What Markov state models can and cannot do: Correlation versus path-based observables in protein folding models

Ernesto Suárez,[†,‖] Rafal P. Wiewiora,[‡,‖] Chris Wehmeyer,[¶] Frank Noé,[¶] John D. Chodera,[*,‡] and Daniel M. Zuckerman[*,§]

Advanced Biomedical Computational Science, Frederick National Laboratory for Cancer Research, Frederick, MD 21702, Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, Freie Universität Berlin, Germany, and Department of Biomedical Engineering, Oregon Health and Science University, Portland, OR 97239

E-mail: john.chodera@choderalab.org; zuckermd@ohsu.edu

Supporting Information

MSM scoring MSMs at a lag time of 100 ns were constructed using discrete microstate trajectories from the training set and scored on the test set trajectories. For comparison, we also performed scoring at a short MSM lag time of 10 ns, hypothesizing this could better

[*]To whom correspondence should be addressed

[†]Advanced Biomedical Computational Science, Frederick National Laboratory for Cancer Research, Frederick, MD 21702

[‡]Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065

[¶]Freie Universität Berlin, Germany

[§]Department of Biomedical Engineering, Oregon Health and Science University, Portland, OR 97239

[‖]These authors contributed equally to this work

optimize the reproduction of kinetics at short lag times for comparison with haMSMs. We used the 100 ns lag time top models for all analysis, and compared the results to the 10 ns lag time models in SI figures.

We used a 50:50 shuffle-split cross-validation scheme to find the optimal set of hyperparameters while avoiding overfitting. In this scheme, 2 μs long fragments of the trajectories (i.e., the original fragments in which the datasets are provided by DESRES) are randomly split into training and test sets of approximately equal sizes. To obtain standard deviations indicative of out-of-sample model performance, this shuffle-split model evaluation procedure was repeated 10 times with different random divisions of the dataset into training and test sets. Scoring was based on the sum of squared-eigenvalues of the transition matrix (VAMP-2 score[1]), as this particular score is physically interpretable as 'kinetic content'.

To choose the best number of top eigenvalues to score the models with, we initially performed scoring separately at each number of top eigenvalues between 2 and 50. We then chose the number of eigenvalues for which the score of the top scoring model was closest to 50% of the number of eigenvalues (i.e., the highest possible score), in order to maximize the signal from the true dynamical processes and increase the resolution of the scores, while minimizing the noise from spurious eigenvalues (Figure S1, Figure S2 shows the analogical results for the 10 ns scoring lag time). Hence 3 (4 at 10 ns) top eigenvalues were used for chignolin, 5 (12 at 10 ns) eigenvalues for villin, 6 (17 at 10 ns) eigenvalues for Trp-cage, and 5 (38 at 10 ns) eigenvalues for NTL9. To evaluate a large set of hyperparameters, reduced datasets subsampled to 10 ns/frame (for NTL9: 10 ns/frame at 10 ns scoring lag time, increased to 50 ns/frame at 100 ns scoring lag time) were used for computational feasibility, except for chignolin, which remained at 0.2 ns/frame intervals due to its small size. The datasets were featurized with all minimal residue–residue distances (calculated as the closest distance between the heavy atoms of two residues separated in sequence by at least two neighboring residues). For consistency in interpretation and computational feasibility, this featurization choice was made without variational scoring. The datasets were

projected into a kinetically relevant space using tICA,[2,3] at lag time 10 ns (50 ns for NTL9 at the 100 ns MSM scoring lag due to higher subsampling; for chignolin lag times 1 ns and 5 ns were also possible due to no subsampling), with either kinetic[4] or commute[5] mapping, retaining the following numbers of tICs (maximum number of which depends on the number of features and hence size of the protein): 2 or the number of tICs corresponding to 95% of total kinetic variance/content ("95%") for chignolin; 2, 10, 50, 100, 150, or 95% for Trp-cage; 2, 10, 50, 100, 300, 500, or 95% for villin and NTL9. Each of the tICA outputs was discretized using k-means clustering into 50, 100, 300, 500, 800, or 1000 microstate clusters (except for chignolin, where 200, 400, 600, or 900 microstate clusters were also tried for better scoring resolution, due to the smaller number of features and hence number of tICs retained). Table S1 summarizes all hyperparameter options assessed.

Figure S3 shows the results of the scoring at the 100 ns lag time, and Figure S4 at the 10 ns lag time. Model scores are reported below as means with standard deviations over 10 shuffle-splits. As there were no statistically significant differences between the scores for chignolin at the 100 ns lag time, we used the top model scored at 10 ns for all analyses. We also note the top scoring models for villin were identical at both lag times.

The following top scoring models were selected using the 100 ns scoring lag time: chignolin: all models statistically the same; villin: commute tICA mapping, 10 ns tICA lag time, 10 tICs, 100 microstates, score 2.43 (5 eigenvalues, SD: 0.31); Trp-cage: kinetic tICA mapping, 10 ns tICA lag time, 50 tICs, 100 microstates, score 3.13 (SD: 0.24, 6 eigenvalues); NTL9: commute tICA mapping, 50 ns tICA lag time, 50 tICs, 200 microstates, score 2.34 (SD: 0.26, 5 eigenvalues).

The following top scoring models were selected using the 10 ns scoring lag time: chignolin: kinetic tICA mapping, 1 ns tICA lag time, 15 tICs (95% kinetic variance), 100 microstates, score 2.05 (4 eigenvalues, SD: 0.05); villin: commute tICA mapping, 10 ns tICA lag time, 10 tICs, 100 microstates, score 5.96 (12 eigenvalues, SD: 0.58); Trp-cage: commute tICA mapping, 10 ns tICA lag time, 100 tICs, 50 microstates, score 8.51 (SD: 0.42, 17 eigenvalues);

NTL9: commute tICA mapping, 10 ns tICA lag time, 10 tICs, 200 microstates, score 19.11 (SD: 0.89, 38 eigenvalues).

Finally, to construct the discrete microstate trajectories used in this work, the modeling process was repeated with full datasets (with no additional striding, i.e. at 0.2 ns/frame intervals and with no train-test splitting) using the top scoring parameters for the repeated tICA and k-means calculations.

Coarse-graining into macrostates PCCA++ villin coarse-graining. Our villin MSM identified a microsecond timescale ($\sim$1.1 μs), which was not present in a previously published MSM of this dataset,[6] likely due to our use of the residue-residue distances featurization combined with tICA, compared to the minRMSD metric in.[6] By coarse-graining the MSM into two macrostates, we identified this longer timescale as corresponding to the transition between the "folded – unfolded" and "misfolded" macrostates. The "misfolded" macrostate shows formation of short-lived helicity between residues ASN60 and LEU63 (Figures S6, S7). The previously[6] identified folding timescale of $\sim$ 400 ns is the second slowest timescale in our MSM, and a 3 macrostate coarse-graining was necessary to obtain separation between the "folded" and "unfolded" macrostates. All folding kinetics are considered only between the "folded" and "unfolded" macrostates, with no regard to the "misfolded" macrostate.

Defining macrostates based on kinetic clustering. As an alternative coarse-graining procedure, we use a hierarchical (or progressive) clustering based on a cutoff t_{cut}. If the round-trip time (t_{ij}) between any two states is less that t_{cut} then we merge the states. The procedure is as follows:

1. Compute MFPT matrix M and add it to M^T to obtain the round-trip times $\{t_{ij}\}$
2. While $\min(\{t_{ij}\}) < t_{cut}$:
    - Merge the corresponding states

- Recompute $\{t_{ij}\}$ (step 1) for merged states

3. Increase t_{cut} until clustering results in only one macrostate. Plot the highest $\{t_{ij}\}$ vs. t_{cut}, identify the longest plateau, and take macrostates at t_{cut} in the middle of the plateau. The following t_{cut} values were identified: 436 ns for chignolin, 277.6 ns for Trp-cage, 449.2 ns for villin, and 817.6 ns for NTL9.

Issues with DESRES NTL9 dataset **Thermodynamics and kinetics of trajectory NTL9-2 are inconsistent with the other trajectories.** While the folded content of the other three NTL9 trajectories is ∼90%, in agreement with the MSM results, trajectory NTL9-2 has only ∼50% folded frames. This can be clearly seen in the RMSD plots of the 'Individual proteins' section of the SI of[7] (NTL9-2 is the 3rd trajectory). The kinetics are also much faster: using macrostates defined from an MSM computed including NTL9-2, this trajectory has 185 folded-unfolded transitions, compared to just 7 in trajectory NTL9-3 of similar length. This is especially pronounced at the beginning of the trajectory—removal of the first 80 microseconds (mean of the folding and unfolding MFPTs in our converged MSM) leaves only 18 transitions left, though still many more than just 6 in the same length of NTL9-3. We removed trajectory NTL9-2 from the dataset analyzed in this paper.

Topology of trajectory NTL9-1 is different from the other trajectories. The dataset is provided with one NTL9.pdb topology file, as well as .mae files for each trajectory, to use for loading the topology-less .dcd trajectory files. However, using the NTL9.pdb file with the NTL9-1 trajectory provides clearly nonsensical results, e.g. by visual inspection in PyMOL[8] or lack of MSM convergence if using features defined from this file. Inspection of the .mae files reveals the arrangement of the hydrogen atoms in the topologies are different between NTL9-1 (within each residue, except for the first residue, some of whose hydrogens are at the end of the file) and the other trajectories (at the end of file). Importantly, conversion of the .mae file for NTL9-1 to a .pdb file with PyMOL[8] or Maestro[9] FAILS to preserve the original order of hydrogens, while VMD[10] preserves it and was used by us for the conversion.

The results were verified by manual inspection in PyMOL.[8]

Table S1: **All of the model hyperparameters assessed combinatorially.**

	Featurization	**Number of tICs retained**	**tICA lag time**	**tICA mapping**	**Number of microstates**	**MSM lag time**
Chignolin	res.–res. min. distances	2, 95% kin. var./cont.	1 ns, 5 ns, 10 ns	kinetic, commute	50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000	10 ns
Villin	res.–res. min. distances	2, 10, 50, 100, 300, 500, 95% kin. var./cont.	10 ns	kinetic, commute	50, 100, 300, 500, 800, 1000	10 ns
Trp-cage	res.–res. min. distances	2, 10, 50, 100, 150, 95% kin. var./cont.	10 ns	kinetic, commute	50, 100, 300, 500, 800, 1000	10 ns
NTL9	res.–res. min. distances	2, 10, 50, 100, 300, 500, 95% kin. var./cont.	10 ns	kinetic, commute	50, 100, 300, 500, 800, 1000	10 ns

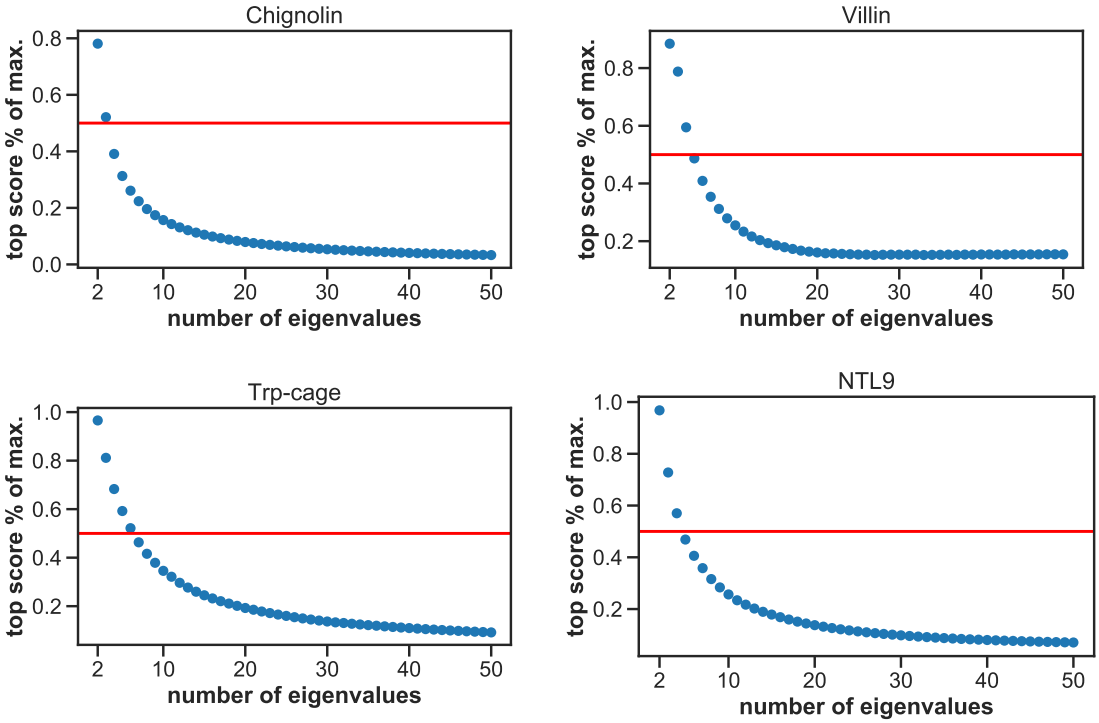

Figure S1: **Selection of the number of top eigenvalues for scoring, at a 100 ns lag time.** Markov state models (MSMs) were VAMP-2 scored separately at each number of top eigenvalues between 2 and 50 included in the scoring. The ratios of the top scoring models at each choice of numbers of eigenvalues and that number of eigenvalues (i.e. the highest possible score) are plotted. For selection of the final model, we chose the number of eigenvalues for which the ratio was closest to 0.5, marked by the red horizontal line.

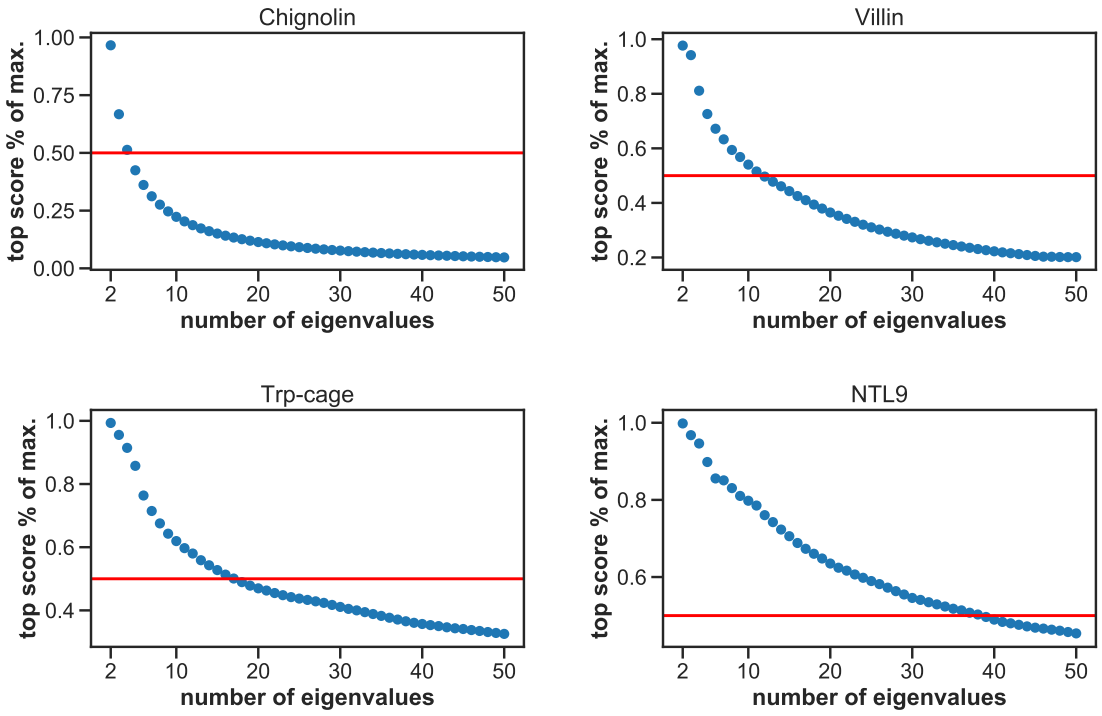

Figure S2: **Selection of the number of top eigenvalues for scoring, at a 10 ns lag time.** Markov state models (MSMs) were VAMP-2 scored separately at each number of top eigenvalues between 2 and 50 included in the scoring. The ratios of the top scoring models at each choice of numbers of eigenvalues and that number of eigenvalues (i.e. the highest possible score) are plotted. For selection of the final model, we chose the number of eigenvalues for which the ratio was closest to 0.5, marked by the red horizontal line.

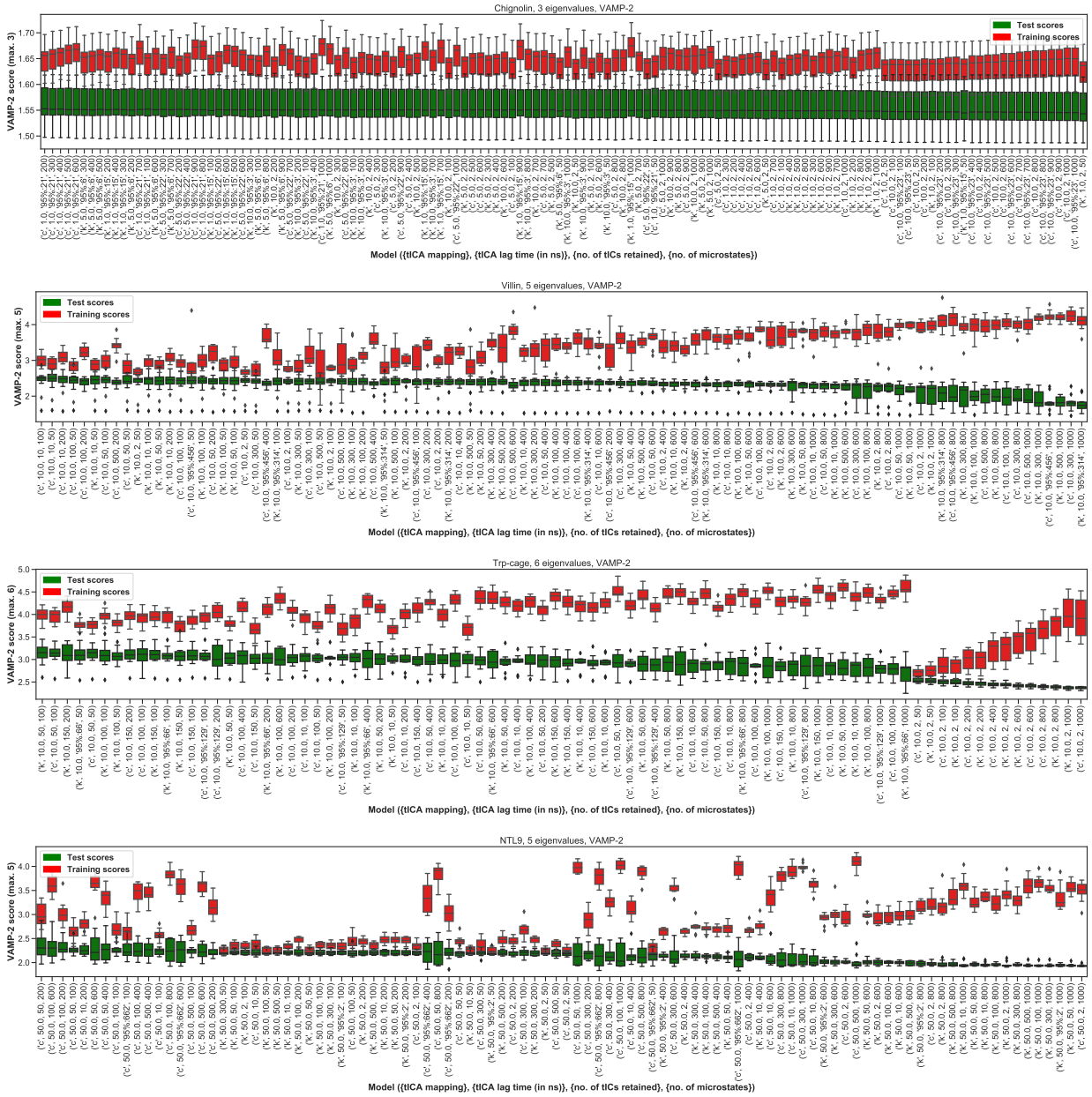

Figure S3: **VAMP-2 scoring results for optimal hyperparameter choice, at a 100 ns lag time.** The distributions of the VAMP-2 scores of ten shuffle-splits of the data for each individual set of hyperparameters (model) are shown as box-and-whisker plots. Bands of boxes show the first, second, and third quartiles, while whisker ends represent the lowest and highest scores still within 1.5 of the interquartile range from the first and third quartiles respectively. Scores lying outside of that range are shown as diamonds. The models are denoted as ([tICA mapping], [tICA lag time (in ns)], [number of tICs retained], [number of microstates]). Test scores are shown in green and training scores in red.

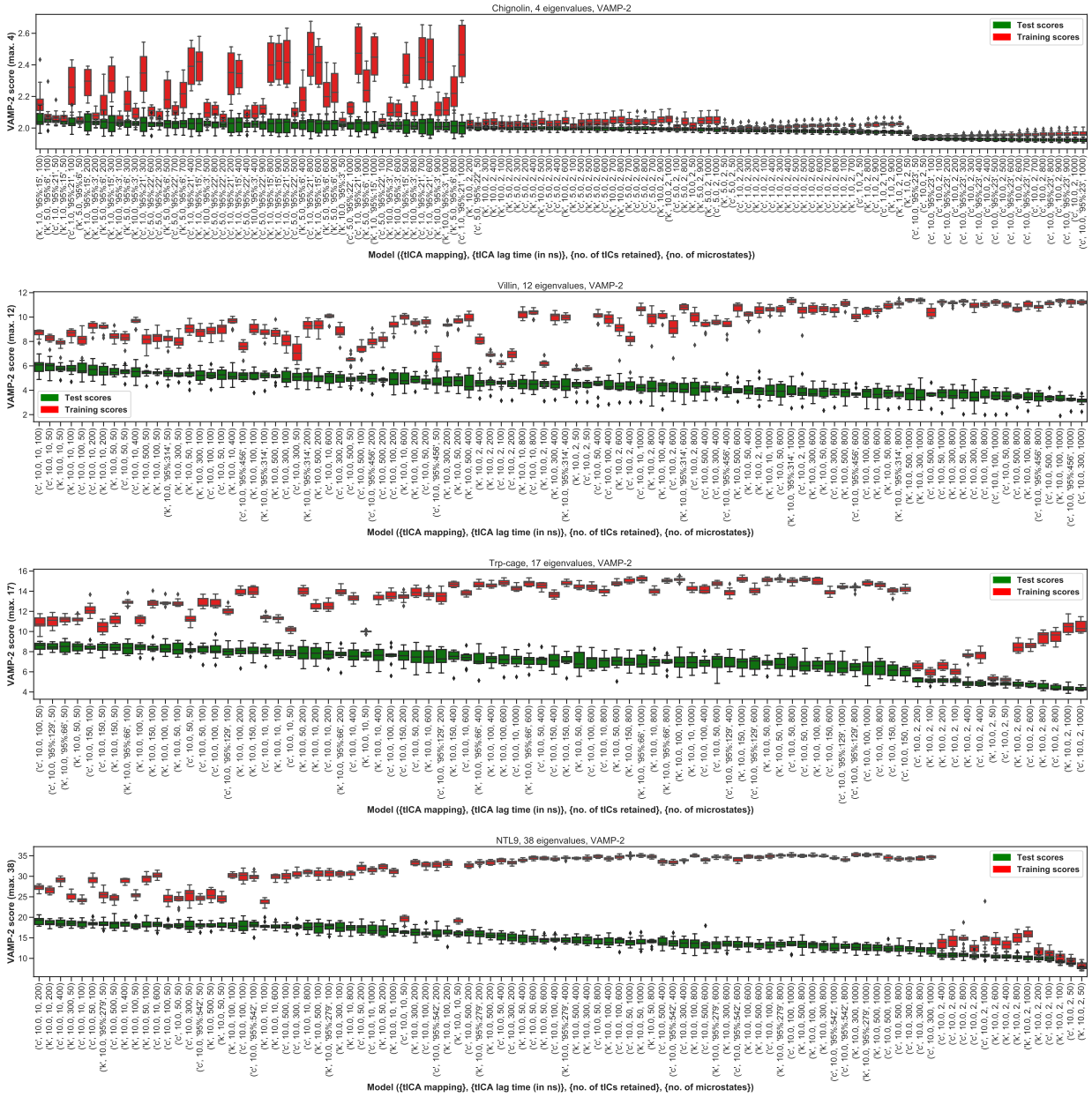

Figure S4: **VAMP-2 scoring results for optimal hyperparameter choice, at a 10 ns lag time.** The distributions of the VAMP-2 scores of ten shuffle-splits of the data for each individual set of hyperparameters (model) are shown as box-and-whisker plots. Bands of boxes show the first, second, and third quartiles, while whisker ends represent the lowest and highest scores still within 1.5 of the interquartile range from the first and third quartiles respectively. Scores lying outside of that range are shown as diamonds. The models are denoted as ([tICA mapping], [tICA lag time (in ns)], [number of tICs retained], [number of microstates]). Test scores are shown in green and training scores in red.

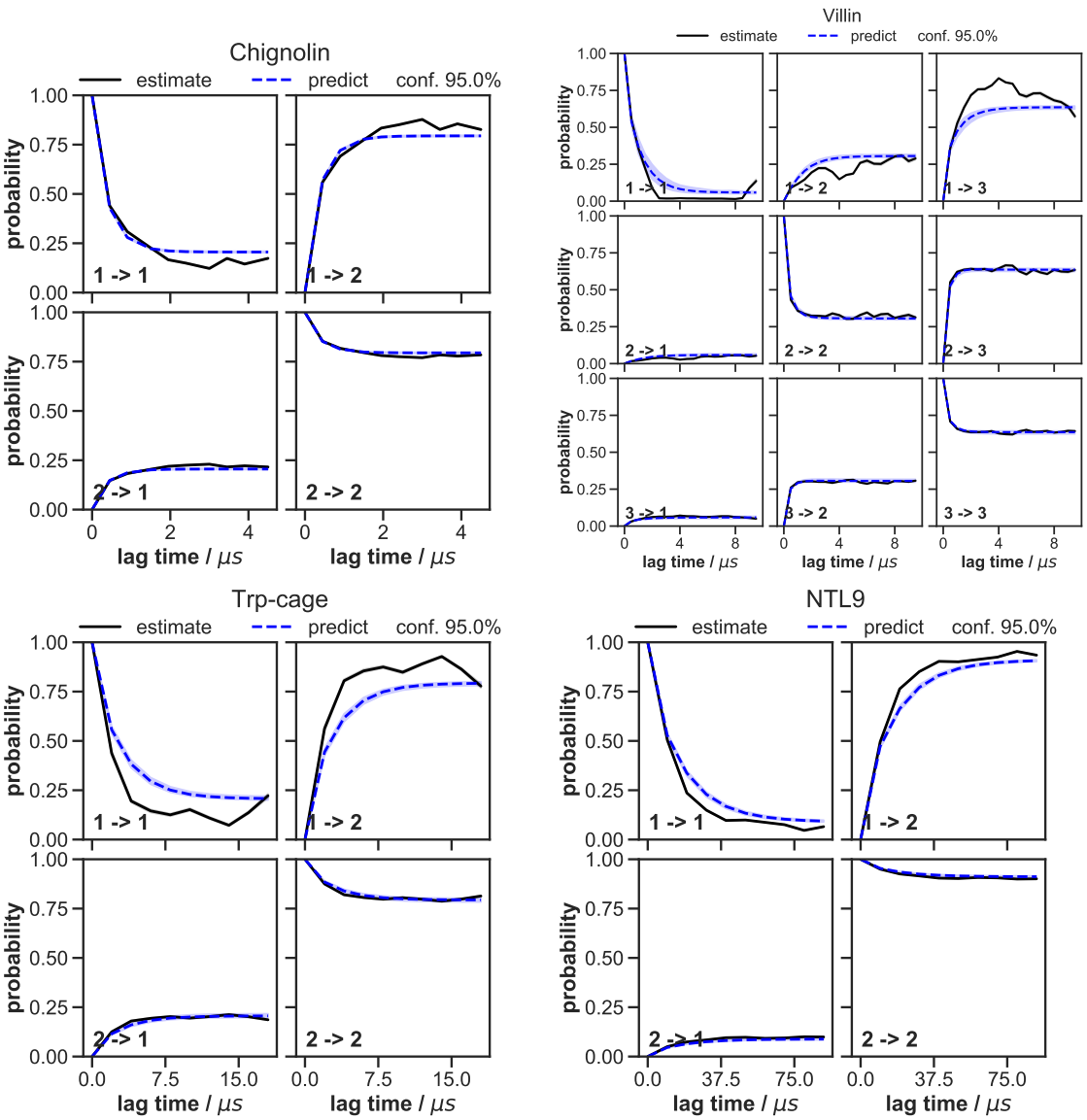

Figure S5: **Chapman-Kolmogorov tests of the Bayesian Markov state models constructed for each system.** The objective of the Chapman-Kolmogorov test is to assess the kinetic self-consistency of the MSM, i.e., whether the predictions of longer time behavior made from the BMSM being tested match the estimates made from BMSMs generated at longer lag times. For each macrostate, probability density is assigned to the BMSM microstates according to their metastable memberships to the given macrostate and evolution of the probability in time in the tested BMSM is plotted in blue ("predictions"). At those same longer lag times new BMSMs are estimated and their probability densities of being in the given macrostate after one lag time are plotted in black ("estimates"). The shaded regions correspond to the 95% confidence intervals of the mean of the predictions and estimates (the estimate confidence intervals are very narrow).

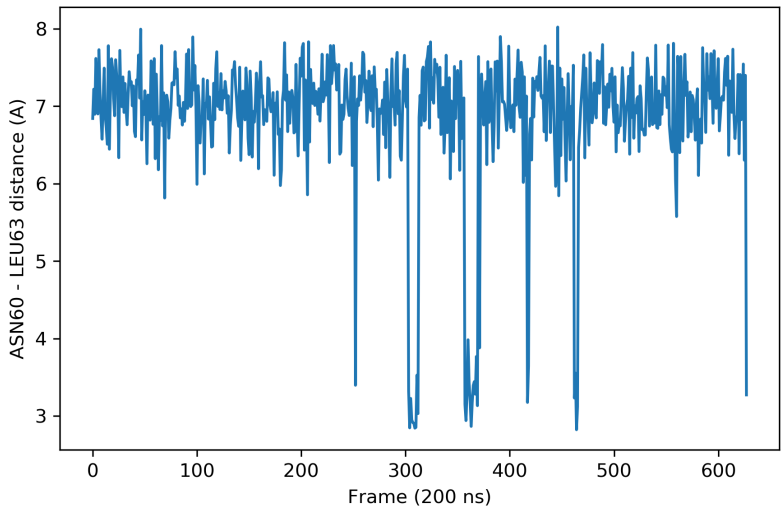

Figure S6: **ASN60 - LEU63 minimum distance along the villin trajectory.** Short-lived "misfolded" states can be seen, explaining the appearance of a very long timescale that does not correspond to folding in the MSM (see the 'PCCA++ villin coarse-graining' SI section for details).

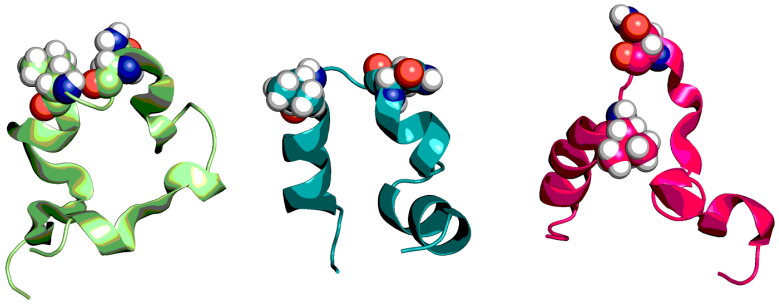

Figure S7: **Sample frames from the three macrostates of the villin MSM.** From left to right, cartoon representations of "misfolded", "folded", and "unfolded" frames are shown. ASN60 and LEU63 are also shown in spheres. The presence of the "misfolded" states explains the appearance of a very long timescale that does not correspond to folding in the MSM (see the 'PCCA++ villin coarse-graining' SI section for details).

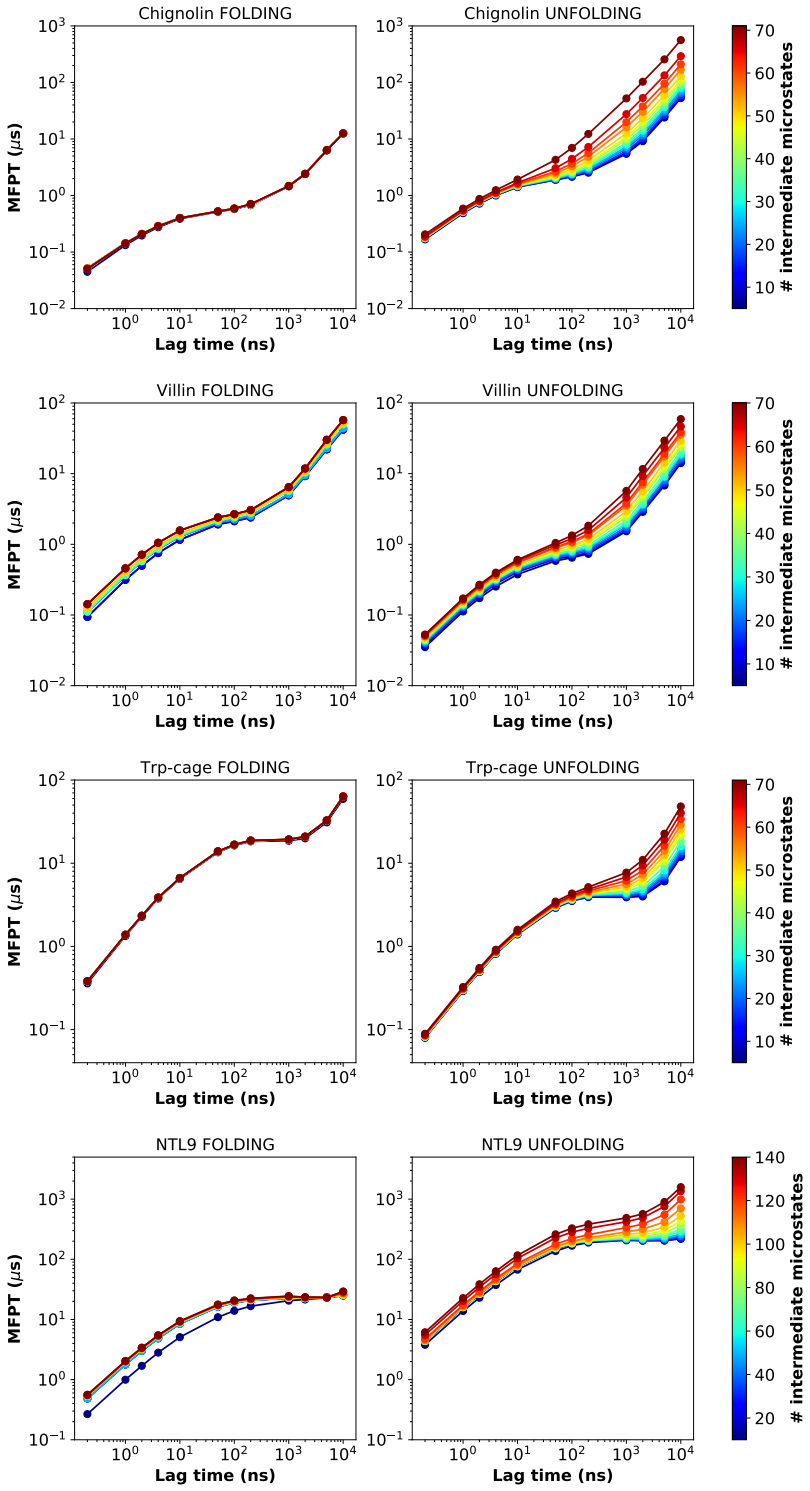

Figure S8: **Sensitivity of MFPTs calculated from Markov state models to the core-likeness of the macrostates.** The MFPT for both folding and unfolding calculated from MSMs is plotted as a function of lag time. The curves are colored by the number of microstates defined as the intermediate region - the larger the intermediate, the more core-like the macrostates become. The gray area signifies the region where MFPTs become equal to or smaller than the lag time and can no longer be resolved.

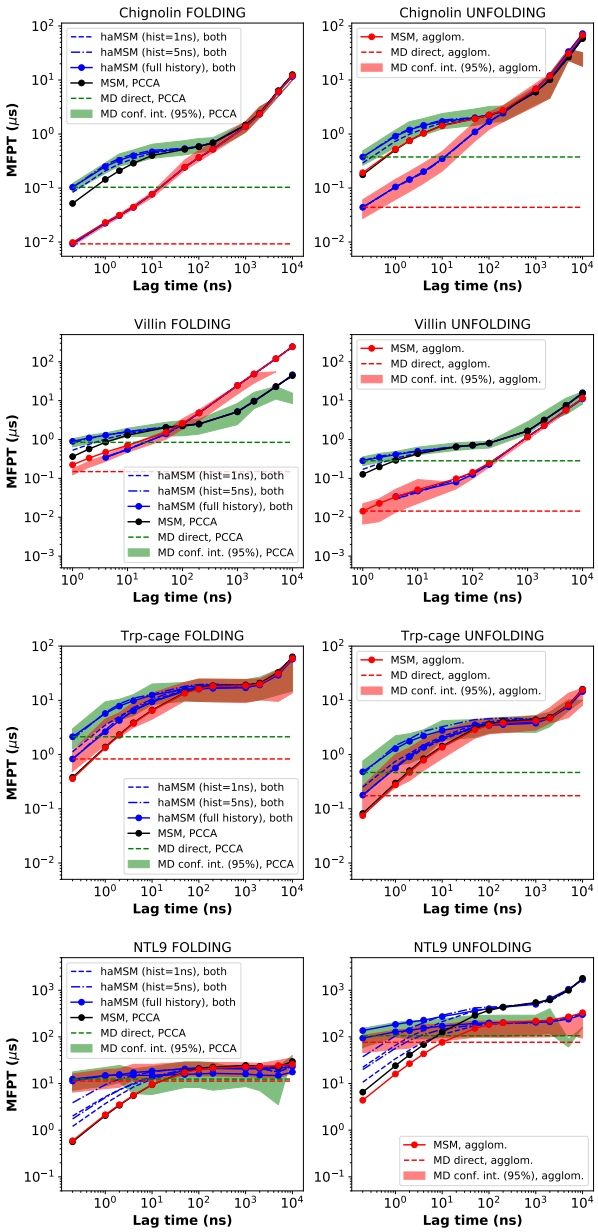

Figure S9: **Comparison of the MFPT dependence on lag time for macrostates defined by PCCA++ or agglomerative clustering.** The MFPT for both folding and unfolding is plotted as a function of lag time. Reference MD data is shown as the 95% confidence interval (green bands for PCCA and red bands for agglom.), which can be compared to validated MSM data (black lines for PCCA and red lines for agglom.) and haMSM values with full history (solid blue lines for both methods) and partial history (dashed blue lines for both methods). The PCCA macrostates are defined using a cutoff such that the resulting intermediate is of the same size as with agglomerative clustering. The gray area signifies the region where MFPTs become equal to or smaller than the lag time and can no longer be resolved. The MD confidence intervals missing for some final data points are due to no more transition events seen at those very long lag times. Two shortest lag time haMSM data points missing for villin with agglomerative clustering are due to no stationary solutions found.

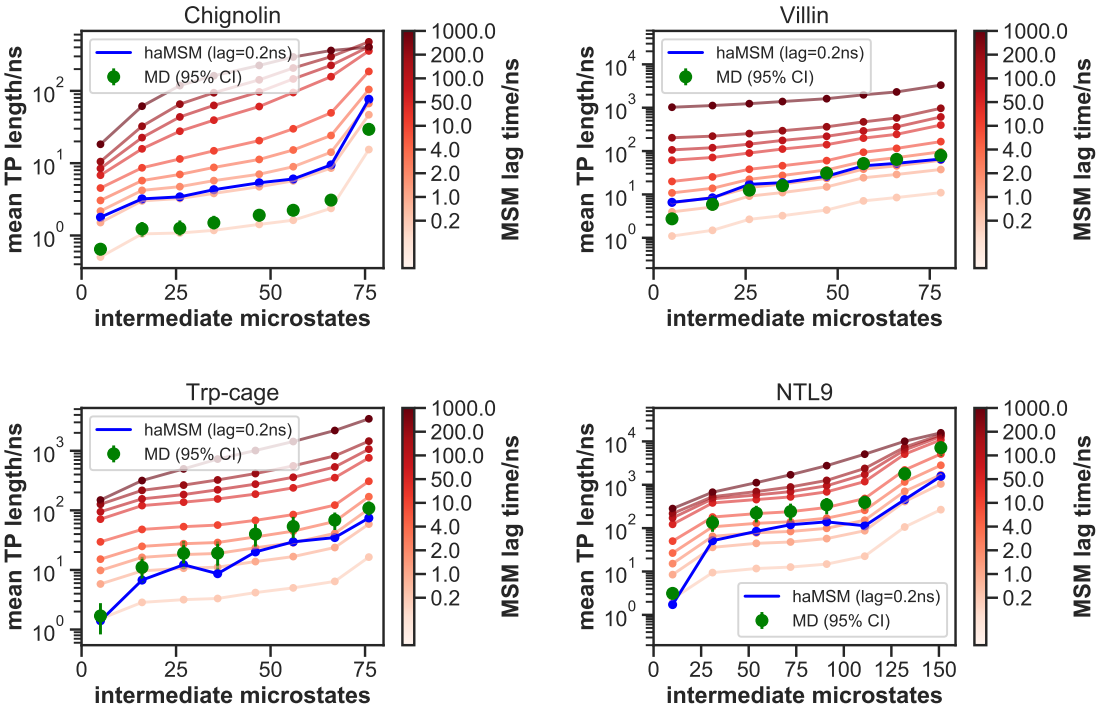

Figure S10: **Simple mechanism comparison of MD, MSMs, and haMSMs using the mean lengths of transition paths.** For the given number of intermediate microstates, ensembles of discretized transition trajectories were analyzed to determine the mean of the distribution of all transition paths. Differently from the configurational analysis of indirect paths, we include the direct paths here; the transition paths include the last frame in the origin macrostate and the first frame in the destination macrostate, i.e. the length of a direct path is 2 x the lag time.

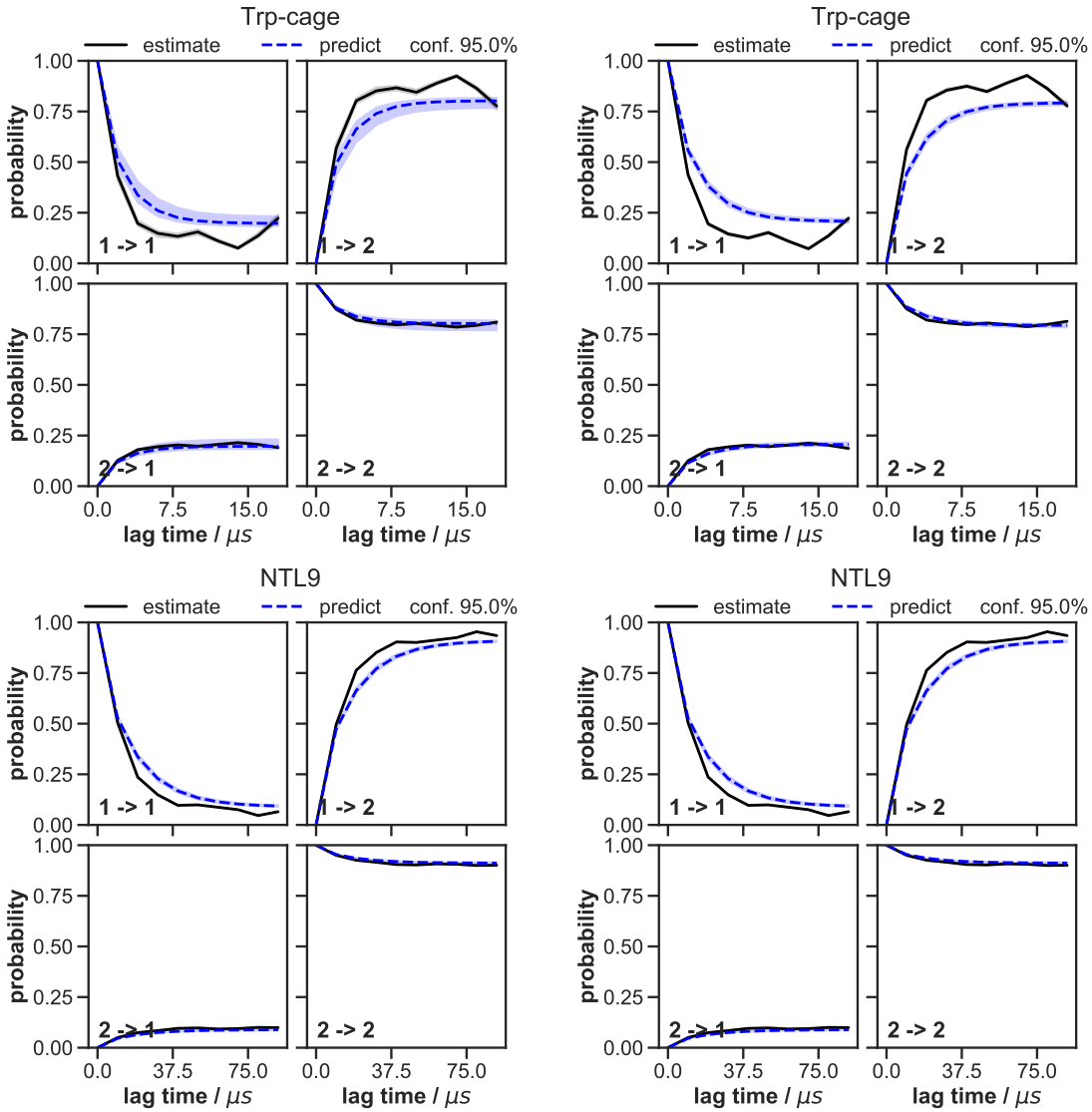

Figure S11: **Comparison of top models selected at scoring lag times 100 ns (top) and 10 ns (bottom), for Trp-cage and NTL9: Chapman-Kolmogorov tests of the Bayesian Markov state models constructed for each system.** The objective of the Chapman-Kolmogorov test is to assess the kinetic self-consistency of the MSM, i.e., whether the predictions of longer time behavior made from the BMSM being tested match the estimates made from BMSMs generated at longer lag times. For each macrostate, probability density is assigned to the BMSM microstates according to their metastable memberships to the given macrostate and evolution of the probability in time in the tested BMSM is plotted in blue ("predictions"). At those same longer lag times new BMSMs are estimated and their probability densities of being in the given macrostate after one lag time are plotted in black ("estimates"). The shaded regions correspond to the 95% confidence intervals of the mean of the predictions and estimates (the estimate confidence intervals are very narrow).

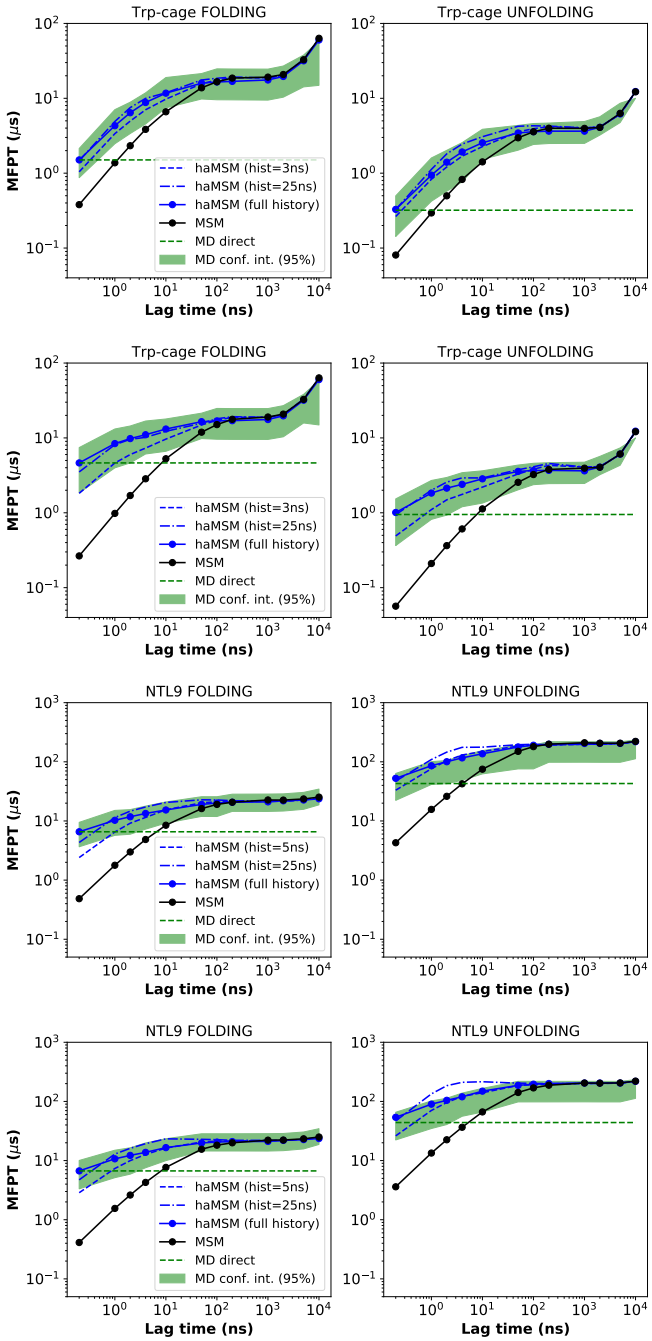

Figure S12: **Comparison of top models selected at scoring lag times 100 ns (top) and 10 ns (bottom), for Trp-cage and NTL9: MFPT estimates compared among MD, MSMs, and haMSMs.** The MFPT for both folding and unfolding is plotted as a function of lag time. Reference MD data is shown as the 95% confidence interval (green band), which can be compared to validated MSM data (black lines) and haMSM values with full history (solid blue lines) and partial history (dashed blue lines). The gray area signifies the region where MFPTs become equal to or smaller than the lag time and can no longer be resolved. The MD confidence intervals missing for the final data points of chignolin and villin are due to no more transition events seen at those very long lag times.

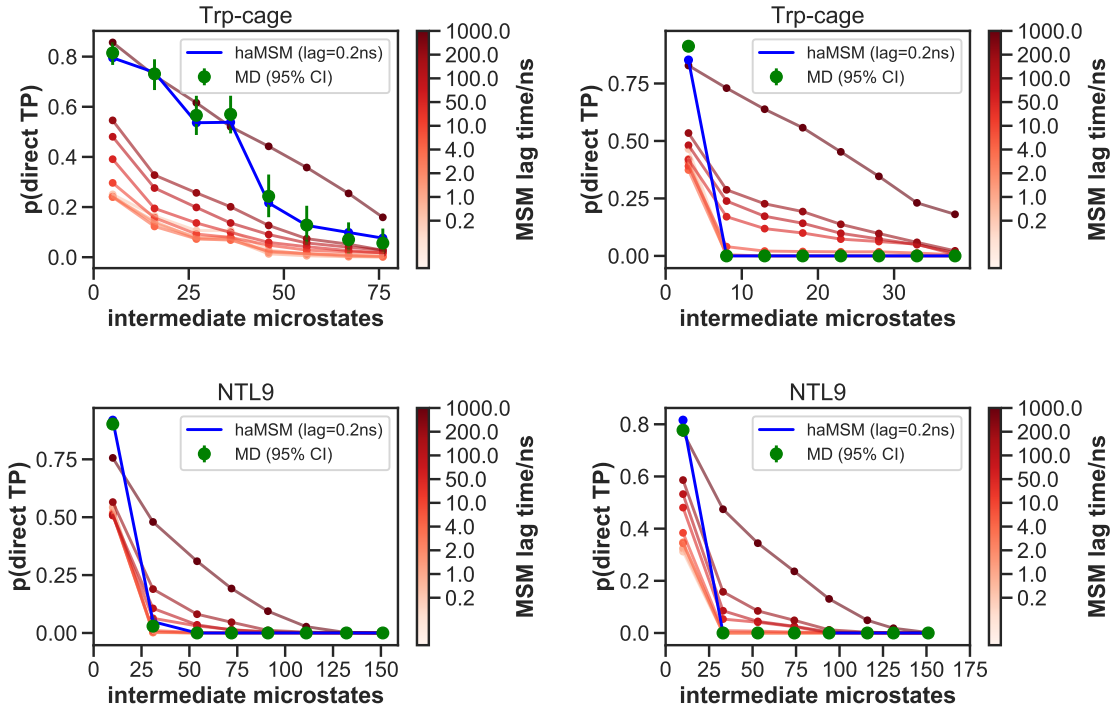

Figure S13: **Comparison of top models selected at scoring lag times 100 ns (left side) and 10 ns (right side), for Trp-cage and NTL9: simple mechanism comparison of MD, MSMs, and haMSMs using the fraction of direct folding pathways.** For the given number of intermediate microstates, ensembles of discretized transition trajectories were analyzed to determine the fraction which directly 'hopped over' the intermediate region based on either the MD discretization time Δt, also used for haMSM modeling, or else the indicated MSM lag time. A greater number of intermediate microstates indicates relatively smaller macrostates and accounts for the monotonic decrease of direct transitions. Two haMSMs data points are missing for NTL9 (10 ns) due to numerical problems.

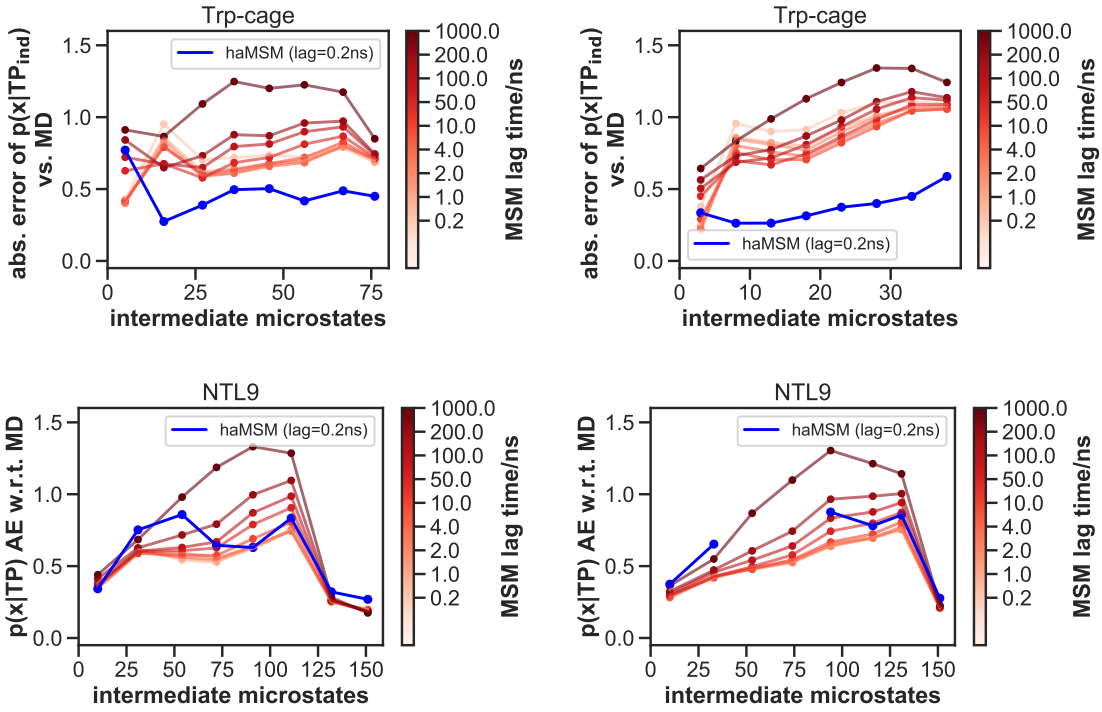

Figure S14: **Comparison of top models selected at scoring lag times 100 ns (left side) and 10 ns (right side), for Trp-cage and NTL9: mechanism comparison of MSMs and haMSMs to MD using the configurational distributions of transition path ensemble.** Each panel plots the summed *absolute error, as compared to MD,* for intermediate microstate probabilities calculated for the transition path ensembles, i.e., $p(x|TP_{\text{ind}})$, for the given number of intermediate microstates. Importantly, the "ind" subscript indicates that direct pathways analyzed in Figure 6 (main manuscript) were excluded from the ensembles prior to computation of errors; had they been included, the MSM errors would be substantially larger. Two haMSMs data points are missing for NTL9 (10 ns) due to numerical problems.

References

(1) Wu, H.; Noé, F. *arXiv preprint arXiv:1707.04659* **2017**,

(2) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. *J. Chem. Phys.* **2013**, *139*, 015102.

(3) Schwantes, C. R.; Pande, V. S. *J. Chem. Theory Comput.* **2013**, *9*, 2000–2009.

(4) Noé, F.; Clementi, C. *J. Chem. Theory Comput.* **2015**, *11*, 5002–5011.

(5) Noé, F.; Banisch, R.; Clementi, C. *J. Chem. Theory Comput.* **2016**, *12*, 5620–5630.

(6) Beauchamp, K. A.; McGibbon, R.; Lin, Y.-S.; Pande, V. S. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 17807–17813.

(7) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517–520.

(8) Schrödinger, L. The PyMOL Molecular Graphics System. 2020.

(9) Schrödinger, LLC, Maestro. 2020.

(10) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.