Published in final edited form as:
Curr Opin Struct Biol. 2013 February ; 23(1): 58–65. doi:10.1016/j.sbi.2012.11.002.

To Milliseconds and Beyond: Challenges in the Simulation of Protein Folding

Thomas J. Lane[†], **Diwakar Shukla**[†,‡], **Kyle A. Beauchamp**[§], and **Vijay S. Pande**[†,§,°,*]

[†]Department of Chemistry, Stanford University

[‡]Department of Bioengineering, Stanford University

[§]Biophysics Program, Stanford University

[°]Department of Computer Science, Stanford University

Abstract

Quantitatively accurate all-atom molecular dynamics (MD) simulations of protein folding have long been considered a holy grail of computational biology. Due to the large system sizes and long timescales involved, such a pursuit was for many years computationally intractable. Further, sufficiently accurate forcefields needed to be developed in order to realistically model folding. This decade, however, saw the first reports of folding simulations describing kinetics on the order of milliseconds, placing many proteins firmly within reach of these methods. Progress in sampling and forcefield accuracy, however, presents a new challenge: how to turn huge MD datasets into scientific understanding. Here, we review recent progress in MD simulation techniques and show how the vast datasets generated by such techniques present new challenges for analysis. We critically discuss the state of the art, including reaction coordinate and Markov state model (MSM) methods, and provide a perspective for the future.

Introduction

Understanding protein folding via molecular simulation has been an aspiration of computational chemists ever since Anfinsen uncovered the surprising fact that proteins folded to a unique structure[1–3]. Applying simulation to folding appeals for many reasons. Folding is rapid and complex, requiring atomic-level resolution at nanosecond timescales for a complete detailed picture – outside the hard limits of temporal and spatial resolution of most experimental techniques [4]. Furthermore, the complexity of protein native states and the inherent physical heterogeneity in the folding process have frustrated the search for microscopic physical theories of folding, though some advanced phenomenological approaches have been proposed [5–9]. Thus atomic-level simulations of folding, possessing intrinsically high resolution, have been aggressively pursued with the hope of surmounting these difficulties.

Three main problems must be overcome to achieve useful simulations of protein folding: accurate models (forcefields), sufficient sampling, and robust data analysis. Forcefield

[*]pande@stanford.edu.

development has received much attention from the field, and has been extensively discussed [10–15]. Although more remains to be done to build and validate even more accurate models, forcefields capable of folding proteins in good agreement with experiment already exist (Fig. 1). Instead, the major challenge in producing reliable simulations of folding has been harnessing enough computer power to produce sufficient sampling to study folding. Because classical simulations must integrate Newton's equations of motion with femtosecond timesteps (10^{-15} s), folding simulations require ~10^{12} timesteps to reach millisecond timescales. This expense is compounded by large system sizes (~10^5 atoms for explicit solvent simulations) and the need to witness many events for statistical confidence, making the computational effort required to study folding via simulation enormous.

Very recently, advances in hardware, software, and sampling techniques have made millisecond simulations possible. In 2010, using an aggregate of 1.5 ms of data, Voelz *et. al.* reported the first simulation describing the folding of a millisecond folder in implicit solvent, from data generated by the distributed computing network Folding@home [16,17]. Later that year, Bowman *et. al.* reported a millisecond timescale in the folding dynamics of lambda repressor in explicit solvent [18]. More recently, using 30 milliseconds of aggregate data, Voelz *et. al.* studied the folding of ACBP on the 10 millisecond timescale, revealing that an experimentally observed folding intermediate was in fact a complex, heterogeneous ensemble of structures [19]. Finally, with the advent of ANTON, a computer specialized for protein simulation, the first single trajectory of millisecond length was reported near the end of 2010 [20], making it possible to predict folding times of up to 100 μs from a single trajectory.

While challenging, generating enough sampling in an accurate forcefield does not constitute the end of the road for a folding simulation. Another major challenge is gaining scientific insight from the simulation – turning data into knowledge. Insights gained from simulations have already begun to shape the protein folding field, through connection to experiment and analysis of the simulations themselves [18,19,21–25]. These preliminary studies have revealed that the analysis of simulation data is difficult, with cases where certain techniques have led researchers to believe results inconsistent with their raw simulation data [20,26,27]. The simulation community needs to develop general, robust, and easy to use data analysis tools to continue towards the goal of understanding folding.

In this review, we briefly explain how the sampling problem has been overcome, and why we can expect the future to yield even longer simulations more efficiently. As sampling becomes less of an issue, a new challenge in folding simulations raises its head – given the massive amount of data extensive sampling provides, how does one make sense of it all? MD simulations are a high-dimensional time series, and therefore present a "Big Data" challenge [28,29]. We expect the techniques of data analysis will be the new limiting factor in the quest to understand folding through molecular simulation.

Overcoming the Barrier to Sampling

Over the past decade, the system sizes and time scales accessible to protein simulations have grown exponentially (Fig. 2). This gain has been achieved through progress on three main fronts: efficient parallelization of MD codes, specialized hardware, and statistical analysis of multiple independent trajectories. While there are many techniques designed to accelerate dynamic sampling via biasing the system in some way (*e.g.* replica exchange, metadynamics, aMD, string method, *etc.* [30–34]) each has advantages and disadvantages, and detailed discussion of these methods is beyond the scope of this review. Here, we focus on unbiased molecular dynamics simulations capable of describing realistic system kinetics.

Traditionally, such calculations were accelerated by dividing up single, long simulations across many processors. This tactic is inherently challenging, since each individual machine must communicate with one another, and as the number of machines grows so does the necessary time spent communicating, rather than actually simulating. While significant advances in mitigating communication costs have been made [35–38], such costs nonetheless place a hard upper limit on the efficiency of such techniques.

A second avenue pursued to enhance sampling has been hardware development. GPU technology, adapted from the video gaming industry, has resulted in tremendous acceleration of simulations at very low cost [39–46]. GPU simulations, like their CPU counterparts, are bounded by an upper limit of possible parallelism. Shaw and co-workers overcame these difficulties with ANTON, a special purpose machine, that has generated single trajectories one hundred times longer than previously reported simulations [47]. ANTON combines specialized computer chips and a fast network that allow it to generate simulations with unprecedented speed.

A different approach, Markov state models (MSMs), have recently become a practical alternative to overcoming computational difficulties associated with traditional single trajectory simulations [48–52]. In this paradigm, independent short simulations are generated and then aggregated in a statistical fashion, resulting in a complete model of the system dynamics. The MSM effectively pieces together this complete model from independent parts (trajectories), with each trajectory describing one small part of protein phase space – similar to how a complete picture in a jigsaw puzzle emerges by connecting many individual pieces. When combined in an MSM, one can predict kinetic phenomena on timescales much longer than the individual trajectories used to build the model.

Through this mechanism, the MSM facilitates efficient use of resources, since machine-level parallelism can be employed until it becomes inefficient due to communication costs, and then further parallelism can be gained by running independent trajectories on different machines. Furthermore, because the MSM framework naturally partitions the simulation into many independent parts, and it can provide feedback about which areas of protein phase space are undersampled. Simulations can then be intelligently placed to increase sampling in the areas where it is most needed, through a process called *adaptive sampling*, avoided wasteful simulation of processes that have already been witnessed with statistical confidence [53–55].

Analysis: The Final Challenge

The advances in sampling techniques, along with historical exponential increases in achievable sampling methods (Fig. 2), make us hopeful that protein folding simulations will become routine calculations for commodity hardware within a decade. Indeed, one can now simulate the folding of small proteins in explicit solvent at a rate of up to 100ns/day/GPU [44–46], such that a cluster of 100 GPUs can produce MSMs with the ability to predict the millisecond time scales in only three months. Given such extensive sampling, the next challenge simulators must face is that of data analysis – the process of turning information into knowledge. A successful analysis method should be able to reduce simulation data to its essential scientific features, without oversimplifying. It should let the data tell the story, discovering things that the investigator didn't think to look for and mitigating any biases she might have. Currently, the analysis techniques employed by the simulation community fall into two broad classes: 1) methods focused on finding reaction coordinates and associated transition states and 2) Markov State Models (Fig. 3).

In the reaction coordinate paradigm, one looks for a single coordinate capable of describing the progress from unfolded to folded structures, and builds a model for kinetics along that

coordinate [56–59]. This usually culminates in finding one or more transition state ensembles, usually defined to be the structures along the coordinate that have a 50% probability of proceeding to fold or unfold [60]. These structures are usually assumed to be kinetically relevant in the same way transition states in physical organic chemistry are, such that their geometry reflects the kinetics of the process. These methods are appealing because they reduce information down to a single coordinate – discarding orthogonal degrees of freedom – and identify specific structures (the transition states) to investigate.

Markov state models (MSMs), on the other hand, represent folding as first-order kinetics between a set of discrete states. Automatic methods exist to employ simulation to identify a set of states and parameterize discrete-time master equation describing dynamics on that state space [61–63]. MSMs simplify the simulation analysis by discarding very fast dynamics, below the so-called *lag time*. Because the lag time is tunable, this allows for multiple levels of resolution, from fine (order nanoseconds) to coarse (microseconds or longer). This tunable nature allows the same model to be simultaneously quantitatively accurate (high resolution) and comprehendible (low resolution), all within a common theoretical framework.

These two techniques do not always yield the same results; there has been significant disagreement between investigators studying the same simulation datasets with different techniques. One such disagreement is at the very heart of understanding how proteins fold – the question of whether folding occurs via a single, dominant pathway or many independent routes. This is one of the most basic questions in the study of folding, and has been actively debated for almost two decades [20,26,64–69].

For instance, in the folding of a WW domain, Shaw *et. al.* employed a state of the art technique to construct a reaction coordinate for the folding process [20,57]. From that coordinate, they concluded that the folding of the WW domain was mechanistically homogenous, always beginning with the first hairpin – this was in contrast to the previous WW simulations of Noé *et. al.* [70]. While the coordinate reproduced the correct folding time and was validated by a committor analysis (gave the correct probability that a give structure would fold before completely unfolding) [71,72], it failed to detect a parallel pathway, where the second hairpin of the WW domain folded first [26,27]. Later, the same technique was employed to analyze the folding of 12 small proteins, once again leading to the conclusion of predominantly single, serial folding pathways [67,68]. However, when MSM analysis was applied to these 12 simulations, a richer picture emerged [73] that suggested two-state models were inappropriate for half of the simulated systems.

In the above examples, the reaction coordinate scheme Shaw *et. al.* employed was capable only finding the highest-flux folding path, and ignored the others. In fact, reaction coordinates have great difficulties dealing with many of the features that make simulations appealing, such as their ability to elucidate parallel paths. Simulations provide a way to get information about the entire folding process, while reaction coordinates lead one to focus just on transition state(s). Further, reaction coordinates, by construction, discard potentially interesting dynamics orthogonal to the coordinate.

MSMs are not susceptible to these drawbacks, and are able to capture either simple or complex phenomena. While MSM techniques are well developed, some challenges do remain. In particular, the optimal manner in which to partition configuration space [74,75], choose a lag time, and perform adaptive sampling remain unknown. Effective heuristics are currently available for these problems [51,62,63], but systematic errors are often observed in MSMs, and improvements are certainly possible. Answering these challenges is the next

step in analysis methods development, and should go a long way towards generalizing MSM techniques beyond protein folding.

What Has and Can Be Learned from Simulations of Folding

Given that the sampling at millisecond timescales has been possible for only two years, and analysis methodology is still immature, unambiguous scientific results learned from atomic simulation have thus far been modest. It will be a major challenge in the next five years to turn advances in sampling and accuracy into scientific insight about how proteins fold.

Despite this relative immaturity, atomistic simulation has already begun to influence our view of protein folding. Detailed comparisons to experiment have been performed for many specific proteins, including villin, NTL9, WW domains, lambda repressor, ACBP, and the 12 fast folding proteins studied by Shaw [16,18–20,22,26,68,70,76,77]. Universally accepted generalities amongst these specific protein simulations have not yet emerged, though some have been suggested, for instance that folding kinetics might be hub-like [78–80], that folding proceeds via parallel paths [26,73], and that the unfolded state of proteins is compact [21]. These hypotheses are by no means completely vetted, and require validation through additional simulations and experiment.

The list of potential questions the folding field might hope to address through simulation is long. A few of the most exciting include

- Can we build models allowing for the detailed comparison of simulations to experiment in order to both test simulations and aid in the interpretation of experiments [81]? Further, simulations might be able to direct the design of future experiments, suggesting those with the greatest impact.
- With a detailed comparison of experiment in hand as tests of simulation accuracy, can we answer how do particular proteins fold? Why do so many proteins appear to fold in a two-state manner? What is the nature of “downhill” folding? Can we describe these in microscopic, physical terms?
- With the knowledge *the mechanism* of how particular proteins fold, we can learn how this mechanism is encoded in the inherent physical interaction of the amino acids in a given protein sequence?
- With the knowledge of how many individual proteins fold, can simulations help reveal general features of protein folding amongst broad groups of proteins (or ideally some general properties for all proteins)?

Much effort has been poured into advancing molecular simulation, and in this decade the fruits of that effort are coming to bear. Hopefully with continued progress in sampling and forcefields, combined with powerful analysis techniques, simulation can play a key role – alongside experiment and theory – in discovering how proteins fold.

Acknowledgments

DS and VSP acknowledge support from the Simbios NIH Center for Biomedical Computation (NIH U54 Roadmap GM072970), VSP acknowledges NIH (R01GM62828). TJL was supported by an NSF GRF, KAB was supported by a Stanford Graduate Fellowship.

References

1. Anfinsen C. Principles that Govern the Folding of Protein Chains. Science. 1973; 181:223–230. [PubMed: 4124164]

2. McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. Nature. 1977; 267:585–590. [PubMed: 301613]

3. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. Nat. Struct. Biol. 2002; 9:646–652. [PubMed: 12198485]

4. Dror RO, Dirks RM, Grossman JP, Xu H, Shaw DE. Biomolecular Simulation: A Computational Microscope for Molecular Biology. Ann Rev Biophys. 2012; 41:429–452. [PubMed: 22577825]

5. Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. P Natl Acad Sci Usa. 1987; 84:7524–7528.

6. Onuchic JN, Wolynes PG. Theory of protein folding. Curr. Op. Struct. Biol. 2004; 14:70–75.

7. Ghosh K, Ozkan SB, Dill KA. The Ultimate Speed Limit to Protein Folding Is Conformational Searching. J. Am. Chem. Soc. 2007; 129:11920–11927. [PubMed: 17824609]

8. Deechongkit S, Nguyen H, Jager M, Powers E, Gruebele M, Kelly J. β-Sheet folding mechanisms from perturbation energetics. Curr. Op. Struct. Biol. 2006; 16:94–101.

9. Plaxco KW, Simons KT, Ruczinski I, Baker D. Topology, Stability, Sequence, and Length: Defining the Determinants of Two-State Protein Folding Kinetics. Biochemistry. 2000; 39:11177–11183. [PubMed: 10985762]

10. Kollman P. Advances and continuing challenges in achieving realistic and predictive simulations of the properties of organic and biological molecules. Acc. Chem. Res. 1996; 29:461–469.

11. Ponder J, Case D. Force fields for protein simulations. Adv Protein Chem. 2003; 66:27–84. [PubMed: 14631816]

12. Best RB, Buchete N-V, Hummer G. Are Current Molecular Dynamics Force Fields too Helical? Biophysical J. 2008; 95:L07–L09.

13. Nerenberg PS, Head-Gordon T. Optimizing Protein–Solvent Force Fields to Reproduce Intrinsic Conformational Preferences of Model Peptides. J. Chem. Theory Comput. 2011; 7:1220–1230.

14. Beauchamp KA, Lin Y-S, Das R, Pande VS. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. J. Chem. Theory Comput. 2012; 8:1409–1414. [PubMed: 22754404]

15. Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, Shaw DE. Systematic Validation of Protein Force Fields against Experimental Data. PLoS ONE. 2012; 7:e32131. [PubMed: 22384157]

16. Voelz VA, Bowman GR, Beauchamp K, Pande VS. Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9(1–39). J. Am. Chem. Soc. 2010; 132:1526–1528. [PubMed: 20070076] Voelz and colleagues use distributed GPU computing to simulate the millisecond-folder NTL9, marking the first simulation study capable of describing millisecond timescales. An MSM analysis finds that the folding proceeds along multiple pathways, via discrete states.

17. Shirts M, Pande VS. Screen Savers of the World Unite! Science. 2000; 290:1903–1904. [PubMed: 17742054]

18. Bowman GR, Voelz VA, Pande VS. Atomistic folding simulations of the five-helix bundle protein λ(6–85). J. Am. Chem. Soc. 2011; 133:664–667. [PubMed: 21174461]

19. Voelz VA, Jager M, Yao S, Chen Y, Zhu L, Waldauer SA, Bowman GR, Friedrichs M, Bakajin O, Lapidus LJ, et al. Slow Unfolded-State Structuring in Acyl-CoA Binding Protein Folding Revealed by Simulation and Experiment. J. Am. Chem. Soc. 2012 Studying an MSM predictive of ACPB's 10 ms folding process, the authors were able to show that an experimentally observed folding intermediate is in fact a heterogeneous ensemble of structures. Represents the longest timescale MD study of folding to date.

20. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, et al. Atomic-Level Characterization of the Structural Dynamics of Proteins. Science. 2010; 330:341–346. [PubMed: 20947758] The ANTON machine was used to simulate the reversible folding of the model proteins HP35 and FiP35. In addition, the near-native dynamics of BPTI was studied using an unprecedented millisecond trajectory.

21. Voelz VA, Singh VR, Wedemeyer WJ, Lapidus LJ, Pande VS. Unfolded-State Dynamics and Structure of Protein L Characterized by Simulation and Experiment. J. Am. Chem. Soc. 2010; 132:4702–4709. [PubMed: 20218718]

22. Morcos F, Chatterjee S, McClendon CL, Brenner PR, López-Rendón R, Zintsmaster J, Ercsey-Ravasz M, Sweet CR, Jacobson MP, Peng JW, et al. Modeling Conformational Ensembles of Slow Functional Motions in Pin1-WW. PLoS Comput Biol. 2010; 6:e1001015. [PubMed: 21152000]
23. Prigozhin MB, Gruebele M. The Fast and the Slow: Folding and Trapping of λ 6–85. J. Am. Chem. Soc. 2011; 133:19338–19341. [PubMed: 22066714]
24. Piana S, Sarkar K, Lindorff-Larsen K, Guo M, Gruebele M, Shaw DE. Computational Design and Experimental Testing of the Fastest-Folding β-Sheet Protein. J. Mol. Biol. 2011; 405:43–48. [PubMed: 20974152]
25. Chung HS, McHale K, Louis JM, Eaton WA. Single-Molecule Fluorescence Experiments Determine Protein Folding Transition Path Times. Science. 2012; 335:981–984. [PubMed: 22363011]
26. Lane TJ, Bowman GR, Beauchamp K, Voelz VA, Pande VS. Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. J. Am. Chem. Soc. 2011; 133:18413–18419. [PubMed: 21988563] The authors show, using an MSM, that the single pathway hypothesis advanced in Ref. 20 was inconsistent with the raw data.
27. Krivov SV. The Free Energy Landscape Analysis of Protein (FIP35) Folding Dynamics. J. Phys. Chem. B. 2011; 115:12315–12324. [PubMed: 21902225]
28. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Computational solutions to largescale data management and analysis. Nat. Rev. Genet. 2010; 11:647–657. [PubMed: 20717155]
29. Stone, J.; Vandivort, K.; Schulten, K. Lecture Notes in Computer Science. Springer; 2011. Immersive out-of-core visualization of large-size and longtimescale molecular dynamics trajectories; p. 1-12.
30. E WW, Vanden-Eijnden EE. Transition-path theory and path-finding algorithms for the study of rare events. Annu. Rev. Phys. Chem. 2010; 61:391–420. [PubMed: 18999998]
31. Hamelberg D, Mongan J, McCammon JA. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. J. Chem. Phys. 2004; 120:11919. [PubMed: 15268227]
32. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. Chem. Phys. Lett. 1999; 314:141–151.
33. Paschek D, Nymeyer H, García AE. Replica exchange simulation of reversible folding/unfolding of the Trp-cage miniprotein in explicit solvent: On the structure and possible role of internal water. J Struct Biol. 2007; 157:524–533. [PubMed: 17293125]
34. Piana S, Laio A. A Bias-Exchange Approach to Protein Folding. J. Phys. Chem. B. 2007; 111:4553–4559. [PubMed: 17419610]
35. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J. Chem. Theory Comput. 2008; 4:435–447.
36. Chow E, Rendleman CA, Bowers KJ, Dror RO, Hughes DH, Gullingsrud J, Sacerdoti FD, Shaw DE. Desmond performance on a cluster of multicore processors. 2008
37. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kalé L, Schulten K. Scalable molecular dynamics with NAMD. J. Comput. Chem. 2005; 26:1781–1802. [PubMed: 16222654]
38. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. J. Comput. Chem. 2005; 26:1668–1688. [PubMed: 16200636]
39. Stone JE, Phillips JC, Freddolino PL, Hardy DJ, Trabuco LG, Schulten K. Accelerating molecular modeling applications with graphics processors. J. Comput. Chem. 2007; 28:2618–2640. [PubMed: 17894371]
40. Phillips JC, Stone JE. Probing biomolecular machines with graphics processors. Commun Acm. 2009; 52:34.
41. Friedrichs MS, Eastman P, Vaidyanathan V, Houston M, Legrand S, Beberg AL, Ensign DL, Bruns CM, Pande VS. Accelerating molecular dynamic simulation on graphics processing units. J. Comput. Chem. 2009; 30:864–872. [PubMed: 19191337]
42. Eastman P, Pande VS. OpenMM: A Hardware-Independent Framework for Molecular Simulations. Comput Sci Eng. 2010; 12:34–39.

43. Stone JE, Hardy DJ, Ufimtsev IS, Schulten K. GPU-accelerated molecular modeling coming of age. J Mol Graph Model. 2010; 29:116–125. [PubMed: 20675161] A review of how GPU-accelerated codes are changing the possibilities for molecular simulation.
44. Götz AW, Williamson MJ, Xu D, Poole D, Le Grand S, Walker RC. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. J. Chem. Theory Comput. 2012; 8:1542–1555. [PubMed: 22582031]
45. Harvey MJ, Giupponi G, De Fabritiis G. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. J. Chem. Theory Comput. 2009; 5:1632–1639.
46. Harvey MJ, De Fabritiis G. An Implementation of the Smooth Particle Mesh Ewald Method on GPU Hardware. J. Chem. Theory Comput. 2009; 5:2371–2377.
47. Shaw DE, Deneroff MM, Dror RO, Kuskin JS, Larson RH, Salmon JK, Young C, Batson B, Bowers KJ, Chao JC, et al. Anton, a special-purpose machine for molecular dynamics simulation. Commun Acm. 2008; 51:91–97. A technical description of the ANTON machine, a piece of specialized hardware capable of generating MD trajectories of unprecedented length.
48. Noé F, Fischer S. Transition networks for modeling the kinetics of conformational change in macromolecules. Curr. Op. Struct. Biol. 2008; 18:154–162.
49. Bowman GR, Beauchamp KA, Boxer G, Pande VS. Progress and challenges in the automated construction of Markov state models for full protein systems. J. Chem. Phys. 2009; 131:124101. [PubMed: 19791846]
50. Pande VS, Beauchamp K, Bowman GR. Everything you wanted to know about Markov State Models but were afraid to ask. Methods. 2010; 52:99–105. [PubMed: 20570730]
51. Prinz J-H, Keller B, Noé F. Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables. Phys. Chem. Chem. Phys. 2011; 13:16912–16927. [PubMed: 21858310]
52. Prinz J-H, Wu H, Sarich M, Keller B, Senne M, Held M, Chodera JD, Schütte C, Noé F. Markov models of molecular kinetics: Generation and validation. J. Chem. Phys. 2011; 134:174105. [PubMed: 21548671] Noe and colleagues describe the Markov State Model paradigm for analyzing molecular simulations. They outline the underlying theory, state decomposition techniques, and methods for model validation.
53. Singhal N, Pande VS. Error analysis and efficient sampling in Markovian state models for molecular dynamics. J. Chem. Phys. 2005; 123:204909. [PubMed: 16351319]
54. Bowman GR, Ensign DL, Pande VS. Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models. J. Chem. Theory Comput. 2010; 6:787–794. [PubMed: 23626502]
55. Weber JK, Pande VS. Characterization and Rapid Sampling of Protein Folding Markov State Model Topologies. J. Chem. Theory Comput. 2011; 7:3405–3411. [PubMed: 22140370]
56. van Kampen N. Stochastic processes in physics and chemistry. 2007
57. Best RB, Hummer G. Reaction coordinates and rates from transition paths. P Natl Acad Sci Usa. 2005; 102:6732–6737. The authors present a Bayesian formalism for optimizing a single reaction coordinate. They apply the formalism to understand the transition state of a Go-model of protein folding. This method was later used extensively by Shaw to analyze long trajectories of folding.
58. Best RB, Hummer G. Coordinate-dependent diffusion in protein folding. P Natl Acad Sci Usa. 2010; 107:1088–1093.
59. Best RB, Hummer G. Diffusion models of protein folding. Phys. Chem. Chem. Phys. 2011; 13:16902. [PubMed: 21842082]
60. Hummer G. From transition paths to transition states and rate coefficients. J. Chem. Phys. 2004; 120:516–523. [PubMed: 15267886]
61. Bowman GR, Huang X, Pande VS. Using generalized ensemble simulations and Markov state models to identify conformational states. Methods. 2009; 49:197–201. [PubMed: 19410002]
62. Beauchamp KA, Bowman GR, Lane TJ, Maibaum L, Haque IS, Pande VS. MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. J. Chem. Theory Comput. 2011; 7:3412–3419. [PubMed: 22125474]
63. Senne M, Trendelkamp-Schroer B, Mey ASJS, Schütte C, Noé F. EMMA - A software package for Markov model building and analysis. J. Chem. Theory Comput. 2012

64. Sali A, Shakhnovich E, Karplus M. How does a protein fold? Nature. 1994; 369:248–251. [PubMed: 7710478]
65. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nat. Struct. Biol. 1997; 4:10–19. [PubMed: 8989315]
66. Baldwin RL. The nature of protein folding pathways: the classical versus the new view. J Biomol Nmr. 1995; 5:103–109. [PubMed: 7703696]
67. Sosnick TR, Hinshaw JR. How Proteins Fold. Science. 2011; 334:464–465. [PubMed: 22034424]
68. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How Fast-Folding Proteins Fold. Science. 2011; 334:517–520. [PubMed: 22034434] Using the ANTON machine, Shaw and colleagues simulate the reversible folding of 12 proteins whose folding times range from the hundred-nanosecond to hundred-microsecond regimes.
69. Englander SW, Mayne L, Krishna MMG. Protein folding and misfolding: mechanism and principles. Quart. Rev. Biophys. 2008; 40
70. Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. P Natl Acad Sci Usa. 2009; 106:19011–19016. In this work, Noe et al use 180 short MD simulations to model the folding of the Pin WW protein. They describe a modeling framework that quantifies mechanistic questions such as order of events. They find that the folding of Pin WW proceeds along parallel pathways, and that the folding is slowed by the presence of register-shifted trap states.
71. Du R, Pande VS, Grosberg AY, Tanaka T, Shakhnovich ES. On the transition coordinate for protein folding. J. Chem. Phys. 1998; 108:334.
72. Chodera J, Pande V. Splitting Probabilities as a Test of Reaction Coordinate Choice in Single-Molecule Experiments. Phys. Rev. Lett. 2011; 107
73. Beauchamp KA, McGibbon R, Lin Y-S, Pande VS. Simple few-state models reveal hidden complexity in protein folding. P Natl Acad Sci Usa. 2012
74. Buchete N-V, Hummer G. Coarse Master Equations for Peptide Folding Dynamics. J. Phys. Chem. B. 2008; 112:6057–6069. [PubMed: 18232681]
75. Schuette C, Noé F, Lu J, Sarich M, Vanden-Eijnden E. Markov state models based on milestoning. J. Chem. Phys. 2011; 134:204105. [PubMed: 21639422]
76. Beauchamp KA, Ensign DL, Das R, Pande VS. Quantitative comparison of villin headpiece subdomain simulations and triplet–triplet energy transfer experiments. P Natl Acad Sci Usa. 2011; 108:12734.
77. Liu Y, Strümpfer J, Freddolino PL, Gruebele M, Schulten K. Structural Characterization of λ-Repressor Folding from All-Atom Molecular Dynamics Simulations. J. Phys. Chem. Lett. 2012; 3:1117–1123. [PubMed: 22737279]
78. Rao F, Caflisch A. The Protein Folding Network. J. Mol. Biol. 2004; 342:299–306. [PubMed: 15313625]
79. Bowman GR, Pande VS. Protein folded states are kinetic hubs. P Natl Acad Sci Usa. 2010; 107:10890–10895.
80. Lane TJ, Pande VS. A simple model predicts experimental folding rates and a hub-like topology. J. Phys. Chem. B. 2012; 116:6764–6774. [PubMed: 22452581]
81. Noé F, Doose S, Daidone I, Löllmann M, Sauer M, Chodera JD, Smith JC. Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments. P Natl Acad Sci Usa. 2011; 108:4822.
82. Pande V, Baker I, Chapman J, Elmer S, Khaliq S, Larson S, Rhee Y, Shirts M, Snow C, Sorin E, et al. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. Biopolymers. 2003; 68:91–109. [PubMed: 12579582]
83. Snow CD, Qiu L, Du D, Gai F, Hagen SJ, Pande VS. Trp zipper folding kinetics by molecular dynamics and temperature-jump spectroscopy. P Natl Acad Sci Usa. 2004; 101:4077.
84. Snow CD, Nguyen H, Pande VS, Gruebele M. Absolute comparison of simulated and experimental protein-folding dynamics. Nature. 2002; 420:102–106. [PubMed: 12422224]
85. Snow CD, Zagrovic B, Pande VS. The Trp Cage: Folding Kinetics and Unfolded State Topology via Molecular Dynamics Simulations. J. Am. Chem. Soc. 2002; 124:14548–14549. [PubMed: 12465960]

86. Zagrovic B, Snow CD, Shirts MR, Pande VS. Simulation of Folding of a Small Alpha-helical Protein in Atomistic Detail using Worldwide-distributed Computing. J. Mol. Biol. 2002; 323:927–937. [PubMed: 12417204]

87. Ensign DL, Kasson PM, Pande VS. Heterogeneity Even at the Speed Limit of Folding: Largescale Molecular Dynamics Study of a Fast-folding Variant of the Villin Headpiece. J. Mol. Biol. 2007; 374:806–816. [PubMed: 17950314]

88. Ensign DL, Pande VS. The Fip35 WW domain folds with structural and mechanistic heterogeneity in molecular dynamics simulations. Biophys J. 2009; 96:L53–L55. [PubMed: 19383445]

89. Freddolino PL, Schulten K. Common Structural Transitions in Explicit-Solvent Simulations of Villin Headpiece Folding. Biophys J. 2009; 97:2338–2347. [PubMed: 19843466]

90. Freddolino PL, Liu F, Gruebele M, Schulten K. Ten-Microsecond Molecular Dynamics Simulation of a Fast-Folding WW Domain. Biophys J. 2008; 94:L75–L77. [PubMed: 18339748]

91. Freddolino PL, Park S, Roux B, Schulten K. Force Field Bias in Protein Folding Simulations. Biophys J. 2009; 96:3772–3780. [PubMed: 19413983]

92. Rhee YM, Sorin EJ, Jayachandran G, Lindahl E, Pande VS. Simulations of the role of water in the protein-folding mechanism. Proceedings of the National Academy of Sciences. 2004; 101:6456–6461. A pioneering study of MD protein folding simulations. The authors employ a supercomputer to witness the earliest stages in the folding of villin, marking the first time atomic simulation was employed to study protein folding directly.

93. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1- microsecond simulation in aqueous solution. Science. 1998; 282:740–744. [PubMed: 9784131]

Highlights

- Millisecond simulations
- Hardware and software advances
- Simulation to understanding

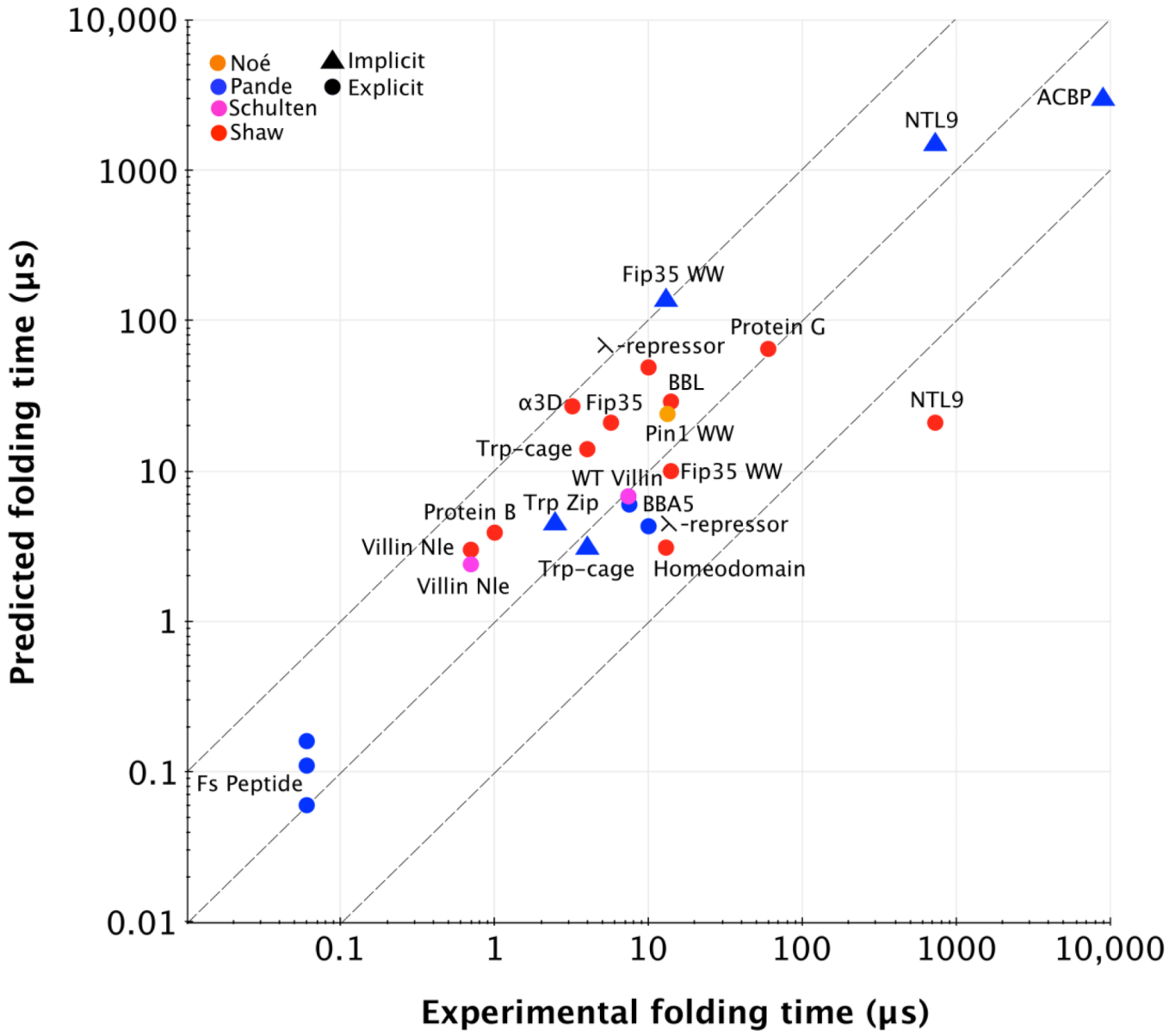

Figure 1.

Comparison of predicted and experimentally measured folding times. Central dashed line indicates perfect agreement, outside lines are within one order of magnitude of perfect agreement. Given that experimental folding times can vary over more than an order of magnitude given different conditions (temperature, salt, pH, etc.), as well as uncertainties associated with measuring experimental and simulated folding times, an order of magnitude agreement is close to the upper limit of accuracy one might expect. Data from [16,18–20,68,77,82–92].

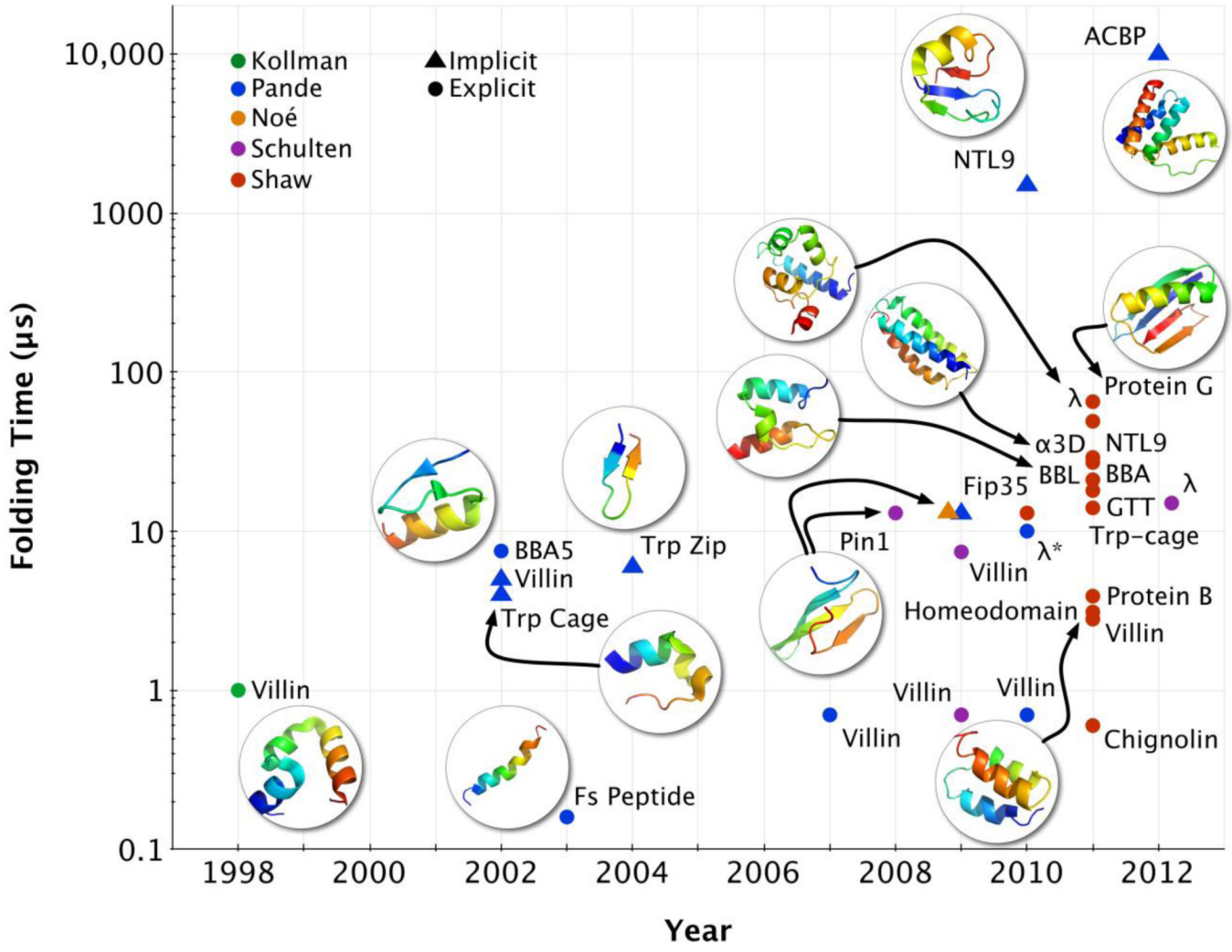

Figure 2.

The folding times accessible by simulation have increased exponentially over the past decade. Shown are all protein folding simulations conducted using unbiased, all-atom MD in empirical forcefields reported in the literature. Some folding times for the same protein differ, due to various mutations. For lambda marked with a (*), the longest timescale seen in that simulation, which was not the folding time, occurred on the order of 10 ms [18,23]. Data are same as Figure 1, with [93].

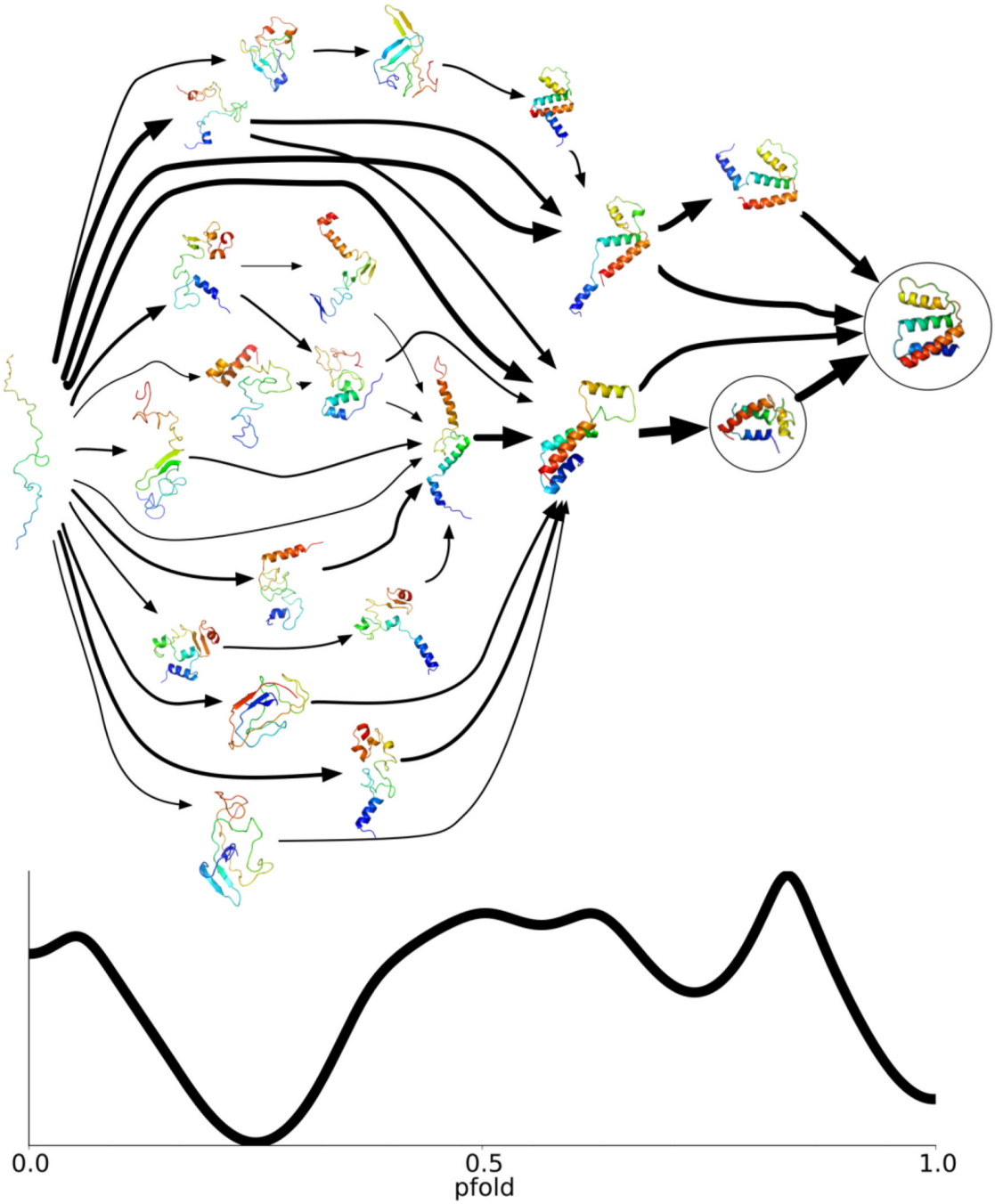

Figure 3.
Two methods of data analysis, the MSM (top) and reaction coordinate (bottom), shown for the same system (ACBP) [19]. The MSM represents folding as interconversion between structurally similar states, and can be illustrated as flow through a network. Reaction coordinates attempt to depict folding as progress along a single degree of freedom, such as the committors (P_{fold}, shown). The MSM picture is more detailed, can capture parallel paths, has tunable resolution, and connects naturally to experiment – all advantages over the coordinate-based approach.