Supporting Information for Modeling the Mechanism of CLN025 Beta-Hairpin Formation

Keri A. McKiernan,[1, a)] Brooke E. Husic,[1, a)] and Vijay S. Pande[1]

Department of Chemistry, Stanford University, Stanford CA 94305, USA

[a)]K.A.M. and B.E.H. contributed equally to this work.

CLN025

CLN025 (PDB: 5AWL) is a 10-residue protein designed by Honda *et al.*[1] based on the structure of chignolin.[2] CLN025 has the sequence sequence TYR-TYR-ASP-PRO-GLU-THR-GLY-THR-TRP-TYR. It is interchangeably referred to as chignolin or CLN025 in the literature.

CHARMM22* SIMULATION DATA

The baseline CLN025 dataset was generated by Lindorff-Larsen *et al.*[3] The simulation was performed at 340K in the NVT ensemble using the CHARMM22* force field[4] and the CHARMM-compatible TIP3P water model[5]. The dataset comprises one 106 μs simulation that contains 39 folding events and 38 unfolding events.

PREPARATION OF FOLDING@HOME SIMULATION DATA

The CLN025 crystal structure was first solvated (3961 water molecules) and neutralized (2 Na^+ ions). This system was denatured via simulation at 600 K until fully extended, using a ff99SB-ILDN parameterization. This configuration was then ported to a range of new force fields, and equilibrated for 1 ns in the new force field at the melting temperature. After equilibration, 100 instances of this positional configuration were written as initial conditions for a unique Folding@home simulation, each with a unique velocity distribution. The following OpenMM[6] script was used to convert the denatured system to the set of Folding@home initial states as a function of input protein and water force fields.

```
from __future__  import print_function
from simtk.openmm import app
import simtk.openmm as mm
from simtk import unit
from sys import stdout
import os
import argparse

parser = argparse.ArgumentParser(description="equilibrate
  denatured protein via input ff")
parser.add_argument("--pdb", type=str, help="pdb")
parser.add_argument("--pff", type=str, help="protein ff")
parser.add_argument("--wff", type=str, help="water ff")
args = parser.parse_args()

# read in system pdb
ref = args.pdb
pdb = app.PDBFile(ref)
topology = pdb.getTopology()
positions = pdb.getPositions()
# set force field
pff = args.pff
wff = args.wff

ff = app.ForceField(pff, wff)
platform = mm.Platform.getPlatformByName("CUDA")
properties = {"CudaPrecision": "mixed"}
integrator = mm.LangevinIntegrator(t_eq*unit.kelvin,
    1.0/unit.picoseconds, 2.0*unit.femtoseconds)
integrator.setConstraintTolerance(0.00001)
system = ff.createSystem(topology, nonbondedMethod=app.PME,
    nonbondedCutoff=1.0*unit.nanometers, constraints=app.HBonds,
    rigidWater=True, ewaldErrorTolerance=0.0005,
    vdwCutoff=1.2*unit.nanometer)
simulation = app.Simulation(topology, system, integrator,
        platform, properties)
simulation.context.setPositions(positions)

# minimize
simulation.minimizeEnergy(maxIterations=50)
# set temperature
t_eq = 340
simulation.context.setVelocitiesToTemperature(t_eq*unit.kelvin)
# eq for 1 ns
simulation.step(500000)

# randomize velocities, save state
intial_conditions = []
for i in range(100):
  simulation.context.setVelocitiesToTemperature(t_eq*unit.kelvin)
  state = simulation.context.getState(getPositions=True, getVelocities=True,
      getForces=True, getEnergy=True, getParameters=True,
      enforcePeriodicBox=True)
  intial_conditions.append(state)
```

MODEL SELECTION

For the baseline model, the CHARMM22* dataset was featurized into the sines and cosines of the α dihedral angles (i.e. the dihedrals along the α-carbon backbone) and transformed using the tICA algorithm[7,8] with a lag time of 128 ns and the kinetic mapping[9] weighting scheme. All 14 components of the tICA solution were retained and clustered into 704 microstates using mini-batch k-means. A Markov state model (MSM) was constructed on the entire dataset with a MSM lag time of 50 ns based on the lag time reported by Beauchamp *et al.*[10] for the same dataset. The tICA lag time, whether or not to use kinetic mapping, and the number of clusters were optimized by randomly searching relevant parameter spaces and then evaluating the average GMRQ[11,12] for the top two dynamical timescales of the resulting MSM over five iterations of shuffle-split cross-validation with 50% of the data used for the validation set. All tICs were included based on previous analysis showing that retaining all tICs when using kinetic mapping does not decrease the GMRQ (Husic *et al.*[13], supplementary materials). These optimizations were performed for the simulation strided at 2 ns. The search space is given in Table S1.

TABLE S1. Search space for CHARMM22* model parameters.

Parameter	Min	Max	Scale
tICA lag time (ns)	20	500	Log
Number of microstates	10	1000	Uniform

MODEL PROJECTION

The resulting baseline MSM was fit on the CHARMM22* dataset (strided at 1 ns). For all other datasets, the baseline model was used to predict the location of each frame along the tICs defined by the baseline model and then the microstate assignment of that frame based on its tICA coordinates. Each projected MSM was built from these predicted assignments. By using projected MSMs, the tICs and microstates are the same for all models. For each model, 100 bootstrapped Markov state models were created to determine model statistics.

MSM lag time can be validated for all models by observing that the implied timescales have leveled off at a lag time of 50 ns, indicating that at a lag time of 50 ns the model is Markovian (Fig. S1). Note that when the plots overlap the identity function this signifies that the longest dynamical process is shorter than the selected lag time.

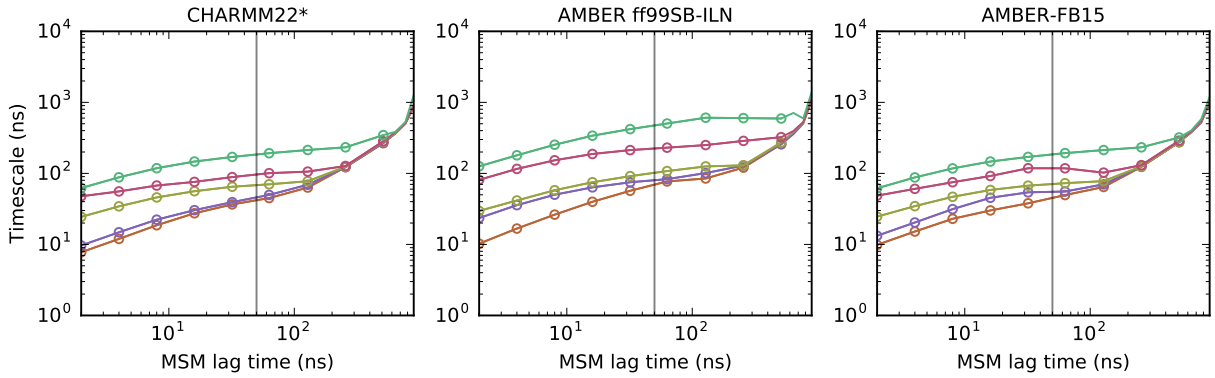

FIG. S1. Implied timescales for three protein and water force field combinations.

MODEL VALIDATION

To evaluate the self-consistency of a model it is standard to assess whether it adheres to the Chapman-Kolmogorov property.[14–16] This test is an evaluation of whether $[\hat{\boldsymbol{T}}(\tau)]^k \approx \hat{\boldsymbol{T}}(k\tau)$, i.e. whether the transition probabilities determined from raw trajectory data approximately match the corresponding probabilities produced from the model. To perform the Chapman-Kolmogorov test for our models we chose two macrostates and then evaluated the probabilities of finding the system in each macrotate at time $k\tau$ as predicted by (1) a model created at lag time τ and (2) an independent model created with lag time $k\tau$. Figures S2-4 show the Chapman-Kolmogorov test for three models. 95% confidence intervals were estimated by constructing Bayesian MSMs.[17] This analysis was performed using the PyEMMA software package.[18]

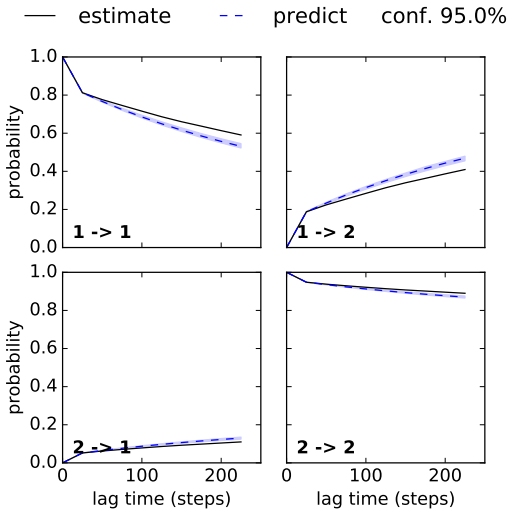

FIG. S2. Chapman-Kolmorgorov test for the CHARMM22*-m3p model.

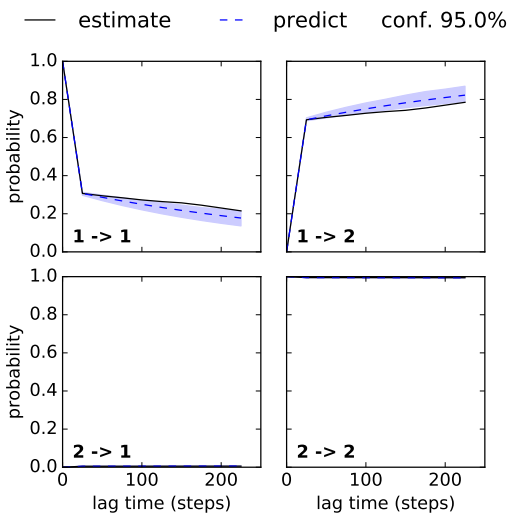

FIG. S3. Chapman-Kolmorgorov test for the ILDN-3p model.

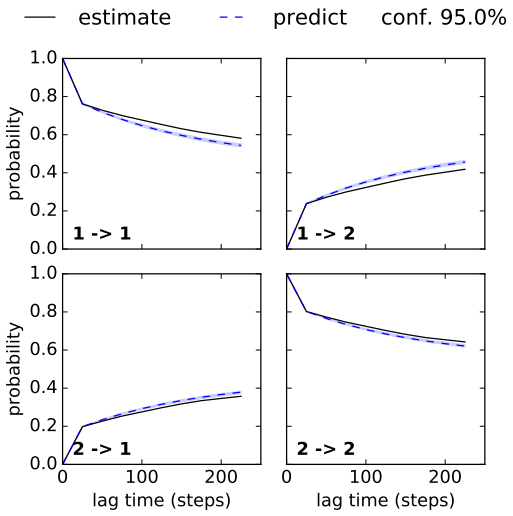

FIG. S4. Chapman-Kolmorgorov test for the FB15-3pfb model.

ALTERNATIVE BASELINE MODELS

The baseline MSM used in this study was fit on the CHARMM22* dataset as described above. The qualitative conclusions are independent of the basis chosen. In Fig. S5, we show the 1-dimensional free energy plot from Fig. 1 in the main text using the AMBER ff99SB-ILDN and AMBER-FB15 bases (left and right, respectively). The tICA lag time was 20 ns and the MSM lag time was 50 ns. 200 models were made with these parameters to optimize the number of microstates based on 10 iterations of shuffle-split cross-validation for each model. 663 microstates were used for the AMBER ff99SB-ILDN model and 157 microstates were used for the AMBER-FB15 model based on these results. The same shapes and relative basin depths are observed when using either AMBER basis set when compared to the CHARMM22* basis set analyzed in the main text. We expect that this correspondence is related to the choice of α dihedral angles as the features, since the backbone dynamics are expected to be the same across force fields.

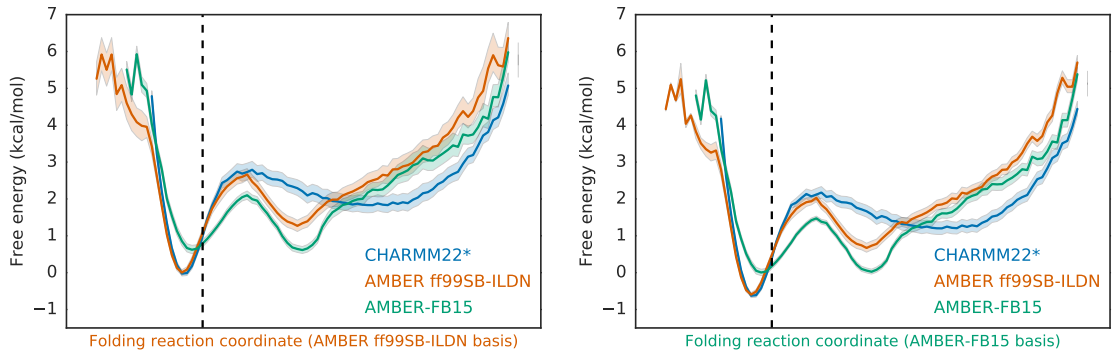

FIG. S5. 1-dimensional free energy landscapes for CLN025 folding when using AMBER ff99SB-ILDN (left) or AMBER-FB15 (right) for the baseline model. Each model is zeroed to the energetic minimum of the baseline model, so the heights are offset when comparing the baseline model free energy landscapes.

MEAN FIRST PASSAGE TIME ANALYSIS

The mean first passage time (MFPT) analysis was performed using the projected models described above. MFPTs can be compared across models because the states are the same but the populations and pairwise transition probabilities differ. To calculate the MFPT to a macrostate, the microstates characterizing that macrostate must be selected. To determine the folded state, all microstate free energies were offset by the minimum free energy in the (baseline) CHARMM22* model. After this offset, the minimum free energy in the CHARMM22* model is 0. Next, the 267 microstates with free energies of 1 kcal/mol or smaller in the CHARMM22* model were selected as the folded state. These microstates were used to define the folded state for all models, and the MFPT to the folded state for each model was calculated using these 267 states as sinks.

MFPT analysis was also performed to the metastable unfolded state corresponding to free energy basins in the 1D reaction coordinate in the main text. The unfolded state was initially characterized unfolded state free energy basin in the (projected) AMBER-FB15 model. The 59 microstates with free energies lower than 1 kcal/mol (after zeroing with the CHARMM22* model) were selected as the unfolded state. These microstates were also chosen to comprise the AMBER ff99SB-ILDN unfolded state since they overlap the smaller free energy basin present in the AMBER ffSB-ILDN model. For the CHARMM22* model, the maximum free energy of the AMBER ff99SB-ILDN microstates composing the unfolded state was used as a cutoff to characterize the CHARMM22* unfolded state. All 123 microstates with free energies lower than the cutoff established by the AMBER ff99SB-ILDN unfolded basin were used to define the CHARMM22* unfolded state. We note that MFPT results and trends are robust to other sensible definitions of the folded and unfolded states, and that this strategy was motivated by choosing the most similar folded and unfolded states possible given the projected MSM framework. The microstates selected for the folded and unfolded microstates are depicted in Fig. S6.

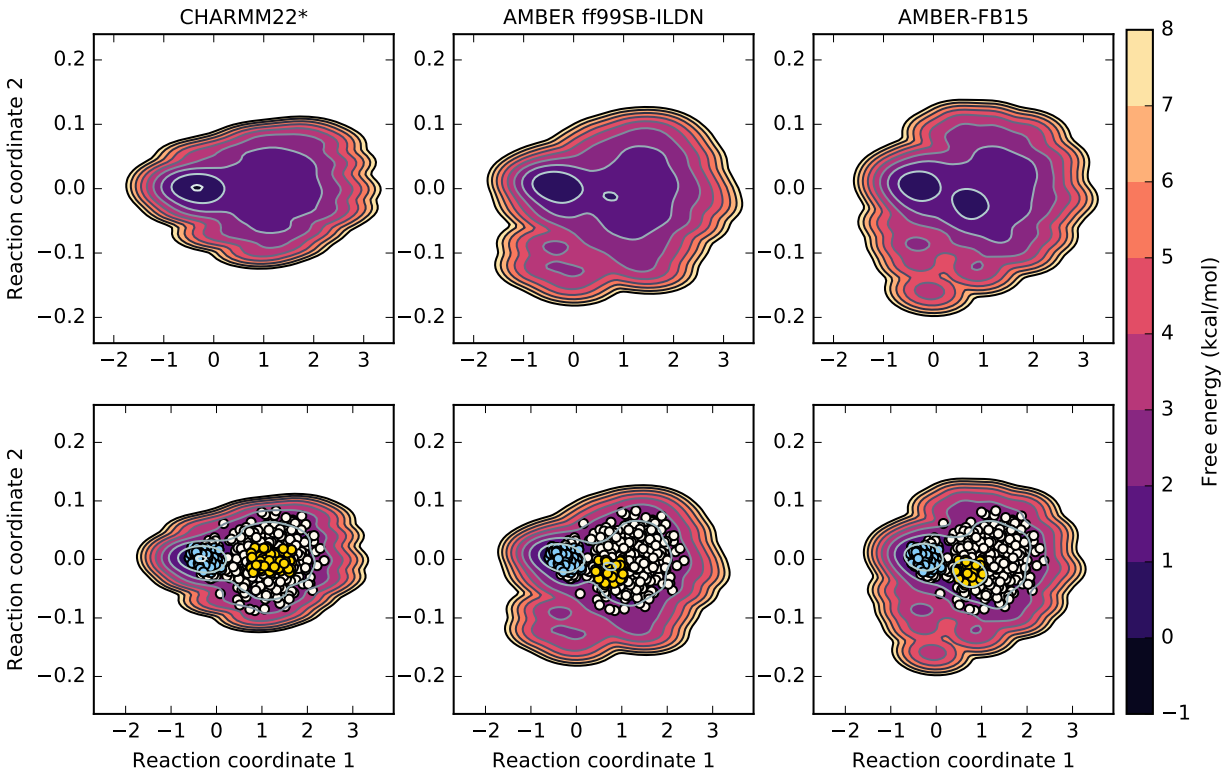

FIG. S6. Top: Free energy surfaces for each protein force field from the baseline (CHARMM22*) and projected (AMBER datasets) models. Bottom: The same free energy surfaces with all microstates shown. The cyan microstates indicate the folded state and are the same in all three datasets. The gold microstates represent the unfolded state. The unfolded states are the same for the AMBER datasets.

HAND-SELECTED FEATURES FOR DYNAMICAL SUBPROCESSES

Structural characterization of the collapsed state was accomplished by monitoring the radius of gyration of two terminal residues. Characterization of the turn and beta sheet zip motifs were analyzed using distances between pairs of hydrogen bonds found in the 5AWL crystal structure. These features are summarized in Table S2. For the hydrogen bond pairs, the listed residue number, atom name, and atom numbers correspond to those listed in the 5AWL crystal structure.

TABLE S2. Table of features used to represent dynamical subprocesses in the folding of CLN025.

Hydrophobic collapse	Turn formation	Beta sheet zip
R_g(TYR1)	ASP3 O (48) - GLY7 N (100)	TYR1 N (1) - TYR10 OXT (157)
R_g(TYR10)	ASP3 OD1 (51) - GLU5 N (71)	TYR1 O (4) - TYR10 N(145)
	ASP3 OD1 (51) - THR6 N (86)	ASP3 N (45) - THR8 O (110)
	ASP3 OD2 (32) - GLU5 N (71) [a]	
	ASP3 OD2 (32) - THR6 N (86) [b]	

[a] Pair symmetrically equivalent to native bond
[b] Pair symmetrically equivalent to native bond

We expect that these feature sets are not completely independent and that there exists some degree of coupling between the collapse and turn processes. Furthermore, only certain varieties of turn geometries may lead to beta sheet formation[19]. Note that although the ASP3 OD1 (51) - GLU5 N (17) bond corresponds to the experimental fast rate (related to hydrophobic collapse), this bond is present in the turn region of the crystal structure. We do not expect our feature sets to completely isolate each subprocess. Rather, they are structurally motivated and informed by the native state contacts.

REPRESENTATIVE TRAJECTORY SELECTION

Representative trajectories selected for mechanism analysis were obtained using transition path theory (TPT)[15,20,21]. The microstates characterizing the folded state selected for the MFPT analysis above were used as sinks. For each dataset, the most extended state according to the sum of the radii of gyration of all 10 residues was used as the source. Then, using TPT, the top 4 paths from the extended state to any of the folded states were enumerated for the CHARMM22* and AMBER ff99SB-ILDN datasets and the top 8 paths were enumerated for the AMBER-FB15 dataset. Each trajectory dataset was searched for single trajectories that contained a pathway between the most extended state (for that trajectory) and the sink identified from the TPT analysis. This analysis found 43 paths for CHARMM22*, 40 paths for AMBER ff99SB-ILDN, and 13 paths for AMBER-FB15. It is likely that fewer paths were found for AMBER-FB15 due to population of the most extended state in the projected MSM. Representative trajectories for Fig. 3 in the main text were chosen to contain a single folding event and were truncated soon after folding if the single folded microstate specified was not reached within 10-20 ns of folding (as long as the system remained folded until it reached the specified folded microstate). The first fifteen paths for CHARMM22* and AMBER ffSB-ILDN and all thirteen paths for AMBER-FB15 are shown below (Figs. S7-S9). The pathways included in the main text are indicated with an asterisk. Multiple asterisks mean the same path was identified for more than one pair of start and end states.

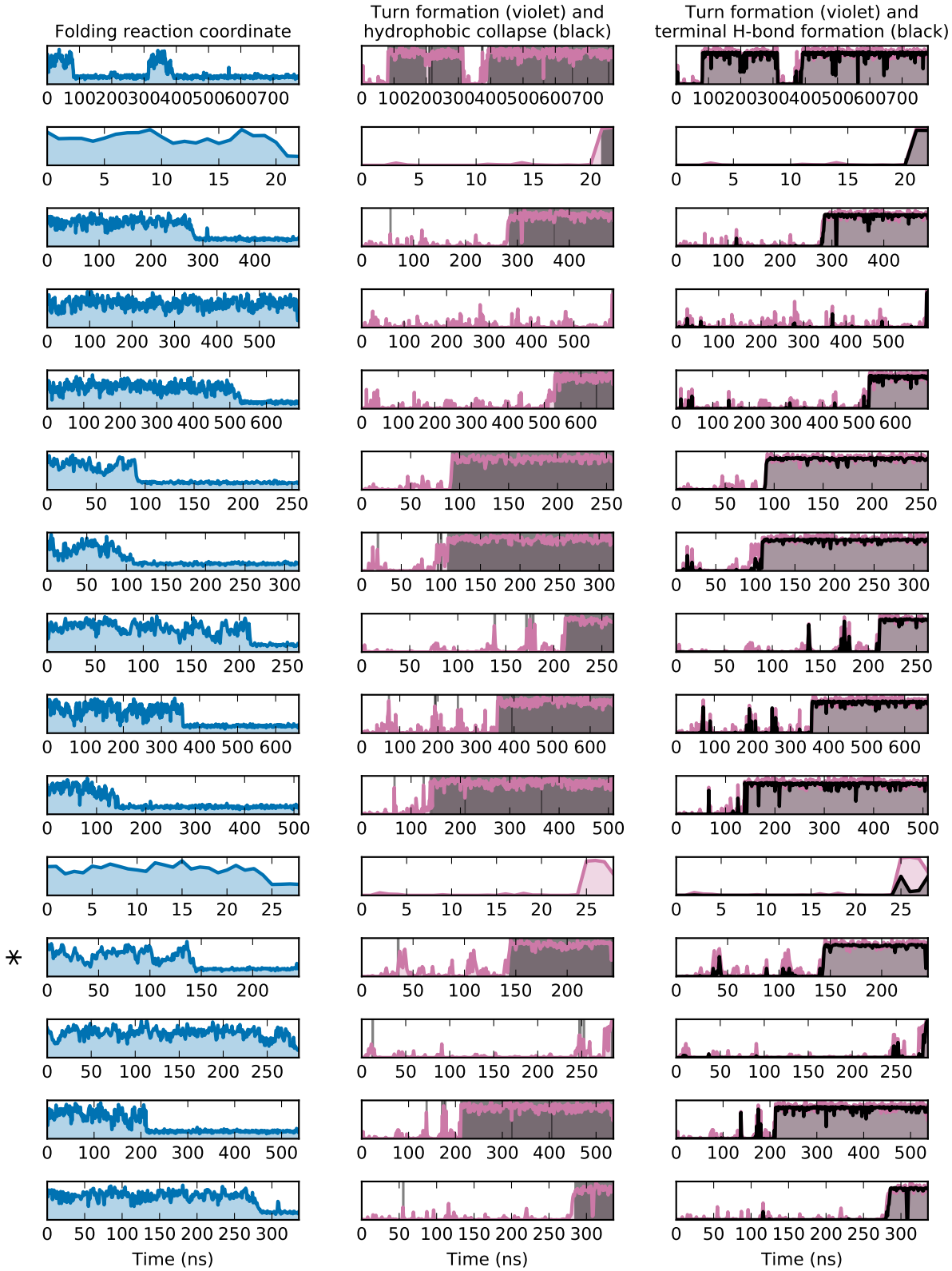

FIG. S7. 15 trajectories for CHARMM22* from the top 3 paths.

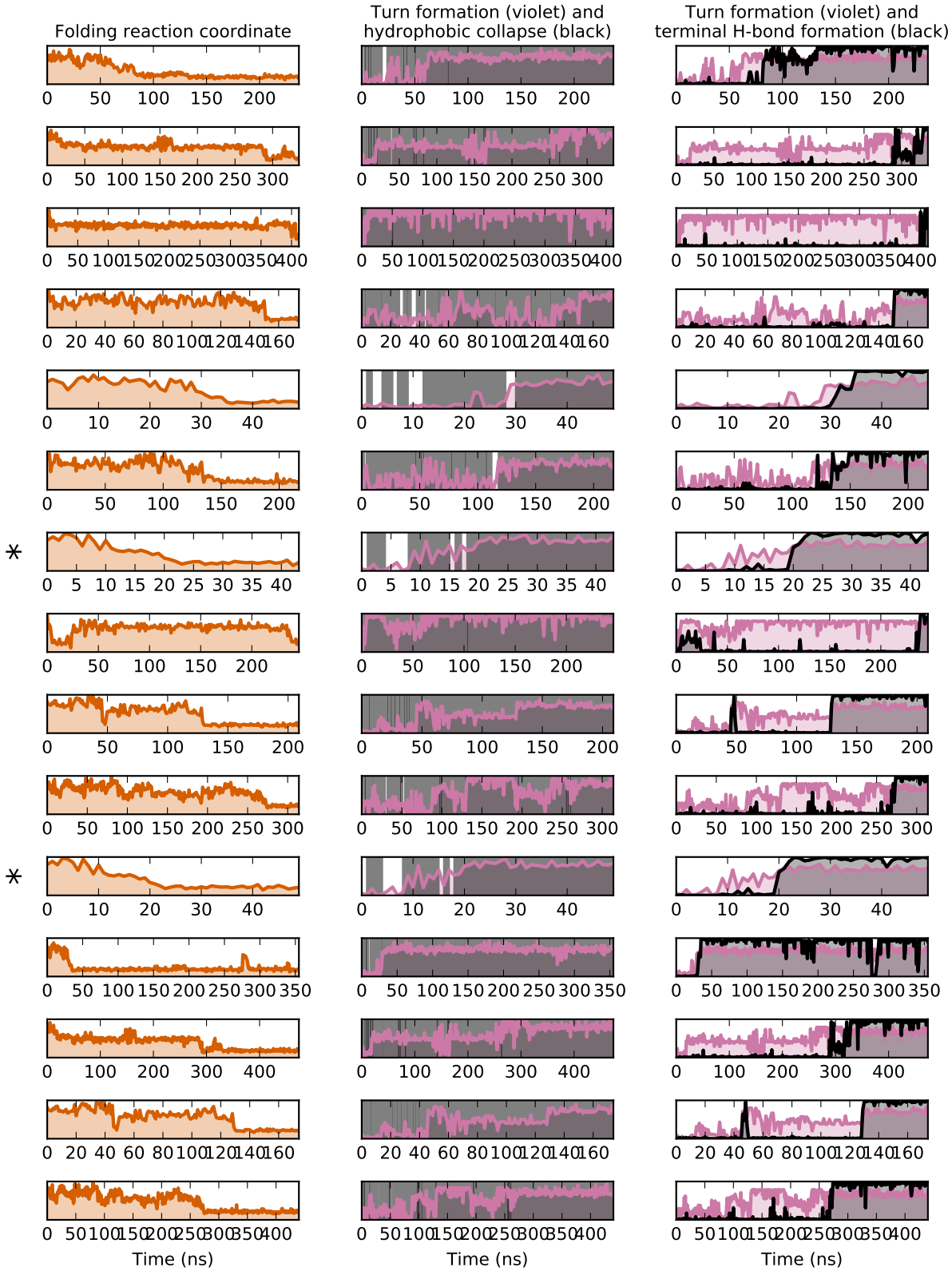

FIG. S8. 15 trajectories for AMBER ff99SB-ILDN from the top 2 paths.

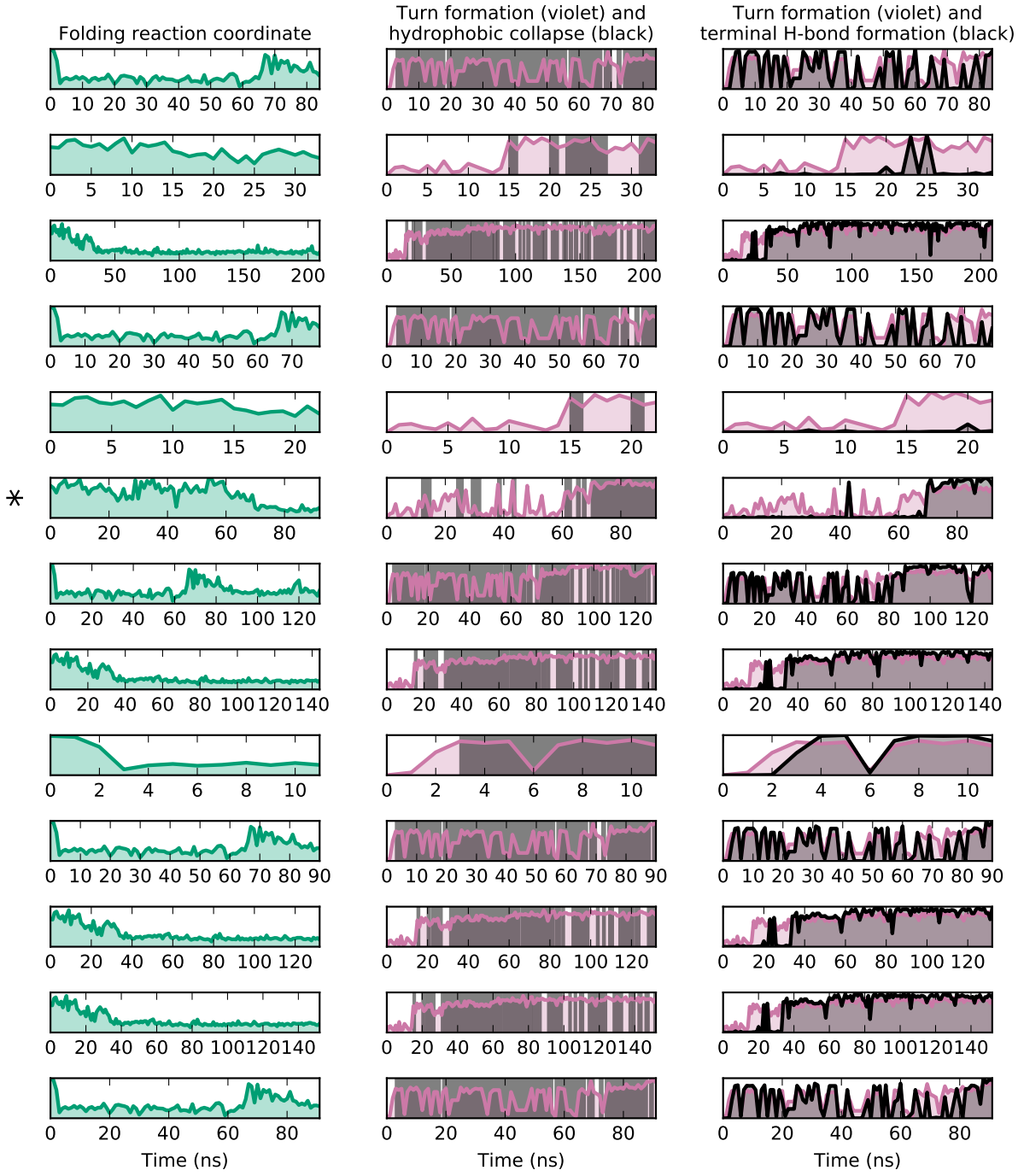

FIG. S9. All 13 trajectories for AMBER-FB15 from the top 8 paths.

MSMS FOR RELAXATION TIMESCALES

When constructing a MSM each frame of each trajectory in a dataset is "featurized" from its raw Cartesian coordinates into informative vectors containing such information as dihedral angles and/or pairwise contact distances. If the modeler wishes to analyze the dynamical processes of the trajectory it is crucial to use features that optimally encode the degrees of freedom of the system.[13] However, a MSM can also be used to interrogate a specific dynamical process by deliberately choosing features that isolate the relevant degrees of freedom. The latter methodology was utilized to determine the relaxation timescales corresponding to turn formation and hydrophobic collapse in the mechanism of CLN025 folding. The MSM eigenfunctions correspond to the dynamical processes present in the MD dataset, and their associated timescales indicate the time it takes for the process to decay to the stationary distribution that characterizes the MSM. The relative ordering of MSM timescales are expected to correspond to the relative ordering of relaxation timescales reported in the T-jump experiments performed by Davis *et al.*[22].

In order to capture the degrees of freedom related to turn formation, the five distances between the hydrogen bond contacts of the native state turn region reported in Table S2 and their associated dihedral angles were measured. In order to isolate the degrees of freedom that track the hydrophobic collapse of CLN025, the radii of gyration of the two terminal, hydrophobic tyrosines were used as the only two features. The model specifications were

TABLE S3. Optimized model parameters for process-isolating MSMs.

Protein force field	Water model	tICA lag time (ns)	Number of microstates
Turn formation			
CHARMM22*	mTIP3P	57.6	30
AMBER ff99SB-ILDN	TIP3P	12	923
AMBER-FB15	TIP3P-FB	16.4	53
Hydrophobic collapse			
CHARMM22*	mTIP3P	3	751
AMBER ff99SB-ILDN	TIP3P	14.6	208
AMBER-FB15	TIP3P-FB	1.6	141

optimized using the average GMRQ over 10 cross-validation iterations using shuffle split for 500 choices of tICA lag time and microstate number pairs. Models were optimized at 200 ps strides, retained all tICA components, used the kinetic mapping weighting scheme,[9] were defined at a lag time of 50 ns, and were evaluated on the slowest 5 timescales. The optimized model parameters are provided in Table S3. The dependence of the model score on the optimized parameters is shown in Figs. S10 and S11. The optimization search space is provided in Table S4.

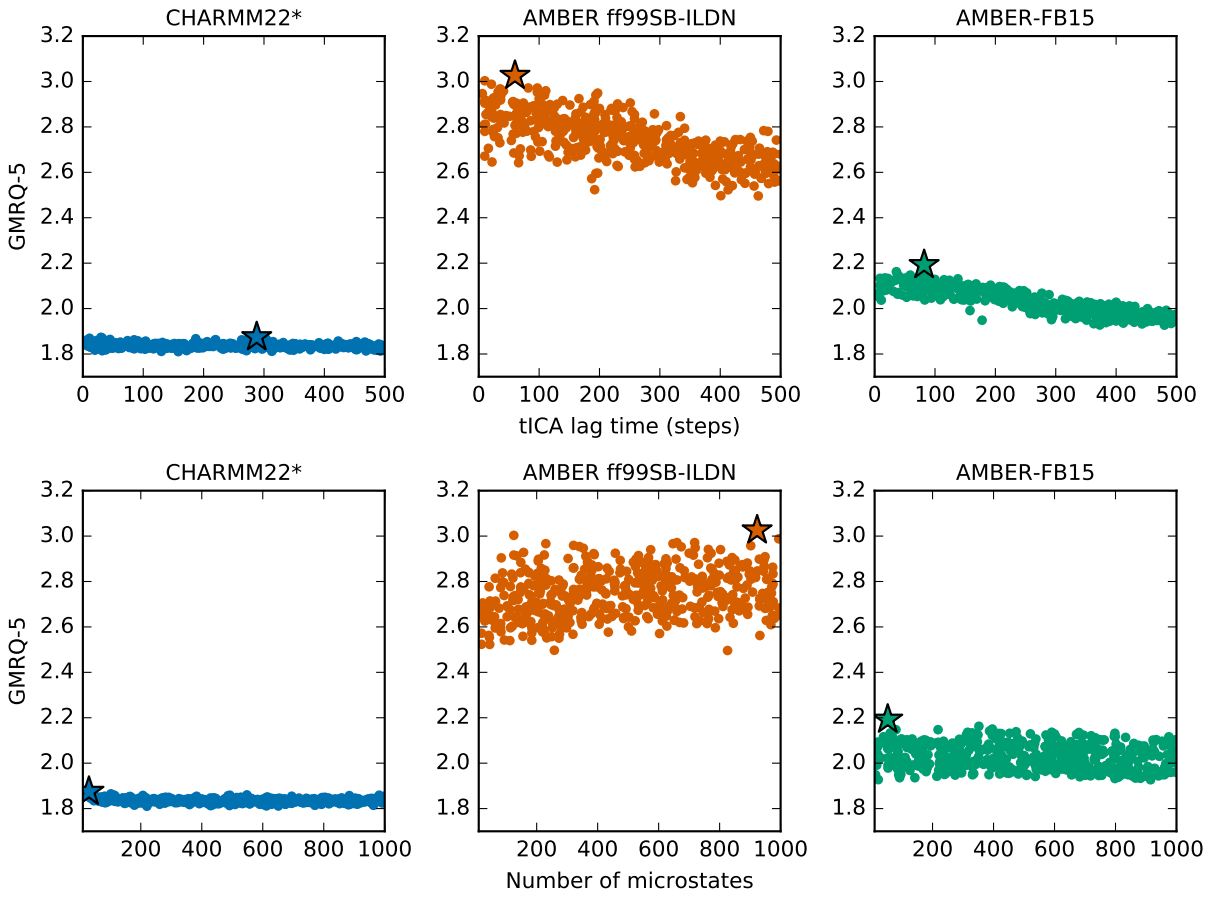

FIG. S10. The effect of tICA lag time and number of microstates on the turn formation model GMRQ score. Steps are 200 ps.

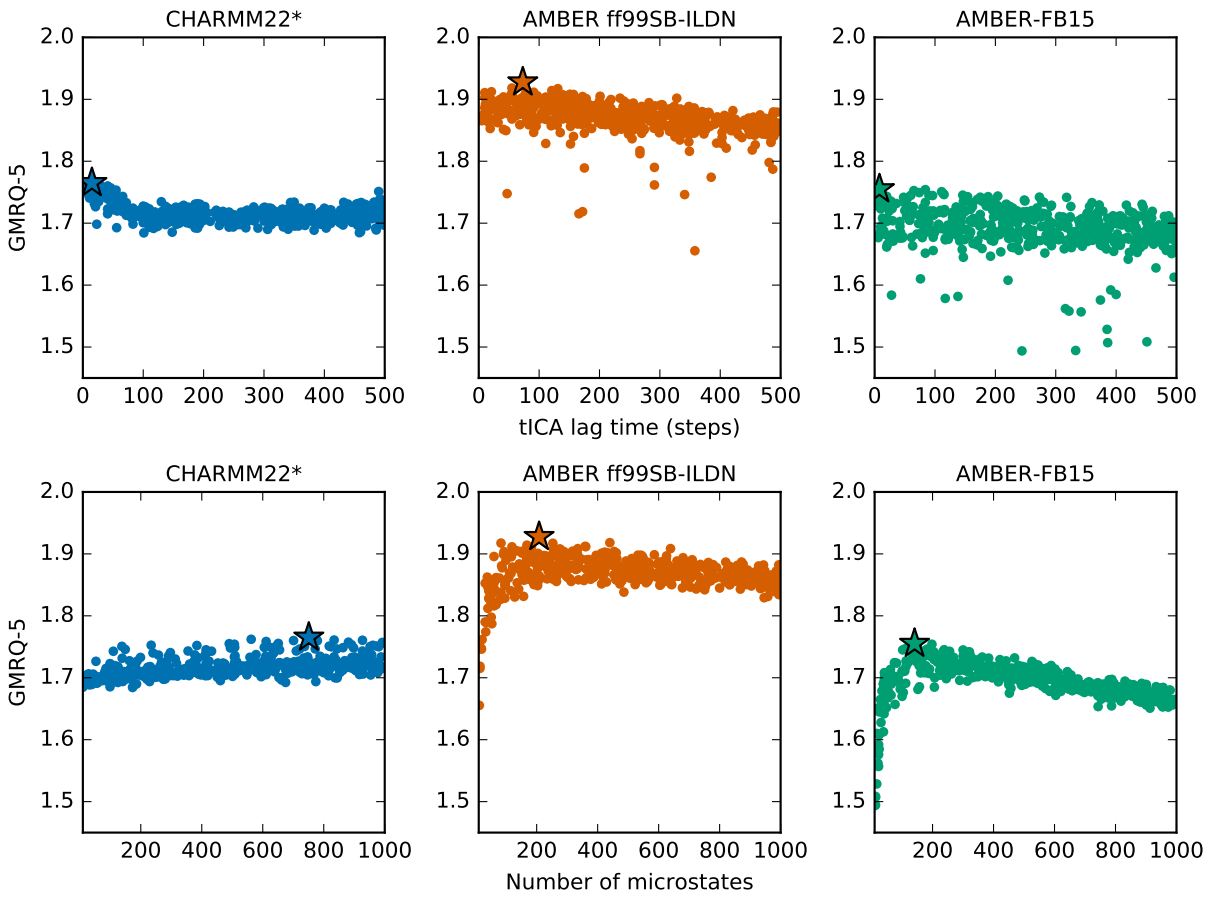

FIG. S11. The effect of tICA lag time and number of microstates on the hydrophobic collapse model GMRQ score. Steps are 200 ps.

TABLE S4. Search space for relaxation timescale model parameters. The same search space was used to optimize both turn formation and hydrophobic collapse MSMs.

Parameter	Min	Max	Scale
tICA lag time (ns)	1	100	Uniform
Number of microstates	10	1000	Uniform

Analysis of implied timescales was performed on these models. The implied timescale plots in Fig. S12 show the flattening of the timescale curves. A lag time of 50 ns was used for further analysis based on the lag time reported by Beauchamp *et al.*[10]. A longer choice would not be appropriate since the timescales of interest are on the same order of magnitude. The conclusions in our study are based on the relative duration of the turn and hydrophobic collapse timescales for each model.

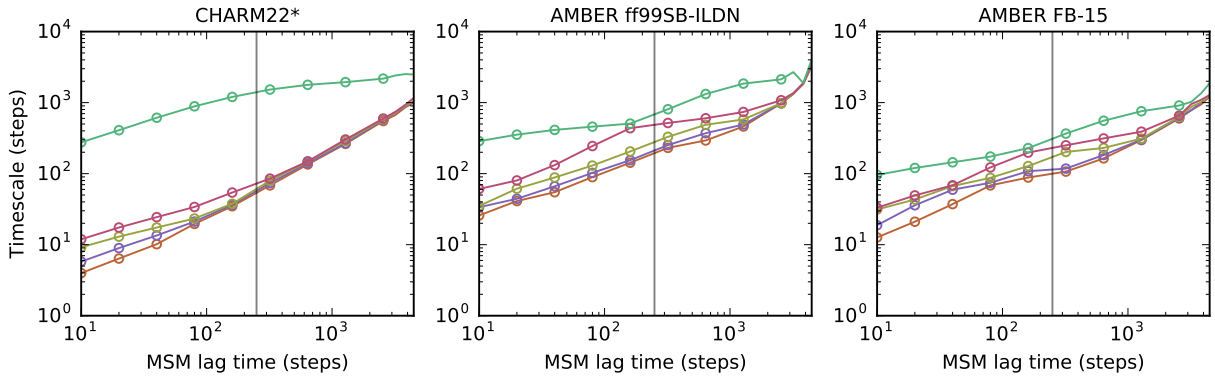

FIG. S12. Implied timescales for turn formation. The line at 50 ns indicates the lag time at which the model was optimized and used for analysis. Steps are 200 ps.

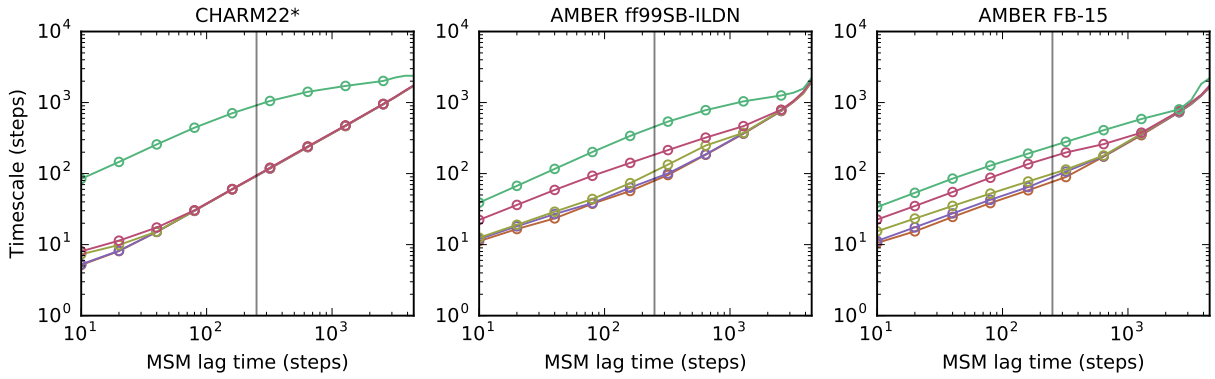

FIG. S13. Implied timescales for hydrophobic collapse. The line at 50 ns indicates the lag time at which the model was optimized and used for analysis. Steps are 200 ps.

BOOTSTRAPPED MSM OBJECTS

Bootstrapped MSMs were created using MSMBuilder 3.8[23]. These objects have been included in the SI require the MSMBuilder software and its dependencies (see `http://msmbuilder.org`). Table S5 contains filenames and information about each MSM. To load a MSM, the code required is:

```
from msmbuilder.msm import MarkovStateModel
from msmbuilder.utils import load

bmsm = load('./bootstrap_msm_baseline_charmm22star.pkl')
msm = bmsm.mle_
```

where `bmsm` is the boostrapped MSM object and `msm` is the model created from the original data. The MSM contains such attributes as state populations, transition probabilities, eigenvalues, and eigenfunctions associated with the MSM. Please see the MSMBuilder documentation for more information regarding the use of these model objects and their attributes.

TABLE S5. Filename key for bootstrapped MSM objects. All model specifications have been described in the relevant sections of the SI.

Filename	Relevant section
`bootstrap_msm_baseline_charmm22star.pkl`	Thermodynamics and kinetics of folding
	Mechanism of beta-hairpin formation
`bootstrap_msm_projection_amberfb15.pkl`	Thermodynamics and kinetics of folding
	Mechanism of beta-hairpin formation
`bootstrap_msm_projection_amberff99sbildn.pkl`	Thermodynamics and kinetics of folding
	Mechanism of beta-hairpin formation
`bootstrap_msm_hc_charmm22star.pkl`	Rate-determining process
`bootstrap_msm_hc_amberff99sbildn.pkl`	Rate-determining process
`bootstrap_msm_hc_amberfb15.pkl`	Rate-determining process
`bootstrap_msm_turn_charmm22star.pkl`	Rate-determining process
`bootstrap_msm_turn_amberff99sbildn.pkl`	Rate-determining process
`bootstrap_msm_turn_amberfb15.pkl`	Rate-determining process

REFERENCES

[1]S. Honda, T. Akiba, Y. S. Kato, Y. Sawada, M. Sekijima, M. Ishimura, A. Ooishi, H. Watanabe, T. Odahara, and K. Harata, J. Am. Chem. Soc. **130**, 15327 (2008).

[2]S. Honda, K. Yamasaki, Y. Sawada, and H. Morii, Structure **12**, 1507 (2004).

[3]K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, Science **334**, 517 (2011).

[4]S. Piana, K. Lindorff-Larsen, and D. E. Shaw, Biophys. J. **100**, L47 (2011).

[5]A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, J. Phys. Chem. B **102**, 3586 (1998).

[6]P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts, and V. S. Pande, J. Chem. Theory Comput. **9**, 461 (2013).

[7]C. R. Schwantes and V. S. Pande, J. Chem. Theory Comput. **9**, 2000 (2013).

[8]G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, J. Chem. Phys. **139**, 015102 (2013).

[9]F. Noé and C. Clementi, J. Chem. Theory Comput. **11**, 5002 (2015).

[10]K. A. Beauchamp, R. McGibbon, Y.-S. Lin, and V. S. Pande, Proc. Natl. Acad. Sci. **109**, 17807 (2012).

[11]F. Noé and F. Nüske, Multiscale Model. Simul. **11**, 635 (2013).

[12]R. T. McGibbon and V. S. Pande, J. Chem. Phys. **142**, 124105 (2015).

[13]B. E. Husic, R. T. McGibbon, M. M. Sultan, and V. S. Pande, J. Chem. Phys. **145**, 194103 (2016).

[14]J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, J. Chem. Phys. **134**, 174105 (2011).

[15]F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. **106**, 19011 (2009).

[16]G. R. Bowman, V. S. Pande, and F. Noé (Springer, 2014).

[17]B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé, J. Chem. Phys. **143**, 174101 (2015).

[18]M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann,

N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, J. Chem. Theory Comput. **11**, 5525 (2015).

[19]T. O. Street, N. C. Fitzkee, L. L. Perskie, and G. D. Rose, Protein Sci. **16**, 1720 (2007).

[20]P. Metzner, C. Schütte, and E. Vanden-Eijnden, Multiscale Model. Simul. **7**, 1192 (2009).

[21]A. Berezhkovskii, G. Hummer, and A. Szabo, J. Chem. Phys. **130**, 205102 (2009).

[22]C. M. Davis, S. Xiao, D. P. Raleigh, and R. B. Dyer, J. Am. Chem. Soc. **134**, 14476 (2012).

[23]M. P. Harrigan, M. M. Sultan, C. X. Hernández, B. E. Husic, P. Eastman, C. R. Schwantes, K. A. Beauchamp, R. T. McGibbon, and V. S. Pande, Biophys. J. **112**, 10 (2017).