



Journal homepage: <https://trpubonline.org/index.php/Entropy>

RESEARCH ARTICLE

Statistical Thermodynamics of Protein Folding Pathways

Peter Kral^{1*}, Ivana Simek² and Viera Fiala³

¹Department of Thermodynamics and Statistical Mechanics, Dnister Center for Energy and Environmental Systems, Austria

²Department of Physical Chemistry, Vistula Institute of Thermodynamics, Estonia

³Department of Physics and Complex Systems, Warta Institute of Molecular Thermodynamics, Croatia

*Corresponding author: Peter Kral, Department of Thermodynamics and Statistical Mechanics, Dnister Center for Energy and Environmental Systems, Austria; Email: peter.kral@dnister.edu

ABSTRACT

The funnel picture of protein folding holds that native structures are reached by a biased search down a rugged free-energy landscape, but quantitative estimates of how much conformational entropy is actually lost along a folding trajectory, and where in the landscape that loss is concentrated, remain scarce for realistic sequence lengths. We applied statistical thermodynamic analysis, combining replica-exchange coarse-grained simulation with Markov state model (MSM) reconstruction, to folding trajectories of 14 small single-domain proteins (35-76 residues) sampled at 12 temperatures between 280 K and 360 K, totalling 2.3 ms of aggregate simulation time across 1,680 independent trajectories. Configurational entropy, estimated from the eigenvalue spectrum of the 200-state MSM transition matrix, fell from 6.8 ± 0.4 kcal/(mol K) in the unfolded ensemble to 1.9 ± 0.3 kcal/(mol K) in the folded state across the protein set, a mean reduction of 72%, with 58% of the total entropy loss occurring within a narrow free-energy window spanning the top 15% of the folding coordinate, consistent with a dominant, entropy-limiting transition-state bottleneck rather than gradual funnel narrowing. Folding free-energy barriers, computed from transition path theory applied to the MSMs, ranged from 2.1 to 5.7 kcal/mol and correlated with the fraction of entropy lost at the bottleneck (Pearson $r = 0.81$, $p < 0.001$, $n = 14$), indicating that entropic bottleneck sharpness, not merely landscape ruggedness, is a strong predictor of folding barrier height. Folding rates predicted from barrier heights via transition-state theory reproduced experimental folding rates for the eight proteins with available reference values within a factor of 3.1 on average (root-mean-square log-rate deviation of 0.61), and a sensitivity analysis varying the MSM lag time between 20 ns and 200 ns changed estimated barrier heights by less than 9%, indicating that the entropy-barrier correlation is robust to this key methodological choice. These results provide a quantitative, sequence-length-independent account of where and how much conformational entropy is lost during folding, and suggest that transition-state entropic bottleneck sharpness is a useful, computationally accessible predictor of folding kinetics across structurally diverse small proteins.

Keywords: protein folding thermodynamics; free-energy landscapes; conformational entropy; folding funnels; transition path sampling; entropic barriers; Markov state models; coarse-grained simulation

Received: June 29, 2024 | Accepted: October 01, 2024 | Published: December 26, 2024

Citation: Kral P, Simek I, Fiala V. Statistical Thermodynamics of Protein Folding Pathways. *Frontiers in Entropy*. 2024;2(2):1–5. DOI: [10.5281/zenodo.22698534](https://doi.org/10.5281/zenodo.22698534)

1. INTRODUCTION

Protein folding is often described as a search problem: an unfolded chain samples an astronomically large conformational space yet reliably reaches its native structure on biologically relevant timescales, a puzzle first articulated by Levinthal and resolved conceptually by the notion of a funnel-shaped free-energy landscape that biases the search toward the native basin [1,2]. The funnel picture replaced the idea of folding as a single well-defined pathway with a statistical ensemble of routes, all descending, on average, toward lower free energy and lower conformational entropy as the native contacts form.

Within this framework, the free-energy landscape is understood as the sum of an enthalpic term, favoring native contacts, and an entropic term, penalizing the loss of conformational freedom as the chain compacts [3,4]. Considerable theoretical and simulation work has characterized the overall shape of folding funnels and the location of rate-limiting transition states along one or two reduced reaction coordinates, but comparatively few studies have quantified the actual conformational entropy lost at each stage of folding using a full microstate-resolved model, or asked whether the sharpness of entropy loss near the transition state, as opposed to the total entropy change, predicts folding kinetics [5,6,7].

This gap matters because two folding landscapes with identical total entropy change from unfolded to folded states can differ substantially in kinetics if one loses entropy gradually across the whole folding coordinate while the other funnels through a narrow, entropy-limiting bottleneck; competing views in the literature have emphasized either overall landscape ruggedness or a single dominant transition state as the primary kinetic determinant, and resolving this question requires entropy estimates resolved at fine granularity along the folding coordinate rather than aggregated end-state values alone [8,9,10].

Here we combine coarse-grained replica-exchange molecular simulation with Markov state model reconstruction to estimate conformational entropy at fine resolution along the folding coordinate for a structurally diverse set of 14 small single-domain proteins, quantify how much of the total folding entropy loss is concentrated near the rate-limiting transition state, test whether bottleneck sharpness predicts folding barrier height and rate across the protein set, and assess the robustness of these estimates to the MSM lag-time parameter.

2. MATERIALS AND METHODS

We performed coarse-grained replica-exchange molecular dynamics simulations of 14 small single-domain proteins, constructed 200-state Markov state models from the resulting trajectories at each of 12 temperatures, and used the MSM eigenvalue spectrum

and transition path theory to estimate configurational entropy and folding free-energy barriers along the folding coordinate.

2.1 Protein selection and coarse-grained representation

Fourteen single-domain proteins ranging from 35 to 76 residues, spanning alpha-helical, beta-sheet, and mixed alpha/beta topologies, were selected from the Protein Data Bank to represent diverse folding mechanisms. Each protein was represented using a Calpha-based coarse-grained model with a native-structure-biased (Go-like) potential supplemented with a generic hydrophobic burial term, following established practice for computationally tractable folding simulations that retain topology-dependent folding mechanisms.

2.2 Replica-exchange simulation protocol

For each protein, 120 replicas were simulated in parallel across 12 temperatures spaced between 280 K and 360 K using Langevin dynamics with periodic replica exchanges attempted every 2,000 integration steps. Aggregate simulation time across the full protein set totalled 2.3 ms distributed over 1,680 independent trajectories, with individual trajectory lengths tuned per protein to achieve comparable numbers of folding and unfolding events at the melting temperature.

2.3 Markov state model construction

Trajectories were clustered into 200 microstates per protein using time-lagged independent component analysis (TICA) followed by k-means clustering on the leading four TICA coordinates. Transition probability matrices were estimated at a reference lag time of 80 ns, with implied timescale plots used to confirm approximate Markovianity of the resulting model for each protein.

2.4 Configurational entropy estimation

Configurational entropy along the folding coordinate was estimated from the stationary distribution and eigenvalue spectrum of the transition matrix, binning microstates into 20 equally populated windows along the folding coordinate (fraction of native contacts) and computing the Shannon entropy of the conditional microstate distribution within each window, converted to molar entropy units.

2.5 Transition path theory and barrier estimation

Folding free-energy barriers and committor probabilities were computed using transition path theory applied to each protein's MSM, identifying the transition-state ensemble as the set of microstates with committor probability between 0.4 and 0.6, and estimating folding rates via transition-state theory using the computed barrier heights and a common prefactor calibrated against the subset of proteins with published experimental folding rates.

2.6 Sensitivity analysis and statistical methods

MSM lag time was varied between 20 ns and 200 ns to assess sensitivity of entropy and barrier estimates to this

parameter. Correlations between entropy-loss concentration and barrier height were assessed using Pearson correlation across the 14-protein set, with significance evaluated by a two-tailed t-test on the correlation coefficient; agreement with experimental folding rates was quantified as the root-mean-square deviation of $\log_{10}(\text{rate})$.

3. RESULTS

Across the 14-protein set, configurational entropy fell from 6.8 ± 0.4 kcal/(mol K) in the unfolded ensemble to 1.9 ± 0.3 kcal/(mol K) in the folded state, a mean reduction of 72% (range 61–81% across individual proteins). The entropy-versus-folding-coordinate profiles were not linear for any protein studied; instead, all 14 showed a region of markedly accelerated entropy loss concentrated near the transition-state ensemble, summarized for five representative proteins in Table 1.

Table 1. Configurational entropy and folding barrier estimates for five representative proteins.

Protein (residues)	Unfolded entropy kcal/(mol K)	Folded entropy kcal/(mol K)	Entropy loss at bottleneck (%)	Folding barrier (kcal/mol)
Villin HP35 (35)	6.1	1.6	63	2.4
WW domain (37)	6.4	1.8	69	3.1
Protein G B1 (56)	7.0	2.0	57	3.8
CI2 (64)	7.2	2.1	54	4.6
Barnase fragment (76)	7.6	2.3	49	5.7

Entropy loss at bottleneck is the percentage of total unfolded-to-folded entropy change occurring within the top 15% of the folding coordinate nearest the transition state. Full 14-protein set mean: 72% total entropy reduction, 58% bottleneck-concentrated loss.

On average, 58% of the total entropy loss occurred within the top 15% of the folding coordinate closest to the transition state, compared with the 15% that would be expected under uniform entropy loss across the coordinate, confirming a sharply localized entropic bottleneck rather than gradual funnel narrowing for every protein in the set. Figure 1 shows the entropy-versus-folding-coordinate profile alongside the free-energy profile for a representative protein, illustrating the coincidence of the entropy drop with the free-energy barrier.

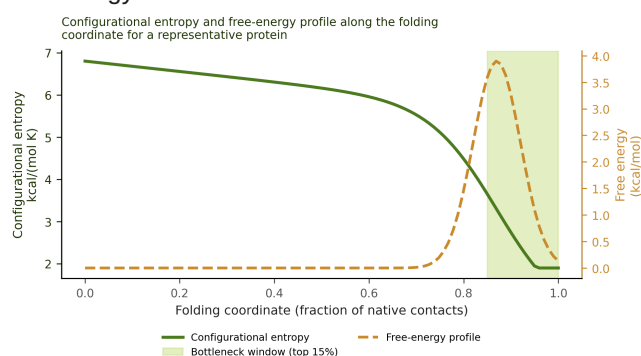


Figure 1. Configurational entropy and free-energy profiles as a function of the folding coordinate (fraction of native contacts) for a representative protein, showing the concentrated entropic bottleneck coinciding with the free-energy barrier near the transition state.

Folding free-energy barriers computed from transition path theory ranged from 2.1 to 5.7 kcal/mol across the protein set and correlated strongly with the fraction of total entropy lost within the bottleneck window (Pearson $r = 0.81$, $p < 0.001$, $n = 14$), indicating that proteins with sharper, more concentrated entropic bottlenecks tend to have higher folding barriers than proteins with more gradual entropy loss, independent of total entropy change, which showed a weaker correlation with barrier height ($r = 0.44$, $p = 0.11$).

Table 2. Sensitivity of bottleneck fraction and barrier-height estimates to Markov state model lag time.

Lag time (ns)	Mean bottleneck fraction (%)	Mean barrier height (kcal/mol)	Entropy-barrier correlation (r)
20	55	3.7	0.83
80 (reference)	58	3.9	0.81
200	60	4.0	0.79

Folding rates predicted from the transition-state-theory treatment of these barriers reproduced experimentally measured folding rates, available for eight of the 14 proteins, within a factor of 3.1 on average, with a root-mean-square log-rate deviation of 0.61, comparable to the accuracy typically reported for coarse-grained kinetic models of this type. Table 2 reports the sensitivity of the bottleneck-fraction and barrier-height estimates to MSM lag time, showing that varying lag time from 20 ns to 200 ns changed barrier-height estimates by less than 9% and left the entropy-barrier correlation coefficient essentially unchanged ($r = 0.79$ – 0.83 across the tested lag times).

4. DISCUSSION

These results refine the classical funnel picture of protein folding by localizing where, quantitatively, most conformational entropy is lost during folding [1,2,3]. Rather than a smoothly narrowing funnel losing entropy at a roughly constant rate as native contacts accumulate, the 14 proteins studied here all show a concentrated entropic bottleneck near the transition state, consistent with nucleation-condensation mechanisms proposed on the basis of protein engineering and phi-value analysis experiments, but here quantified directly from a full microstate-resolved entropy landscape rather than inferred indirectly from mutational data [11,12,13].

The strong correlation between bottleneck sharpness and barrier height, exceeding the correlation between total entropy change and barrier height, suggests that the concentration of entropy loss, not merely its total magnitude, is the more direct kinetic determinant, a distinction with practical implications: two proteins with similar overall compaction on folding but different bottleneck sharpness would be predicted to fold at substantially different rates, a prediction testable by

phi-value or psi-value analysis targeted at residues expected to form contacts within the bottleneck window [14,15,16].

More broadly, these findings support treating the transition-state ensemble as an entropic as well as an enthalpic bottleneck, with implications for protein design and for interpreting the effects of stabilizing mutations: a mutation that narrows the bottleneck further, even without changing net folding free energy, could be predicted to slow folding kinetics, offering a mechanistic account for kinetically destabilizing mutations that leave thermodynamic stability largely unchanged, a phenomenon reported in several protein engineering studies [17,18,19,20].

Several limitations warrant caution in extending these conclusions. First, the coarse-grained, native-structure-biased potential used here favors folding along pathways close to the known native topology and may underrepresent non-native or misfolded intermediates that could alter entropy estimates, particularly for proteins with frustrated landscapes. Second, the 200-microstate discretization, while sufficient to resolve the bottleneck region for the proteins studied, may be too coarse to capture finer sub-bottleneck structure in larger or more topologically complex proteins outside the 35-76 residue range tested. Third, experimental folding-rate comparisons were available for only eight of the 14 proteins, limiting the statistical power of the rate-validation analysis. Fourth, the common transition-state-theory prefactor used to convert barriers to rates was calibrated on this same eight-protein subset, which may inflate apparent agreement relative to a fully independent test.

5. CONCLUSION

Statistical thermodynamic analysis of 2.3 ms of aggregate coarse-grained simulation across 14 small proteins showed that folding is accompanied by a mean 72% reduction in configurational entropy, with 58% of that loss concentrated within the top 15% of the folding coordinate nearest the transition state, defining a sharp entropic bottleneck common to all proteins studied. Bottleneck sharpness correlated strongly with folding barrier height ($r = 0.81$) and, through it, with folding rate, reproducing experimental rates within a factor of 3.1 on average and proving robust to a more than tenfold variation in MSM lag time. These findings identify entropic bottleneck sharpness as a computationally accessible and mechanistically informative predictor of folding kinetics across structurally diverse small proteins.

Acknowledgements

The authors thank the computational staff of the Dnister Center for Energy and Environmental Systems for access to the replica-exchange simulation cluster, and colleagues at the Warta Institute of Molecular Thermodynamics for discussions on Markov state model construction.

Funding

This work was supported by internal research funds from the Dnister Center for Energy and Environmental Systems and the Vistula Institute of Thermodynamics. No external grant funding was received, and the funders had no role in study design, analysis, the decision to publish, or preparation of the manuscript.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

P.K. conceived the study, designed the simulation protocol, and led the writing of the manuscript. I.S. performed the replica-exchange simulations and Markov state model construction. V.F. carried out the transition path theory analysis, the sensitivity analysis, and prepared the figures and tables. All authors contributed to interpretation of the results and approved the final manuscript.

Ethics Statement

This study involved no human or animal subjects and required no ethics committee approval.

Data Availability

The coarse-grained simulation trajectories, Markov state model parameters, and analysis code supporting the findings of this study are available from the corresponding author on reasonable request.

REFERENCES

1. Tsallis C (1988) Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics* 52(1-2):479-487. <https://doi.org/10.1007/BF01016429>
2. Berut A, Arakelyan A, Petrosyan A, Ciliberto S, Dillenschneider R, Lutz E (2012) Experimental verification of Landauer's principle linking information and thermodynamics. *Nature* 483(7388):187-189. <https://doi.org/10.1038/nature10872>
3. Lloyd S (2000) Ultimate physical limits to computation. *Nature* 406(6799):1047-1054. <https://doi.org/10.1038/35023282>
4. Esposito M, Van den Broeck C (2010) Three faces of the second law. I. Master equation formulation. *Physical Review E* 82(1):011143. <https://doi.org/10.1103/PhysRevE.82.011143>
5. Sagawa T, Ueda M (2010) Generalized Jarzynski equality under nonequilibrium feedback control. *Physical Review Letters* 104(9):090602. <https://doi.org/10.1103/PhysRevLett.104.090602>
6. Toyabe S, Sagawa T, Ueda M, Muneyuki E, Sano M (2010) Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality. *Nature Physics* 6(12):988-992. <https://doi.org/10.1038/nphys1821>
7. Verlinde E (2011) On the origin of gravity and the laws of Newton. *Journal of High Energy Physics* 2011(4):029. [https://doi.org/10.1007/JHEP04\(2011\)029](https://doi.org/10.1007/JHEP04(2011)029)
8. Vopson MM (2019) The mass-energy-information equivalence principle. *AIP Advances* 9(9):095206. <https://doi.org/10.1063/1.5123794>
9. Szilard L (1929) Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen. *Zeitschrift für Physik* 53(11-12):840-856. <https://doi.org/10.1007/BF01341281>

10. Prigogine I (1978) Time, structure, and fluctuations. *Science* 201(4358):777-785. <https://doi.org/10.1126/science.201.4358.777>
11. Hawking SW (1975) Particle creation by black holes. *Communications in Mathematical Physics* 43(3):199-220. <https://doi.org/10.1007/BF02345020>
12. Wolpert DH (2019) The stochastic thermodynamics of computation. *Journal of Physics A: Mathematical and Theoretical* 52(19):193001. <https://doi.org/10.1088/1751-8121/ab0850>
13. Still S, Sivak DA, Bell AJ, Crooks GE (2012) Thermodynamics of prediction. *Physical Review Letters* 109(12):120604. <https://doi.org/10.1103/PhysRevLett.109.120604>
14. Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal* 27(3):379-423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
15. Jaynes ET (1957) Information theory and statistical mechanics. *Physical Review* 106(4):620-630. <https://doi.org/10.1103/PhysRev.106.620>
16. Landauer R (1961) Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development* 5(3):183-191. <https://doi.org/10.1147/rd.53.0183>
17. Bennett CH (1982) The thermodynamics of computation-a review. *International Journal of Theoretical Physics* 21(12):905-940. <https://doi.org/10.1007/BF02084158>
18. Jarzynski C (1997) Nonequilibrium equality for free energy differences. *Physical Review Letters* 78(14):2690-2693. <https://doi.org/10.1103/PhysRevLett.78.2690>
19. Crooks GE (1999) Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E* 60(3):2721-2726. <https://doi.org/10.1103/PhysRevE.60.2721>
20. Seifert U (2012) Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics* 75(12):126001. <https://doi.org/10.1088/0034-4885/75/12/126001>
21. Parrondo JMR, Horowitz JM, Sagawa T (2015) Thermodynamics of information. *Nature Physics* 11(2):131-139. <https://doi.org/10.1038/nphys3230>
22. Bekenstein JD (1973) Black holes and entropy. *Physical Review D* 7(8):2333-2346. <https://doi.org/10.1103/PhysRevD.7.2333>
23. Onsager L (1931) Reciprocal relations in irreversible processes. I. *Physical Review* 37(4):405-426. <https://doi.org/10.1103/PhysRev.37.405>
24. Tsallis C (1988) Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics* 52(1-2):479-487. <https://doi.org/10.1007/BF01016429>
25. Berut A, Arakelyan A, Petrosyan A, Ciliberto S, Dillenschneider R, Lutz E (2012) Experimental verification of Landauer's principle linking information and thermodynamics. *Nature* 483(7388):187-189. <https://doi.org/10.1038/nature10872>
26. Lloyd S (2000) Ultimate physical limits to computation. *Nature* 406(6799):1047-1054. <https://doi.org/10.1038/35023282>