Supporting Information - Hierarchical Markov State Model Building to Describe Molecular Processes

David K. Wolfe[†#], Joseph R. Persichetti[†#], Ajeet K. Sharma[†,‖], Phillip S. Hudson[⊥,▽], H. Lee Woodcock[⊥], Edward P. O'Brien[†,‡, §,*]

[†]Department of Chemistry, [‡]Bioinformatics and Genomics Graduate Program, [§]Institute for Computational and Data Sciences, Penn State University, University Park, Pennsylvania 16802, United States

[‖]Department of Physics, Indian Institute of Technology, Jammu 181221, India

[⊥]Department of Chemistry, University of South Florida, Tampa, Florida 33620, United States

[#] Authors contributed equally.

* Corresponding author: epo2@psu.edu

[▽] Present address: Laboratory of Computational Biology, National Heart, Lung, and Blood Institute (NHLBI), National Institutes of Health (NIH), Bethesda, Maryland 20892, United States

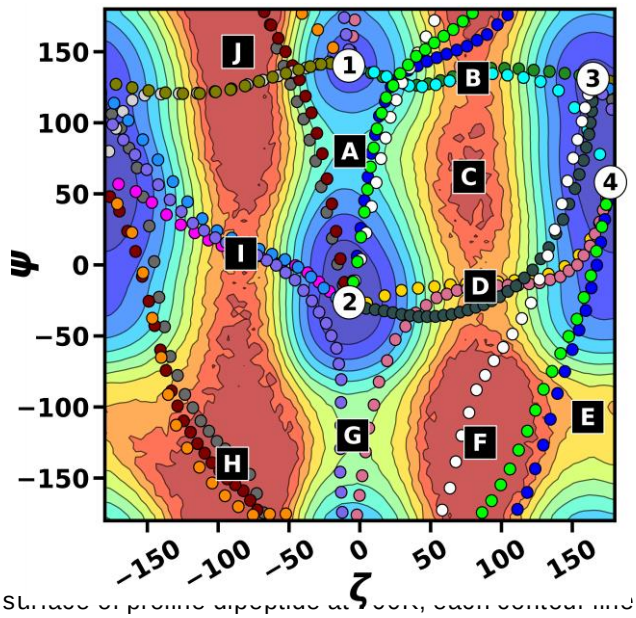


Figure S1. Free-energy surface of proline dipeptide at 300K, each contour line represents 1.5 $k_B T$. The positions of all states are numbered and all transition states considered are lettered. The pathway colors are consistent with the flux legend (Figure S3).

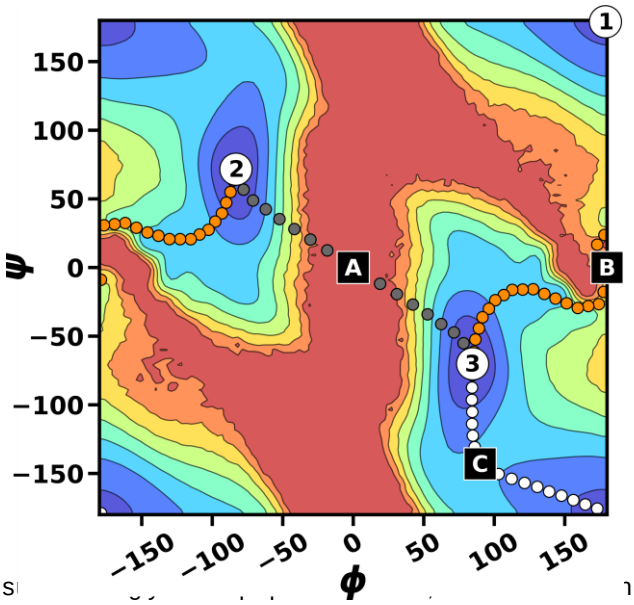


Figure S2. Free-energy s ıe represents 1 $k_B T$. The positions of all states are numbered and all transition states considered are lettered. The following pathway colors apply: 2A3: grey, 1C3: white, and 2B3: orange.

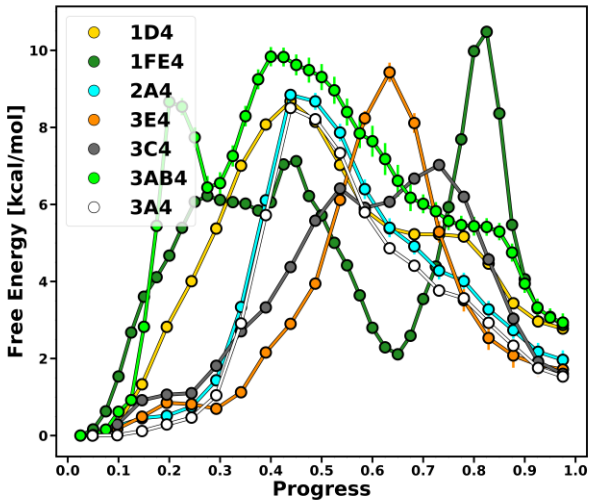


Figure S3. Free-energy profiles for all pathways for alanine dipeptide at 500K. Error bars represent the 95% confidence intervals obtained from block averaging the umbrella sampling simulations into 5 evenly sized blocks.

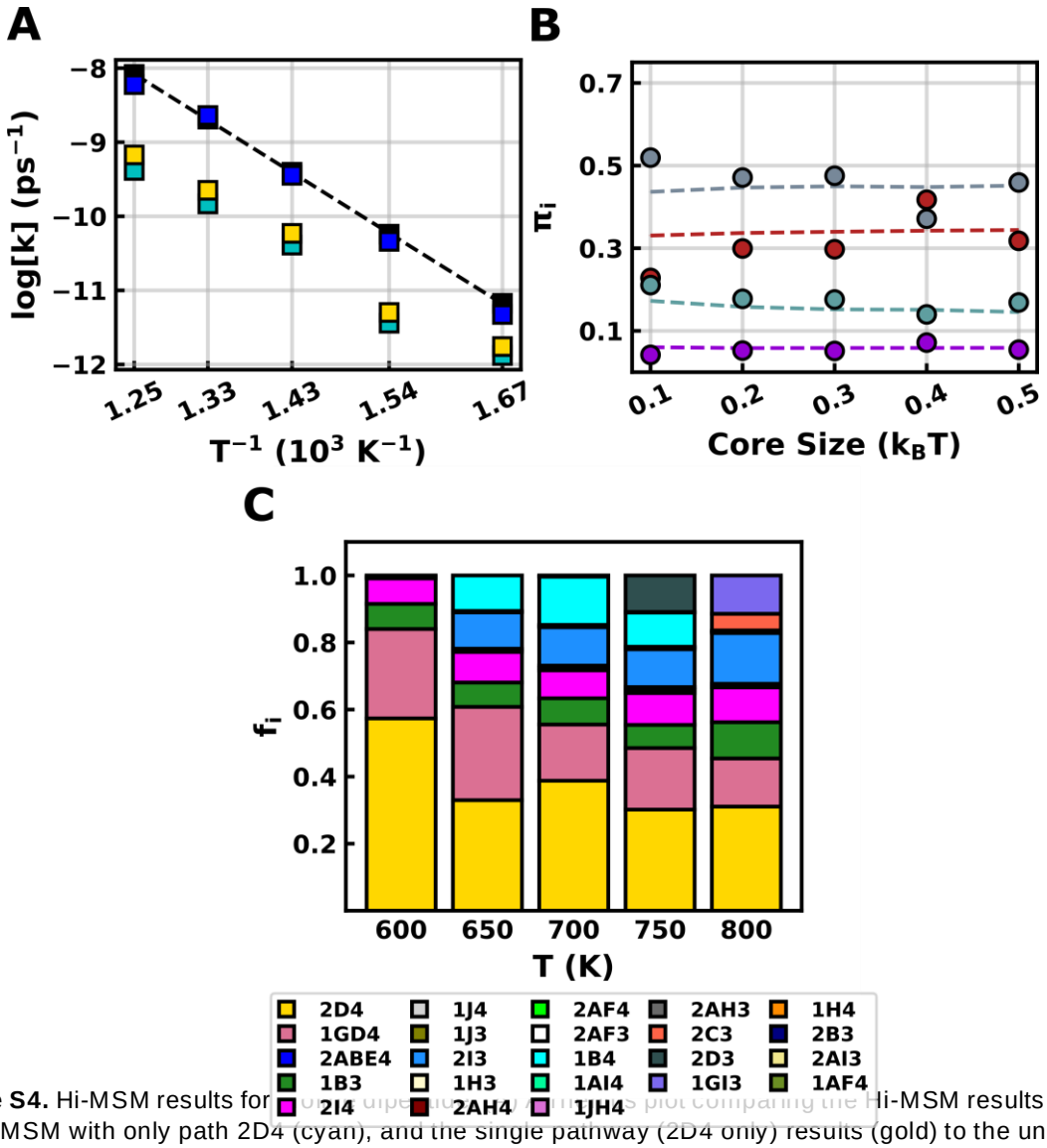


Figure S4. Hi-MSM results for [illegible] Hi-MSM results (blue), the Hi-MSM with only path 2D4 (cyan), and the single pathway (2D4 only) results (gold) to the unbiased simulations (black). The results shown are for the Hi-MSM constructed with a core size of $0.3\ k_BT$. (B) The probability of occupying states 1-4 at equilibrium calculated with the Hi-MSM (circles) compared to the simulation results (dashed lines added to guide the eye). The color scheme for states 1-4 matches those in Figure 3B. Results shown for the model constructed at 700K. (C) Fraction of the flux from $\Lambda_1 \rightarrow \Lambda_2$ from the Hi-MSM for individual pathways. The results are displayed for the models constructed at a core size of 0.3 k_BT. Due to the pruning procedure for the pathways, not all paths appearing in the legend are present at every temperature. Pathway labels correspond to the Markov and transition state labels shown in Figure S1. The error bars, which are too small to be visible, represent the 95% confidence intervals obtained from bootstrapping the underlying rate matrices.

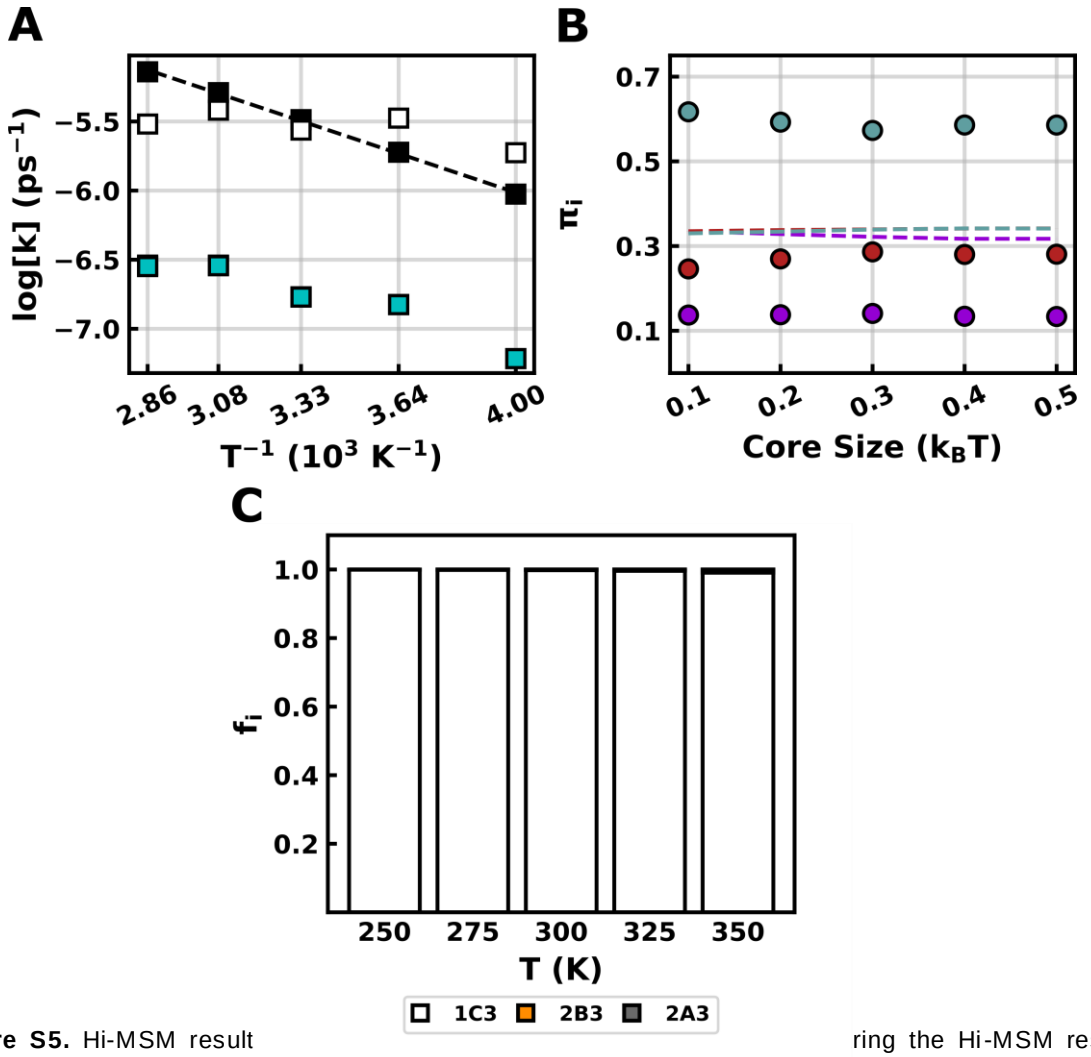


Figure S5. Hi-MSM result ring the Hi-MSM results (blue), the Hi-MSM with only path 1C3 (cyan), and the single pathway (1C3 only) results (white) to the unbiased simulations (black). The blue and cyan traces overlap. The results shown are for the Hi-MSM constructed with a core size of $0.3\,k_{\mathrm{B}}T$. (B) The probability of occupying states 1-3 at equilibrium calculated with the Hi-MSM (circles) compared to the simulation results (dashed lines added to guide the eye). The color scheme for states 1-3 matches those in Figure 3B. Results are shown for the model constructed at 300K. (C) Fraction of the flux from $\Lambda_1 \rightarrow \Lambda_2$ from the Hi-MSM for individual pathways. The results are displayed for the models constructed at 0.3 $k_{\mathrm{B}}T$. Pathway labels correspond to the Markov and transition state labels shown in Figure S2. The error bars, which are too small to be visible, represent the 95% confidence intervals obtained from bootstrapping the underlying rate matrices.

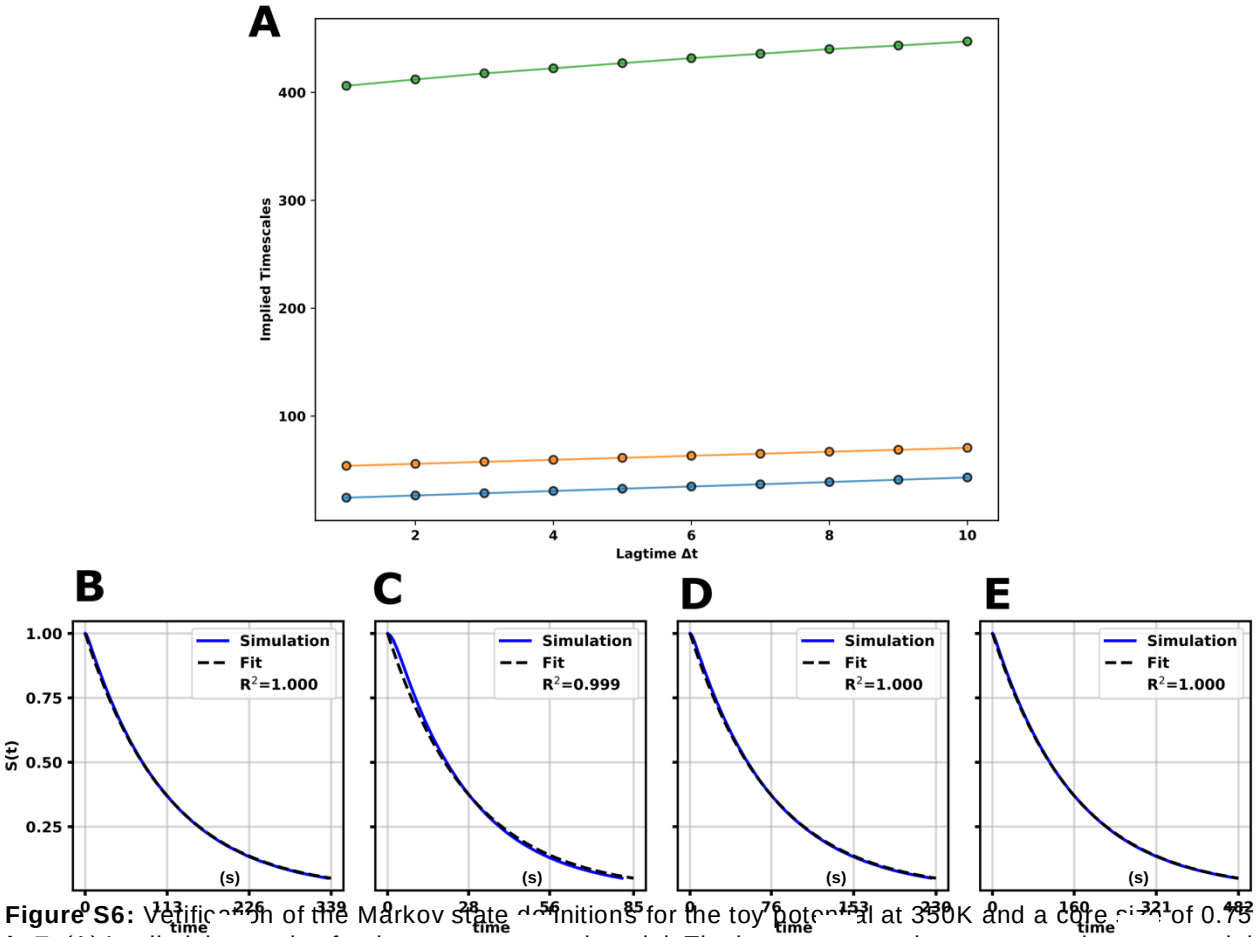


Figure S6: Verification of the Markov state definitions for the toy potential at 350K and a core size of 0.75 $k_B T$. (A) Implied timescales for the coarse-grained model. The largest timescale represents the inter-model transition in the Hi-MSM. Solid lines are added to guide the eye. (B) Survival probability in state 1 fit to a single-exponential function. (C) same as (B) for state 2. (D) state 3. (E) state 4. Note: the survival probabilities were calculated from the simulations at the printing frequency of the simulations. The excellent fit to a single exponential indicates these transitions are Markovian.

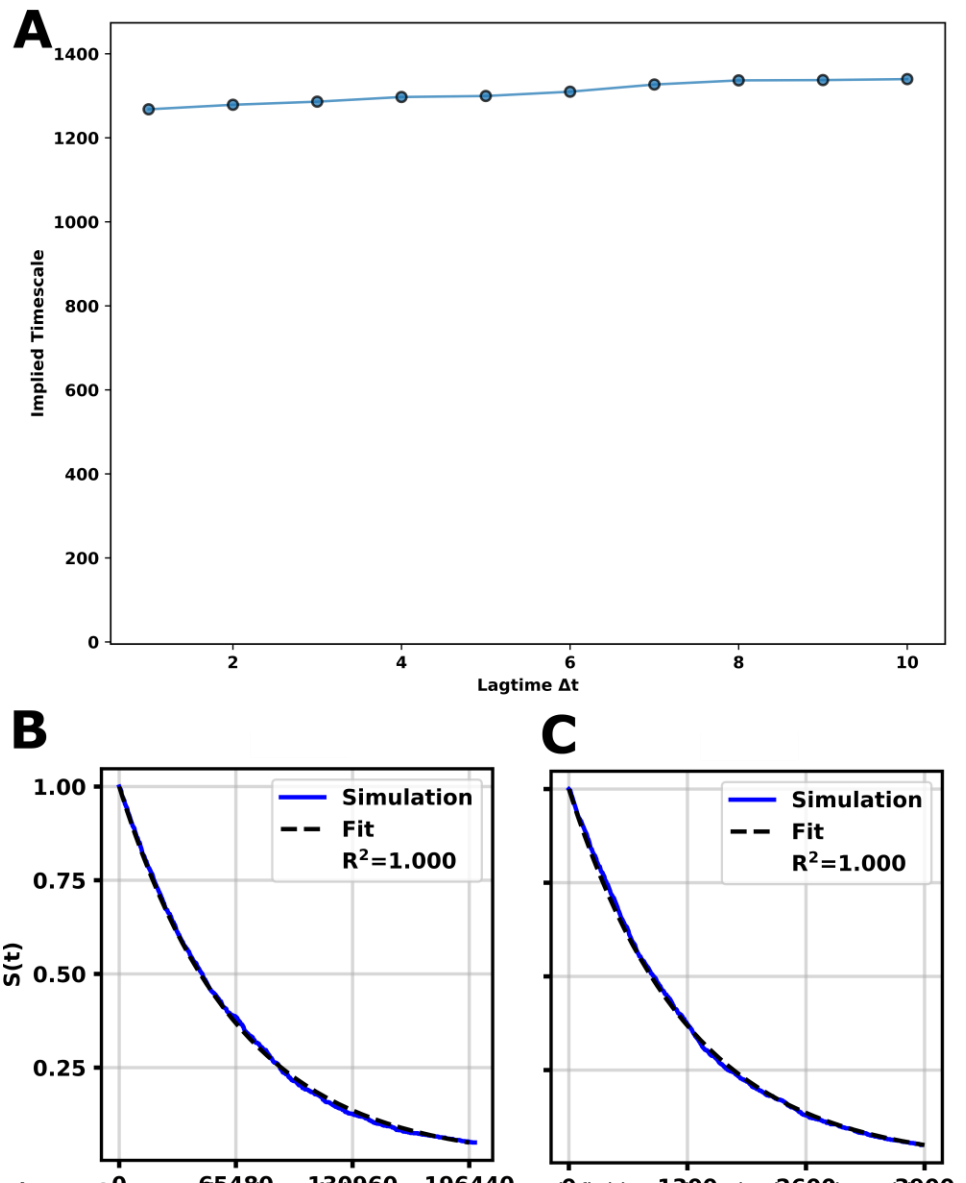


Figure S7: Verification of the Markov state definitions for alanine dipeptide at 400K and a core size of 0.5 $k_B T$. (A) Implied timescale for the coarse-grained model. This timescale represents the transition between local MSMs. (B) Survival probability in state 1 fit to a single-exponential function. (C) same as (B) for state 4. Note: the survival probability was calculated using the printing frequency of the simulations. The excellent fit to a single exponential indicates these transitions are Markovian.

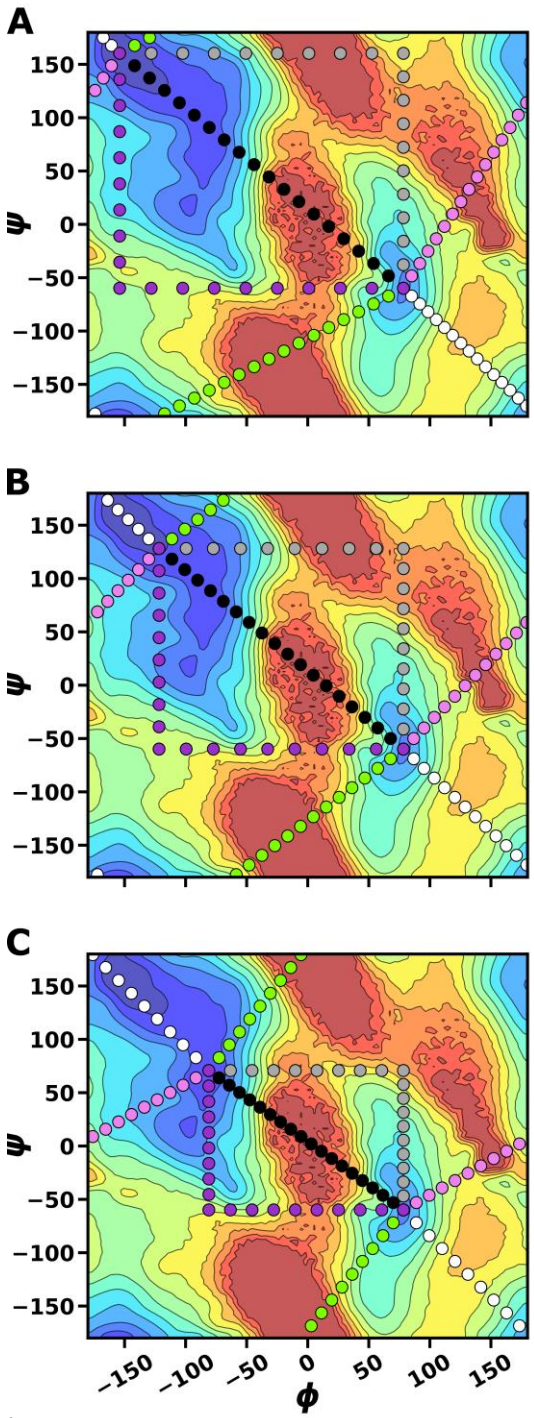
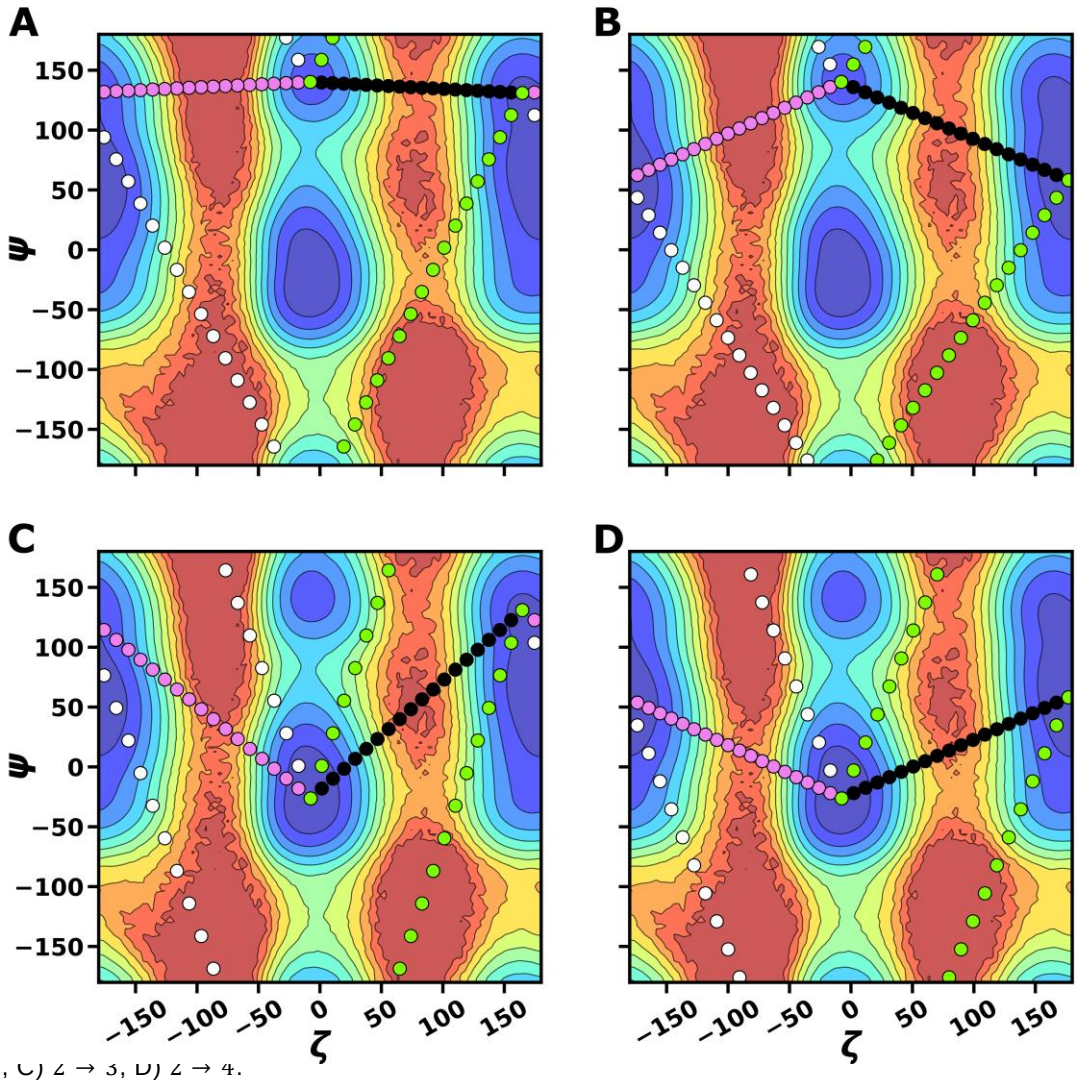


Figure S8. Initial guesses for FTSM of alanine dipeptide for the following endpoint states: A) $1 \rightarrow 4$, B) $2 \rightarrow 4$, C) $3 \rightarrow 4$.



Figur 3, B) 1 → 4, C) 2 → 3, D) 2 → 4.

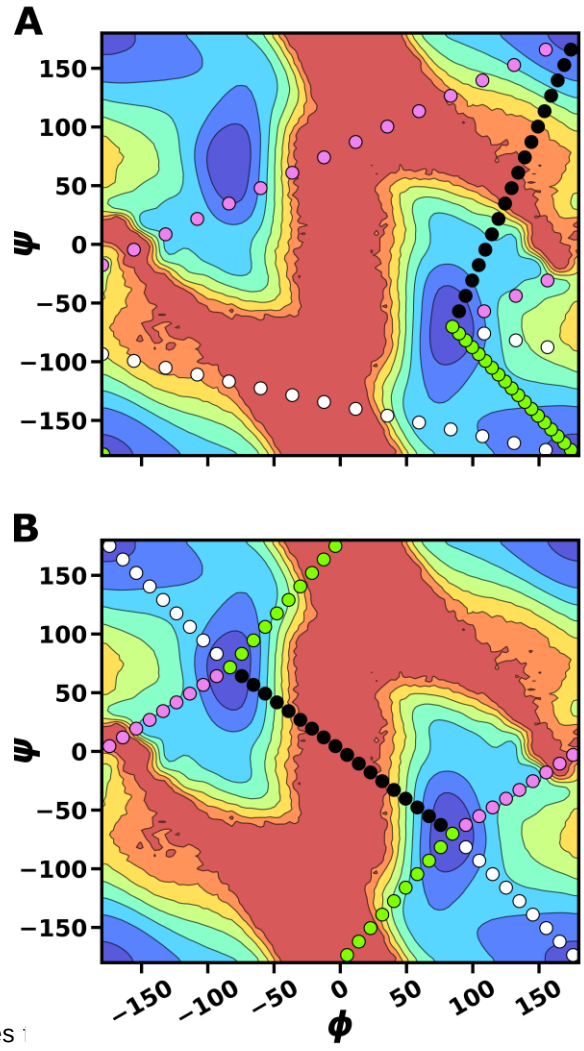


Figure S10. Initial guesses f ndpoint states: A) 1 → 3, B) 2 → 3.

Table S1. Comparison of the Average Percent Errors in Predicted Quantities using the Hi-MSM and Single-Path Methods for Proline Dipeptide[d]

	<Error k_{obs}>	<Error π_i>[a]	<Error f_{obs}>[b]	Error m[c]
Hi-MSM	9.0 ± 0.5	11.6 ± 0.8	16.5 ± 0.4	2.60
Hi-MSM 1 path	64.8 ± 0.5	60.5 ± 0.8	89.9 ± 0.6	14.8
Single Path	58.7 ± 0.5	-	86.4 ± 0.4	12.1

[a] π_i is the probability of being in state i
[b] f_{obs} is the total observed flux from $\Lambda_1 \rightarrow \Lambda_2$
[c] m is the slope of the Arrhenius plot
[d] Reported uncertainties represent the 95% confidence interval.

Table S2. Comparison of the Average Percent Errors in Predicted Quantities using the Hi-MSM and Single Path Methods for Glycine Dipeptide[d]

	<Error k_{obs}>	<Error π_i>[a]	<Error f_{obs}>[b]	Error m[c]
Hi-MSM	71.03 ± 0.05	49.25 ± 0.10	81.94 ± 0.00	24.3
Hi-MSM 1 path	71.14 ± 0.04	49.48 ± 0.10	82.06 ± 0.00	25.3
Single Path	22.73 ± 0.06	-	72.30 ± 0.00	77.0

[a] π_i is the probability of being in state i
[b] f_{obs} is the total observed flux from $\Lambda_1 \rightarrow \Lambda_2$
[c] m is the slope of the Arrhenius plot
[d] Reported uncertainties represent the 95% confidence interval.

Table S3. Percent Errors in Predicted Rates versus Core Size for Proline Dipeptide at 700K. Reported Uncertainties Represent the 95% Confidence Interval.

core size ($k_B T$)	model		
	Hi-MSM	Hi-MSM 1 path	single path
0.1	13.21 ± 0.49	67.11 ± 0.47	61.44 ± 0.48
0.2	18.92 ± 0.51	73.85 ± 0.48	69.35 ± 0.49
0.3	9.00 ± 0.50	64.78 ± 0.48	58.73 ± 0.49
0.4	21.03 ± 0.51	67.00 ± 0.47	61.29 ± 0.48
0.5	20.99 ± 0.51	67.68 ± 0.48	62.07 ± 0.48

Table S4. Percent Errors in Predicted Rates versus Core Size for Glycine Dipeptide at 300K. Reported Uncertainties Represent the 95% Confidence Interval.

core size ($k_B T$)	model		
	Hi-MSM	**Hi-MSM 1 path**	**single path**
0.1	61.53 ± 0.05	61.68 ± 0.04	21.49 ± 0.06
0.2	68.67 ± 0.05	68.78 ± 0.04	20.43 ± 0.06
0.3	71.03 ± 0.05	71.14 ± 0.04	22.73 ± 0.06
0.4	71.48 ± 0.05	71.58 ± 0.04	22.87 ± 0.06
0.5	73.01 ± 0.05	73.12 ± 0.04	15.14 ± 0.05

Table S5. Alanine Dipeptide FTSM Parameters at 400K

Temperature (K)	Cutoff ($k_B T$)	Path Label	kpar [Å]	kprp [Å]	dprp [Å]
400	0.1	1C4	7979	6839	0.1
400	0.1	1A4	9867	3524	0.12
400	0.1	1D4	7979	6839	0.1
400	0.1	1E4	9867	3524	0.12
400	0.2	1A4	9867	3524	0.12
400	0.2	1FE4	12500	5000	0.1
400	0.2	1C4	7979	6839	0.1
400	0.2	1E4	7979	6839	0.1
400	0.2	1D4	7979	6839	0.1
400	0.3	1E4	7979	6839	0.1
400	0.3	1C4	7979	6839	0.1
400	0.3	1FE4	12500	5000	0.1
400	0.3	1D4	7979	6839	0.1
400	0.3	1A4	9867	3524	0.12
400	0.4	1D4	7979	6839	0.1
400	0.4	1C4	7979	6839	0.1
400	0.4	1E4	7979	6839	0.1
400	0.4	1FE4	12500	5000	0.1
400	0.4	1A4	9867	3524	0.12

Table S6. Alanine Dipeptide FTSM Parameters at 450K

Temperature (K)	Cutoff (k_BT)	Path Label	kpar [Å]	kprp [Å]	dprp [Å]
450	0.1	1FE4	12500	5000	0.1
450	0.1	1A4	9867	5000	0.12
450	0.1	3C4	4252	3307	0.18
450	0.1	3A4	8128	4000	0.1
450	0.1	1D4	10000	6839	0.1
450	0.1	3AB4	15000	10000	0.1
450	0.2	3C4	4252	3307	0.18
450	0.2	3AB4	15000	10000	0.1
450	0.2	1FE4	12500	5000	0.1
450	0.2	3E4	9867	5000	0.12
450	0.2	1D4	10000	6839	0.1
450	0.2	3A4	8128	4000	0.1
450	0.2	1A4	9867	5000	0.12
450	0.3	1A4	9867	5000	0.12
450	0.3	3E4	8128	4000	0.1
450	0.3	3AB4	15000	10000	0.1
450	0.3	1FE4	12500	5000	0.1
450	0.3	3C4	4252	3307	0.18
450	0.3	3A4	8128	4000	0.1
450	0.3	1D4	10000	6839	0.1
450	0.4	1FE4	12500	5000	0.1
450	0.4	3C4	4252	3307	0.18
450	0.4	3AB4	15000	10000	0.1
450	0.4	1D4	10000	6839	0.1
450	0.4	3E4	9867	5000	0.12
450	0.4	1A4	9867	5000	0.12
450	0.4	3A4	8128	4000	0.1

Table S7. Alanine Dipeptide FTSM Parameters at 500K

Temperature (K)	Cutoff (k_BT)	Path Label	kpar [Å]	kprp [Å]	dprp [Å]
500	0.1	3C4	4252	3307	0.18
500	0.1	1D4	10000	6839	0.1
500	0.1	3AB4	12000	5000	0.12
500	0.1	3E4	8128	4000	0.12
500	0.1	2A4	8128	4000	0.12
500	0.1	1FE4	18000	5000	0.13
500	0.2	1FE4	18000	5000	0.13
500	0.2	1D4	10000	6839	0.1
500	0.2	3E4	8128	4000	0.12
500	0.2	3C4	4252	3307	0.18
500	0.2	3A4	8128	4000	0.12
500	0.2	2A4	8128	4000	0.12
500	0.2	3AB4	12000	5000	0.12
500	0.3	1FE4	18000	5000	0.13
500	0.3	3AB4	12000	5000	0.12
500	0.3	1D4	10000	6839	0.1
500	0.3	3C4	4252	3307	0.18
500	0.3	2A4	8128	4000	0.12
500	0.3	3E4	8128	4000	0.12
500	0.3	3A4	8128	4000	0.12
500	0.4	3AB4	12000	5000	0.12
500	0.4	3E4	8128	4000	0.12
500	0.4	1FE4	18000	5000	0.13
500	0.4	3A4	8128	4000	0.12
500	0.4	3C4	4252	3307	0.18
500	0.4	2A4	8128	4000	0.12
500	0.4	1D4	10000	6839	0.1

Table S8. Alanine Dipeptide FTSM Parameters at 550K

Temperature (K)	Cutoff (k_BT)	Path Label	kpar [Å]	kprp [Å]	dprp [Å]
550	0.1	3E4	8128	4000	0.12
550	0.1	3AB4	30000	10000	0.1
550	0.1	3C4	8000	4000	0.12
550	0.1	3A4	8128	4000	0.1
550	0.1	1D4	10000	6839	0.1
550	0.1	1FE4	10000	6839	0.1
550	0.2	3C4	8000	4000	0.12
550	0.2	3AB4	30000	10000	0.1
550	0.2	3A4	8128	4000	0.1
550	0.2	1D4	10000	6839	0.1
550	0.2	3E4	8000	4000	0.12
550	0.2	2A4	8000	5000	0.12
550	0.2	1FE4	10000	6839	0.1
550	0.3	3E4	8128	4000	0.12
550	0.3	1D4	10000	6839	0.1
550	0.3	3A4	8128	4000	0.1
550	0.3	2A4	8000	5000	0.12
550	0.3	3AB4	30000	10000	0.1
550	0.3	3C4	8000	4000	0.12
550	0.4	3C4	8000	4000	0.12
550	0.4	2A4	8000	5000	0.12
550	0.4	1D4	10000	6839	0.1
550	0.4	3E4	8128	4000	0.12

Table S9. Alanine Dipeptide FTSM Parameters at 600K

Temperature (K)	Cutoff (k_BT)	Path Label	kpar [Å]	kprp [Å]	dprp [Å]
600	0.1	3E4	7935	4534	0.1
600	0.1	3C4	10000	5000	0.1
600	0.1	2A4	8000	4000	0.1
600	0.1	1D4	7979	6839	0.1
600	0.1	3A4	7935	4534	0.1
600	0.2	3C4	10000	5000	0.1
600	0.2	1D4	7979	6839	0.1
600	0.2	3E4	7935	4534	0.1
600	0.2	2A4	8000	4000	0.1
600	0.3	1D4	7979	6839	0.1
600	0.3	3C4	10000	5000	0.1
600	0.3	2A4	8000	4000	0.1
600	0.3	3E4	7935	4534	0.1
600	0.4	2A4	8000	4000	0.1
600	0.4	3C4	10000	5000	0.1
600	0.4	3E4	7935	4534	0.1
600	0.4	1D4	7979	6839	0.1