HHS Public Access

Author manuscript

J Am Chem Soc. Author manuscript; available in PMC 2020 April 24.

Published in final edited form as:

J Am Chem Soc. 2019 April 24; 141(16): 6519–6526. doi:10.1021/jacs.8b10735.

Computational estimation of ms-sec atomistic folding times

Upendra Adhikari[†,1], **Barmak Mostofian**[†,1], **Jeremy Copperman**[†], **Sundar Raman Subramanian**[∞], **Andrew A. Petersen**[‡], **Daniel M. Zuckerman**[†,*]

[†]Department of Biomedical Engineering, School of Medicine, Oregon Health & Science University, Portland, OR 97239

[∞]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15260

[‡]NCSU Data Science Resources, North Carolina State University, Raleigh, NC 27695

Abstract

Despite the development of massively parallel computing hardware including inexpensive graphics processing units (GPUs), it has remained infeasible to simulate the folding of atomistic proteins at room temperature using conventional molecular dynamics (MD) beyond the μs scale. Here we report the folding of atomistic, implicitly solvated protein systems with folding times τ_f ranging from ~10 μs to ~100 ms using the weighted ensemble (WE) strategy in combination with GPU computing. Starting from an initial structure or set of structures, WE organizes an ensemble of GPU-accelerated MD trajectory segments via intermittent pruning and replication events to generate statistically unbiased estimates of rate constants for rare events such as folding; no biasing forces are used. Although the variance among atomistic WE folding runs is significant, multiple independent runs are used to reduce and quantify statistical uncertainty. Folding times are estimated directly from WE probability flux and from history-augmented Markov analysis of the WE data. Three systems were examined: NTL9 at low solvent viscosity (yielding $\tau_f = 0.8 - \mu s$), NTL9 at water-like viscosity ($\tau_f = 0.2 - 2$ ms), and Protein G at low viscosity ($\tau_f = 3 - 200$ ms). In all cases the folding time, uncertainty, and ensemble properties could be estimated from WE simulation; for Protein G, this characterization required significantly less overall computing than would be required to observe a *single folding event* with conventional MD simulations. Our results suggest that the use and calibration of force fields and solvent models for precise estimation of *kinetic* quantities is becoming feasible.

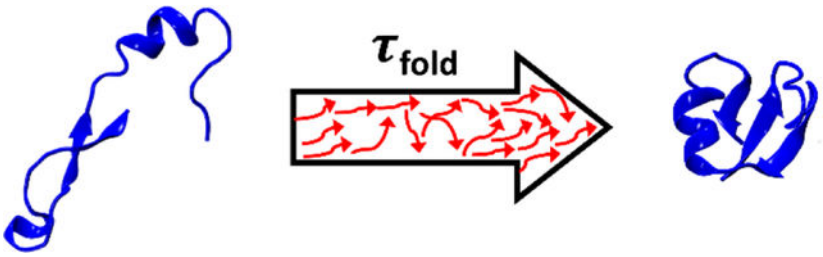
Graphical Abstract

* zuckermd@ohsu.edu.

1. Equal contributions

Supporting Information

Detailed description of all simulation parameters, setup, and analysis, all WE rate constants, flux profiles, and event durations, comparisons to high-temperature brute-force simulations.



Introduction

Elucidating the kinetics and mechanisms of protein folding has been a decades-long focus of molecular biophysics, both experimental and theoretical/computational.[1–19] Significant challenges remain, however, notably whether molecular dynamics (MD) simulations will provide the hoped-for reproducible and atomically detailed folding trajectories.[1, 11, 13–14, 20–23] Despite isolated reports of success,[24–25] MD simulations generally have not produced *room temperature* atomistic folding trajectories beyond the μs timescale even with modern hardware.[26] Promising results have been reported using path-sampling techniques[27–31] but no simulation methodology has emerged as a general-purpose tool for folding, especially for timescales beyond the μs range.

Here we report substantial progress in the application of the weighted ensemble (WE) path sampling method[32–36] to room-temperature folding at the microsecond (μs), millisecond (ms) and second (s) scales, exploiting the power of GPU and cluster computing. We study three atomistic implicitly solvated systems: NTL9 with low and high-friction solvent, as well as Protein G at low friction. These are costly studies, requiring aggregate trajectory totals of 10s to 100s of μs per system, but they enable fairly precise (order-of-magnitude) estimation of folding rate constants. In earlier work, Ensign and Pande[26] were able to estimate the WW-domain folding time of ~100 μs at room temperature using distributed computing with a total cost of 400 – 500 μs per system. To our knowledge, there are no other computations of room-temperature atomistic protein folding rates at the ms scale and beyond. Prior folding-rate calculations of NTL9 and Protein G were conducted at high temperature (355 K[1]/370 K[37], and 350 K[1] respectively) because of the prohibitive room-temperature timescales.

In addition to information about protein folding, the ability to quantify rate constants for slow-timescale biomolecular behavior is a critical step in model (force field) development. Although MD simulation is now a standard tool in structural biology studies,[38–41] the governing parameters of MD force fields have been determined based on energy minima[42–46] whereas energy barriers are expected to govern kinetic behavior. Given the evident importance of dynamic biomolecular phenomena, it is critical to obtain simulation-based rate constants to permit further refinement of force fields. Force fields cannot be assessed fully without the ability to compute kinetic observables, and we report on significant progress in this regard.

The WE method (Fig. 1A) employed in the present report is one of a number of path sampling approaches based on rigorous statistical mechanics[33, 47–52] capable of yielding unbiased rate constants. Although all these methods are theoretically well-grounded, WE

does offer the pragmatic advantage of being fully independent of the dynamics engine employed, which has enabled its application with a wide range of both molecular and cell-scale simulation software.[34, 53–59] This versatility facilitated the integration of the WESTPA software package[60] with the GPU-accelerated version of the AMBER molecular dynamics package[61–63] as employed here. The WE method yields ensembles of fully continuous trajectories from which non-equilibrium observables can be calculated, including kinetic and mechanistic properties. Importantly, the continuous WE trajectories spanning from unfolded to folded macrostates enable folding rate estimation using the history-augmented Markov state model (haMSM) formalism which is unbiased at arbitrary lag times.[56, 64]

Results

The WE procedure takes advantage of running in parallel multiple simulations with well-defined probabilities (or weights) in a conformational space that typically is divided based on pre-defined progress coordinates (see Fig. 1A).[32] The trajectory pruning and replication strategy facilitates progress along the coordinates and guarantees a constant total weight of all trajectories during the WE simulation (see SI Methods for more details). Fig. 1B shows a comparison of a brute-force MD simulation with a typical WE simulation, both starting from the same unfolded NTL9 structure. After ~7 μs of aggregate simulation time, the NTL9 Cα-RMSD in the MD simulation remains > 6 Å, whereas in the WE simulation folded NTL9 structures with Cα-RMSD < 1 Å are sampled. The probability flux of simulations reaching the target state allows estimation of the folding kinetics and the interrogation of continuous trajectories can provide information on folding mechanisms.

Estimates for folding rate constants are derived in two ways from WE data – directly from observed probability flows and using haMSM analysis. In both approaches, all WE simulations were used for a given system. In the direct analysis, Figs. 2–4 show that the probability flux into the folded states, which is an estimator for the rate constant,[65] apparently reaches a steady value in all three atomistic folding systems: NTL9 at low friction, NTL9 at high friction, and Protein G at low friction. The “molecular time”, t_{mol}, shown in Figs. 2–4 represents the time elapsed during individual trajectories. It is noteworthy that the flux reaches a plateau rather abruptly for Protein G, and to a value lower than the experimental value despite the low friction model, in contrast to the NTL9 data. Although the folding flux is dominated by a relatively small fraction of the independent runs, the dominating runs switch during the course of the trajectories (Figs. S1–S3). Nevertheless, the profiles of flux vs. Cα-RMSD (Figs. S4–S6) indicate the Protein G simulations in fact are far from steady state, implying the direct flux value for that system is not reliable despite its apparent plateau as a function of t_{mol}. The flux profiles at true steady state should be constant at all hypersurfaces (e.g., fixed Cα-RMSD values) separating folded from unfolded macrostates.[66–67]

We also employed haMSM analysis, which is unbiased for steady-state flux estimation at arbitrary lag times, and small lag times allow fuller use of the extensive WE data.[56, 64, 68] The approach is of particular interest for Protein G because, in principle, a haMSM can estimate steady-state behavior using trajectories generated in the transient period – i.e., in the approach to steady state. As noted, the flux profile for Protein G indicates those WE

simulations clearly remained in the transient regime. The haMSM results are also shown in Figs. 2–4. For the NTL9 systems, the haMSM rate estimates are consistent with estimates based on direct WE fluxes. For Protein G, the haMSM folding rate estimate is substantially higher than the direct WE estimate, and notably, it slightly exceeds the experimental value as expected for the low-friction solvent model.

WE simulation uses an ensemble of trajectories which all require computing resources, and aggregate simulation times are given in Table 1 (see SI for WE parameters and computing resources). Additional runs for the NTL9 systems were performed with alternative WE protocols to confirm the consistent, unbiased nature of the data: Figs. S7 and S8 show consistent time evolution of the folding flux based on different WE protocols for both low and high-friction systems.

The present study necessarily estimated folding times specific to the chosen force field and solvent model, and also conditioned on the starting structures. The novelty of the results is their relatively high precision and unbiased nature due to the theoretical foundations of the WE and haMSM methods.[35, 56, 64] Hence, although comparison to experimental folding times are shown in Table 1, readers are cautioned that the present study should be considered a first step in assessment of molecular models and initial ensembles. Given these caveats, the rough agreement with experimental values is encouraging but also points to the need for further investigation of solvent modeling and initial ensembles as discussed below.

A comparison of the force field-specific folding times and the aggregate simulation times as given in Table 1 enables assessment of the effectiveness of the WE protocol. In the case where WE exhibits least enhancement of sampling, namely NTL9 at low friction (Fig. 2), the calculated folding time range of 0.8 – 9.0 μs from Bayesian bootstrapping of haMSM estimates employed ~100μs of aggregate simulation. Fig. 2 reveals that much of the computation was used to confirm steady behavior and in fact the folding time could have been inferred from substantially less computation. In principle, similar results could have been obtained via 5–10 independent standard MD runs totaling the same aggregate simulation time. However, given the experimental ms folding time, it is unlikely such MD runs would have been attempted, and WE provided a reliable estimate in an affordable amount of computing effort. The higher-friction NTL9 study, which should be a better mimic of aqueous viscosity,[69–71] reveals a WE-haMSM folding time range of 0.2 – 1.9 ms (Fig. 3) that is essentially prohibitive for harvesting multiple events via conventional MD, even on modern GPU platforms. The value of the WE protocol is unambiguous for the slower Protein G system, where a folding time range of 3.3 – 200 ms is estimated in much less than a ms of aggregate simulation time (Fig. 4). By comparison, the computational cost of rate estimation here is substantially less than the previously reported overall cost of ~500 μs to estimate a ~65 μs room-temperature folding time.[26]

During the WE process, a variety of folding trajectories are simulated, enabling unbiased computation of ensemble properties. The weighted distributions of Cα-RMSD values shown in Figs. S9A, S10A for the NTL9 simulations and in Fig. S11A for the Protein G simulation serve as effective folding free energy profiles, which indicate that NTL9 folding has an energy minimum at Cα-RMSD = ~6 Å and Protein G at Cα-RMSD = ~10 Å. These regions

are separated from the folded state by a free energy barrier, suggesting a definition of the transition region and thus allowing calculation of the transition times (event durations) of the continuous WE folding trajectories. Of growing interest,[72–73] the event duration depends on the exact event starting point and on the solvent viscosity.[74–75] For NTL9, at low viscosity, the distributions of event duration have a peak at 1.5 – 2 ns (Fig. S9B), while at the higher water-like viscosity the peaks occur at slightly larger values ~4–5 ns (Fig. S10B). For Protein G, the event duration peaks are less clearly defined but occur in the range of ~2–7 ns (Fig. S11B).

A visual analysis of representative intermediate structures sheds light on the folding mechanisms. The NTL9 molecular structures shown in Fig. 5A illustrate that during the folding process the α-helix is formed first, followed by the formation of the N-terminal β-hairpin. A putative rate-limiting step of NTL9 folding is characterized by the association of the C-terminal β-strand with the N-terminal β-hairpin through hydrogen bonds. During the final steps (1 Å < Cα-RMSD < 4 Å), the protein reduces its solvent-accessible surface area by ~5 nm^2 when forming the remaining native hydrogen bonds, bending the N-terminal β-hairpin turn, and aligning the α-helix with the β-sheet. Similarly, Protein G (Fig. 5B) folds by first forming the α-helix and both β-hairpins and then bringing them all closer to each other, which appears to define the main free energy barrier, before connecting the two hairpins with hydrogen bonds and establishing the 4-stranded β-sheet. From the initial formation of the secondary structural elements to the fully folded structure (i.e. 1 Å < Cα-RMSD < 10 Å), Protein G reduces its overall surface area by ~ $8nm^2$.

Because some prior folding studies have been performed near the melting temperature, T_m, to improve sampling,[1, 25] it is of interest to investigate the effects of temperature on the folding process. After melting temperatures were estimated approximately (Figs. S12), we performed an additional set of WE simulations for low-friction NTL9 at T = 325 K ~ T_m. Comparison of the two simulation sets shows similar folding kinetics (Fig. S13) which we emphasize were obtained in the context of a single starting structure and implicit solvent. In future work, it will be of interest to compare folding mechanisms when the folding process is modeled more completely and accurately.

Discussion

The data reported here suggest that molecular dynamics calculations may soon be able to measure precisely and regularly a broad array of experimentally relevant timescales characterizing functional motions of biomolecules. Such measurements are necessarily limited by the accuracy of the underlying model equations (i.e., the force field) but understanding and correcting force field mis-calibrations is essential for progress in computational structural biology. These corrections will not be possible without reliable kinetics measurements, and the present data yields roughly order-of-magnitude precision (Table 1). Current force fields can suffer inaccuracies exceeding 1 kcal/mol for free energy *minima*[76–78] and errors at least as large are expected for the barriers which govern kinetics, which have not been part of force field parametrization.[21, 42–43, 79–82] Note for reference that an order-of-magnitude change in an Arrhenius factor $\exp(-\Delta G/RT)$ corresponds to a shift in ΔG of 1.4 kcal/mol; hence uncertainty of only 0.7 kcal/mol corresponds to a tenfold range.

Accuracy in kinetics also depends on the solvent model. Implicit solvation was employed in the present study, i.e., water molecules were not explicitly modeled. Because such models are in common use,[25, 37, 83–87] it is important to assess their kinetic accuracy. Although the overall computational cost for WE-based rate estimation is approximately double at water-like viscosity (γ = 80 ps −1) compared to low viscosity (gamma=5pŝ-1), the estimated folding time is longer by a factor exceeding 10. This difference in folding times is consistent with physical expectations but somewhat at variance with a prior report.[26] Going forward, additional comparison to explicit-solvent folding rate constants will be an important goal.

Another limitation of the present study is also intrinsic to protein folding generally – namely, ambiguity regarding the unfolded state ensemble. Experimentally, proteins are denatured chemically or with temperature,[88–91] each of which should yield a different unfolded ensemble, and the sensitivity of refolding to the denaturing process is an under-explored topic.[92] Given that some folding times are ms-scale or less, measurements may be sensitive to experimental protocols (e.g., mixing, cooling) occurring on the same timescales. Because of these ambiguities, we chose to keep our study as controlled as possible and focused specifically on folding from a *single* initial structure, recognizing the importance of future study of ensemble-initialized folding. Our mechanistic discussion above must be seen as restricted to this condition.

Quantification of statistical uncertainty was a central part of this study, and numerous repeated WE simulations were required to overcome the large variance of the present folding protocol (see Figs. S1–S3). Although a large variance is generally and rightly a cause for concern in data analysis, our ability to perform tens of truly independent simulations distinguishes this work from typical molecular simulation studies. As described elsewhere, neither traditional standard-error analysis nor bootstrapping properly quantify uncertainty in small-size data sets with large log-variance.[93] We therefore employed a Bayesian bootstrapping approach both for direct WE and haMSM flux estimates, which is superior at characterizing precision in such data.[93–94] Nevertheless, no analysis method can correct for insufficient sampling of an unknown distribution, and we estimate that the nominal 95% Bayesian credibility regions reported here empirically correspond to ~60% probability of bracketing the true mean – and such uncertainty in the error analysis is intrinsic to the modest sample sizes.[93] This point is borne out by the apparent ‘false plateau’ of the Protein G direct flux. Future studies will clearly benefit from variance-reduction strategies, which have been proposed.[95–96]

The weighted ensemble method was chosen over other rigorous path sampling approaches[10, 27–31, 47–52] and standard (history-independent) Markov state models (MSMs).[97–98] Compared to other path sampling methods, WE offers fully scalable parallelization and does not require hard-coding within the dynamics engine in order to “catch” trajectories as they cross interfaces.[34] When compared to standard MSMs, WE not only avoids any approximation but also offers continuous trajectories and the fine temporal resolution needed to infer mechanistic details occurring on 5–10 ns timescales (Figs. S9–S11). By contrast, modern well-validated MSMs often require lag times >100 ns.[97–98] The continuous trajectories generated by WE allow application of the history-augmented MSMs at arbitrary

lag times, which are unbiased for estimation of steady-state fluxes.[56, 64] Using short lag-times for haMSMs in turn allows use of all data generated.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We gratefully acknowledge support from the NIH (Grant GM115805) and from the OHSU Center for Spatial Systems Biomedicine. Computing support was provided by the Center for Research Computing at the University of Pittsburgh and by the Advanced Computing Center at the Oregon Health and Science University. Helpful comments on the manuscript were provided by Lillian Chong.

References

1. Lindorff-Larsen K; Piana S; Dror RO; Shaw DE, How Fast-Folding Proteins Fold. Science 2011, 334 (6055), 517–520. [PubMed: 22034434]
2. Englander SW; Mayne L, The nature of protein folding pathways. Proceedings of the National Academy of Sciences of the United States of America 2014, 111 (45), 15873–80. [PubMed: 25326421]
3. Wolynes PG; Eaton WA, The physics of protein folding. Physics World 1999, 12 (9), 39–44.
4. Go N, Theoretical Studies of Protein Folding. Annual Review of Biophysics and Bioengineering 1983, 12 (1), 183–210.
5. Bryngelson JD; Wolynes PG, Spin glasses and the statistical mechanics of protein folding. Proceedings of the National Academy of Sciences of the United States of America 1987, 84 (21), 7524–8. [PubMed: 3478708]
6. Matouschek A; Kellis JT; Serrano L; Fersht AR, Mapping the transition state and pathway of protein folding by protein engineering. Nature 1989, 340 (6229), 122–126. [PubMed: 2739734]
7. Piana S; Lindorff-Larsen K; Shaw DE, Protein folding kinetics and thermodynamics from atomistic simulation. Proceedings of the National Academy of Sciences of the United States of America 2012, 109 (44), 17845–50. [PubMed: 22822217]
8. Dill KA; MacCallum JL, The Protein-Folding Problem, 50 Years On. Science 2012, 338 (6110), 1042–1046. [PubMed: 23180855]
9. Udgaonkar J; Marqusee S, Folding and binding. Current Opinion in Structural Biology 2013, 23 (1), 1–3. [PubMed: 23374590]
10. Bolhuis PG, Two-state protein folding kinetics through all-atom molecular dynamics based sampling. Frontiers in bioscience (Landmark edition) 2009, 14, 2801–28. [PubMed: 19273237]
11. Thirumalai D; O'Brien EP; Morrison G; Hyeon C, Theoretical Perspectives on Protein Folding. Annual Review of Biophysics 2010, 39 (1), 159–183.
12. Baker D; Agard DA, Kinetics versus thermodynamics in protein folding. Biochemistry 1994, 33 (24), 7505–9. [PubMed: 8011615]
13. Ozkan SB; Dill KA; Bahar I, Computing the transition state populations in simple protein models. Biopolymers 2003, 68 (1), 35–46. [PubMed: 12579578]
14. Freddolino PL; Harrison CB; Liu Y; Schulten K, Challenges in protein-folding simulations. Nature Physics 2010, 6 (10), 751–758. [PubMed: 21297873]
15. Eaton WA; Thompson PA; Chan CK; Hage SJ; Hofrichter J, Fast events in protein folding. Structure (London, England : 1993) 1996, 4 (10), 1133–9.
16. Kubelka J; Hofrichter J; Eaton WA, The protein folding 'speed limit'. Current Opinion in Structural Biology 2004, 14 (1), 76–88. [PubMed: 15102453]
17. Levitt M; Warshel A, Computer simulation of protein folding. Nature 1975, 253 (5494), 694–8. [PubMed: 1167625]

18. Rhee YM; Pande VS, Multiplexed-Replica Exchange Molecular Dynamics Method for Protein Folding Simulation. Biophysical Journal 2003, 84 (2), 775–786. [PubMed: 12547762]
19. Paschek D; García AE, Reversible Temperature and Pressure Denaturation of a Protein Fragment: A Replica Exchange Molecular Dynamics Simulation Study. Physical Review Letters 2004, 93 (23), 238105–238105. [PubMed: 15601210]
20. Freddolino PL; Park S; Roux B; Schulten K, Force Field Bias in Protein Folding Simulations. Biophysical Journal 2009, 96 (9), 3772–3780. [PubMed: 19413983]
21. Snow CD; Sorin EJ; Rhee YM; Pande VS, How Well Can Simulation Predict Protein Folding Kinetics and Thermodynamics? Annual Review of Biophysics and Biomolecular Structure 2005, 34 (1), 43–69.
22. Dill KA; Ozkan SB; Weikl TR; Chodera JD; Voelz VA, The protein folding problem: when will it be solved? Current Opinion in Structural Biology 2007, 17 (3), 342–346. [PubMed: 17572080]
23. Zwier MC; Chong LT, Reaching biological timescales with all-atom molecular dynamics simulations. Current Opinion in Pharmacology 2010, 10 (6), 745–752. [PubMed: 20934381]
24. Simmerling C; Strockbine B; Roitberg A, All-Atom Structure Prediction and Folding Simulations of a STable Protein. Journal of the American Chemical Society 2002, 124 (38), 11258–11259. [PubMed: 12236726]
25. Nguyen H; Maier J; Huang H; Perrone V; Simmerling C, Folding Simulations for Proteins with Diverse Topologies Are Accessible in Days with a Physics-Based Force Field and Implicit Solvent. Journal of the American Chemical Society 2014, 136 (40), 13959–13962. [PubMed: 25255057]
26. Ensign DL; Pande VS, Bayesian Single-Exponential Kinetics in Single-Molecule Experiments and Simulations. The Journal of Physical Chemistry B 2009, 113 (36), 12410–12423. [PubMed: 19681587]
27. Cárdenas AE; Elber R, Kinetics of cytochrome C folding: Atomically detailed simulations. Proteins: Structure, Function, and Bioinformatics 2003, 51 (2), 245–257.
28. Kuczera K; Jas GS; Elber R, Kinetics of Helix Unfolding: Molecular Dynamics Simulations with Milestoning. The Journal of Physical Chemistry A 2009, 113 (26), 7461–7473. [PubMed: 19354256]
29. Juraszek J; Bolhuis PG, Rate Constant and Reaction Coordinate of Trp-Cage Folding in Explicit Water. Biophysical Journal 2008, 95 (9), 4246–4257. [PubMed: 18676648]
30. Velez-Vega C; Borrero EE; Escobedo FA, Kinetics and mechanism of the unfolding native-to-loop transition of Trp-cage in explicit solvent via optimized forward flux sampling simulations. The Journal of Chemical Physics 2010, 133 (10), 105103–105103. [PubMed: 20849192]
31. Borrero EE; Escobedo FA, Folding kinetics of a lattice protein via a forward flux sampling approach. The Journal of Chemical Physics 2006, 125 (16), 164904–164904. [PubMed: 17092136]
32. Huber GA; Kim S, Weighted-ensemble Brownian dynamics simulations for protein association reactions. Biophysical Journal 1996, 70 (1), 97–110. [PubMed: 8770190]
33. Chong LT; Saglam AS; Zuckerman DM, Path-sampling strategies for simulating rare events in biomolecular systems. Current Opinion in Structural Biology 2017, 43, 88–94. [PubMed: 27984811]
34. Zuckerman DM; Chong LT, Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. Annual Review of Biophysics 2017, 46 (1), 43–57.
35. Zhang BW; Jasnow D; Zuckerman DM, The “weighted ensemble” path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. The Journal of Chemical Physics 2010, 132 (5), 054107–054107. [PubMed: 20136305]
36. Suárez E; Pratt AJ; Chong LT; Zuckerman DM, Estimating first-passage time distributions from weighted ensemble simulations and non-Markovian analyses. Protein Science 2016, 25 (1), 67–78. [PubMed: 26131764]
37. Voelz VA; Bowman GR; Beauchamp K; Pande VS, Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9(1−39). Journal of the American Chemical Society 2010, 132 (5), 1526–1528. [PubMed: 20070076]
38. Fanning SW; Mayne CG; Dharmarajan V; Carlson KE; Martin TA; Novick SJ; Toy W; Green B; Panchamukhi S; Katzenellenbogen BS; Tajkhorshid E; Griffin PR; Shen Y; Chandarlapaty S;

Katzenellenbogen JA; Greene GL, Estrogen receptor alpha somatic mutations Y537S and D538G confer breast cancer endocrine resistance by stabilizing the activating function-2 binding conformation. eLife 2016, 5, e12792–e12792. [PubMed: 26836308]

39. Speltz TE; Fanning SW; Mayne CG; Fowler C; Tajkhorshid E; Greene GL; Moore TW, Stapled Peptides with γ-Methylated Hydrocarbon Chains for the Estrogen Receptor/Coactivator Interaction. Angewandte Chemie International Edition 2016, 55 (13), 4252–4255. [PubMed: 26928945]
40. Adelman Joshua L.; Sheng Y; Choe S; Abramson J; Wright Ernest M.; Rosenberg John M.; Grabe M, Structural Determinants of Water Permeation through the Sodium-Galactose Transporter vSGLT. Biophysical Journal 2014, 106 (6), 1280–1289. [PubMed: 24655503]
41. Durrant JD; McCammon JA, Molecular dynamics simulations and drug discovery. BMC Biology 2011, 9 (1), 71–71. [PubMed: 22035460]
42. Wang J; Wolf RM; Caldwell JW; Kollman PA; Case DA, Development and testing of a general amber force field. Journal of Computational Chemistry 2004, 25 (9), 1157–1174. [PubMed: 15116359]
43. Hornak V; Abel R; Okur A; Strockbine B; Roitberg A; Simmerling C, Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 2006, 65 (3), 712–25. [PubMed: 16981200]
44. Vanommeslaeghe K; Hatcher E; Acharya C; Kundu S; Zhong S; Shim J; Darian E; Guvench O; Lopes P; Vorobyov I; Mackerell AD, CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. Journal of Computational Chemistry 2009, 31 (4), 671–690.
45. MacKerell AD; Bashford D; Bellott M; Dunbrack RL; Evanseck JD; Field MJ; Fischer S; Gao J; Guo H; Ha S; Joseph-McCarthy D; Kuchnir L; Kuczera K; Lau FTK; Mattos C; Michnick S; Ngo T; Nguyen DT; Prodhom B; Reiher WE; Roux B; Schlenkrich M; Smith JC; Stote R; Straub J; Watanabe M; Wiórkiewicz-Kuczera J; Yin D; Karplus M, All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins The Journal of Physical Chemistry B 1998, 102 (18), 3586–3616. [PubMed: 24889800]
46. William LJ; David S Maxwell A; Tirado-Rives J, Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. J. Am. Chem. Soc 1996, 118 (45), 11225–11236.
47. Bolhuis PG; Chandler D; Dellago C; Geissler PL, Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark. Annual Review of Physical Chemistry 2002, 53 (1), 291–318.
48. van Erp TS; Moroni D; Bolhuis PG, A novel path sampling method for the calculation of rate constants. The Journal of Chemical Physics 2003, 118 (17), 7762–7774.
49. Allen RJ; Warren PB; ten Wolde PR, Sampling Rare Switching Events in Biochemical Networks. Physical Review Letters 2005, 94 (1), 018104–018104. [PubMed: 15698138]
50. Warmflash A; Bhimalapuram P; Dinner AR, Umbrella sampling for nonequilibrium processes. The Journal of Chemical Physics 2007, 127 (15), 154112–154112. [PubMed: 17949137]
51. Faradjian AK; Elber R, Computing time scales from reaction coordinates by milestoning. The Journal of Chemical Physics 2004, 120 (23), 10880–10889. [PubMed: 15268118]
52. Bello-Rivas JM; Elber R, Exact milestoning. The Journal of Chemical Physics 2015, 142 (9), 094102–094102. [PubMed: 25747056]
53. Nunes-Alves A; Zuckerman DM; Arantes GM, Escape of a Small Molecule from Inside T4 Lysozyme by Multiple Pathways. Biophysical journal 2018, 114 (5), 1058–1066. [PubMed: 29539393]
54. Donovan RM; Tapia J-J; Sullivan DP; Faeder JR; Murphy RF; Dittrich M; Zuckerman DM, Unbiased Rare Event Sampling in Spatial Stochastic Systems Biology Models Using a Weighted Ensemble of Trajectories. PLOS Computational Biology 2016, 12 (2), e1004611–e1004611. [PubMed: 26845334]
55. Donovan RM; Sedgewick AJ; Faeder JR; Zuckerman DM, Efficient stochastic simulation of chemical kinetics networks using a weighted ensemble of trajectories. The Journal of Chemical Physics 2013, 139 (11), 115105. [PubMed: 24070313]

56. Suárez E; Lettieri S; Zwier MC; Stringer CA; Subramanian SR; Chong LT; Zuckerman DM, Simultaneous Computation of Dynamical and Equilibrium Information Using a Weighted Ensemble of Trajectories. Journal of Chemical Theory and Computation 2014, 10 (7), 2658–2667. [PubMed: 25246856]
57. Adelman Joshua L.; Dale Amy L.; Zwier Matthew C.; Bhatt D; Chong Lillian T.; Zuckerman Daniel M.; Grabe M, Simulations of the Alternating Access Mechanism of the Sodium Symporter Mhp1. Biophysical Journal 2011, 101 (10), 2399–2407. [PubMed: 22098738]
58. Zwier MC; Pratt AJ; Adelman JL; Kaus JW; Zuckerman DM; Chong LT, Efficient Atomistic Simulation of Pathways and Calculation of Rate Constants for a Protein–Peptide Binding Process: Application to the MDM2 Protein and an Intrinsically Disordered p53 Peptide. The Journal of Physical Chemistry Letters 2016, 7 (17), 3440–3445. [PubMed: 27532687]
59. Saglam AS; Wang DW; Zwier MC; Chong LT, Flexibility vs Preorganization: Direct Comparison of Binding Kinetics for a Disordered Peptide and Its Exact Preorganized Analogues. The Journal of Physical Chemistry B 2017, 121 (43), 10046–10054. [PubMed: 28992700]
60. Zwier MC; Adelman JL; Kaus JW; Pratt AJ; Wong KF; Rego NB; Suárez E; Lettieri S; Wang DW; Grabe M; Zuckerman DM; Chong LT, WESTPA: An Interoperable, Highly Scalable Software Package for Weighted Ensemble Simulation and Analysis. Journal of Chemical Theory and Computation 2015, 11 (2), 800–809. [PubMed: 26392815]
61. Pearlman DA; Case DA; Caldwell JW; Ross WS; Cheatham TE; DeBolt S; Ferguson D; Seibel G; Kollman P, AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Computer Physics Communications 1995, 91 (1–3), 1–41.
62. Case DA; Cheatham TE; Darden T; Gohlke H; Luo R; Merz KM; Onufriev A; Simmerling C; Wang B; Woods RJ; Wang B; Woods RJ, The Amber biomolecular simulation programs. Journal of computational chemistry 2005, 26 (16), 1668–88. [PubMed: 16200636]
63. Case DA, Cerutti DS, Cheatham TE III, Darden TA, Duke RE, Giese TJ, Gohlke H, Goetz AW, Greene D, Homeyer N, Izadi S, Kovalenko A, Lee TS, LeGrand S, Li P, Lin C, Liu J, Luchko T, Luo R, Mermelstein D, Merz KM, Monard G, York HDM, Kollman PA, Amber 2017, University of California, San Francisco 2017.
64. Copperman J; Zuckerman DM, Accelerated estimation of long-timescale kinetics by combining weighted ensemble simulation with Markov model “microstates” using non-Markovian theory. arXiv:1903.04673 2019, 1–7.
65. Hill TL, State Probabilities and Fluxes in Terms of the Rate Constants of the Diagram Springer New York: New York, NY, 1989; pp 39–88.
66. Gardiner C, Stochastic Methods: A Handbook for the Natural and Social Sciences Springer: 2009.
67. Risken H; Frank T, The Fokker-Planck Equation: Methods of Solution and Applications 2nd ed.; Springer: 2011.
68. Suárez E; Adelman JL; Zuckerman DM, Accurate Estimation of Protein Folding and Unfolding Times: Beyond Markov State Models. Journal of Chemical Theory and Computation 2016, 12 (8), 3473–3481. [PubMed: 27340835]
69. Feig M, Kinetics from Implicit Solvent Simulations of Biomolecules as a Function of Viscosity. J Chem Theory Comput 2007, 3 (5), 1734–1748. [PubMed: 26627618]
70. Anandakrishnan R; Drozdetski A; Walker Ross C.; Onufriev Alexey V., Speed of Conformational Change: Comparing Explicit and Implicit Solvent Molecular Dynamics Simulations. Biophysical Journal 2015, 108 (5), 1153–1164. [PubMed: 25762327]
71. Zagrovic B; Pande V, Solvent viscosity dependence of the folding rate of a small protein: Distributed computing study. Journal of Computational Chemistry 2003, 24 (12), 1432–1436. [PubMed: 12868108]
72. Shaw DE; Maragakis P; Lindorff-Larsen K; Piana S; Dror RO; Eastwood MP; Bank JA; Jumper JM; Salmon JK; Shan Y; Wriggers W, Atomic-level characterization of the structural dynamics of proteins. Science (New York, N.Y.) 2010, 330 (6002), 341–6.
73. Chung HS; McHale K; Louis JM; Eaton WA, Single-Molecule Fluorescence Experiments Determine Protein Folding Transition Path Times. Science 2012, 335 (6071), 981–984. [PubMed: 22363011]

74. Zuckerman DM; Woolf TB, Transition events in butane simulations: Similarities across models. The Journal of Chemical Physics 2002, 116 (6), 2586.

75. Zhang BW; Jasnow D; Zuckerman DM, Transition-event durations in one-dimensional activated processes. The Journal of Chemical Physics 2007, 126 (7), 074504. [PubMed: 17328617]

76. Shirts MR; Pitera JW; Swope WC; Pande VS, Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. The Journal of Chemical Physics 2003, 119 (11), 5740–5761.

77. Fujitani H; Tanida Y; Ito M; Jayachandran G; Snow CD; Shirts MR; Sorin EJ; Pande VS, Direct calculation of the binding free energies of FKBP ligands. The Journal of Chemical Physics 2005, 123 (8), 084108–084108. [PubMed: 16164283]

78. Shirts MR; Mobley DL; And JDC; Pande VS, Accurate and Efficient Corrections for Missing Dispersion Interactions in Molecular Simulations. J. Phys. Chem. B 2007, 111 (45), 13052–13063. [PubMed: 17949030]

79. Wang L-P; Martinez TJ; Pande VS, Building Force Fields: An Automatic, Systematic, and Reproducible Approach. The Journal of Physical Chemistry Letters 2014, 5 (11), 1885–1891. [PubMed: 26273869]

80. Beauchamp KA; Lin Y-S; Das R; Pande VS, Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. Journal of Chemical Theory and Computation 2012, 8 (4), 1409–1414. [PubMed: 22754404]

81. Debiec KT; Cerutti DS; Baker LR; Gronenborn AM; Case DA; Chong LT, Further along the Road Less Traveled: AMBER ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. Journal of Chemical Theory and Computation 2016, 12 (8), 3926–3947. [PubMed: 27399642]

82. Maier JA; Martinez C; Kasavajhala K; Wickstrom L; Hauser KE; Simmerling C, ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. Journal of Chemical Theory and Computation 2015, 11 (8), 3696–3713. [PubMed: 26574453]

83. Hua DP; Huang H; Roy A; Post CB, Evaluating the dynamics and electrostatic interactions of folded proteins in implicit solvents. Protein Science 2016, 25 (1), 204–218. [PubMed: 26189497]

84. Cumberworth A; Bui JM; Gsponer J, Free energies of solvation in the context of protein folding: Implications for implicit and explicit solvent models. Journal of Computational Chemistry 2016, 37 (7), 629–640. [PubMed: 26558440]

85. Shao Q; Zhu W, Assessing AMBER force fields for protein folding in an implicit solvent. Physical chemistry chemical physics : PCCP 2018, 20 (10), 7206–7216. [PubMed: 29480910]

86. Nguyen H; Pérez A; Bermeo S; Simmerling C, Refinement of Generalized Born Implicit Solvation Parameters for Nucleic Acids and Their Complexes with Proteins. Journal of Chemical Theory and Computation 2015, 11 (8), 3714–3728. [PubMed: 26574454]

87. Anandakrishnan R; Izadi S; Onufriev AV, Why Computed Protein Folding Landscapes Are Sensitive to the Water Model. Journal of Chemical Theory and Computation 2019, 15 (1), 625–636. [PubMed: 30514080]

88. Kuhlman B; Luisi DL; Evans PA; Raleigh DP, Global analysis of the effects of temperature and denaturant on the folding and unfolding kinetics of the N-terminal domain of the protein L9. Journal of Molecular Biology 1998, 284 (5), 1661–1670. [PubMed: 9878377]

89. Möglich A; Krieger F; Kiefhaber T, Molecular Basis for the Effect of Urea and Guanidinium Chloride on the Dynamics of Unfolded Polypeptide Chains. Journal of Molecular Biology 2005, 345 (1), 153–162. [PubMed: 15567418]

90. Taskent H; Cho J-H; Raleigh DP, Temperature-Dependent Hammond Behavior in a Protein-Folding Reaction: Analysis of Transition-State Movement and Ground-State Effects. Journal of Molecular Biology 2008, 378 (3), 699–706. [PubMed: 18384809]

91. Anil B; Li Y; Cho JH; Raleigh DP, The Unfolded State of NTL9 Is Compact in the Absence of Denaturant. Biochemistry 2006, 45 (33), 10110–10116. [PubMed: 16906769]

92. Guinn EJ; Marqusee S, Exploring the Denatured State Ensemble by Single-Molecule Chemo-Mechanical Unfolding: The Effect of Force, Temperature, and Urea. Journal of Molecular Biology 2018, 430 (4), 450–464. [PubMed: 28782558]

93. Mostofian B; Zuckerman DM, Error analysis for small-sample, high-log variance data: Cautions for bootstrapping and Bayesian bootstrapping. arXiv:1806.01998 2018, 1–15.

94. Rubin DB, The Bayesian Bootstrap. The Annals of Statistics 1981, 9 (1), 130–134.

95. Aristoff D, Analysis and optimization of weighted ensemble sampling. ESAIM: Mathematical Modelling and Numerical Analysis 2017,

96. Aristoff D; Zuckerman DM, Optimizing Weighted Ensemble Sampling of Steady States. arXiv: 1806.00860 2018, 1–19.

97. Lane TJ; Bowman GR; Beauchamp K; Voelz Vincent A.; Pande V, Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. Journal of the American Chemical Society 2011, 133 (45), 18413–18419. [PubMed: 21988563]

98. Schwantes CR; Pande VS, Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. Journal of Chemical Theory and Computation 2013, 9 (4), 2000–2009. [PubMed: 23750122]

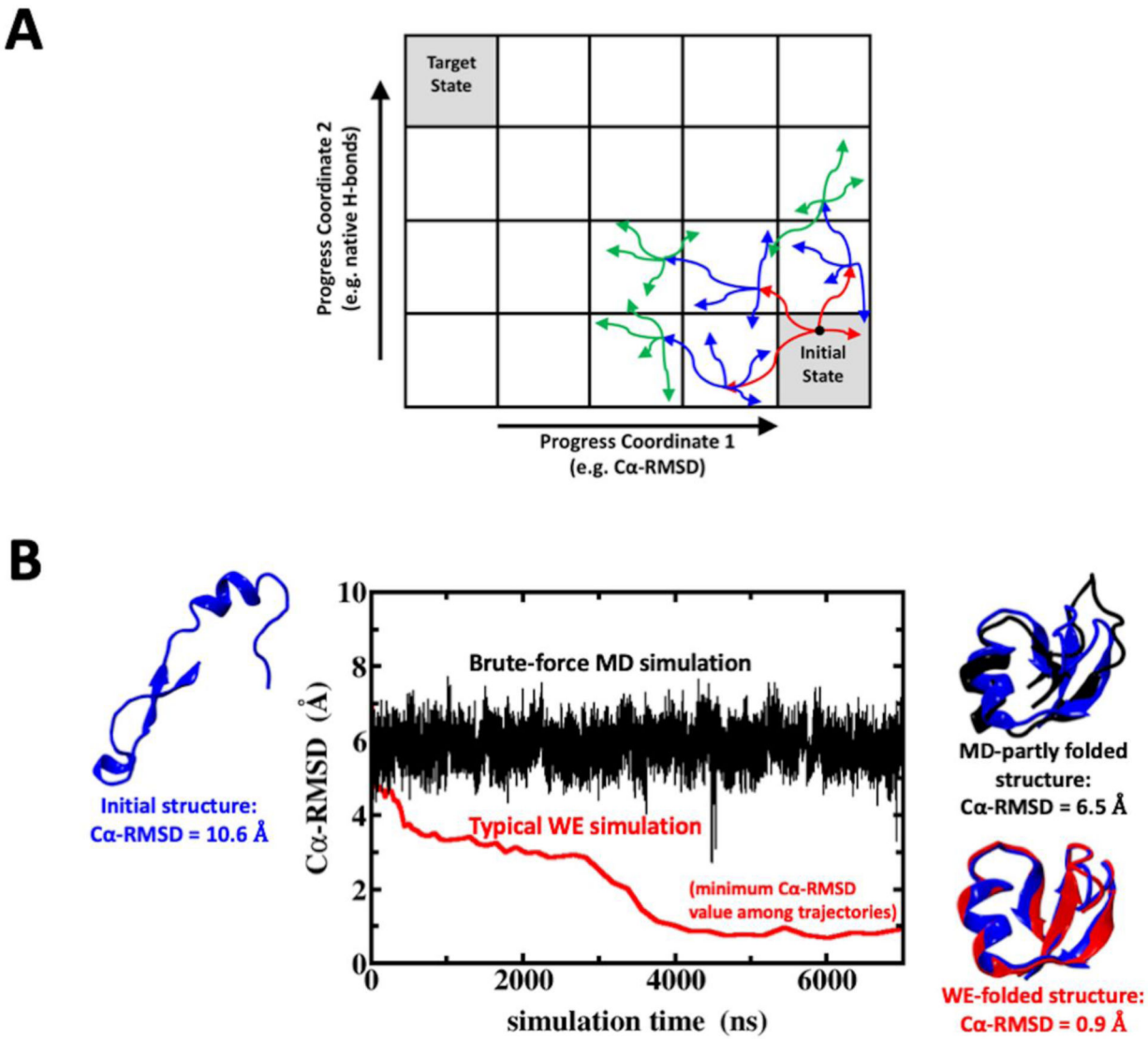


Figure 1:

The WE procedure and comparison to regular MD simulation. (A) A schematic of the WE simulation procedure is shown with two-dimensional binning for protein folding. Three iterations (red, then blue, then green) are shown based on a target number of 4 trajectories per bin, illustrating the “statistical ratcheting” effect which is possible without applying biasing forces. Note that a set of new trajectories is shown only for those parent trajectories that reached a new bin. (B) A brute-force MD simulation of NTL9 leads to the “MD-partly folded” structure (black structure) with a Cα-RMSD of 6.5 Å with respect to the folded crystal structure (blue structures at right) after 7 μs of simulation time. By contrast, a WE simulation starting from the same initial structure (blue structure at left) samples the “WE-folded” NTL9 structure with Cα-RMSD < 1 Å (red structure on the right panel). The WE simulation time is the aggregate time including all trajectory segments, representing a fair comparison using roughly the same amount of computational resources.

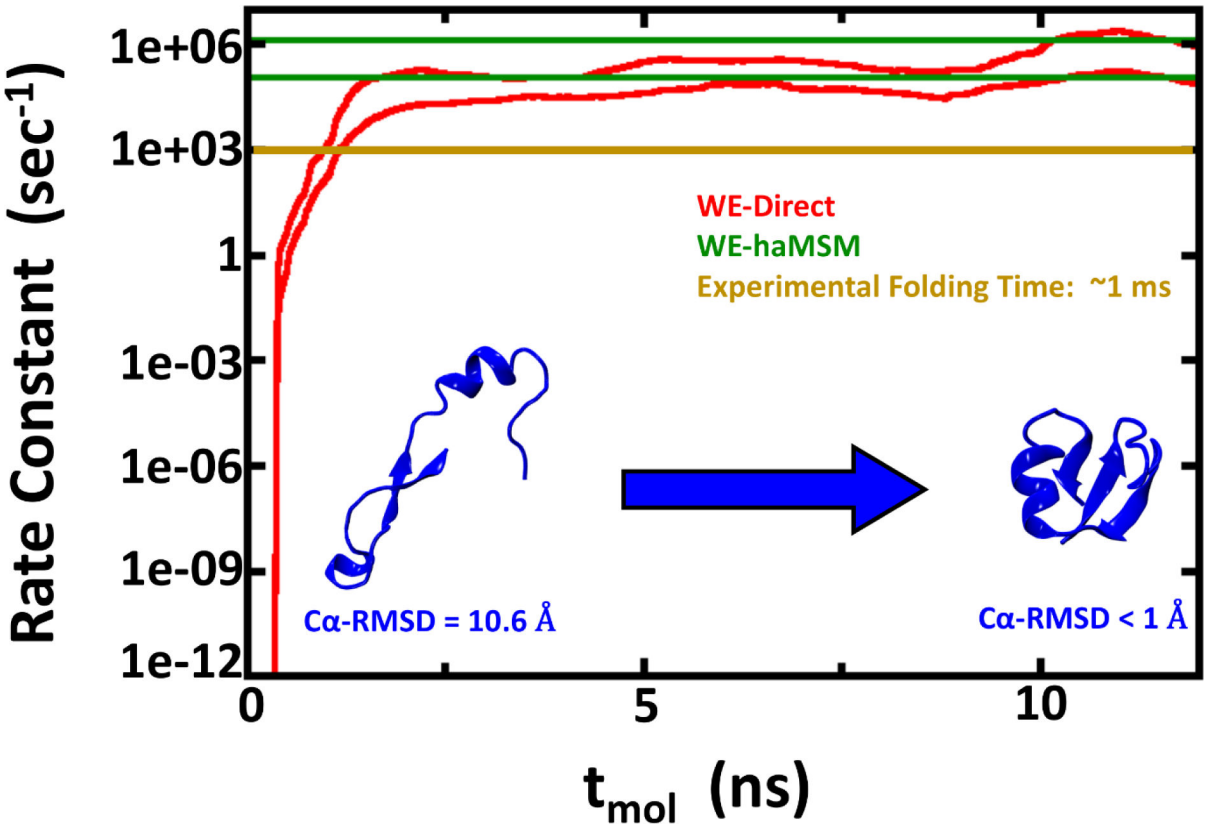


Figure 2:

Rate constant estimations for NTL9 folding using 2D WE method with solvent viscosity (γ) set to 5 ps^{-1}. The red lines show the nominal 95% Credibility Region (CR) as a function of molecular time from Bayesian bootstrapping based on direct WE rate constant estimates, which were windowed averages of the previous 1 ns of molecular time for each of the 10 independent simulations (see Figure S1). The green lines show the 95% CR for rate constants obtained by the haMSM method. The experimental rate constant is shown in gold, but note that the low viscosity used in these simulations is expected to yield overly fast kinetics.

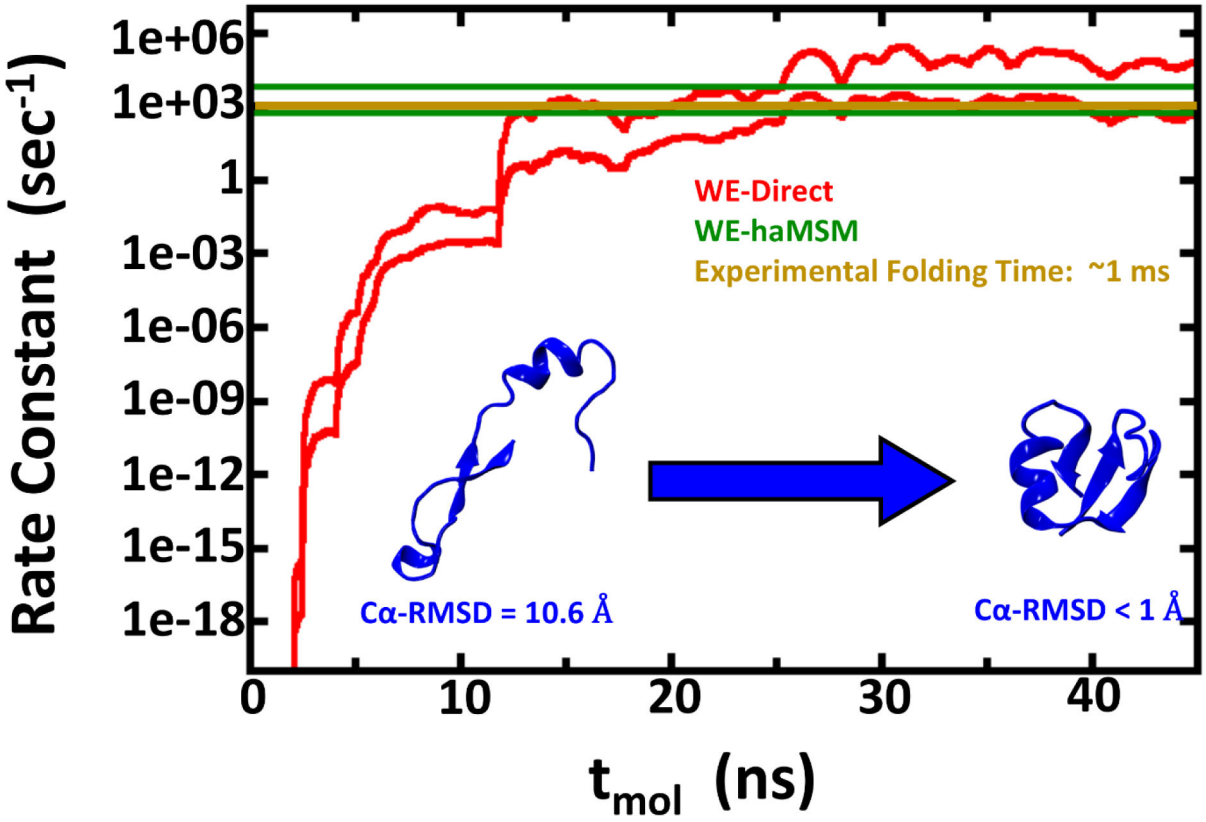


Figure 3:

Rate constant estimations for NTL9 folding using 1D WE method with solvent viscosity (γ) set to 80 ps^{-1}. The red lines show the nominal 95% Credibility Region (CR) as a function of molecular time from Bayesian bootstrapping based on direct WE rate constant estimates, which were windowed averages of the previous 1 ns of molecular time for each of the 30 independent simulations (see Figure S2). The green lines show the 95% CR for rate constants obtained by the haMSM method. The experimental rate constant is shown in gold.

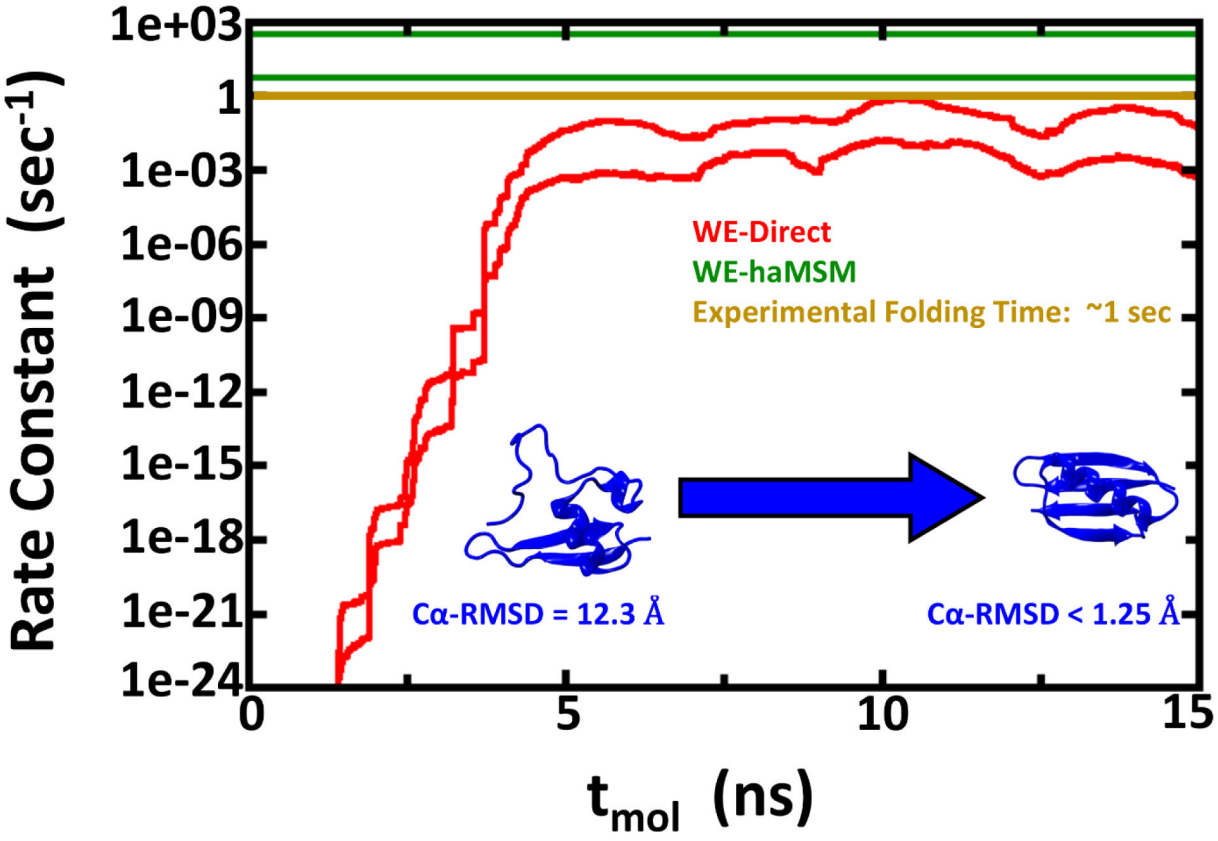


Figure 4:

Rate constant estimations for Protein G folding using 2D WE method with solvent viscosity (γ) set to 5 ps^{-1}. The red lines show the nominal 95% Credibility Region (CR) as a function of molecular time from Bayesian bootstrapping based on direct WE rate constant estimates, which were windowed averages of the previous 1 ns of molecular time for each of the 15 independent simulations (see Figure S3). The green lines show the 95% CR for rate constants obtained by the haMSM method. The experimental rate constant is shown in gold, but note that the low viscosity used in these simulations is expected to yield overly fast kinetics.

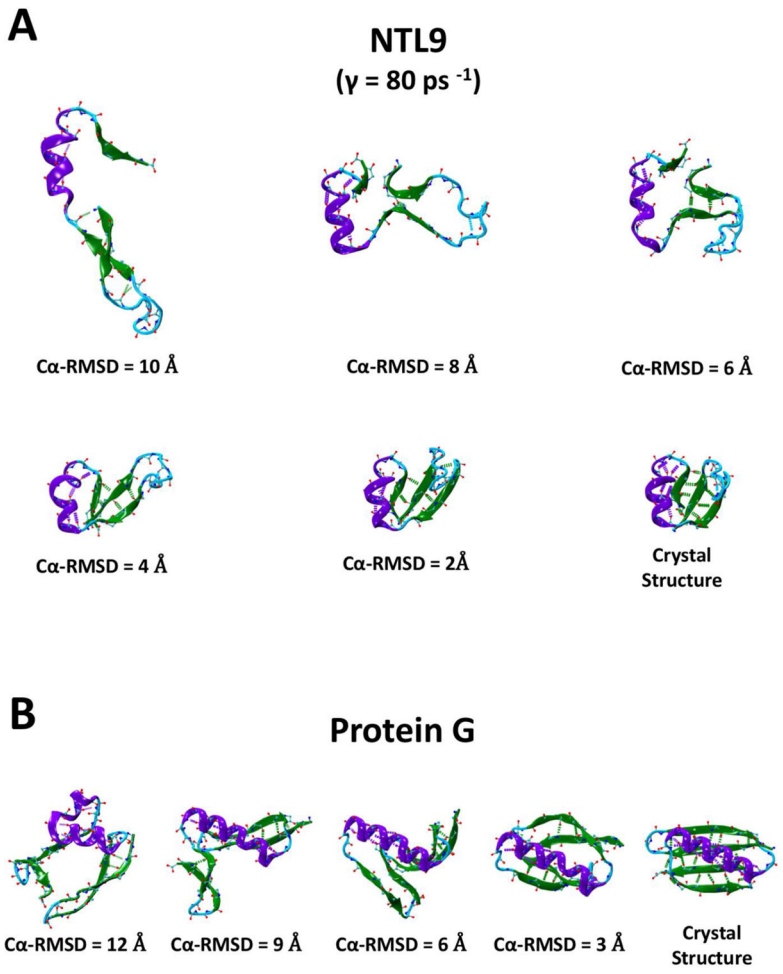


Figure 5:

A set of example NTL9 (A) and Protein G (B) structures with decreasing Cα-RMSDs from left to right obtained from a continuous trajectory along with the folded crystal structure. Residues are colored based on their native secondary structures in violet (α-helix), green (β-sheet), and cyan (loops). Native backbone hydrogen bonds are indicated as dashed lines, if they emerge in the structure shown.

Table 1:

Computational cost and folding rate constants for the three systems studied here (AMBER FF14SB force field).

System	Number of indep. WE simulations	Molec. time (ns)	Aggregate simulation time (μs)	Wall-clock time* (days / simulation)	WE-Direct folding time**	WE-haMSM folding time***	Expt'l folding time
NTL9, $\gamma = 5$ ps^{-1}	10	12	115	22	0.6 – 8 μs	0.8 – 9.0 μs	~ 1 ms
NTL9, $\gamma = 80$ ps^{-1}	30	45	252	20	0.01 – 0.8 ms	0.2 – 2 ms	~ 1 ms
Protein G, $\gamma = 5$ ps^{-1}	15	15	225	31	4 – 200 sec	3 – 200 ms	~ 1 sec

*Based on 1 GPU card or ~48 CPU cores.

**Averaged from molecular times 10–12 ns (NTL9, $\gamma = 5$ ps^{-1}), 30–45 ns (NTL9, $\gamma = 80$ ps^{-1}), and 10–15 ns (Protein G), respectively. Given range is the nominal 95% Bayesian Credibility Region, which corresponds to a ~60% range of uncertainty: see Ref. 92 and SI.

***From haMSMs with 10,000 microstates trained from molecular time 10 –12 ns (NTL9, $\gamma = 5$ ps^{-1}), 30–45 ns (NTL9, $\gamma = 80$ ps^{-1}), and 10–15 ns (Protein G), respectively. The uncertainty range is the nominal 95% Bayesian Credibility Region.