering dynamical properties of interest.[1] Their primary innovation is to discretize high-dimensional molecular conformational space into coarse-grained states, wherein the dynamical interconversions between microstates within a macrostate are fast relative to transitions between macrostates. Accordingly, the macrostate dynamical transitions are approximately memoryless (i.e., Markovian) and can be modeled by a master equation.[2] Protein folding has benefited immensely from developments in MSM methodology which have pushed the limits of recoverable long-term kinetics while simultaneously yielding insight into microscopic quantities.[3,4] Nevertheless, the quality of a MSM is highly dependent on the input features, state space decomposition, and a number of parameters chosen during its construction. This has motivated research into optimizing each stage of the MSM pipeline including theory,[5,6] basis selection,,[7,8] clustering,[9] and validation.[10,11]

The current state of the art in MSM construction involves the use of time-lagged independent component analysis (TICA)[2,7,12] to identify a linearly-optimal combination of input features which maximizes their kinetic variance. Clustering is then performed in this slow subspace to produce the states between which interconversion rates are estimated. TICA has all but superseded structural clustering based on metrics such as minimum root mean square distance (RMSD) that tend to capture motions of high structural variance as opposed to the desired slowest motions.[1,12] A recently proposed alternative to MSMs are VAMPnets, an artificial neural network (ANN) approach that seeks to replace the entire MSM pipeline.[13] VAMPnets are a very promising new technique, but as an end-to-end replacement to MSM

construction cannot yet be interfaced with the extensive machinery and extensions developed for MSMs such as statistical error estimators, rare event sampling techniques, and incorporation of experimental constraints.[13]

In a recent work, we proposed state-free reversible VAMPnets (SRVs)[14] as a deep learning framework based on VAMPnets,[13] which themselves are based on deep canonical correlation analysis (DCCA).[15] Contrary to VAMPnets, SRVs were designed not to approximate MSMs but rather to directly learn nonlinear approximations to the slowest dynamical modes of a molecular system obeying detailed balance. The approach is founded on the variational approach to conformational dynamics (VAC), which defines a variational principle for the slowest eigenfunctions of the transfer operator that propagates state functions through time.[6,16] The essence of our approach is to use twin-lobed neural networks to learn the best nonlinear basis set to pass to the linear variational problem defined by the VAC. The VAC then furnishes the optimal eigenvector approximations of the transfer operator ordered by decreasing implied timescales. Following VAMPnets, we deviate from DCCA in choosing as our loss function the VAMP-2 score informed by the variational approach to Markov processes (VAMP) principle.[13] Contrary to VAMPnets, we modify our network architecture to directly approximate the slow modes of the transfer operator rather than soft metastable state assignments, and employ the variational approach under detailed balance to approximate the slow modes of equilibrium dynamics. (Our prefix "state-free reversible" reflects these two key differences.) SRVs can also be viewed as a multi-dimensional generalization of variational dynamics encoder,[17] a variational analog to time-lagged autoencoders,[18] and are closely related to kernel TICA.[19]

In this work, we demonstrate the utility of employing the slow modes recovered by SRVs as a basis within which to construct MSMs. This study was motivated by the hypothesis that compared to MSMs based on linear TICA approximations to the transfer operator eigenfunctions, MSMs constructed from the nonlinear SRV approximations would permit the use of shorter lagtimes and therefore furnish models with higher kinetic resolution. Whereas

VAMPnets perform nonlinear featurization, slow-mode estimation, and soft clustering into metastable states macrostates,[18] SRVs perform only the first two steps. The final step of MSM construction is performed using standard protocols utilizing the slow modes learned by SRVs rather than TICA coordinates. In this manner, we take advantage of the large body of theoretical work and mature numerical implementations developed for MSM construction[1,2,20,21] where SRVs serve as a modular replacement for TICA. SRV-MSMs are shown to perform better than TICA-MSMs under cross validation, offer more flexibility and robustness in feature selection, and converge implied timescales quicker, allowing for shorter lagtimes and ultimately a higher resolution kinetic model. VAMPnets and SRV-MSMs perform comparably, but, as we will show, the SRV-MSM exhibits slightly faster convergence of the implied timescales and enables access to the statistical error estimators,[22] multi-ensemble approaches,[23] and other extensions developed for MSMs.[13,24]

We demonstrate SRV-MSMs in an application to an ultra-long 208 μs explicit solvent simulation of the K8A mutant of Trp-cage TC10b at 290 K performed by D.E. Shaw Research.[25] Trp-cage is a fast-folding miniprotein that has been the subject of numerous experimental[26,27] and computational studies.[27–32] Despite its status as an archetypal miniprotein for the testing of new computational methods, its kinetic behavior remains incompletely understood. Given the sensitivity of the Trp-cage folding landscape to mutations[26] and termini,[33] a direct comparison of the behavior of different mutants is not possible. The K8A mutant of Trp-cage TC10b considered in this work has been previously studied by Dickson & Brooks,[34] who determined that the Trp-cage unfolded ensemble displays two-state behavior. Suárez et al.[35] analyzed the same data using non-Markovian techniques to determine mean first passage times (MFPT) between the folded and unfolded states. Deng et al. conducted perhaps the most comprehensive study of the the kinetics of this data to date,[36,37] identifying two representative folding mechanisms: the hydrophobic collapse of Trp-cage into a molten globule followed by the formation of the N-terminal α-helix and native core (nucleation-condensation), and the pre-formation of the α-helix in an extended unfolded

state then the joint formation the 3_{10} helix and hydrophobic core (diffusion collision). The diffusion-collision mechanism is identified as the dominant folding pathway with a substantially smaller transit time of 3 ns, compared to 42 ns for nucleation-condensation. The high kinetic resolution of the model furnished by SRV-MSMs in the present work establishes new understanding of the Trp-cage folding mechanism, and demonstrates SRVs as a valuable tool in the construction of high kinetic resolution MSMs.

2 Methods

We now proceed to describe our SRV-MSM construction pipeline, comprising the following six steps: (i) feature selection, (ii) SRV learning of the slow modes, (iii) definition of microstates and microstate transition rates by k-means clustering in the SRV coordinates, (v) definition of MSM macrostates and macrostate transition rates by spectral clustering of the microstate transition matrix, and (vi) comparison of the resulting SRV-MSM with a TICA-MSM and VAMPnets. The molecular simulation we study is a 208 μs explicit solvent K8A mutant of Trp-cage TC10b simulation performed by D.E. Shaw Research.[25] The protein was prepared with the Asp and Arg side chains and N- and C- termini in their charged states and modeled using the CHARMM22* forcefield. The protein was immersed in a cubic box of side length $\sim$3.7 nm along with $\sim$1700 TIP3P water molecules and a number of sodium and chloride ions to neutralize charge and bring the NaCl concentration up to 65 mM. Simulations were equilibrated in the NPT ensemble for 1 ns before being passed to the special-purpose Anton hardware for the 208 μs NVT production run employing a 2.5 fs integration time step and a Nosé-Hoover thermostat with a 1 ps time constant. Short-range electrostatics and Lennard-Jones interactions were treated using a 0.9 nm cutoff and long-range electrostatics treated using the Gaussian Split Ewald (GSE) method over a $32\times32\times32$ grid. Simulation snapshots were saved at intervals of 200 picoseconds to produce a trajectory containing approximately 10^6 frames.

2.1 Molecular feature selection

In order to perform slow variable discovery we must first define the set of features derived from each instantaneous configuration of the molecular system that will be used to represent the trajectory to the learning algorithm. Scherer et al.[38] have recently shown that feature choices can be optimized directly through a variational principle based on VAMP scoring without requiring construction of the entire kinetic model. The scoring method, known as VAMP-2 scoring,[39] is the sum of the squared estimated eigenvalues of the transfer operator. Under this variational approach, larger cross-validated VAMP-2 scores correspond to more kinetically accurate models and the cross-validated test score is bounded from above by the true kinetic model. We employ this method of variational feature selection using backbone and sidechain torsions, Cα pairwise atom distances, a combination of these two features, and the aligned Cartesian coordinates of the entire molecule. Figure 1 shows the result of ten-fold cross-validated VAMP-2 scoring for the aforementioned feature sets at different lagtimes τ using the top ten eigenvalues. It is clear the the combined set of torsions and Cα pairwise distances contain more kinetic variance at all lagtimes considered,[8] and hence should be preferred over the other feature sets. The aligned Cartesian coordinates consistently underperform the other choices. We use the combined set of torsions and Cα pairwise distances for all further analysis unless otherwise stated.

2.2 SRVs outperform TICA-MSMs under cross-validation

TICA-MSM models were built using PyEMMA[20] 2.5.4 following the general protocol outlined in Ref.[40] Using the combined feature set, the number of TICA dimensions, TICA lagtime, and number of cluster centers were optimized under VAMP-2 scoring. Specifically, the VAMP-2 score was maximized under five TICs computed at a lagtime of 20 ns. The simulation trajectory was projected into these five leading TICs and clustered into 200 microstates using k-means clustering. The VAMP-2 score was found to be quite insensitive to the number of selected microstates over the range 20 to 500, which motivated us to select 200. The resulting

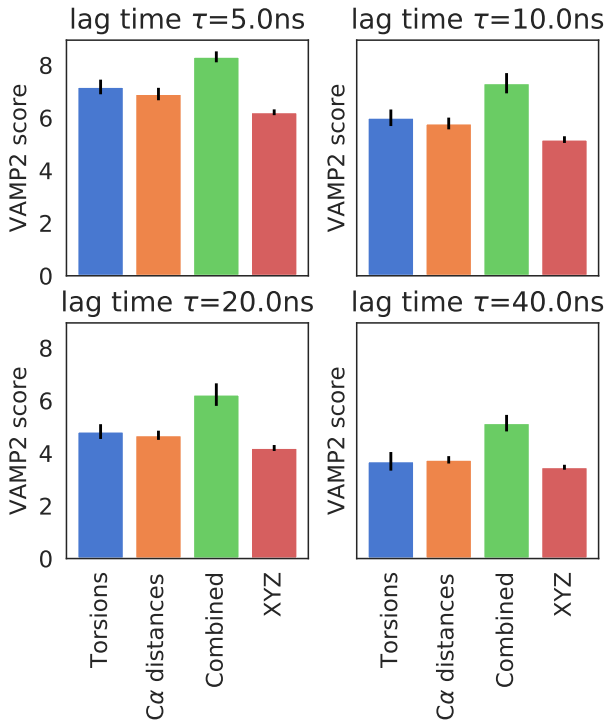


Figure 1: Molecular feature selection. VAMP-2 scores of the five slowest processes for various feature transformations of the Trp-Cage trajectory calculated at a variety of lagtimes τ: backbone and sidechain torsions (torsions), Cα pairwise atom distances (Cα distances), a combination of the previous two features (combined), and the aligned Cartesian coordinates of the entire molecule (XYZ). The combined featurization comprising backbone and sidechain torsions and C_α pairwise distances is superior across all tested lagtimes and is used for all subsequent analysis.

microstate transition matrix computed over these clusters was then diagonalized to identify the leading eigenvectors and associated implied timescales of the kinetic model. Although we do not do so here, the 200 microstates may be coarsened into a far smaller number of macrostates by performing PCCA++ clustering over these leading eight microstate transition matrix eigenvectors. SRVs were trained using the SRV package (`https://github.com/hsidky/srv`) with the default architecture of two hidden layers with 100 neurons each, and tanh activation functions. We specified a batch size of 500,000, a learning rate of 0.01, and employed batch normalization within all hidden layers. No early stopping or weight decay was used to maximize data utilization and avoid having to tune regularization strength.

Instead, we screened for number of training epochs as part of the VAMP-2 optimization. All code used for model screening, selection, and generation of results can be found in the repository `https://github.com/hsidky/srv-trpcage`.

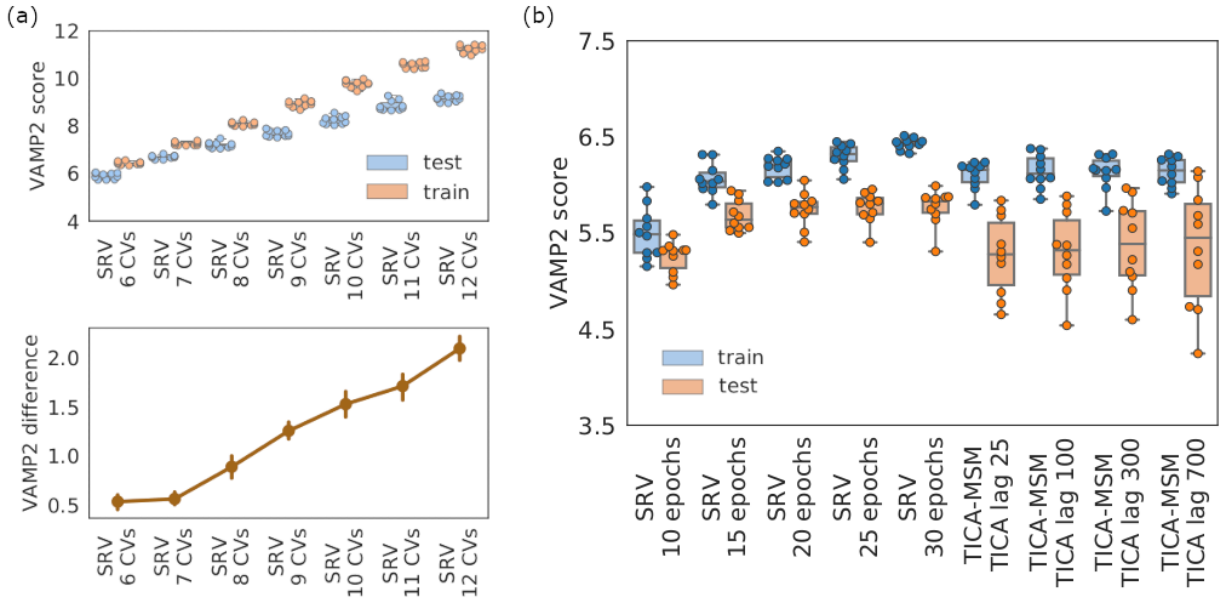


Figure 2: SRV and TICA-MSM model validation. (a) Ten-fold cross-validated VAMP-2 scores for SRV models containing an increasing number of SRV coordinates constructed at a lagtime of 20 ns (upper panel). An increase in the gap between testing and training scores (lower panel) indicates the onset of overfitting and motivates the selection of a seven SRV coordinate model as that best supported by the data. (b) Cross validation of the SRV training epochs and TICA lagtime in steps (5 steps = 1 ns) hyperparameters against the VAMP-2 score demonstrate SRVs have higher train and test scores and narrower distributions, which is indicative of model robustness and generalizability.

We used ten-fold cross-validated VAMP-2 scores to compare the quality of different SRV models and TICA-MSM models. Specifically, to maximize the similarity between the train and test data distributions, we first divided the full 208 μs trajectory into 100 equal segments which are treated as independent trajectories for the purposes of our comparative analysis. The segments are then shuffled and subsampled as part of train-test split procedure for each fold. This approach has the drawback of losing transitions across the individual segments, but it ensures that the conformational distribution over the complete trajectory is well represented in both the training and testing sets. Note that here we choose to compare

the VAMP-2 scores of the TICA-MSMs directly to the SRVs rather than a subsequent SRV-MSM. The primary reason for this is that we want to make clear the contribution of the SRV coordinates themselves to the kinetic content without additional processing. We present a comparison between the TICA-MSM and SRV-MSM implied timescales later on in Section 2.4.

To determine the number of eigenvalues to retain for cross validation, we calculate train and test VAMP-2 scores for SRVs of increasing dimensionality. From Figure 2a, there is a marked increase in the gap between training and testing scores after seven dimensions, which is indicative of overfitting and motivating our choice of the seven SRV eigenvector model as that best supported by the data. The SRVs were trained at a lagtime of 20 ns which is the same lagtime used for TICA-MSM construction and VAMP-2 scoring.

Figure 2b presents the cross-validated VAMP-2 scoring for the SRV and TICA-MSM models. Both classes of models perform well but with some significant differences. While the distributions of training scores are very similar, SRVs display remarkable consistency in test scores compared to the TICA-MSMs. TICA-MSM test scores vary considerably between folds, which is indicative of model sensitivity to training data and characteristic of overfitting. The SRVs show consistent improvement in training scores with the number of training epochs, but the plateau in the testing score and widening gap between the training and test scores after 20 epochs signals overfitting. The 30 epoch model still yields a marginally higher test VAMP-2 score than other epochs, which is our selection, but the difference between 20, 25, and 30 epochs is insignificant. The TICA lagtime does not appear to have much of an impact on train or test score means, although we do note a marked increase in the testing variance for the largest lagtime of 700 steps (140 ns).

For comparison, Figure 3 shows the result of ten-fold cross validation for RMSD-based MSMs for increasing number of microstates k. An RMSD-MSM with $k = 25{,}000$ microstates was previously utilized by Deng et al.[36] in the analysis of the D.E. Shaw 208 μs Trp-cage simulation considered herein. The training VAMP-2 score increases with the number of mi-

crostates, which results higher implied timescales and seemingly better performance. However, the test scores remain approximately constant, with a small decrease at $k = 10{,}000$. This widening gap between testing and training scores is indicative of overfitting, and although the RMSD-MSM training scores are similar to TICA-MSM and SRVs, the test scores are significantly worse for all values of k.

The higher train and test scores of the SRVs and improved variance over TICA-MSMs and RMSD-MSMs indicate that they are more kinetically accurate, capture more information about the system dynamics, and thus present an excellent basis in which to construct kinetic models of the system dynamics.

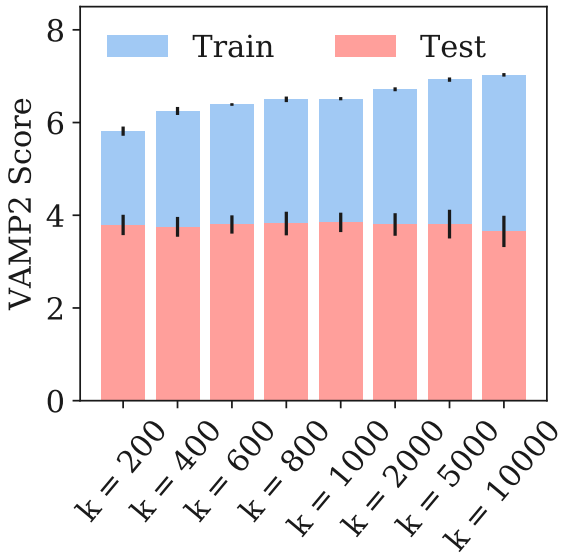


Figure 3: Ten-fold cross validated VAMP-2 scores for RMSD-MSMs. Although the training score increases with number of states, the widening gap between the test and training scores is indicative of overfitting.

2.3 SRVs are robust to the choice of feature set

We showed in Section 2.2 that using the same optimized feature set, SRVs outperform TICA-MSMs under cross-validation. We now address the situation where sub-optimal features are

used to construct both models. Empirical evidence suggests that it may be useful[38] to generate a bank of distance or contact-based features that are nonlinear featurizations of the atomic coordinates to improve MSM quality. Examples of these transformations include reciprocals, logarithms, polynomials, or exponentials of pairwise distances. Improvement is possible since TICA is restricted to discover linear combinations of the input features, and nonlinear feature engineering can introduce nonlinearities into the model. Since SRVs are based on a deep learning architecture, the universal approximation theorem[41,42] asserts that they should, by employing sufficiently many hidden nodes, be capable of discovering nonlinear feature transformations to maximize the kinetic variance from rather poor choices of input feature sets without extensive feature engineering. Here, we test this conjecture by omitting the backbone and sidechain torsions from the feature set.

Figure 4 presents a visualization of the top seven SRV and top seven TICA-MSM eigenvectors constructed over two feature sets: one comprising $C\alpha$ pairwise distances only, and one comprising $C\alpha$ pairwise distances plus backbone and sidechain torsions. To perform this comparison we project the snapshots from the molecular simulation trajectory into the leading TICA coordinates (TIC1-7) obtained in construction of the TICA-MSM under the $C\alpha$ pairwise distances plus backbone and sidechain torsions feature set. We choose to visualize along TICA coordinates since they contain more variance than the SRV or TICA-MSM eigenfunctions, which makes them more suitable for visualization purposes. Each point is colored according to the corresponding value of each of the leading seven eigenvectors computed by a SRV or diagonalization of the TICA-MSM microstate transition matrix under a molecular featurization employing either $C\alpha$ pairwise distances only or $C\alpha$ pairwise distances plus backbone and sidechain torsions. The key difference between the feature sets emerges in the second slow mode (TIC2, second column) learned from the combined $C\alpha$ pairwise distances plus backbone and sidechain torsions, where the SRV constructed using only $C\alpha$ pairwise distances (second row) is able to learn a transition along TIC2 whereas the MSM trained on only $C\alpha$ pairwise distances data (fourth row) fails to do so. Furthermore, the SRV trained

only on $C\alpha$ pairwise distances (second row) successfully discovers the remaining higher-order modes with only a minor degradation in the implied timescales relative to the SRV trained on torsions and $C\alpha$ pairwise distances (first row). The dynamical motion associated with TIC2 has a timescale of $t_1 \approx 1$ μs, and by failing to account for it a significant contribution to the kinetic variance is lost. The nonlinear nature of the SRV enabled it to form nonlinear combinations of the $C\alpha$ pairwise distances input features to discover the dynamical motions associated with torsional angles necessary to resolve this mode. SRVs are therefore able to discover an important slow dynamical mode that is invisible to a TICA-MSM presented with the same data. This capability is particularly valuable in extracting maximal kinetic variance from suboptimal input feature sets, and can be used in concert with VAMP scoring to identify the optimal feature set without extensive manual feature engineering.

2.4 Implied timescales of SRV-MSMs exhibit faster convergence than TICA-MSMs and VAMPnets

We demonstrated in Section 2.2 that the leading SRV eigenvectors present a good basis in which to represent the long time system dynamics, and that cross-validation with respect to the VAMP-2 score showed the kinetic model based on the top seven SRV eigenvectors to be best supported by the data. We now proceed to use these coordinates to construct a SRV-MSM with which we may propagate the long-time evolution of the system and analyze for its macrostate configurational discretization, stationary state occupancy probabilities, dwell times, and transition rates.

The SRV-MSM was constructed using the PyEMMA software package.[20] A microstate transition matrix comprising 100 microstates, where this number was selected by hyperparameter optimization, was constructed by performing k-means clustering of projections of the simulation trajectory into the leading seven SRV eigenvectors. Diagonalization of the microstate transition matrix reveals eight leading timescales followed by a spectral gap, motivating the construction of a nine macrostate SRV-MSM. The nine metastable macrostates

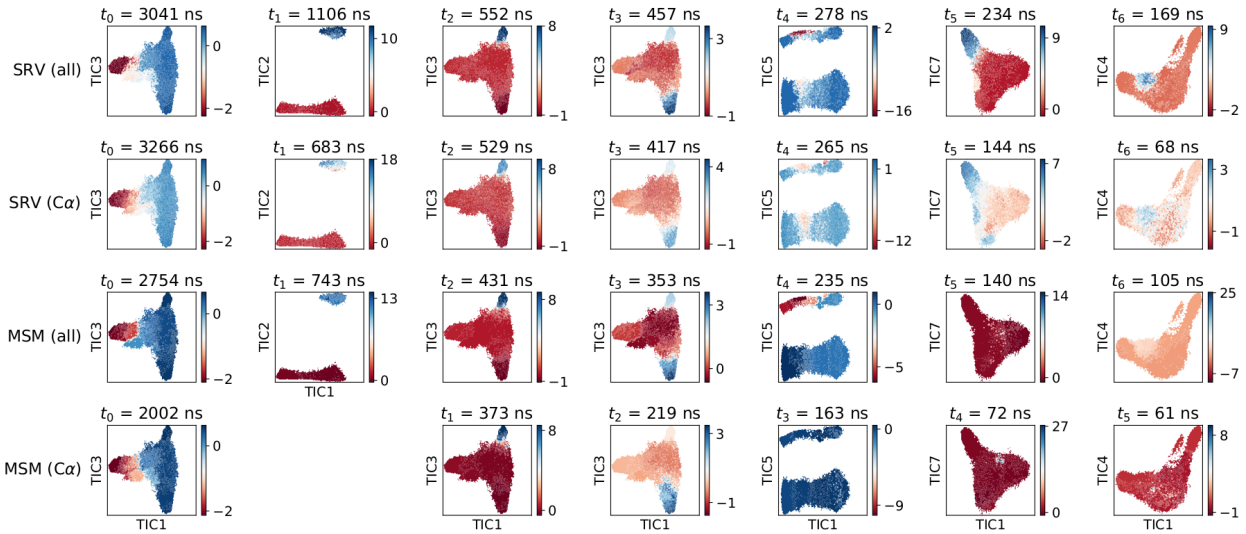


Figure 4: Visualization and comparison of the top seven eigenvectors determined by the SRV and TICA-MSM under two different feature sets: $C\alpha$ pairwise distances only ($C\alpha$), and $C\alpha$ pairwise distances plus backbone and sidechain torsions (all). The snapshots from the molecular simulation trajectory are projected into the leading TICA coordinates TIC1-7 obtained in construction of the TICA-MSM under the $C\alpha$ pairwise distances plus backbone and sidechain torsions feature set. Each point is colored according to the eigenvectors computed by: a SRV employing a feature set comprising $C\alpha$ pairwise distances plus backbone and sidechain torsions (first row), a SRV employing $C\alpha$ pairwise distances only (second row), a TICA-MSM employing $C\alpha$ pairwise distances plus backbone and sidechain torsions (third row), and a TICA-MSM employing $C\alpha$ pairwise distances only (fourth row). The leading seven eigenvectors n=0-6 computed under each SRV or TICA-MSM are presented in decreasing order of slowness in the columns of the plot and the implied timescales t_n associated with each eigenvector is printed above each panel. SRVs (first row) and TICA-MSMs (third row) trained on $C\alpha$ pairwise distances and backbone and sidechain torsions discover the same leading five slowest modes, although the SRV discovers slower timescales. Excluding torsions from the input feature set renders the second leading mode TIC2 (second column, $t_1 \approx 1\ \mu$s) invisible to the TICA-MSM (fourth row), whereas a SRV (second row) can adequately resolve it and the remaining higher-order modes by forming nonlinear combinations of the $C\alpha$ pairwise distances input features.

$\mathcal{S}_{0-8}$ are computed by performing PCCA++ spectral clustering over the leading eight eigenvectors of the microstate transition matrix.[43–45] In this way the SRV coordinates serve as a modular replacement of TICA coordinates in performing microstate clustering within the MSM construction pipeline and, as we shall demonstrate, enable the construction of kinetic models with higher temporal resolution. Figure 5a shows the eight implied timescales to

converge extremely rapidly with lagtime τ, enabling selection of a very short $\tau = 10$ ns lagtime and construction of a high temporal resolution SRV-MSM. To validate the resulting SRV-MSM, we conduct a Chapman-Kolmogorov (CK) test. The CK test compares the transition probabilities between pairs of states $i \rightarrow j$ at a lagtime of $k\tau$ predicted by a model constructed at a lagtime τ and that computed directly from a model constructed at a lagtime $k\tau$. Figure 5b presents the results of the CK test for within state (i.e., $i \rightarrow i$) transitions. We observe that our $\tau = 10$ ns lagtime model performs excellently in predicting the transition probabilities even out to very long times of $k\tau = 200$ ns. This CK analysis demonstrates that the SRV-MSM kinetic model is Markovian for a $\tau = 10$ ns lagtime, where this high temporal resolution is made possible by our use of SRV eigenvectors for microstate clustering.

To compare SRV-MSMs, TICA-MSMs, and VAMPnets, we present in Figure 6 a close-up of the convergence of the implied timescales as a function of lagtime for the optimized TICA-MSM (Section 2.2), SRV-MSM, and an equivalent nine-state VAMPnet constructed at the same $\tau = 10$ ns lagtime. Due to the congested nature of this plot, we choose to plot only the leading six implied timescales for clarity. The SRV-MSM converges the slowest implied timescale at approximately five times shorter lagtimes than the TICA-MSM or VAMPnets. Convergence of the higher-order timescales is similar for VAMPnets and the SRV-MSM, whereas the TICA-MSM fails to converge to the same values even at quite long lagtimes. This trend can be attributed to the fact that the SRV-MSM and VAMPnets are able to learn nonlinear transformations of the input coordinates and therefore better resolve slower processes that are invisible to the inherently linear TICA-MSM (cf. Section 2.3).

In summary, the convergence of the implied timescales and validation of the CK test demonstrates that the SRV-MSM based on seven SRV coordinates (selected by cross-validating the training and testing VAMP-2 scores), nine metastable macrostates (selected by a gap in the microstate eigenvalue spectrum after the eighth non-trivial eigenvalue), and a lagtime of $\tau = 10$ ns (estimated by convergence of implied timescales) presents a good kinetic model

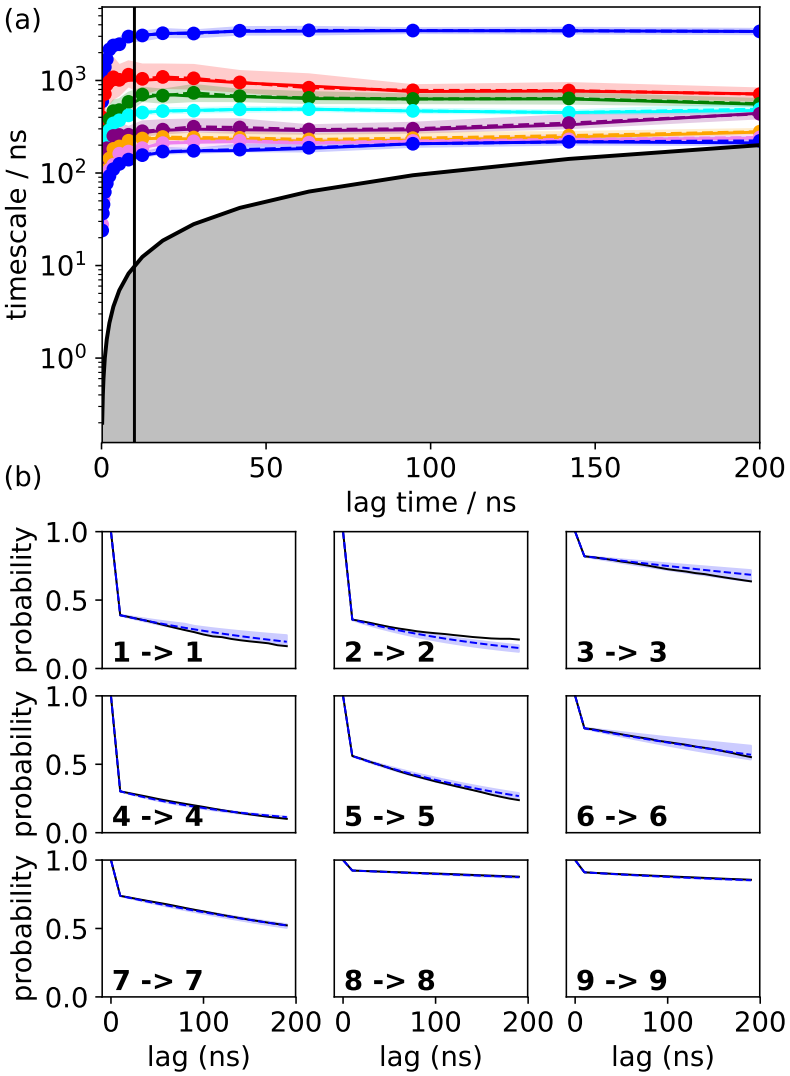


Figure 5: Validation of the SRV-MSM. (a) Convergence of the eight implied timescales of the nine-macrostate SRV-MSM as a function of lagtime. Solid lines indicate maximum likelihood result while dashed lines show the Bayesian ensemble means. The SRV-MSM timescales converge at a lagtime of $\tau = 10$ ns (vertical line). The black solid curve marks equality of the implied timescale and lagtime and delimit the shaded region where the implied timescales are shorter than the lagtime and cannot be resolved. (b) The Chapman-Kolmogorov (CK) test comparing the probabilities of remaining within each of the nine macrostates as a function of lagtime predicted by a SRV-MSM constructed at a $\tau = 10$ ns lagtime (dashed blue line) and those computed from a SRV-MSM constructed at the particular lagtime (solid black line). In both panels the shaded areas represent 95% confidence intervals. Rapid convergence of the implied timescales and agreement of the predicted and computed transition probabilities confirm the dynamic validity of the SRV-MSM.

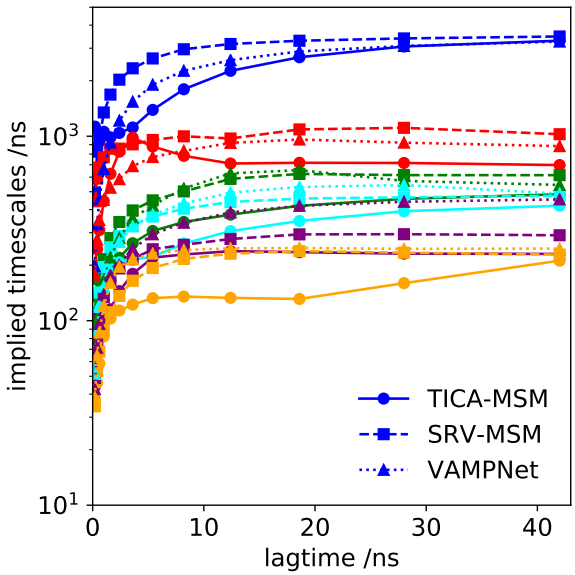


Figure 6: Close-up of the convergence of the leading six implied timescales as a function of lagtime for TICA-MSM (solid line, circles), SRV-MSM (dashed line, squares), and VAMPnets (dotted line, triangles). The SRV-MSM converges the implied timescales at approximately five times shorter lagtimes than VAMPnets or TICA-MSMs, enabling the construction of an extremely high time resolution MSM.

for the long-term system dynamics at a higher temporal resolution than is accessible using a TICA-MSM. This demonstrates the value of a modular replacement of TICA coordinates conventionally used for microstate clustering by SRV coordinates within an MSM pipeline in order to achieve higher temporal resolution MSM models while preserving access to the large body of tools and infrastructure developed for the construction, validation, and analysis of Markov state models.[22–24,46]

3 Results and Discussion

We now commence our analysis of Trp-cage folding dynamics based on the SRV-MSM constructed and validated in Section 2. It is first useful to visualize low-dimensional free energy landscapes illustrating the nine macrostates in order to generate an overview of the relative locations of the metastable macrostates of the model. As is customary,[20] we visualize the free

energy landscapes in the leading TICA coordinates as good high-variance collective variables in which to construct and display the free energy surface. We emphasize that these TICA coordinates are used exclusively as convenient linear collective variables that support good visualizations, whereas the MSM is constructed from the nonlinear SRV coordinates. To obtain more accurate free energy estimates along the TICA coordinates, we reweight each frame of the simulation trajectory by the associated values of the stationary distribution computed from the 100 microstate transition matrix, project these weighted data onto the leading TICA coordinates TIC1-7, and then estimate free energy surfaces from the empirical probability distributions within this space. We display selected 2D projections of the free energy surface within pairs of TICs in the top row of Figure 7, and in the bottom row show the clustering into the nine metastable macrostates $\mathcal{S}_{0-8}$ computed from PCCA++ spectral clustering.[43–45]

We caution against over-interpreting low-dimensional free energy landscape projections, but the gross features of the landscape are a folded state represented by $\mathcal{S}_7$ connected to the large unfolded ensemble of states $\mathcal{S}_{2-6,8}$ by a narrow neck. $\mathcal{S}_{2-6}$ represent structured metastable conformations within the unfolded ensemble. $\mathcal{S}_2$ corresponds to an extended conformation with outwardly rotated prolines, $\mathcal{S}_3$ a crossed conformation with a minor central hairpin, $\mathcal{S}_4$ a braided hairpin-like structure, $\mathcal{S}_5$ a hairpin, and $\mathcal{S}_6$ a configuration with a collapsed N-terminus and an extended C-terminus. We provide a finer-grained molecular-level description and visualization of the states when we discuss the macrostate transition matrix.

The first non-trivial right eigenvector of the macrostate transition matrix describes the transition to and from $\mathcal{S}_7$ which represents the folded state. Based on this definition, there are 12 observed folding and unfolding events in the trajectory, which agrees with the value reported by Lindorff-Larsen et al.,[25] who used a native contacts-based definition of folded and unfolded states. The fraction of native contacts, Q, has been previously shown to accurately characterize the thermodynamics of protein folding in and out of the native state.[47,48]

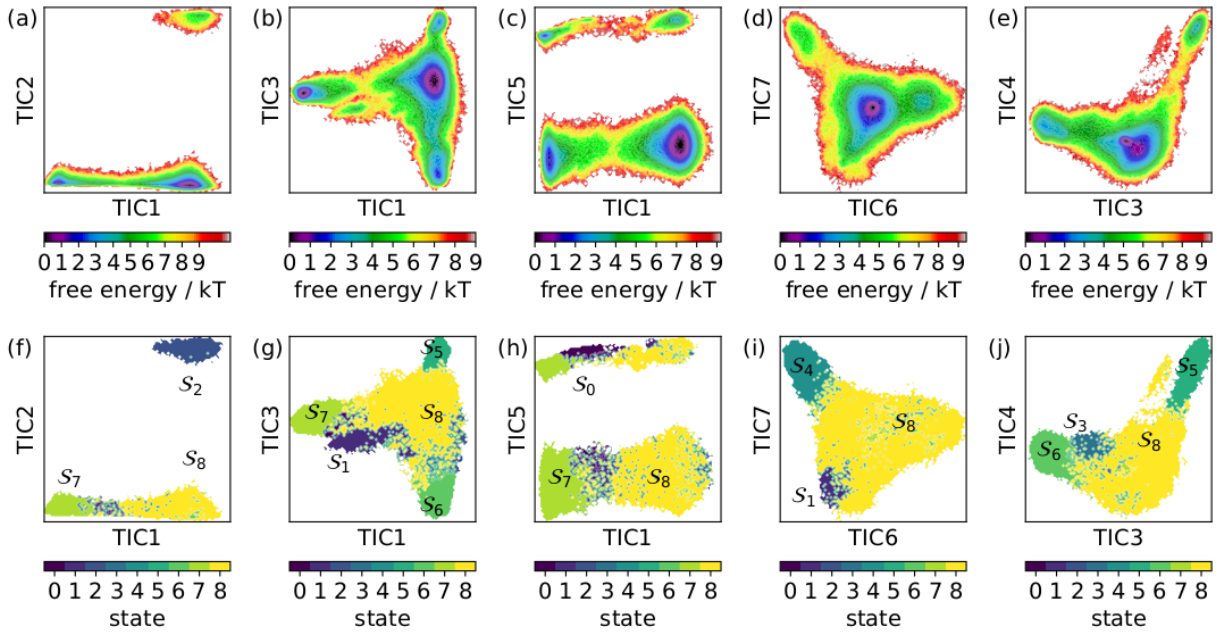


Figure 7: Free energy surfaces and macrostate clustering visualizations projected into the leading TICA coordinates (TIC1-7) for visualization purposes. (a-e) Free energy surface projected onto various pairwise combinations of TICs. The arbitrary additive constant in the free energy was fixed such that the lowest free energy in each plot was set to zero. Accordingly, absolute values of free energy within each plot are not meaningful but relative differences are. It is also not meaningful to compare free energies across the different plots. (f-j) Metastable macrostate assignments $\mathcal{S}_{0-8}$ computed by PCCA++ within the same TIC projections. Macrostate $\mathcal{S}_7$ contains the folded state.

Conversely, this figure is in poor agreement with the 31 folding transitions reported by Deng et al.[36] who use an RMSD-based definition of folding. This choice of an RMSD distance introduces a number of additional rapid folding transitions, and – based on the good agreement between the Q-based and MSM-based definitions of folding – suggests that this structural measure is a poor proxy for kinetic proximity.

We report in Table 1 the mean first passage times (MFPTs) into and out of the the folded state, $\mathcal{S}_7$, with uncertainties estimated using a Bayesian scheme emnploying 50 samples.[40,49,50] Our calculated MFPTs are in good agreement with Lindorff-Larsen et al.[25] who report values of 14.4 μs and 3.1 μs for folding and unfolding respectively. Indeed, coarsening our MSM from nine to two macrostates gives us near perfect agreement with a folding MFPT

of 14.0 μs and an unfolding MFPT of 3.0 μs. The high temporal resolution models produced by the rapid convergence of our implied timescales with lagtime is likely the key reason for the high accuracy MFPT estimates from our SRV-MSM. Suarez et al.[35] analyzed this same data using higher-order Markov approaches to report folding and unfolding MFPTs of 8.4 μs and 1.9 μs, respectively. The discrepancy may be due to different macrostate definitions. Laser temperature-jump spectroscopy measurements on the TC5b Trp-cage mutant conducted by Hagen and co-workers resolve folding and unfolding times of $\sim$4 μs and $\sim$13 μs at 293 K.[51,52] Although this is a different Trp-cage variant from the TC10b mutant studied herein, the folding and unfolding times are of the same order of magnitude as our calculated values.

Table 1: Calculated MFPTs into and out of the folded state $\mathcal{S}_7$.

transition	mean / μs		std / μs
$\mathcal{S}_7 \rightarrow \mathcal{S}_{(0,1,2,3,4,5,6,8)}$	2.9	$\pm$	0.2
$\mathcal{S}_{(0,1,2,3,4,5,6,8)} \rightarrow \mathcal{S}_7$	16.3	$\pm$	0.9

A visualization of the metastable conformational ensembles and the associated transitions of the nine macrostate SRV-MSM is presented in Figure 8. The stationary probabilities $\pi_{\mathcal{S}_i}$ and associated free energies $G_{\mathcal{S}_i}$ of each state $\mathcal{S}_{0-8}$ are listed in Table 2. The native fold resides in $\mathcal{S}_7$ and occupies $\sim$17% of the stationary probability distribution at the $T = 290$ K state point at which the molecular dynamics simulation was conducted.

Table 2: Stationary probabilities $\pi_{\mathcal{S}_i}$ and associated free energies $G_{\mathcal{S}_i}$ of each state $\mathcal{S}_{0-8}$ within the nine macrostate SRV-MSM.

macrostate $\mathcal{S}_i$	$\pi_{\mathcal{S}_i}$	$G_{\mathcal{S}_i}/\mathrm{k_B}T$
$\mathcal{S}_0$	0.004837	5.332
$\mathcal{S}_1$	0.008090	4.817
$\mathcal{S}_2$	0.006681	5.009
$\mathcal{S}_3$	0.016846	4.084
$\mathcal{S}_4$	0.012673	4.368
$\mathcal{S}_5$	0.020058	3.909
$\mathcal{S}_6$	0.075622	2.582
$\mathcal{S}_7$	0.168266	1.782
$\mathcal{S}_8$	0.686928	0.376

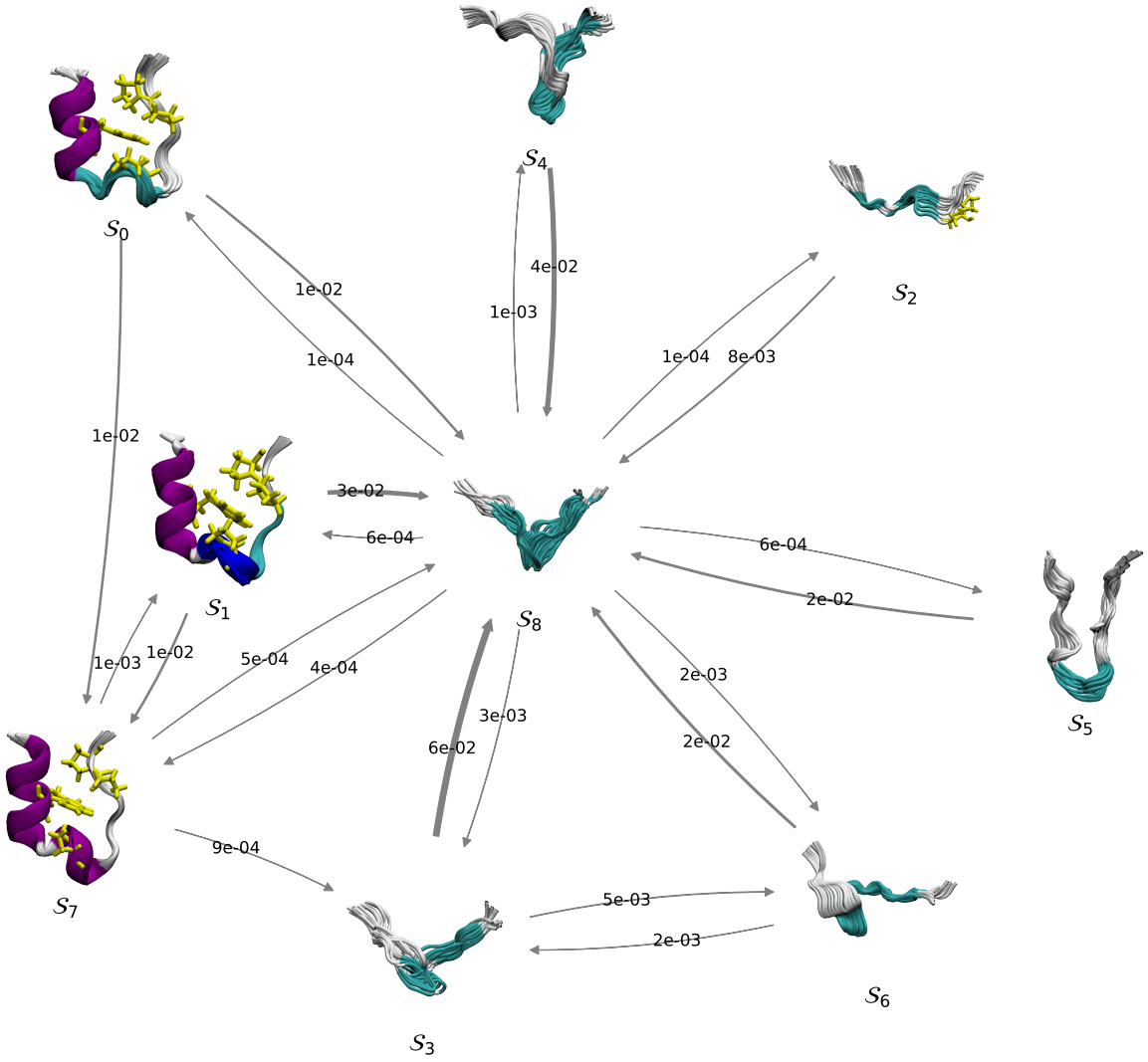


Figure 8: Visualization of the SRV-MSM for Trp-cage. The nine metastable macrostates $\mathcal{S}_{0-8}$ defined by PCCA++ spectral clustering are represented by visualizations of an ensemble of 20 mutually aligned representative molecular states. The equilibrium transitions between states are represented by arrows and annotated by the transition probability corresponding to the associated off-diagonal macrostate transition matrix element. Arrow widths are drawn proportional to probabilities, and arrows corresponding to fluxes smaller than 10^{-5} are not visualized for clarity. The unfolded ensemble comprises $\mathcal{S}_{2,3,4,5,6,8}$ and is characterized by a central molten globule $\mathcal{S}_8$ that rapidly interconverts with the other metastable unfolded states. Folding into the native state $\mathcal{S}_7$ proceeds either directly from $\mathcal{S}_8$, through an intermediate $\mathcal{S}_0$ in which the 3_{10} loop is misfolded but stabilized by the hydrophobic core, or second intermediate $\mathcal{S}_1$ possessing an unfolded 3_{10} loop.

The unfolded ensemble comprises $\mathcal{S}_{2,3,4,5,6,8}$ and accounts for ∼82% of the stationary probability distribution. This ensemble is dominated by a central molten globule $\mathcal{S}_8$ that itself accounts for ∼69% of the stationary distribution and acts as a kinetic hub for interconversions with the other unfolded metastable conformations. Of the remaining unfolded states, state $\mathcal{S}_2$ is particularly interesting and structurally interpretable. Transitions from $\mathcal{S}_8$ to $\mathcal{S}_2$ correspond to transitions along TIC1 in the free energy surface visualization in Figure 7a,f). Structurally, this transition can be identified as the rearrangement of the polyproline II structure (residues 17-20) from an unstructured to an alpha helix-like conformation. In particular, transitions into $\mathcal{S}_2$ are defined by conversions of the Pro18 residue dihedrals from a $P_{||}$ (ϕ = -75°, ψ = 160°) to an α (ϕ = -75°, ψ = -50°) configuration. The native fold in $\mathcal{S}_7$ is stabilized by hydrophobic interactions of the Trp6 side chain with Pro12, Pro18, and Pro19,[53] and transitioning into $\mathcal{S}_2$ prohibits folding because the Pro18 rotates externally, facing away from the hydrophobic core and precluding stacking against the Trp6 side chain. Indeed, Figure 7a,f shows the absence of any pathway along TIC1 from $\mathcal{S}_2$ to $\mathcal{S}_7$ and Figure 8 shows the absence of any significant flux between these states. Instead, in order to fold the conformations in $\mathcal{S}_2$ must first transition into $\mathcal{S}_8$, which effectively "unlocks" the molecule by enabling Trp6-Pro12 hydrophobic stacking.

The remaining unfolded macrostates, $\mathcal{S}_3$, $\mathcal{S}_4$, $\mathcal{S}_5$, and $\mathcal{S}_6$, together account for ∼13% of the stationary probability distribution, and show negligible flux between one another or to the folded state $\mathcal{S}_7$ or intermediates $\mathcal{S}_0$ or $\mathcal{S}_1$. Accordingly, folding is mediated through the compact molten globule $\mathcal{S}_8$ as evinced by the fact that the slowest timescale is associated with transitions from the unfolded to folded ensembles. In other words, mixing of the different unfolded states occurs at faster timescales than folding transitions, the flux of which is gated almost exclusively through $\mathcal{S}_8$. Deng et al.[36] indicate in their analysis of this simulation data that they find no evidence of kinetic partitioning of the unfolded state space, which is consistent with a hub-like scenario. Our observation of folding mediate by the molten globule state is also similar to an observation made by Marinelli et al.[28] in a study of the Trp-cage

TC5b mutant, although they note a lower occupancy probability of the molten globule state. An analysis of the same D.E. Shaw trajectory as studied herein by Dickson and Brooks[34] is also consistent with our model. They calculate a "hub score" for the native state, defining the degree to which it mediates non-native-to-non-native transitions to determine that a substantial number of these transitions are not mediated by the native state.

Our model predicts folding to the native state to proceed either directly from the molten globule kinetic hub $\mathcal{S}_8 \rightarrow \mathcal{S}_7$, or via one of two intermediates: $\mathcal{S}_8 \rightarrow \mathcal{S}_0 \rightarrow \mathcal{S}_7$, or $\mathcal{S}_8 \rightarrow \mathcal{S}_1 \rightarrow \mathcal{S}_7$. The intermediates $\mathcal{S}_0$ and $\mathcal{S}_1$ respectively occupy $\sim$0.48% and $\sim$0.81% of the stationary probability distribution, and bear a great deal of resemblance to both one another and to the native folded state. They are differentiated almost exclusively by the degree of folding of of the 3_{10}-helix (residues 11-14). Figure 9 shows the distributions of the root mean squared deviation (RMSD) from the native fold of these four 3_{10}-helix residues for the simulation snapshots populating states $\mathcal{S}_0$, $\mathcal{S}_1$ and $\mathcal{S}_7$. The distributions for $\mathcal{S}_0$ and $\mathcal{S}_7$ are narrow and normal, indicating locally stable conformations. $\mathcal{S}_1$, on the other hand, displays a much broader non-normal distribution. This may be characteristic of multiple states grouped together which cannot be separated at the temporal resolution of our model, or alternatively of a greater degree of flexibility in the motion of the 3_{10}-helix region due to it being unfolded in this conformation. Focusing on the dihedral angles within the 3_{10}-helix, Figure 10 displays the Ramachandran plots for residues 12 (panels a,c,e) and 14 (panels b,d,f). The unlooping of intermediate $\mathcal{S}_1$ relative to the native fold $\mathcal{S}_7$ can be seen most obviously in Pro12, with this residue transitioning from a native α (ϕ = -75°, ψ = -30°) configuration to a α'' (ϕ = 75°, ψ = 145°) configuration. The distinction between intermediate $\mathcal{S}_0$ and native state $\mathcal{S}_7$ is characterized by largely Ser14 β character (ϕ = -80°, ψ = 155°) in the former, compared to predominantly α_L character (ϕ = 90°, ψ = -10°) in the latter. The Ser14 residue in $\mathcal{S}_1$ shows occupancy β, α_L, and α_R (ϕ = -80°, ψ = -20°) configurations owing to the greater flexibility of the 3_{10} loop in this state.

The state $\mathcal{S}_1$ is a well-known structural metastable state – sometimes referred to as

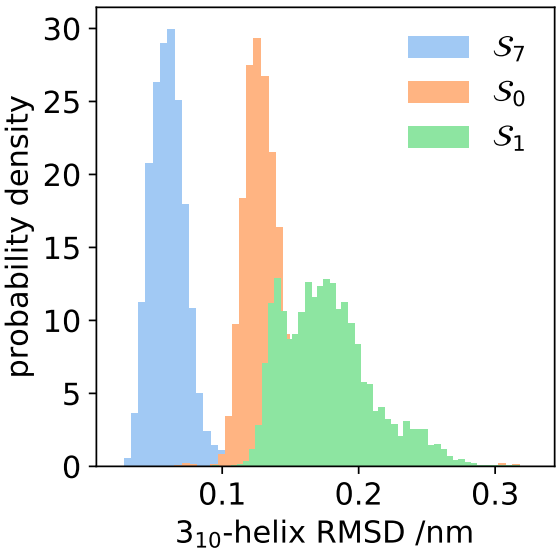


Figure 9: Distribution of the root mean squared deviation (RMSD) relative to the native fold of residues 11-14 comprising the 3_{10}-helix for states $\mathcal{S}_0$, $\mathcal{S}_1$ and $\mathcal{S}_5$. The folded state $\mathcal{S}_7$ and intermediate state $\mathcal{S}_0$ are both normally distributed with means of 0.065 nm and 0.135 nm, respectively. The intermediate $\mathcal{S}_1$ possesses a much broader non-normal distribution with mean 0.176 nm.

the "loop" structure – in close proximity to the native fold but possessing an unfolded 3_{10}-helix.[30–32,54] This state $\mathcal{S}_1$ can be identified as a local minimum in Figure 7b,g existing as a finger protruding below the direct path linking $\mathcal{S}_8$ and $\mathcal{S}_7$ along TIC1. Long-range interactions between the Trp6 core and Pro12 on the 3_{10} loop stabilize the $\mathcal{S}_1$ intermediate, and there is significant flux both into the native state $\mathcal{S}_7$ or back to the molten globule $\mathcal{S}_8$. The state we identify as $\mathcal{S}_0$ does not receive much mention in the literature, likely due to its relative instability, possessing about half the stationary probability distribution compared to $\mathcal{S}_1$ (cf. Table 2) and about one sixth of the flux from the molten globule $\mathcal{S}_8$ (cf. Figure 8). In sum, folding proceeds from the molten globule hub $\mathcal{S}_8$ into $\mathcal{S}_{7,0,1}$ through the formation of the hydrophobic core in which the Trp6 sidechain is "caged" by the Tyr3, Leu7, Gly11, Pro12, Pro18, and Pro19 sidechains, and the N-terminal α-helix (residues 2-8). These structural formation events are either accompanied by complete folding of the 3_{10}-helix (residues 11-14) through a direct transition into the native state $\mathcal{S}_7$, or partial folding of the 3_{10}-helix

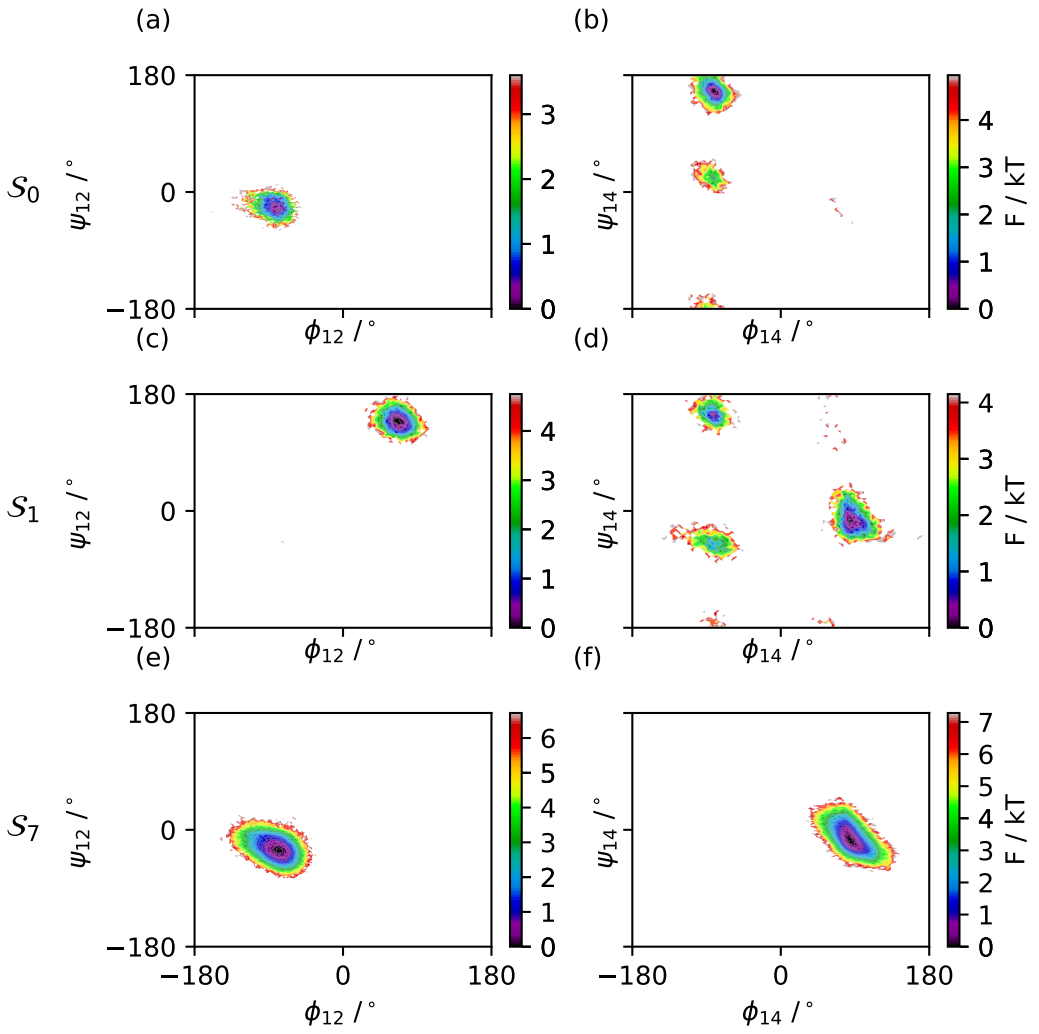


Figure 10: Ramachandran plots of backbone dihedrals of residues Pro12 and Ser14 for states $\mathcal{S}_0$ (a-b), $\mathcal{S}_1$ (c-d), and $\mathcal{S}_7$ (e-f). The recognized "loop structure" $\mathcal{S}_1$ is distinguished by an unfolded 3_{10} loop as reflected in ϕ_{12}-ψ_{12} (c). The folding of the 3_{10} loop in $\mathcal{S}_0$ is disrupted by β or $P_{||}$ character present in ϕ_{14}-ψ_{14} (b) as opposed to the native α_L conformation (f). State $\mathcal{S}_1$ occupies both the β or $P_{||}$ and α_L conformations in addition to α_R owing to the greater degree of flexibility of the 3_{10} loop in this state.

that leads to one or other of the metastable intermediates $\mathcal{S}_0$ or $\mathcal{S}_1$ that require subsequent structural rearrangements of the 3_{10} region to reach the native fold.

In summary, we have demonstrated the use of SRVs to establish a high time resolution SRV-MSM for the K8A mutant of Trp-cage TC10b at 290 K.[25] We carefully selected the model hyperparameters and verified its dynamic validity through cross-validation, spectral

analysis, implied timescale convergence, and the Chapman-Kolmogorov test to present free energy surfaces and a macrostate transition model that sheds new understanding on its folding. In particular, we identify an unfolded ensemble dominated by a hub-like molten globule that mediates transitions to the folded state. Folding proceeds either directly through the simultaneous formation of the hydrophobic core, N-terminal α-helix, and 3_{10}-helix, or indirectly through one of two metastable intermediates that possess misfolded 3_{10}-helices. We note that our results differ from, although not necessarily inconsistent with, the folding model extracted from this data by Deng et al.[36] employing an RMSD-based MSM. Based on that analysis, Trp-cage folding was reported to proceed by two representative parallel paths as proposed by Juraszek and Bolhuis[31] corresponding to two archetypal mechanisms of protein folding:[32,55–57] (i) a nucleation-condensation mechanism wherein formation of a compact molten globule precedes folding of the N-terminal α-helix, 3_{10}-helix, and native packing of the hydrophobic core, and (ii) a diffusion-collision mechanism wherein pre-formation of the α-helix precedes formation of the hydrophobic core and 3_{10}-helix. The high resolution SRV-MSM established in this work establishes the dominance of a molten globule kinetic hub state that mediates folding. However, this statistical portrait is limited to the resolution of the 10 ns lagtime, and the fastest implied timescale resulting from PCCA++ macrostate clustering is $\sim$100 ns (Figure 5a).

Notwithstanding, our results present three important adjustments to the picture of the conformational and kinetic landscape. First, the presentation of two independent folding pathways, starting either from a molten globule or an extended conformation with a pre-formed helix can be limiting. The molten globule conformation serves as the gateway for folding and, within the statistical resolution supported by the data and our model, acts as a source for both folding pathways. Second, the unfolded ensemble possesses structural and kinetic richness centered upon this molten globule kinetic hub. Third, folding proceeds either directly to the native state, or through two non-native folded intermediates, which differ in the nativeness of the 3_{10}-helix. The high kinetic resolution MSM enabled by the

replacement of TICA by SRVs reveals a new intermediate $\mathcal{S}_0$ as an important metastable intermediate for folding. We note that it is not possible to resolve further structural details of the folding process by introducing additional macrostates into the SRV-MSM since analysis of the microstate transition matrix eigenvalue spectrum shows the simulation data to support no more than nine statistically robust macrostates. Resolution of finer-scale folding mechanisms and pathways from $\mathcal{S}_8$ to $\mathcal{S}_{7,0,1}$ would require a more detailed analysis of the microstate transition matrix, as previously studied by Deng et al.,[36] and/or path sampling calculations, which are beyond the scope of this work.

4 Conclusions

We have presented SRVs as viable and promising modular replacement for TICA in the construction of Markov state models for protein folding. In an application to an ultra-long 208 μs explicit solvent simulation of the K8A mutant of Trp-cage TC10b conducted by D.E. Shaw Research.,[25] we showed SRV coordinates to outperform TICA-MSMs under cross validation by displaying higher test VAMP-2 scores with lower variance, and also to be more robust to input feature choices than TICA due to their capacity to learn nonlinear transformations of the input features. Employing SRVs as a basis set MSM construction produced a superior convergence rate of implied timescales with respect to lagtime, enabling the construction of extremely high resolution kinetic models. The resulting SRV-MSM revealed new understanding and insight into the kinetics and mechanisms of Trp-cage folding. A compact molten globular state acts as a kinetic hub for the unfolded ensemble and serves as the gateway for transitions into the folded state. The dominant folding pathway proceeds by formation of the hydrophobic core and N-terminal α-helix either directly into native state or via one of two intermediates that possess imperfectly folded 3_{10}-helices. The high time resolution MSMs enabled by SRVs represent a valuable new addition to the MSM construction pipeline that can help squeeze the most out of the simulation data used to parameterize the models

and produce high-temporal resolution kinetic models to understand and predict biomolecular folding.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. CHE-1841805. H.S. acknowledges support from the Molecular Software Sciences Institute (MolSSI) Software Fellows program (NSF grant ACI-1547580).[58,59] We are grateful to D.E. Shaw Research for sharing the Trp-cage simulation trajectories.

References

(1) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**, *140*, 2386–2396.

(2) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything You Wanted to Know About Markov State Models But Were Afraid to Ask. *Methods* **2010**, *52*, 99–105.

(3) Prinz, J.-H.; Keller, B.; Noé, F. Probing Molecular Kinetics with Markov Models: Metastable States, Transition Pathways and Spectroscopic Observables. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16912.

(4) 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. *Nat. Chem.* **2017**, *9*, 1005–1011.

(5) 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. *J. Chem. Phys.* **2011**, *134*, 174105.

(6) Noé, F.; Nuske, F. A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems. *Multiscale Model. Sim.* **2013**, *11*, 635–655.

(7) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9*, 2000–2009.

(8) Noé, F.; Clementi, C. Kinetic Distance and Kinetic Maps from Molecular Dynamics Simulation. *J. Chem. Theory Comput.* **2015**, *11*, 5002–5011.

(9) Husic, B. E.; Pande, V. S. Ward Clustering Improves Cross-Validated Markov State Models of Protein Folding. *J. Chem. Theory Comput.* **2017**, *13*, 963–967.

(10) McGibbon, R. T.; Pande, V. S. Variational Cross-Validation of Slow Dynamical Modes in Molecular Kinetics. *J. Chem. Phys.* **2015**, *142*, 124105.

(11) Husic, B. E.; McGibbon, R. T.; Sultan, M. M.; Pande, V. S. Optimized Parameter Selection Reveals Trends in Markov State Models for protein Folding. *J. Chem. Phys.* **2016**, *145*, 194103.

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

(13) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for Deep Learning of Molecular Kinetics. *Nat. Commun.* **2018**, *9*, 5.

(14) Chen, W.; Sidky, H.; Ferguson, A. L. Nonlinear Discovery of Slow Molecular Modes Using State-Free Reversible VAMPnets. *J. Chem. Phys.* **2019**, *150*, 214114.

(15) Andrew, G.; Arora, R.; Bilmes, J.; Livescu, K. In *Proceedings of Machine Learning Research*; Dasgupta, S., McAllester, D., Eds.; PMLR: Atlanta, Georgia, USA, 2013; Vol. 28; pp 1247–1255.

(16) Nüske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S. J. S.; Noé, F. Variational Approach to Molecular Kinetics. *J. Chem. Theory Comput.* **2014**, *10*, 1739–1752.

(17) Hernández, C. X.; Wayment-Steele, H. K.; Sultan, M. M.; Husic, B. E.; Pande, V. S. Variational Encoding of Complex Dynamics. *Phys. Rev. E* **2018**, *97*, 062412.

(18) Wehmeyer, C.; Noé, F. Time-Lagged Autoencoders: Deep Learning of Slow Collective Variables for Molecular Kinetics. *J. Chem. Phys.* **2018**, *148*, 241703.

(19) Schwantes, C. R.; Pande, V. S. Modeling Molecular Kinetics with tICA and the Kernel Trick. *J. Chem. Theory Comput.* **2015**, *11*, 600–608.

(20) 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. *J. Chem. Theory Comput.* **2015**, *11*, 5525–5542.

(21) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7*, 3412–3419.

(22) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105.

(23) Wu, H.; Paul, F.; Wehmeyer, C.; Noé, F. Multiensemble Markov Models of Molecular Thermodynamics and Kinetics. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, E3221–E3230.

(24) Prinz, J.-H.; Chodera, J. D.; Pande, V. S.; Swope, W. C.; Smith, J. C.; Noé, F. Optimal Use of Data in Parallel Tempering Simulations for the Construction of Discrete-State Markov Models of Biomolecular Dynamics. *J. Chem. Phys.* **2011**, *134*, 244108.

(25) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334*, 517–520.

(26) Barua, B.; Lin, J. C.; Williams, V. D.; Kummler, P.; Neidigh, J. W.; Andersen, N. H. The Trp-Cage: Optimizing the Stability of a Globular Miniprotein. *Protein Engineering Design and Selection* **2008**, *21*, 171–185.

(27) Meuzelaar, H.; Marino, K. A.; Huerta-Viga, A.; Panman, M. R.; Smeenk, L. E. J.; Kettelarij, A. J.; van Maarseveen, J. H.; Timmerman, P.; Bolhuis, P. G.; Woutersen, S. Folding Dynamics of the Trp-Cage Miniprotein: Evidence for a Native-Like Intermediate from Combined Time-Resolved Vibrational Spectroscopy and Molecular Dynamics Simulations. *J. Phys. Chem. B* **2013**, *117*, 11490–11501.

(28) Marinelli, F.; Pietrucci, F.; Laio, A.; Piana, S. A Kinetic Model of Trp-Cage Folding from Multiple Biased Molecular Dynamics Simulations. *PLOS Comput. Biol.* **2009**, *5*, e1000452.

(29) Zhou, R.; Berne, B. J.; Germain, R. The Free Energy Landscape for Hairpin Folding in Explicit Water. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931–14936.

(30) Zhou, R. Trp-cage: Folding Free Energy Landscape in Explicit Water. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13280–13285.

(31) Juraszek, J.; Bolhuis, P. G. Sampling the Multiple Folding Mechanisms of Trp-Cage in Explicit Solvent. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 15859–15864.

(32) Kim, S. B.; Dsilva, C. J.; Kevrekidis, I. G.; Debenedetti, P. G. Systematic Characterization of Protein Folding Pathways Using Diffusion Maps: Application to Trp-Cage Miniprotein. *J. Chem. Phys.* **2015**, *142*, 085101.

(33) English, C. A.; García, A. E. Charged Termini on the Trp-Cage Roughen the Folding Energy Landscape. *J. Phys. Chem. B* **2015**, *119*, 7874–7881.

(34) Dickson, A.; Brooks, C. L. Native States of Fast-Folding Proteins are Kinetic Traps. *J. Am. Chem. Soc.* **2013**, *135*, 4729–4734.

(35) Suárez, E.; Adelman, J. L.; Zuckerman, D. M. Accurate Estimation of Protein Folding and Unfolding Times: Beyond Markov State Models. *J. Chem. Theory Comput.* **2016**, *12*, 3473–3481.

(36) Deng, N.-J.; Dai, W.; Levy, R. M. How Kinetics within the Unfolded State Affects Protein Folding: An Analysis Based on Markov State Models and an Ultra-Long MD Trajectory. *J. Phys. Chem. B* **2013**, *117*, 12787–12799.

(37) Levy, R. M.; Dai, W.; Deng, N.-J.; Makarov, D. E. How Long Does it Take to Equilibrate the Unfolded State of a Protein? *Protein Science* **2013**, *22*, 1459–1465.

(38) Scherer, M. K.; Husic, B. E.; Hoffmann, M.; Paul, F.; Wu, H.; Noé, F. Variational Selection of Features for Molecular Kinetics. *J. Chem. Phys.* **2019**, *150*, 194108.

(39) Wu, H.; Noé, F. Variational Approach for Learning Markov Processes from Time Series Data. *arXiv:1707.04659* **2017**, 1–30.

(40) Wehmeyer, C.; Husic, B. E.; Hempel, T.; Scherer, M. K.; Noé, F.; Olsson, S. Introduction to Markov State Modeling with the PyEMMA Software [Article v1.0]. *Living Journal of Computational Molecular Science* **2019**, *1*, 5965.

(41) Hassoun, M. H. *Fundamentals of Artificial Neural Networks*; MIT Press: Cambridge, USA, 1995.

(42) Chen, T.; Chen, H. Universal Approximation to Nonlinear Operators by Neural Networks with Arbitrary Activation Functions and its Application to Dynamical Systems. *IEEE Trans. Neural Netw.* **1995**, *6*, 911–917.

(43) Röblitz, S.; Weber, M. Fuzzy Spectral Clustering by PCCA+: Application to Markov State Models and Data Classification. *Advances in Data Analysis and Classification* **2013**, *7*, 147–179.

(44) Deuflhard, P.; Weber, M. Robust Perron Cluster Analysis in Conformation Dynamics. *Linear Algebra and its Applications* **2005**, *398*, 161–184.

(45) Kube, S.; Weber, M. A Coarse Graining Method for the Identification of Transition Rates Between Molecular Conformations. *J. Chem. Phys.* **2007**, *126*, 024103.

(46) Bowman, G. R.; Pande, V. S. Protein Folded States are Kinetic Hubs. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 10890–10895.

(47) Best, R. B.; Hummer, G.; Eaton, W. A. Native Contacts Determine Protein Folding Mechanisms in Atomistic Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 17874–17879.

(48) Meshkin, H.; Zhu, F. Thermodynamics of Protein Folding Studied by Umbrella Sampling along a Reaction Coordinate of Native Contacts. *J. Chem. Theory Comput.* **2017**, *13*, 2086–2097.

(49) Noé, F. Probability Distributions of Molecular Observables Computed from Markov Models. *J. Chem. Phys.* **2008**, *128*, 244103.

(50) Trendelkamp-Schroer, B.; Wu, H.; Paul, F.; Noé, F. Estimation and Uncertainty of Reversible Markov Models. *J. Chem. Phys.* **2015**, *143*, 174101.

(51) Qiu, L.; Hagen, S. J. A Limiting Speed for Protein Folding at Low Solvent Viscosity. *Journal of the American Chemical Society* **2004**, *126*, 3398–3399.

(52) Qiu, L.; Pabit, S. A.; Roitberg, A. E.; Hagen, S. J. Smaller and Faster: The 20-Residue Trp-Cage Protein Folds in 4 μs. *Journal of the American Chemical Society* **2002**, *124*, 12952–12953.

(53) Hałabis, A.; Żmudzińska, W.; Liwo, A.; Ołdziej, S. Conformational Dynamics of the Trp-Cage Miniprotein at Its Folding Temperature. *J. Phys. Chem. B* **2012**, *116*, 6898–6907.

(54) Wang, J.; Ferguson, A. L. Recovery of Protein Folding Funnels from Single-Molecule Time Series by Delay Embeddings and Manifold Learning. *J. Phys. Chem. B* **2018**, *122*, 11931–11952.

(55) Karplus, M.; Weaver, D. L. Protein-Folding Dynamics. *Nature* **1976**, *260*, 404–406.

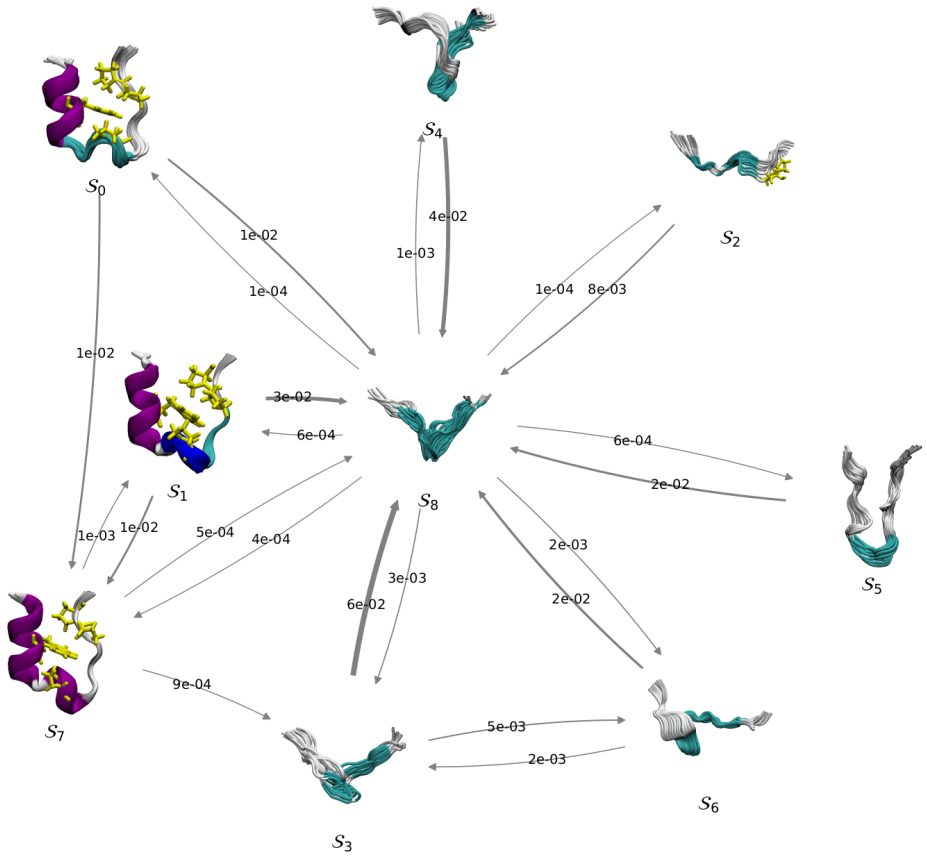
(56) Abkevich, V. I.; Gutin, A. M.; Shakhnovich, E. I. Specific Nucleus as the Transition State for Protein Folding: Evidence From the Lattice Model. *Biochemistry* **1994**, *33*, 10026–10036.

(57) Gianni, S.; Guydosh, N. R.; Khan, F.; Caldas, T. D.; Mayor, U.; White, G. W.; DeMarco, M. L.; Daggett, V.; Fersht, A. R. Unifying Features in Protein-Folding Mechanisms. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13286–13291.

(58) Krylov, A.; Windus, T. L.; Barnes, T.; Marin-Rimoldi, E.; Nash, J. A.; Pritchard, B.; Smith, D. G.; Altarawy, D.; Saxe, P.; Clementi, C. et al. Perspective: Computational Chemistry Software and its Advancement as Illustrated Through Three Grand Challenge Cases for Molecular Science. *J. Chem. Phys.* **2018**, *149*, 180901.

(59) Wilkins-Diehr, N.; Crawford, T. D. NSF's Inaugural Software Institutes: The Science Gateways Community Institute and the Molecular Sciences Software Institute. *Comput. Sci. Eng.* **2018**, *20*, 26–38.

TOC Graphic


S_0
S_1
S_2
S_3
S_4
S_5
S_6
S_7
S_8
1e-02
1e-04
4e-02
1e-03
1e-04
8e-03
1e-02
3e-02
6e-04
6e-04
2e-02
1e-03
1e-02
5e-04
4e-04
2e-03
3e-03
6e-02
2e-02
9e-04
5e-03
2e-03