Using seeding methods for efficient and accurate simulation of folding dynamics

Hongbin Wan

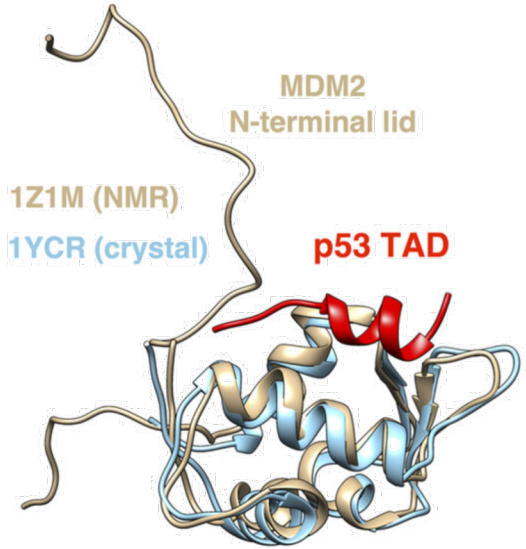
Vincent Voelz lab

Temple University

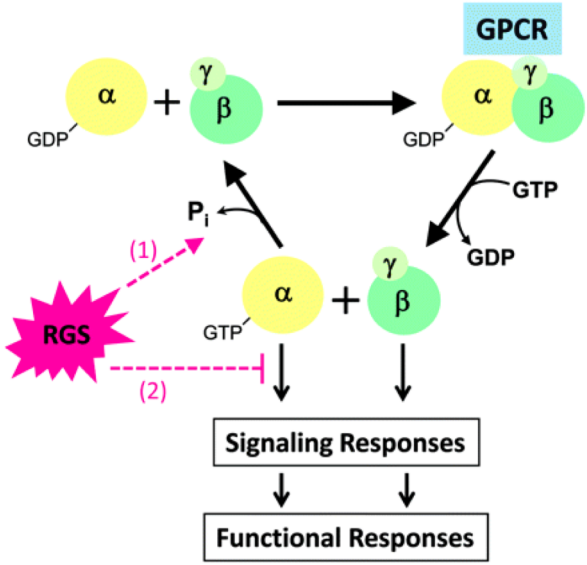
Protein Folding Consortium Workshop, 2018

Protein conformational dynamics are directly linked to protein function in many important biological processes

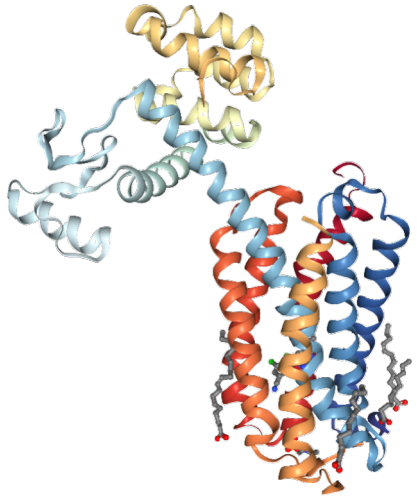
- protein-protein interaction



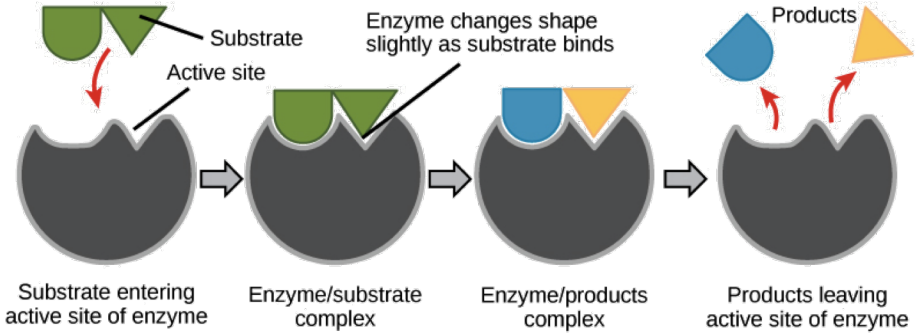
- Cell signaling



- allosteric regulation

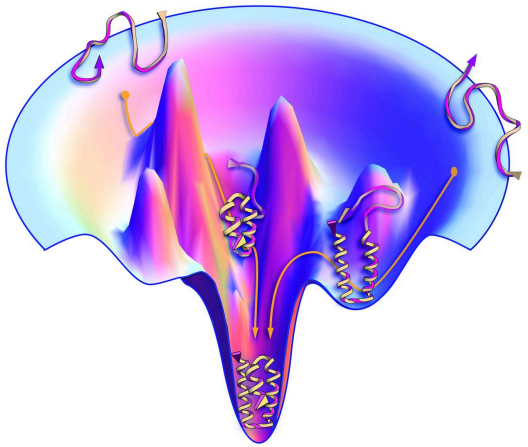


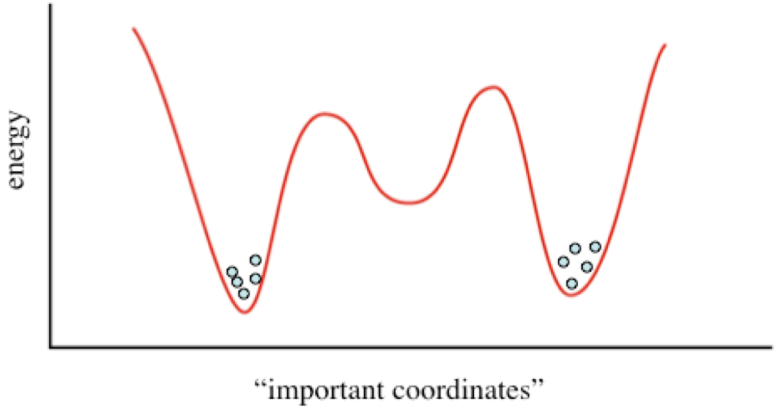
- Enzyme mechanism



Mukherjee, S., Pantelopulos, G. A., & Voelz, V. A. (2016). Markov models of the apo-MDM2 lid region reveal diffuse yet two-state binding dynamics and receptor poses for computational docking. *Nature Publishing Group*, 1–10.
Zhang, P., & Mende, U. (2011). Regulators of G-protein signaling in the heart and their potential as therapeutic targets. *Circulation Research*, *109*(3), 320–333.
https://www.khanacademy.org/science/biology/energy-and-enzymes/introduction-to-enzymes/a/enzymes-and-the-active-site

It is challenging to simulate conformational dynamics using MD



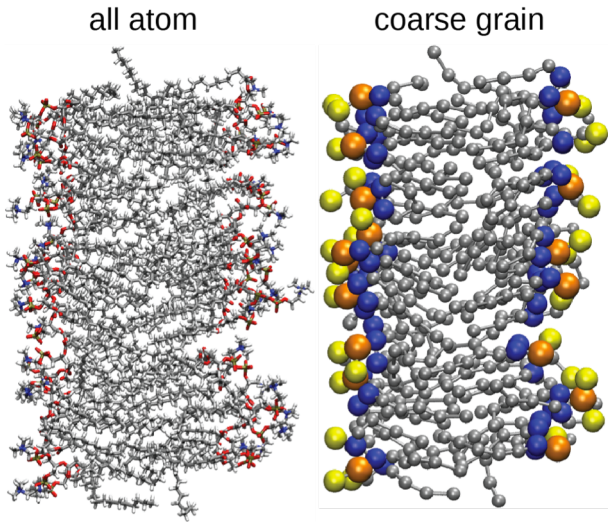


- Long timescale is needed to capture biologically important processes.
- Energy landscapes are rugged and high-dimensional.
- Important degrees of freedom are hard to define ahead of time.
- Needs lots of samplings (time, computational resources) to explore all reaction coordinates.

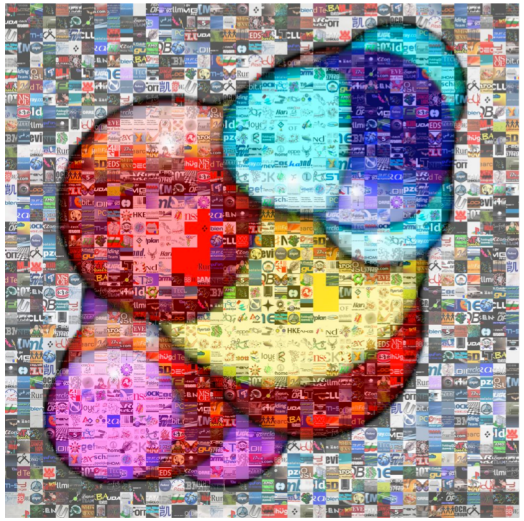
Q: how could we deal with these challenges?

Dill, K. A., & MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science (New York, N.Y.)*, *338*(6110), 1042–1046.

Many methods exist to address these sampling problems



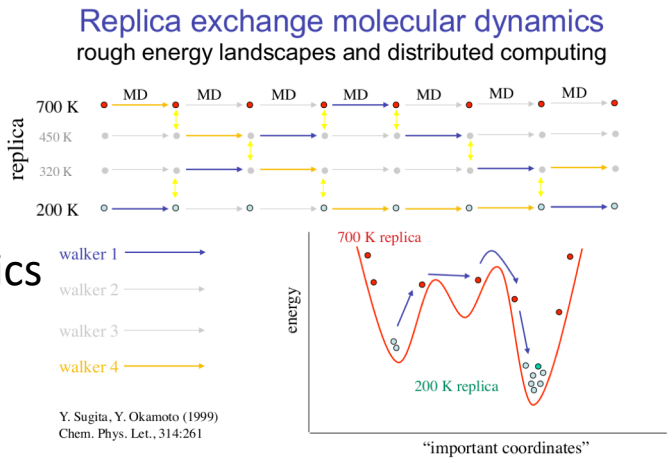
Coarse-grain model



Special purpose hardware

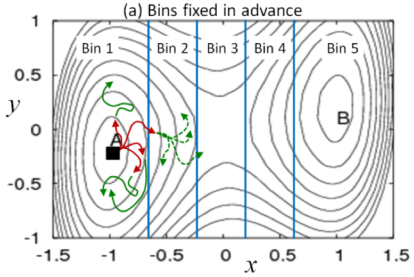
Replica-exchange

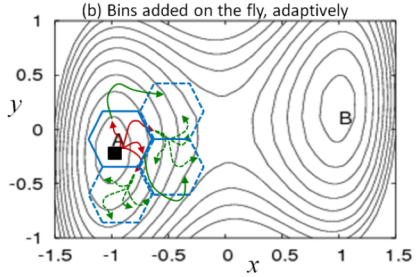
- Mainly thermodynamics



Weighted Ensemble

- Mainly kinetics

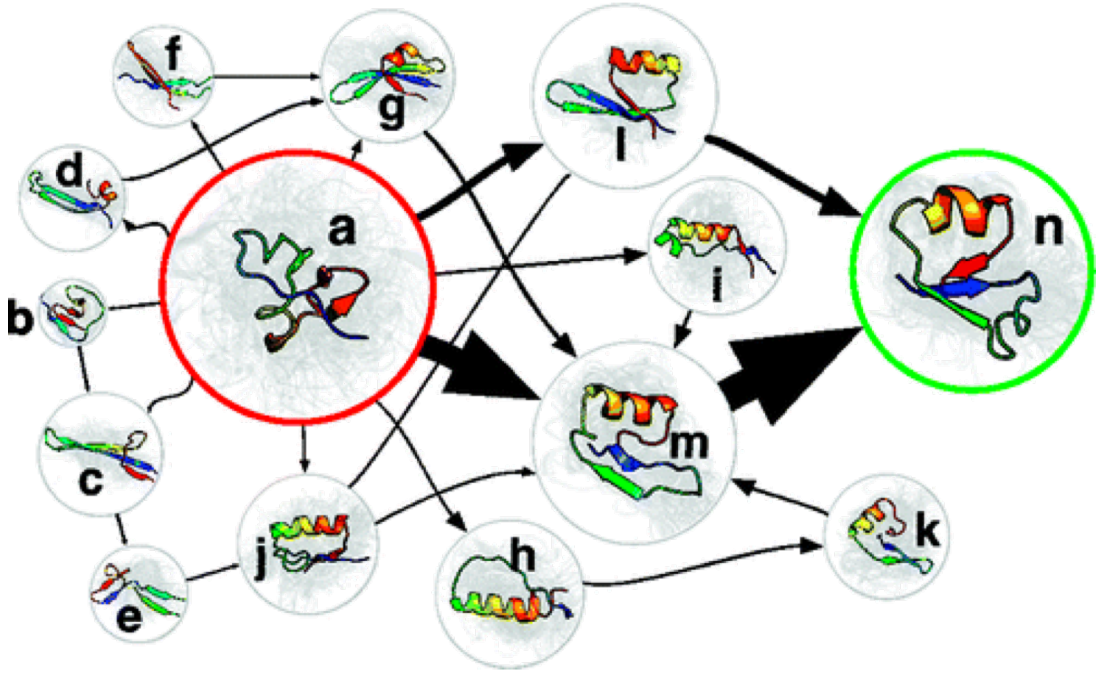




Enhanced sampling methods

Markov State models are a good way to capture both thermodynamics and kinetics

- Top 10 folding pathways of MSM of N-terminal domain of the crystal structure of ribosomal protein L9(NTL9)



- **MSMs estimate long time dynamics from large numbers of short trajectories (local equilibrium).**

Voelz, V. A., Bowman, G. R., Beauchamp, K., & Pande, V. S. (2010). Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *Journal of the American Chemical Society*, *132*(5), 1526–1528.
Husic, B. E., & Pande, V. S. (2018). Markov State Models: From an Art to a Science. *Journal of the American Chemical Society*, *140*(7), 2386–2396.

Suppose we already have constructed an MSM.

Can we leverage this information to build new MSMs?

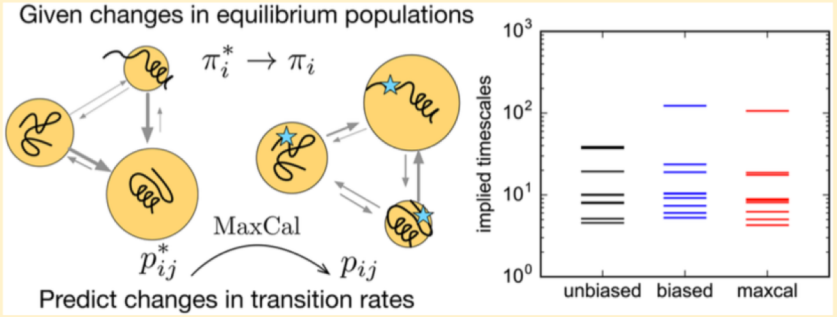
One approach: Using **maximum caliber** (MaxCal) methods to infer perturbation effects

A Maximum-Caliber Approach to Predicting Perturbed Folding Kinetics Due to Mutations

Hongbin Wan, Guangfeng Zhou, and Vincent A. Voelz*

Department of Chemistry, Temple University, Philadelphia, Pennsylvania 19122, United States

Supporting Information

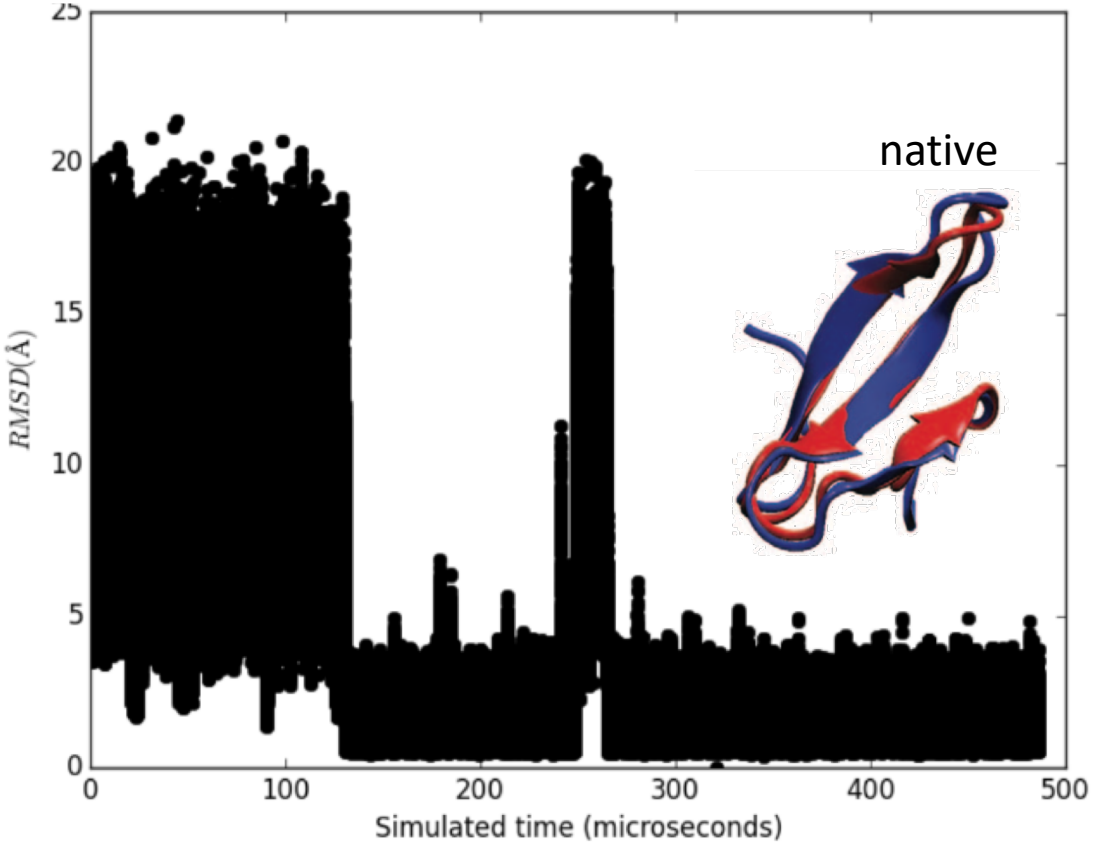


ABSTRACT: We present a maximum-caliber method for inferring transition rates of a Markov state model (MSM) with perturbed equilibrium populations given estimates of state populations and rates for an unperturbed MSM. It is similar in spirit to previous approaches, but given the inclusion of prior information, it is more robust and simple to implement. We examine its performance in simple biased diffusion models of kinetics and then apply the method to predicting changes in folding rates for several highly nontrivial protein folding systems for which non-native interactions play a significant role, including (1) tryptophan variants of the GB1 hairpin, (2) salt-bridge mutations of the Fs peptide helix, and (3) MSMs built from ultralong folding trajectories of FiP35 and GTT variants of the WW domain. In all cases, the method correctly predicts changes in folding rates, suggesting the wide applicability of maximum-caliber approaches to efficiently predict how mutations perturb protein conformational dynamics.

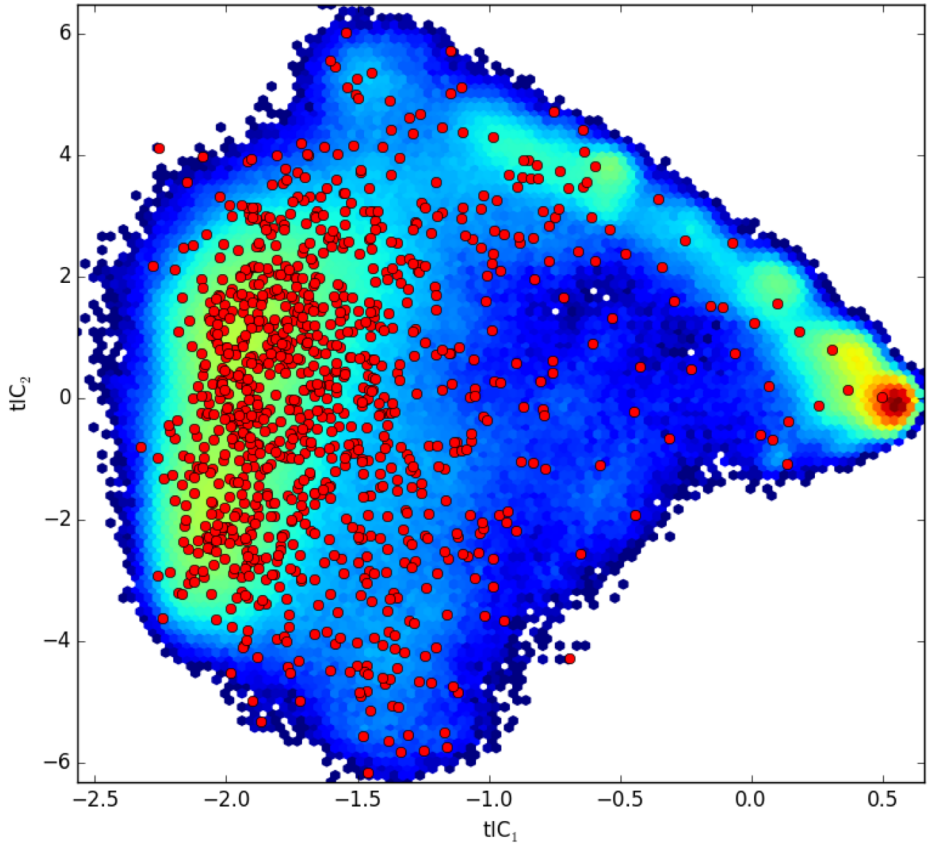
WAN, H., Zhou, G., & Voelz, V. A. (2016). A Maximum-Caliber Approach to Predicting Perturbed Folding Kinetics Due to Mutations. *Journal of Chemical Theory and Computation*, acs.jctc.6b00938.

Another approach: adaptive seeding

Ultra-long trajectory of GTT ww-domain folding



MSM built from the trajectory

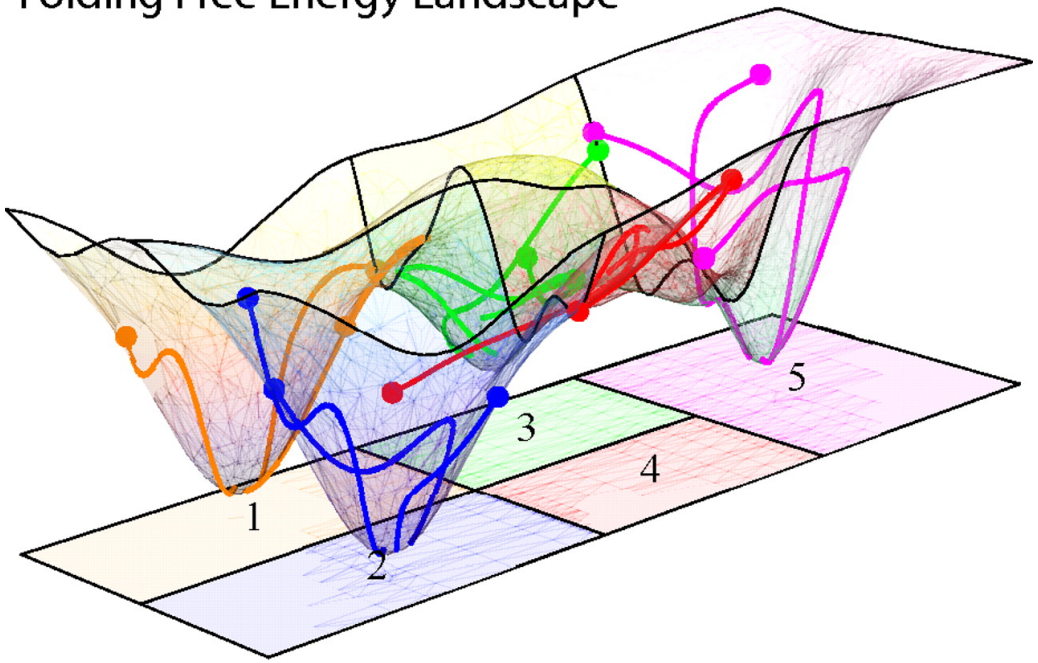


- Long trajectories spend 100s of microseconds in the folded state.
- Can we seed ensembles of short-trajectories from each MSM microstate to sample transitions?

Lindorff-Larsen, K., Piana, S., Dror, R. O., & Shaw, D. E. (2011). How Fast-Folding Proteins Fold. *Science (New York, N.Y.)*, *334*(6055), 517–520.
WAN, H., Zhou, G., & Voelz, V. A. (2016). A Maximum-Caliber Approach to Predicting Perturbed Folding Kinetics Due to Mutations. *Journal of Chemical Theory and Computation*, acs.jctc.6b00938.
Huang, X., Bowman, G. R., Bacallado, S., & Pande, V. S. (2009). Rapid equilibrium sampling initiated from nonequilibrium data. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(47), 19765–19769.

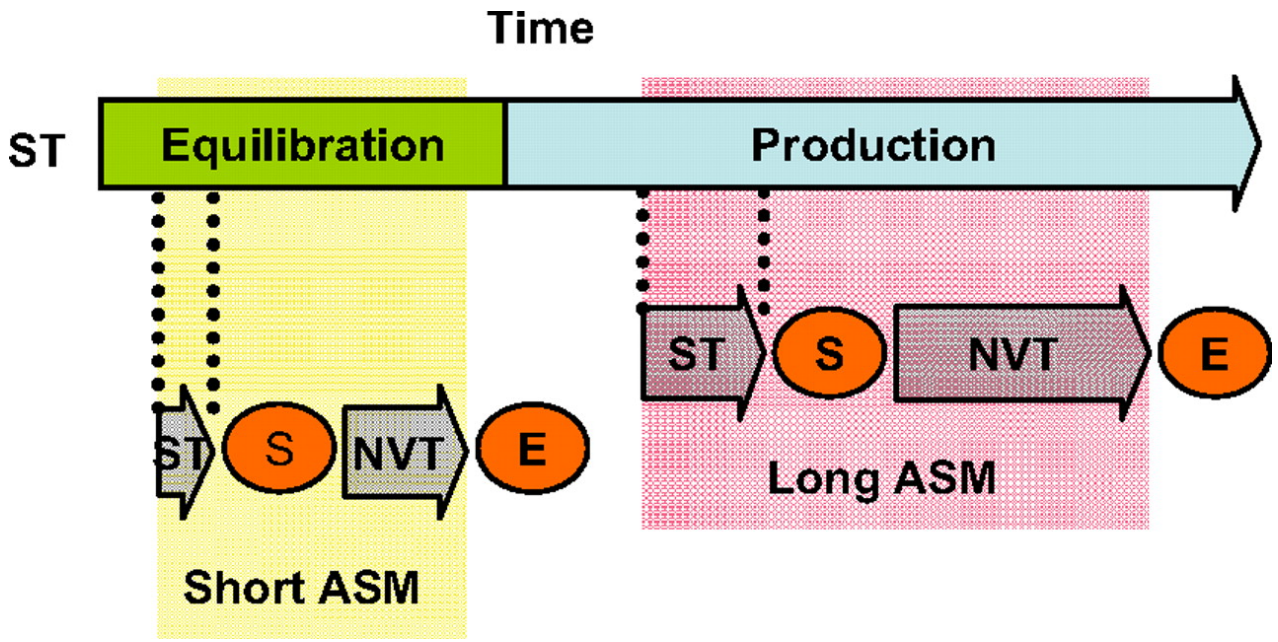
Adaptive seeding method (ASM)

Folding Free Energy Landscape



Metastable States
from Markov State Models

Schematic of the adaptive seeding scheme



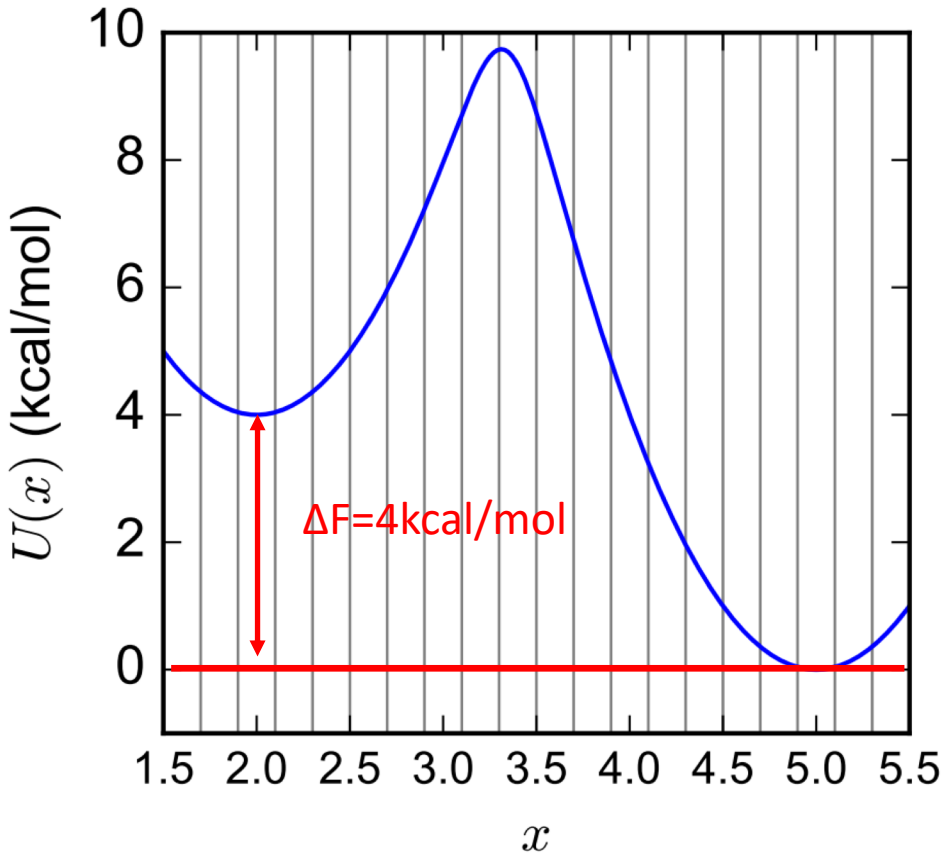
Huang, X., Bowman, G. R., Bacallado, S., & Pande, V. S. (2009). Rapid equilibrium sampling initiated from nonequilibrium data. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(47), 19765–19769.

Adaptive seeding of a 1-D two-state potential

$$U(x) = -\frac{2k_BT}{0.596}\ln[e^{-2(x-2)^2-2} + e^{-2(x-5)^2}]$$

- Monte Carlo random-walk with step-size=0.05 simulations.
- Seeds: 20bins(microstates) * 20trajs/bin.
- Slowest relaxation time ~5*10^6steps

Shoot 20 trajectories from each bin

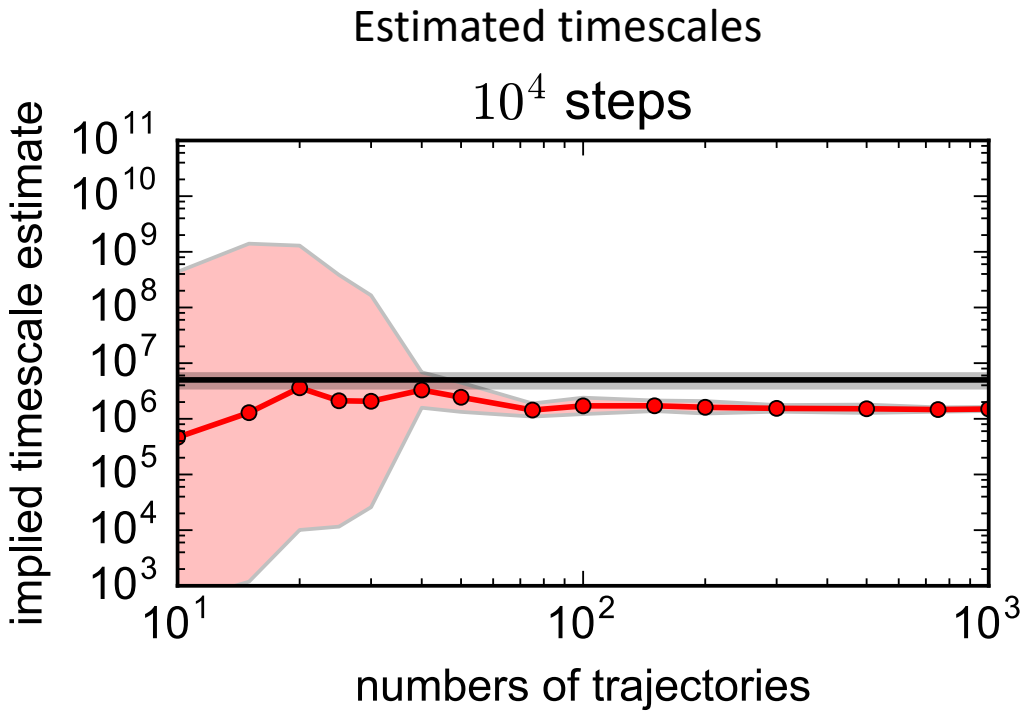


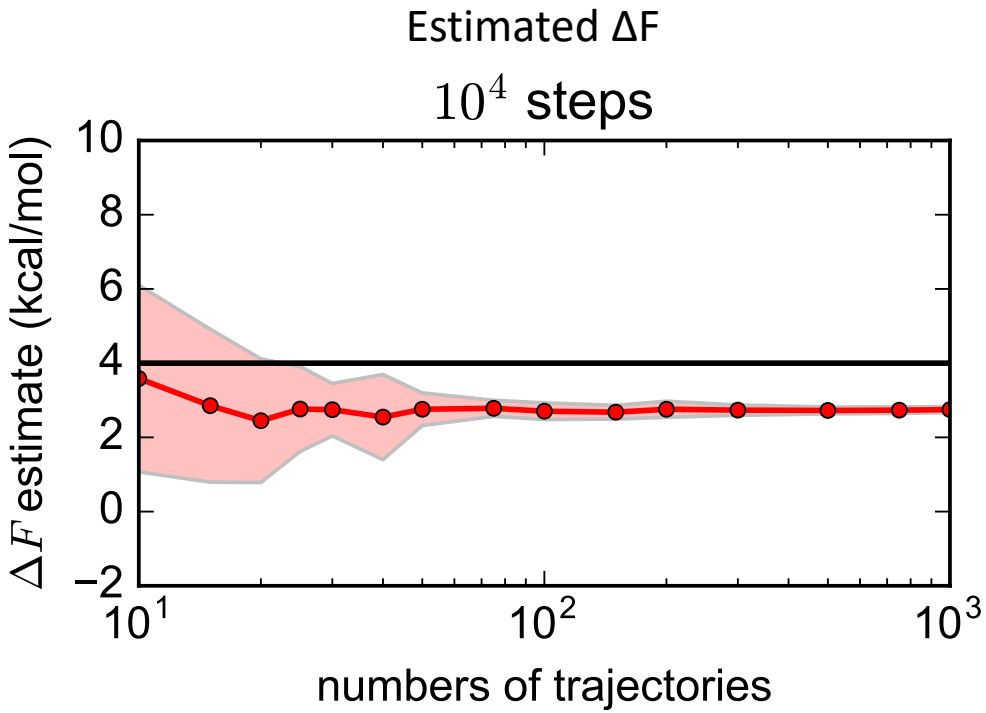
Problem

- The samples are not drawn from equilibrium
- But most MSM estimators assume that samples ARE drawn from equilibrium

Stelzl, L. S., Kells, A., Rosta, E., & Hummer, G. (2017). Dynamic Histogram Analysis To Determine Free Energies and Rates from Biased Simulations. *Journal of Chemical Theory and Computation*, *13*(12), 6328–6342.

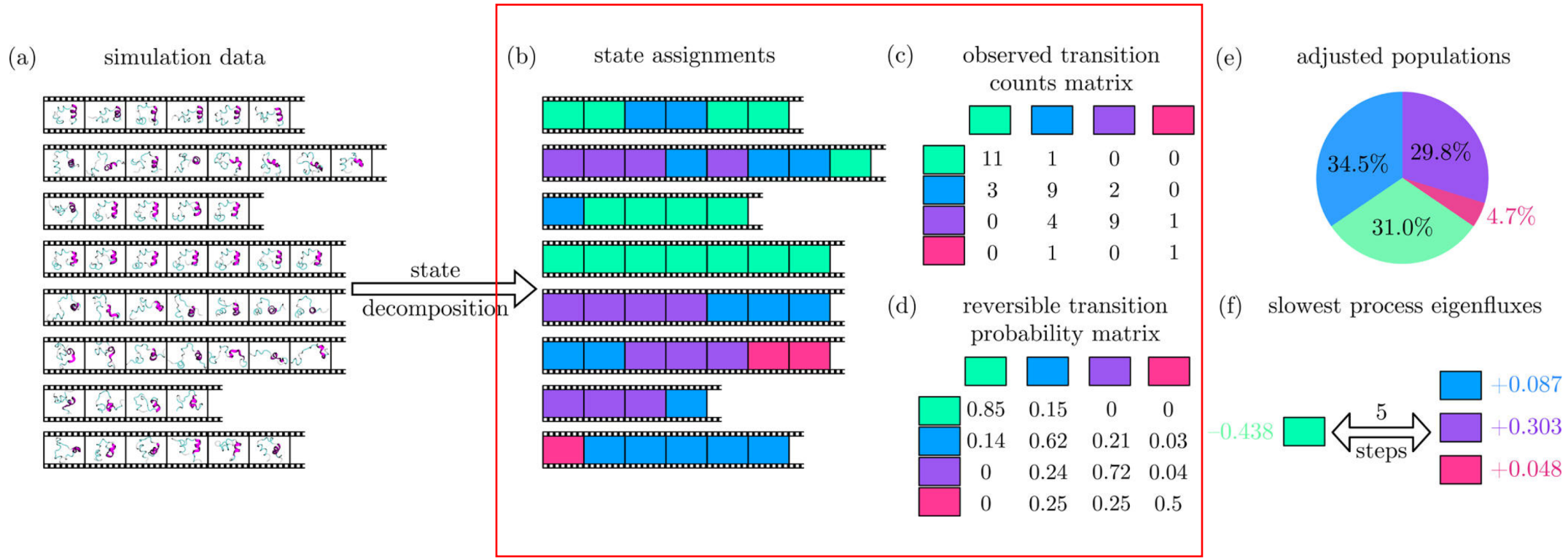
Standard MSM methods under-estimate timescales and free energies





Are there better estimators we can use?

MSMs are built by estimating transition probabilities from observed counts



Husic, B. E., & Pande, V. S. (2018). Markov State Models: From an Art to a Science. *Journal of the American Chemical Society*, *140*(7), 2386–2396.

Maximum-likelihood estimator (MLE)

$$L(p) = \prod_{i,j}^{k} p_{ij}^{c_{ij}}$$

Likelihood function

Detailed balance

$$\pi_i p_{ij} = \pi_j p_{ji}$$

Constraints

Maximize L

$$\pi_i = \sum_j \frac{c_{ij} + c_{ji}}{\frac{N_j}{\pi_j} + \frac{N_i}{\pi_i}}$$

Self-consistent iteration to solve for the equilibrium populations π_i

Then the transitions rates are:

$$p_{ij} = \frac{(c_{ij} + c_{ji})\pi_j}{N_j\pi_i + N_i\pi_j}$$

- c_{ij} is the number of counts from state *i* to *j* (in lag time τ)
- p_{ij} are transition probabilities from state *i* to *j*
- π_i are the equilibrium populations of state *i*

Prinz, J.-H., et al. Markov models of molecular kinetics: Generation and validation J. Chem. Phys. 134.17 (2011): 174105.
Metzner, P., F. Noe, and C. Schutte. Estimating the sampling error: Distribution of transition matrices and functions of transition matrices for given trajectory data. Phys. Rev. E 80.2 (2009): 021106.

Maximum-likelihood estimator (MLE) if populations π_i are known

$$L(p) = \prod_{i,j}^{k} p_{ij}^{c_{ij}}$$

Likelihood function

Detailed balance

$$\pi_i p_{ij} = \pi_j p_{ji}$$

Constraints

Maximize L

$$\lambda_i = \sum_j \frac{(c_{ij} + c_{ji})\pi_j \lambda_i}{\lambda_j \pi_i + \lambda_i \pi_j}$$

Self-consistent iteration to solve for the Lagrange multipliers λ_i

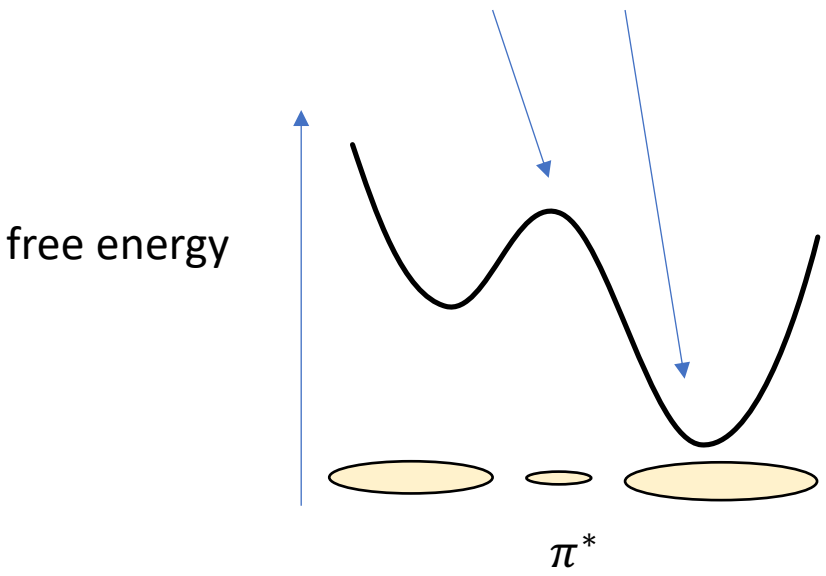
Then the transitions rates are:

$$p_{ij} = \frac{(c_{ij} + c_{ji})\pi_j}{\lambda_j \pi_i + \lambda_i \pi_j}$$

- c_{ij} is the number of counts from state *i* to *j* (in lag time τ)
- p_{ij} are transition probabilities from state *i* to *j*
- π_i are the equilibrium populations of state *i*

Maximum-likelihood estimator (MLE) with π_i^* -weighted trajectory counts

Suppose 100 trajectories are seeded at each state



some a priori estimate of the state populations at equilibrium

IDEA: weight the counts from each trajectory to better resemble counts collected at equilibrium

$$c'_{ij} = \sum_k \pi^{*(k)} \, c_{ji}^{(k)}$$

Then use the MLE estimator with detailed balance constraint from before

$$L(p) = \prod_{i,j}^{k} p_{ij}^{c'_{ij}}$$

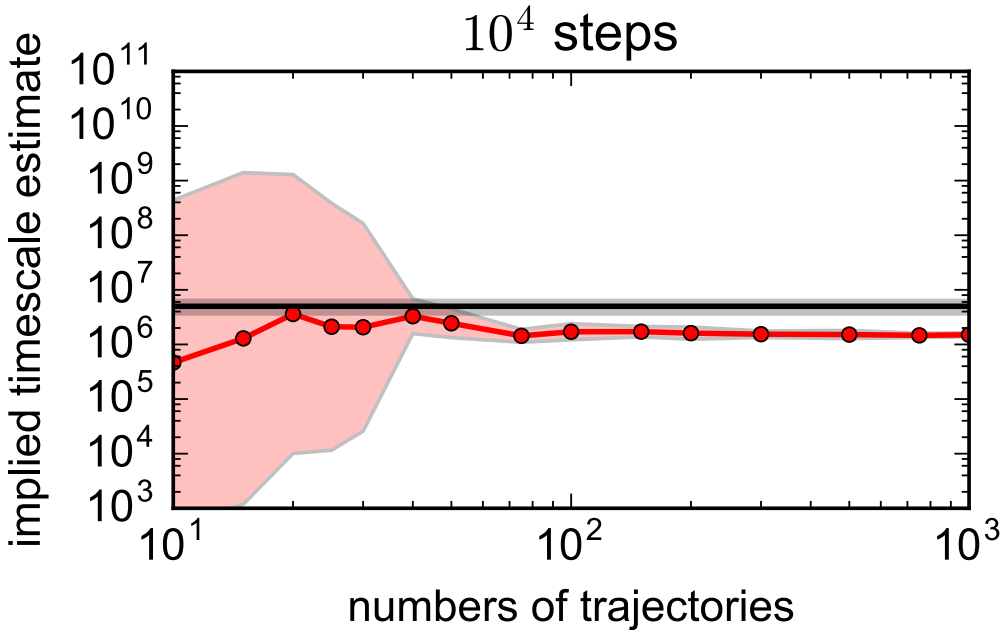
- $c_{ji}^{(k)}$ are counts from trajectory *k*.
- $\pi^{*(k)}$ is the weight of trajectory *k*

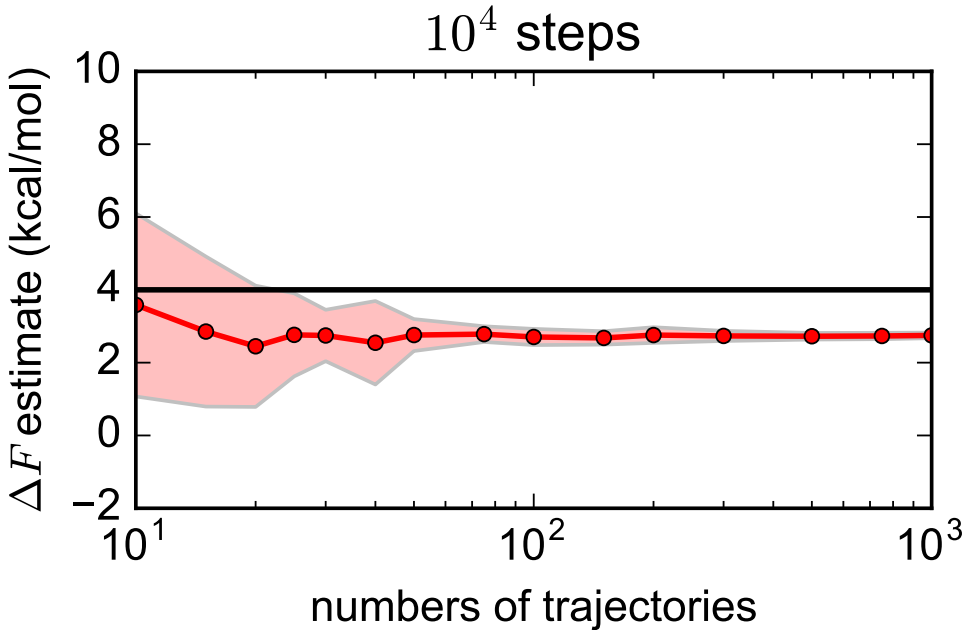
Row-normalized counts estimator

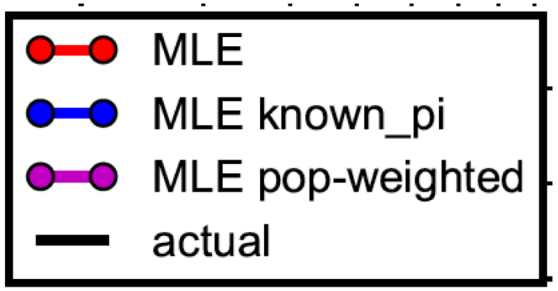
$$p_{ij} = \frac{c_{ij}}{\sum_j c_{ij}}$$

- Ignores detailed balance (!)
- Works only if $\sum_j c_{ij} \neq 0$.
- Hard for short trajectories with a larger lag time.

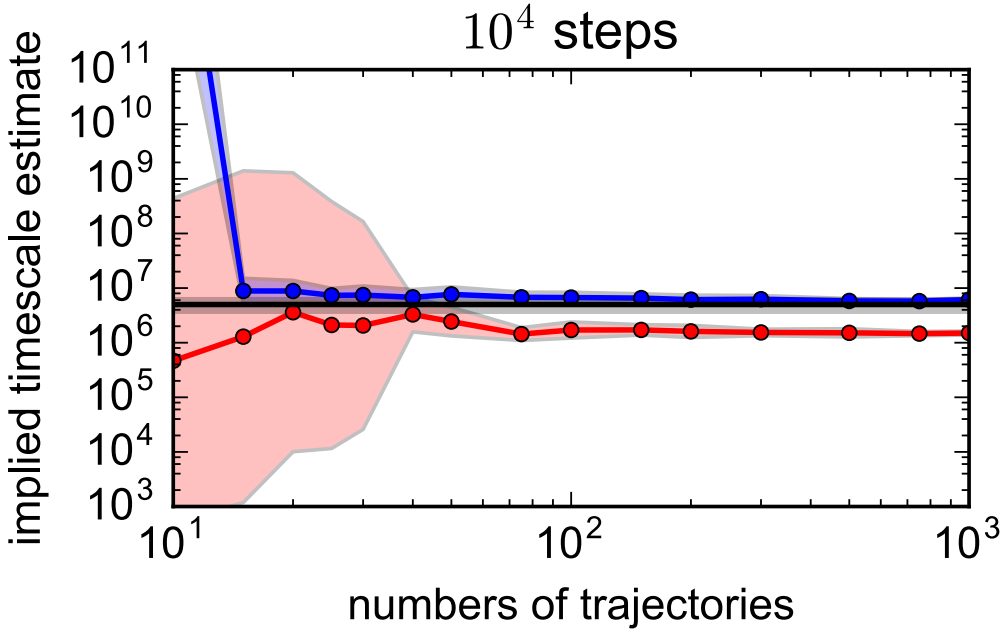
MLE(red) under-estimates timescales and free energies

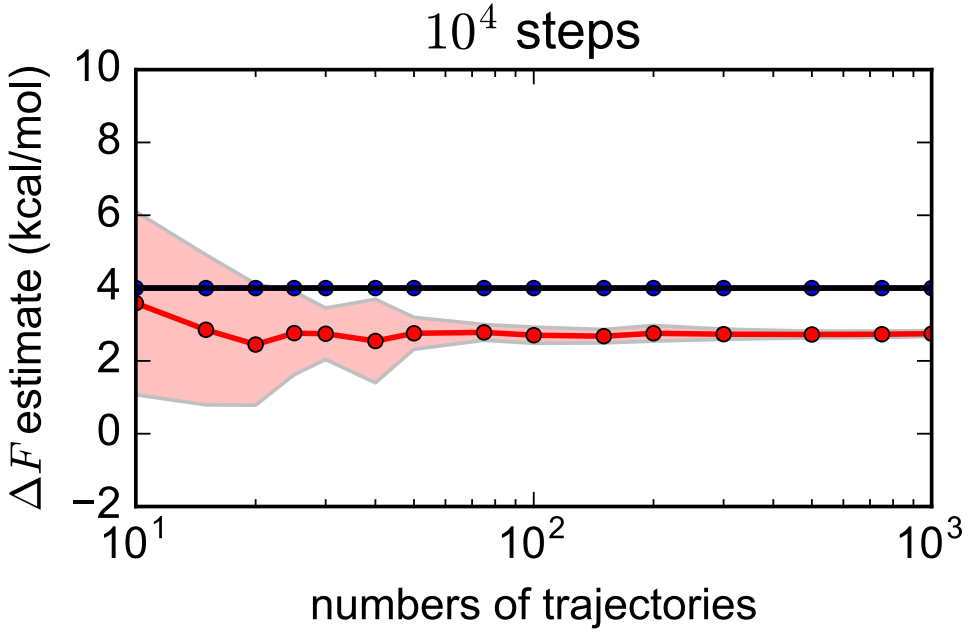


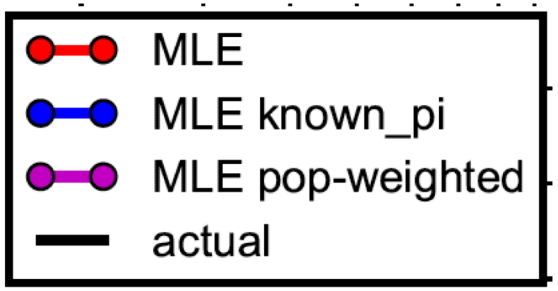




MLE-known_population (blue) converged quickly on both rate and free-energy estimations.

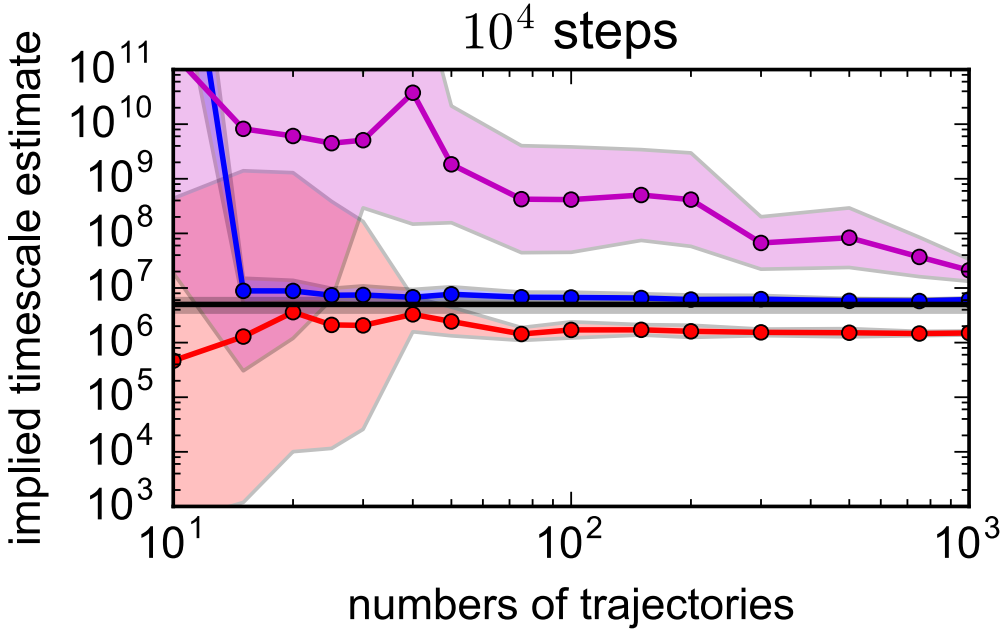


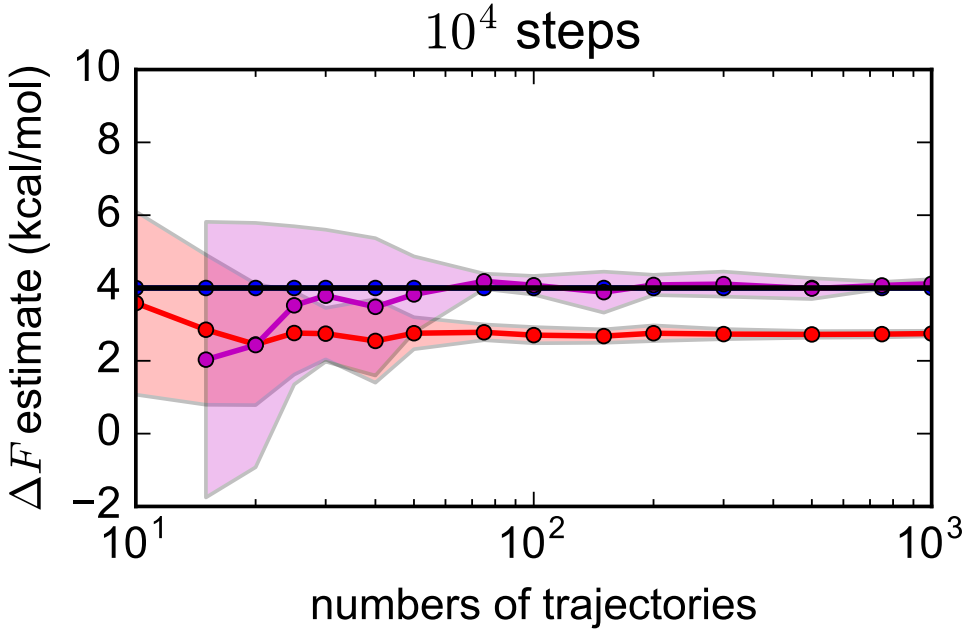


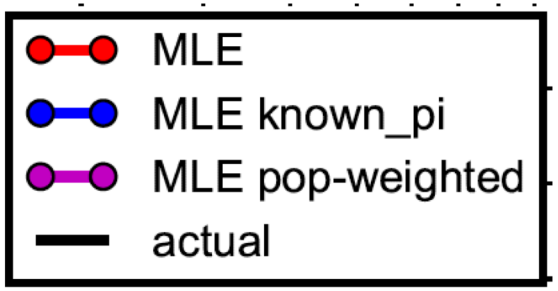


Trendelkamp-Schroer, B., & Noé, F. (2016). Efficient estimation of rare-event kinetics. *Physical Review X*.

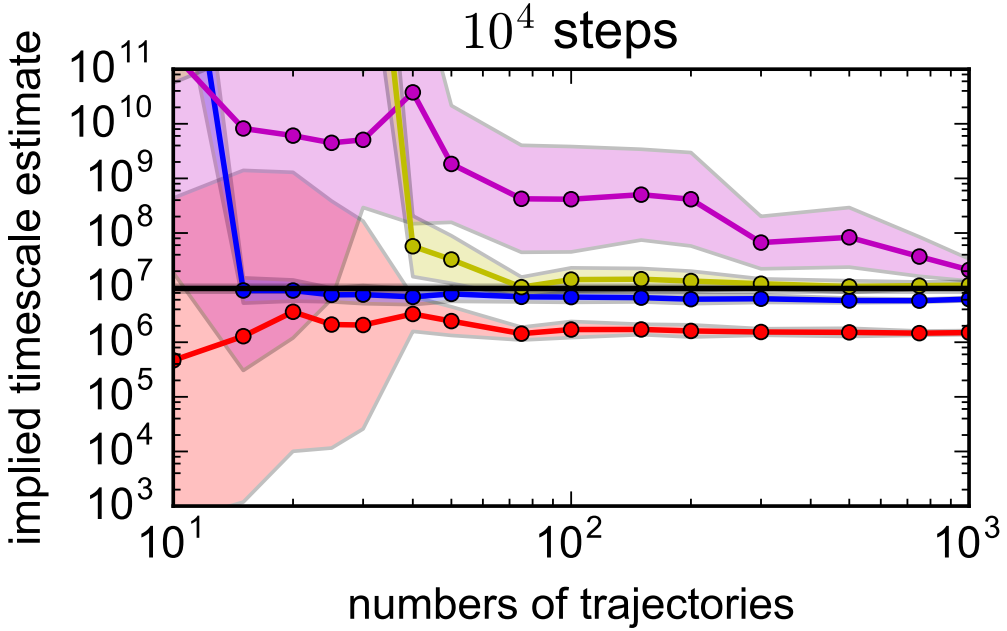
MLE-pop-reweight(magenta) converges as more sampling is collected

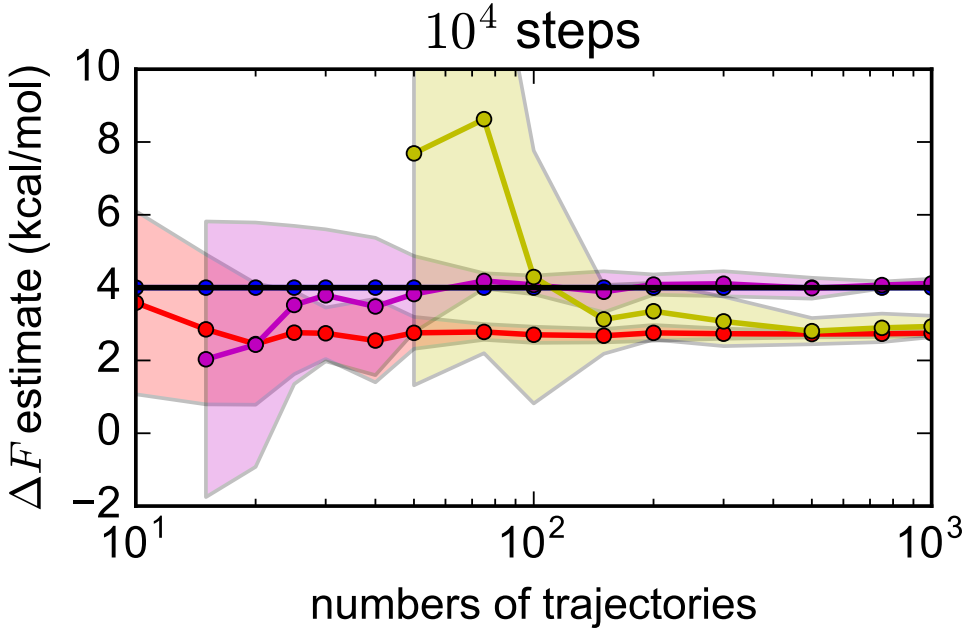


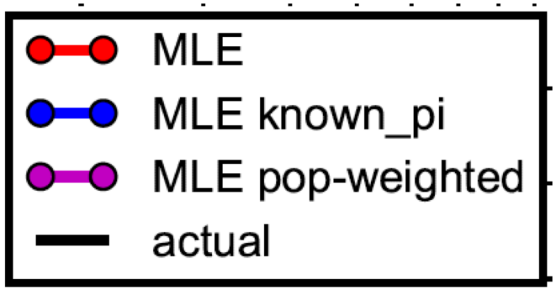




Row-norm needs a lot of trajectories to work well.

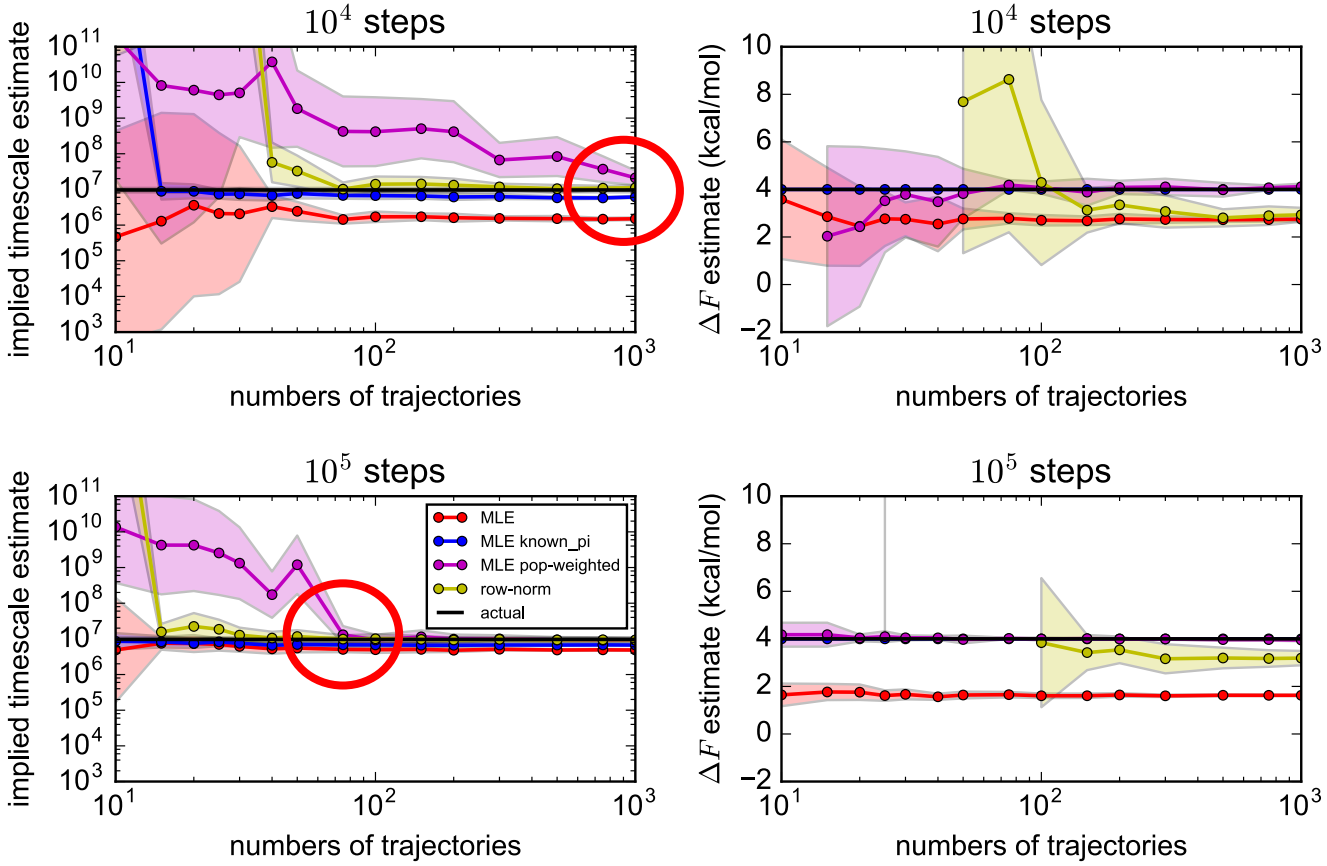




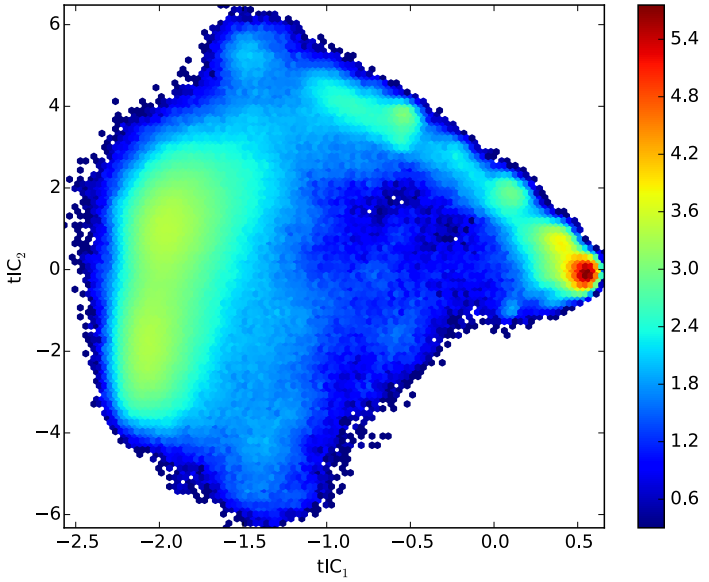


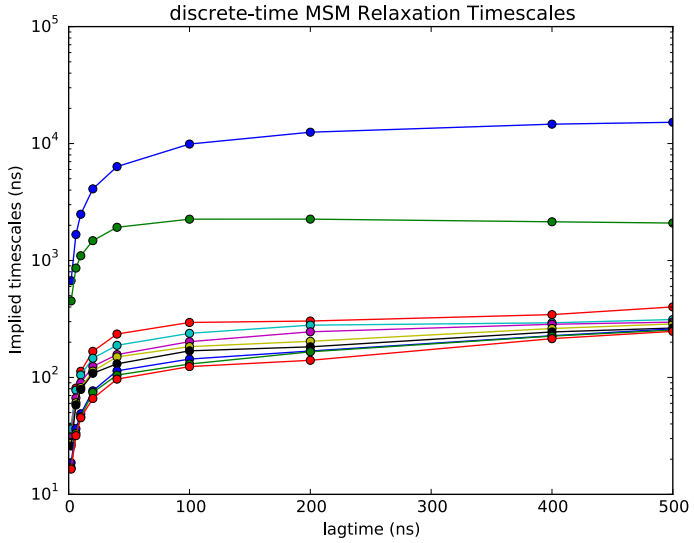
MLE-pop-reweight(magenta) converges if enough sampling is collected

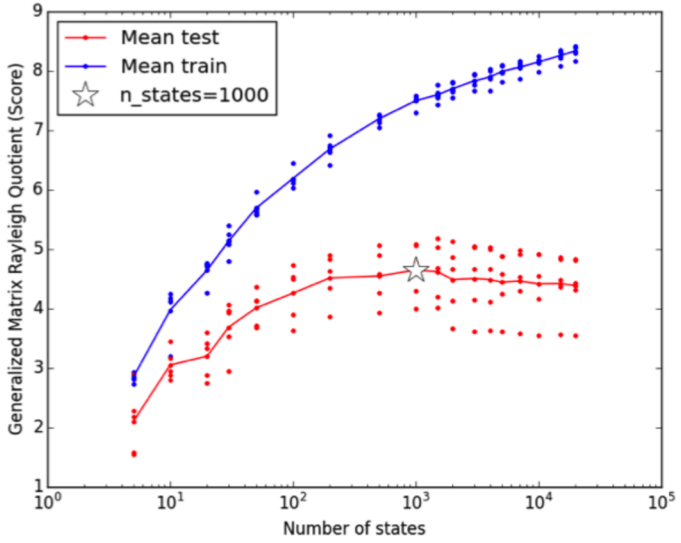
- N_bins * 10^7 (slowest motion) steps(sampling) are needed for MLE-pop-reweights to get converged.



GTT at 360K with charmm22-star forcefield



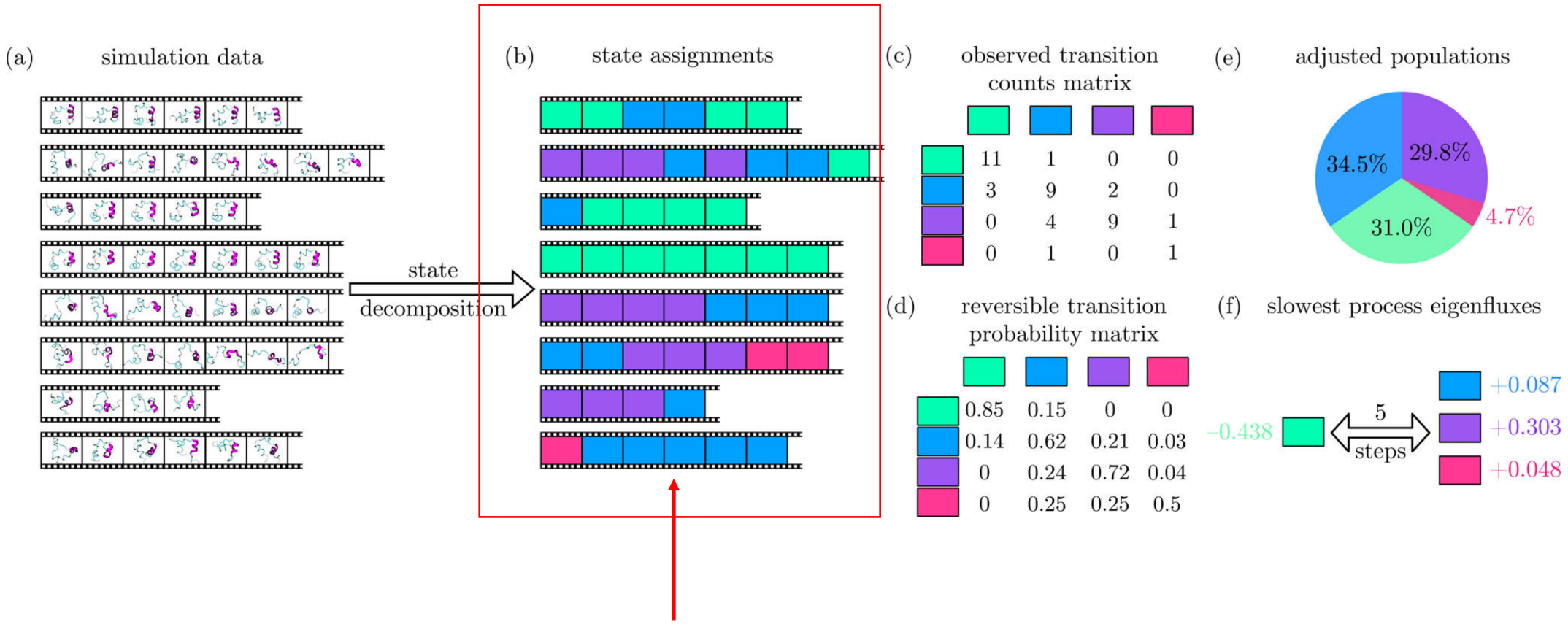




- Two independent simulation trajs with lengths of 651us and 486us.
- tICA with 10ns lagtime + k-center clustering algorithm.
- $C_\alpha + C_\beta$ pairwise distances was used as the input coordinates for tICA.
- GMRQ was used to determine that 1000 micro-states and 8 tICAs were optimal to capture the dynamics.
- MSM was built with a lag-time of 100ns.

WAN, H., Zhou, G., & Voelz, V. A. (2016). A Maximum-Caliber Approach to Predicting Perturbed Folding Kinetics Due to Mutations. *Journal of Chemical Theory and Computation*.

Making-up short trajectories from an existing dataset

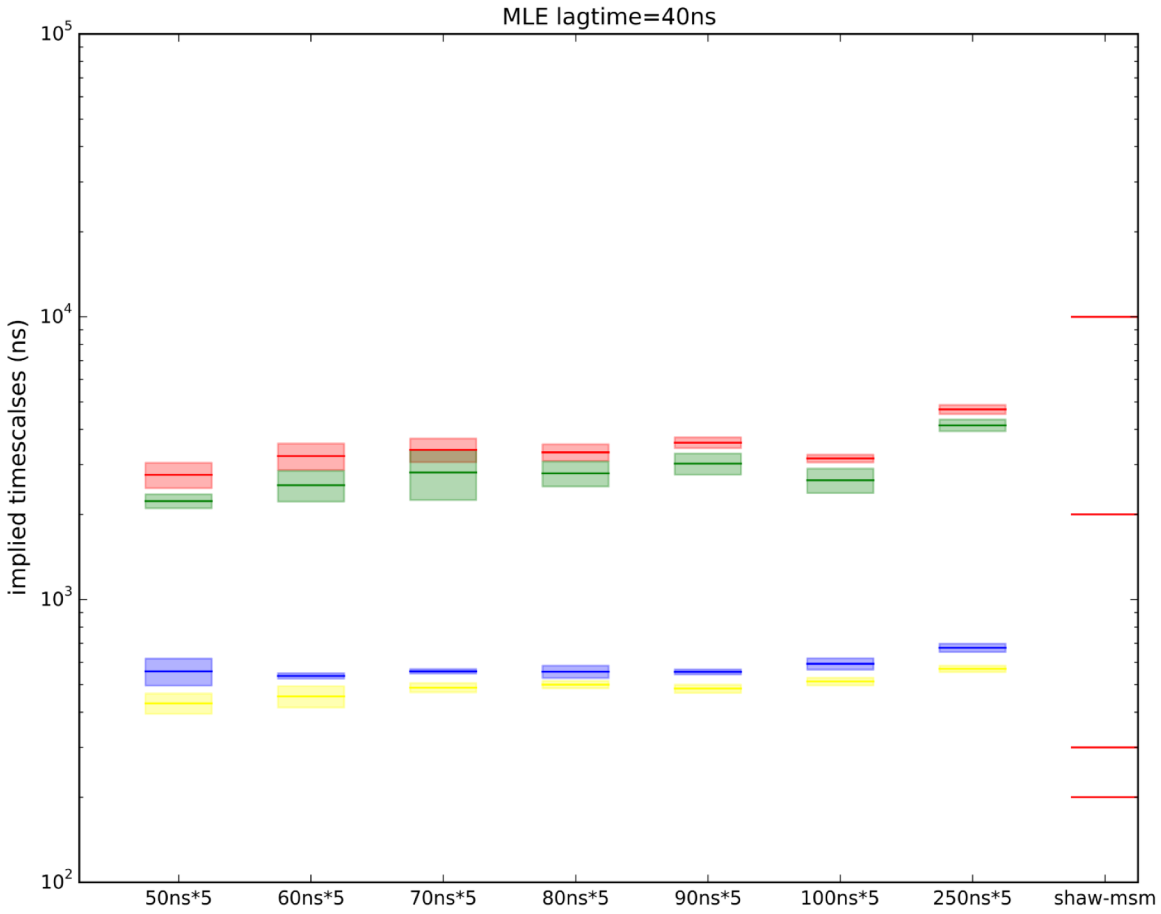


- Randomly draw short trajectories starting from a particular state

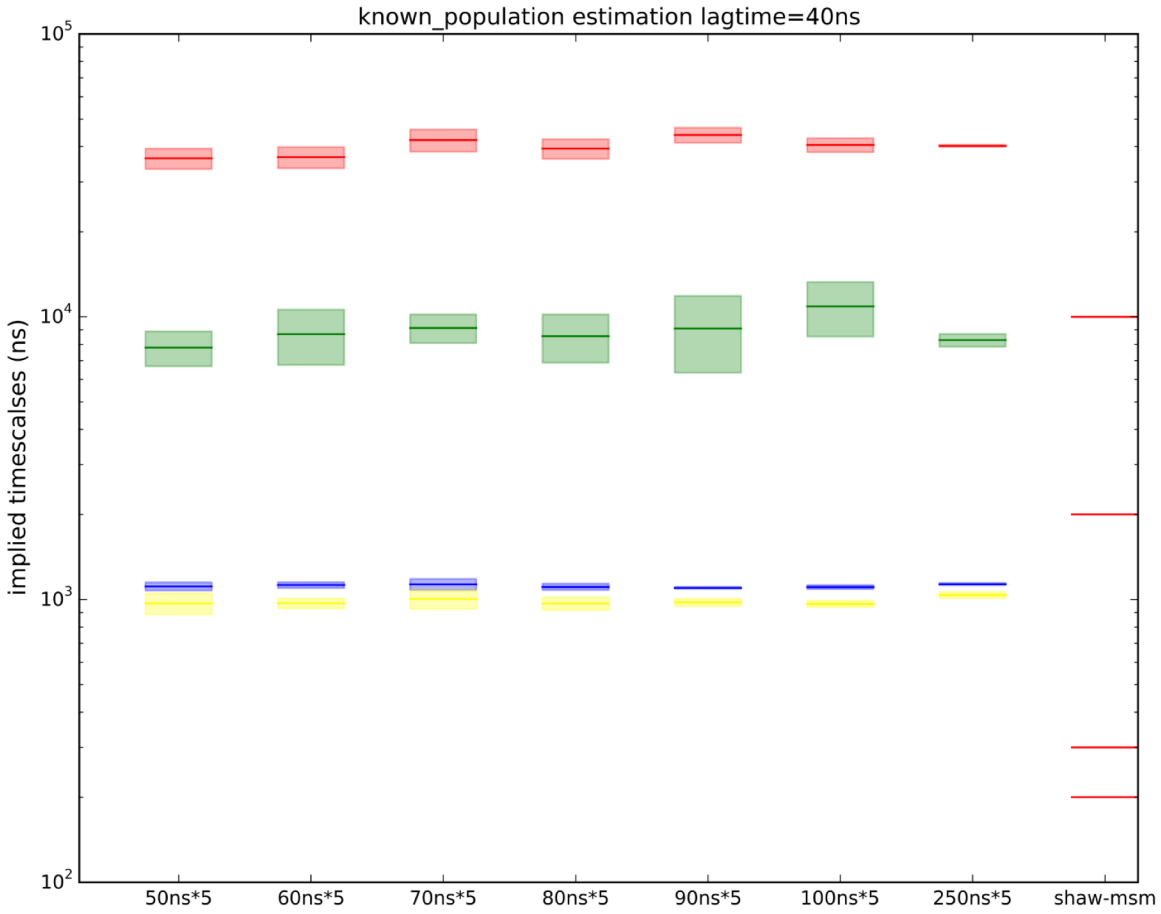
Expectation: it should re-produce both kinetic and thermodynamic properties

Husic, B. E., & Pande, V. S. (2018). Markov State Models: From an Art to a Science. *Journal of the American Chemical Society*, *140*(7), 2386–2396.

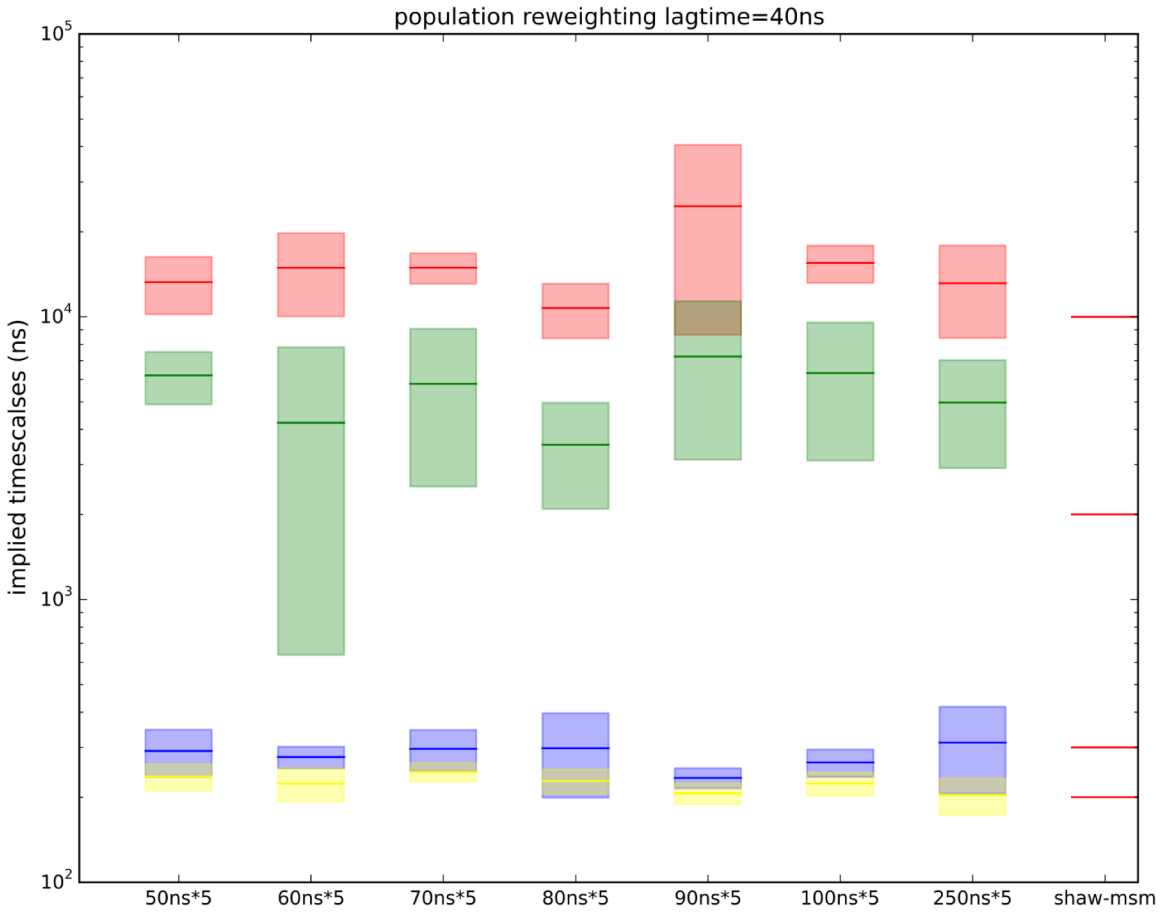
MLE underestimates folding rate



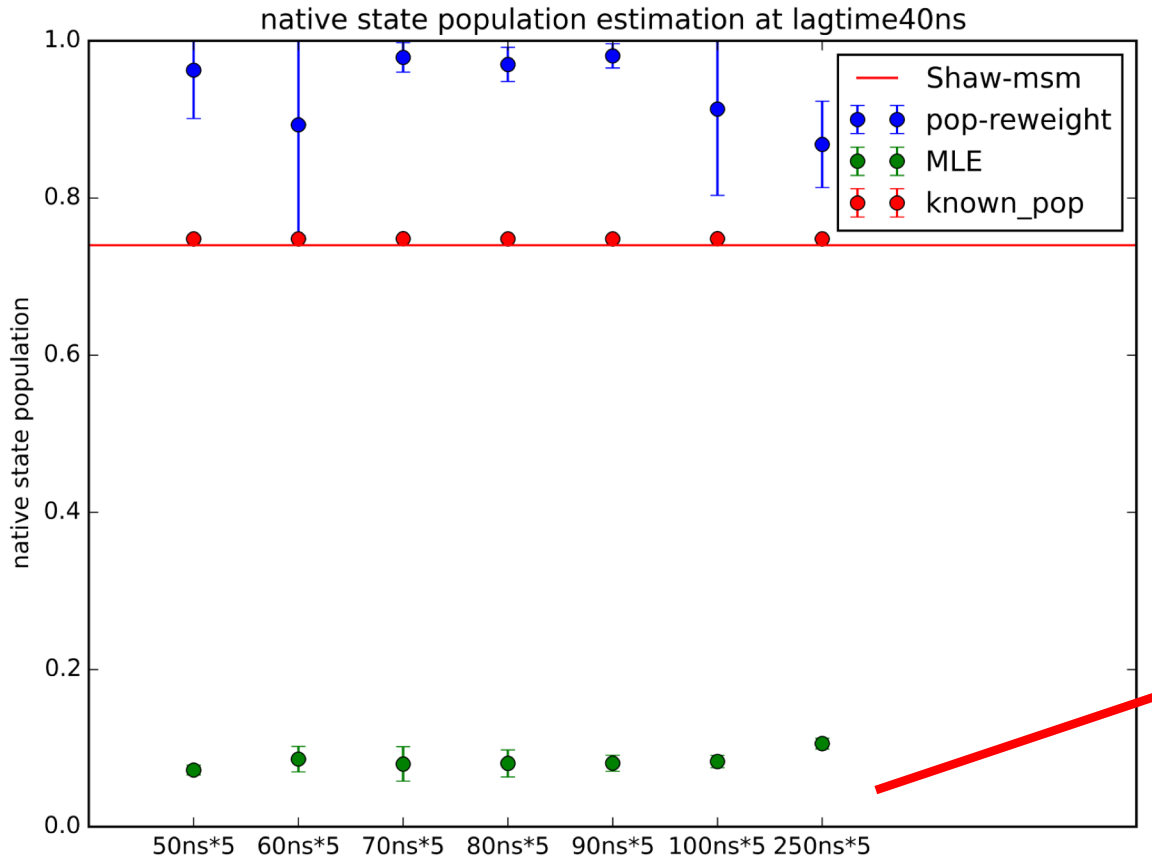
MLE-known_population overestimates folding time



MLE-pop-reweight converges in this large dataset



The native population estimation from MLE-pop-reweight(blue) gets better as more sampling collected

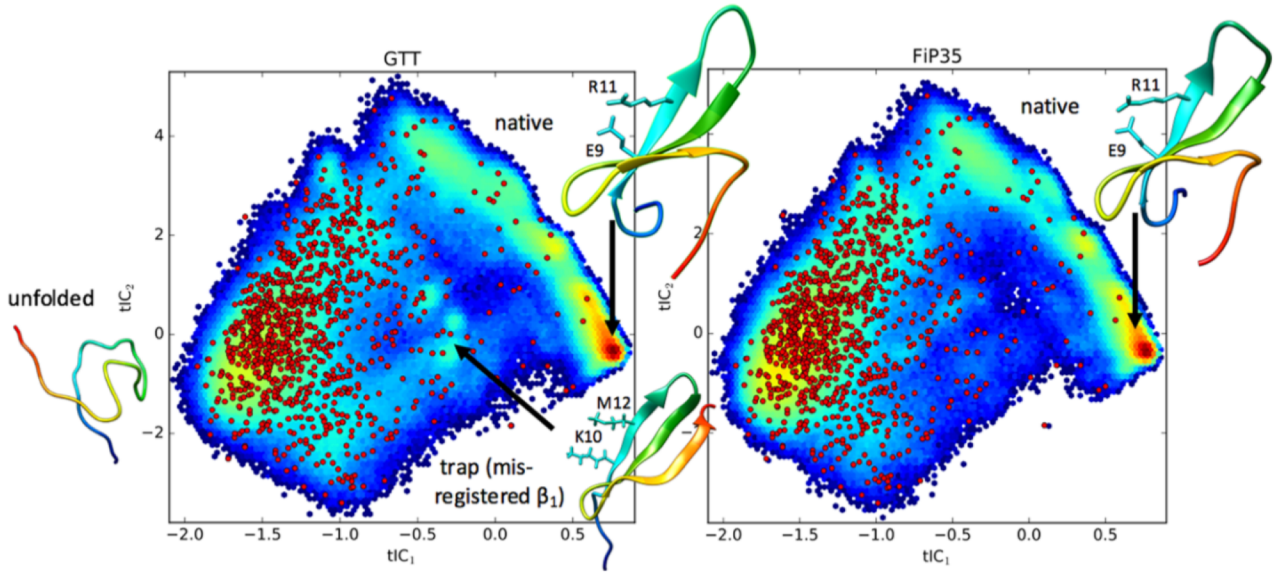


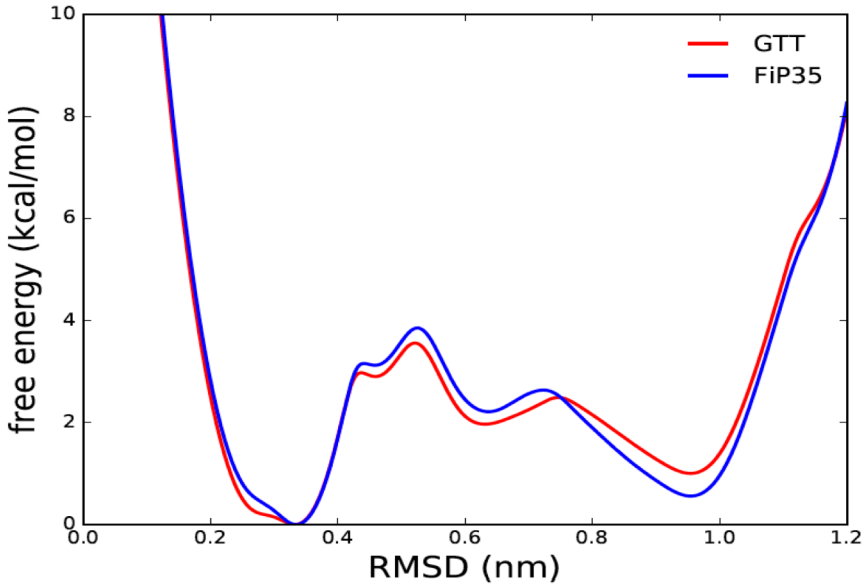
Beyond the limit

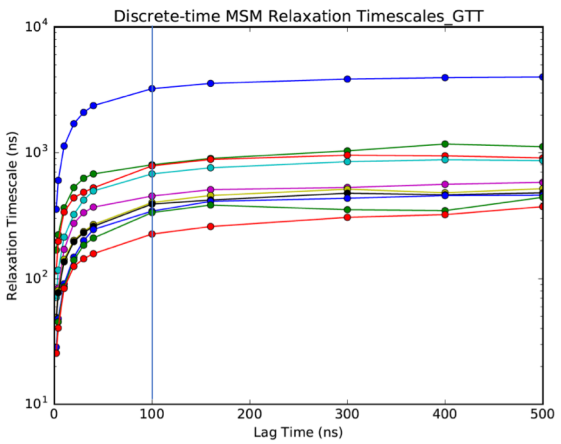
1000 microstates*250*5ns ≈1.25ms

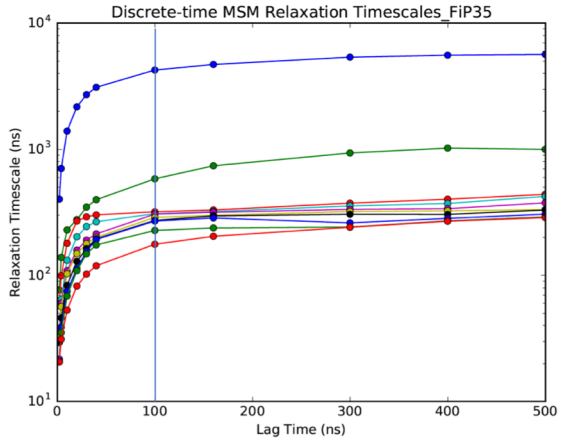
- N_bins * slowest motion ≈ 1000 microstates*10μs=10ms

WW domain FiP35 and GTT







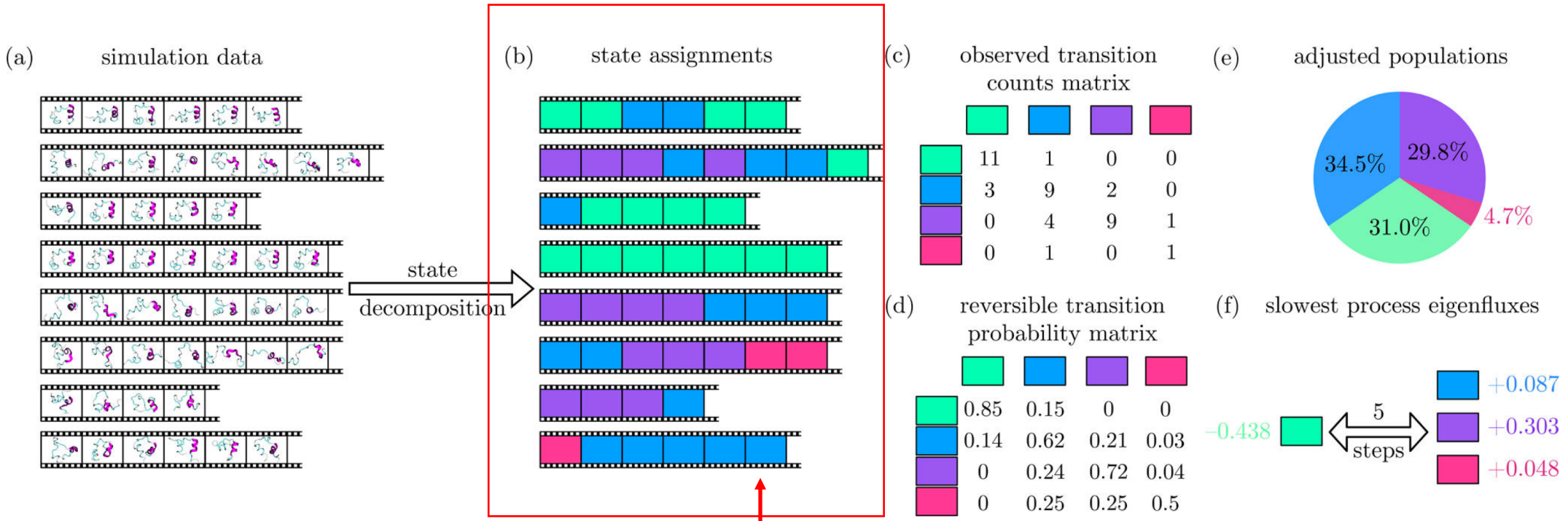


- GTT-mutation folds faster than FiP35 and stabilizes the native state.
- These two simulations are at same temperature and used same forcefield.

Q: to see if we can accurately predict FiP35 folding behaviors from seeding simulations of GTT(ref)?

WAN, H., Zhou, G., & Voelz, V. A. (2016). A Maximum-Caliber Approach to Predicting Perturbed Folding Kinetics Due to Mutations. *Journal of Chemical Theory and Computation*.
Piana, S., Sarkar, K., Lindorff-Larsen, K., Guo, M., Gruebele, M., & Shaw, D. E. (2011). Computational design and experimental testing of the fastest-folding β-sheet protein. *Journal of Molecular Biology*, *405*(1), 43–48.

Making-up short trajectories from an existing dataset

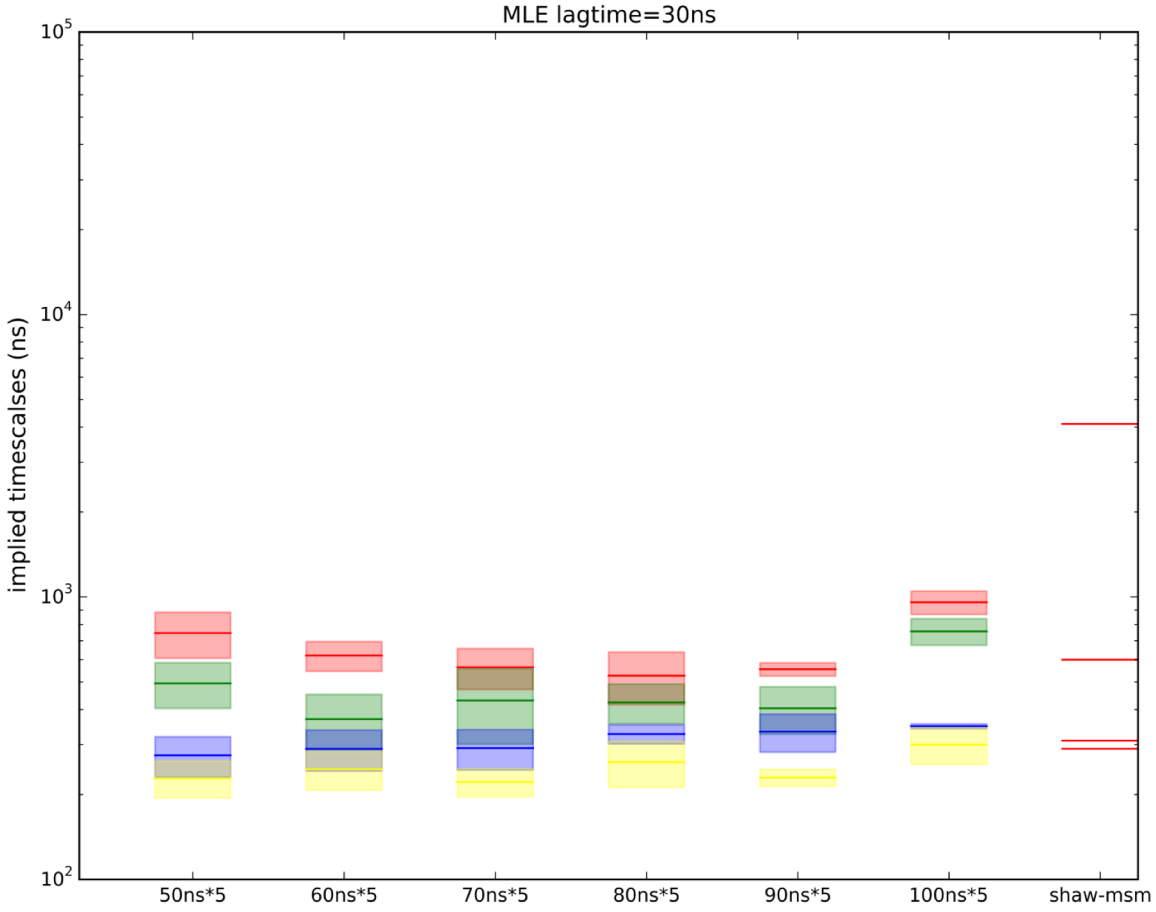


- Cluster them together
- Finding the overlapped states
  - 982/1000
- Randomly draw short trajectories starting from a particular state

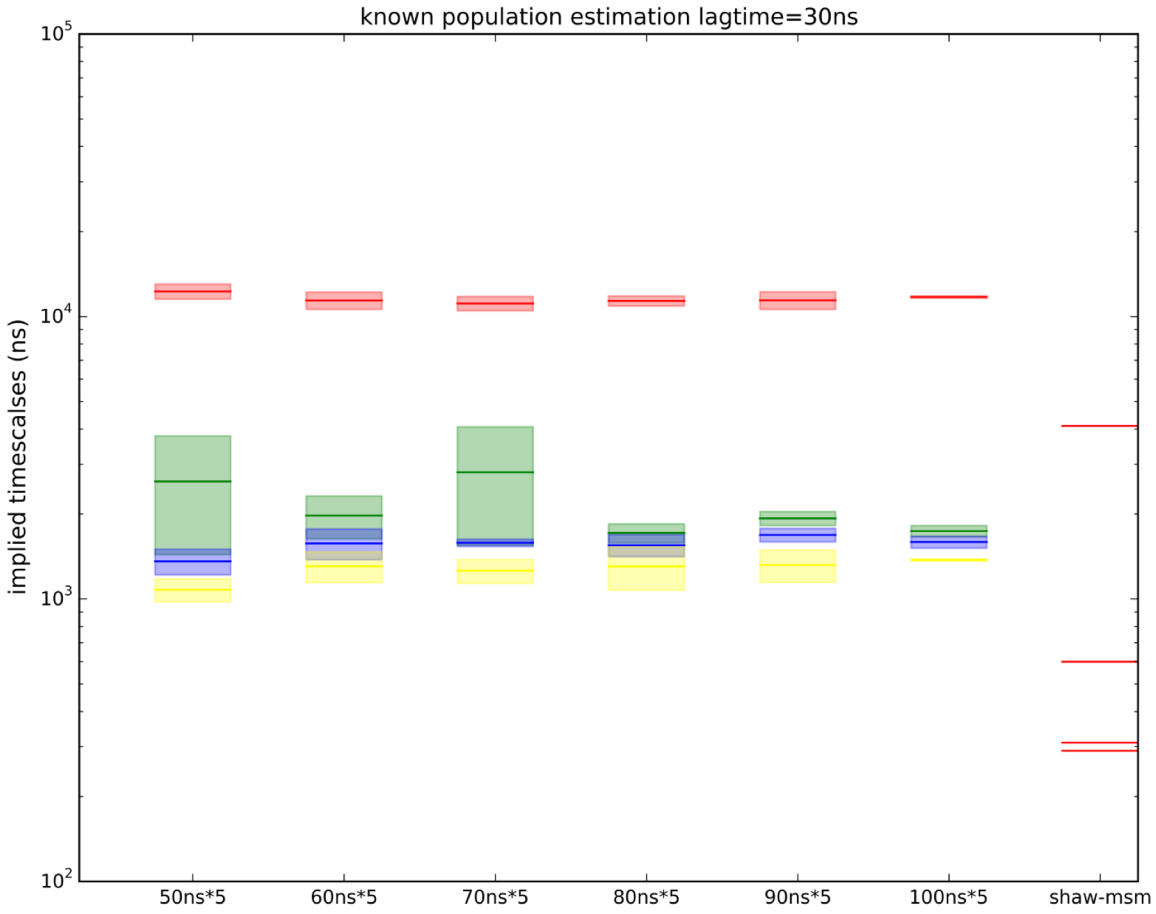
Expectation: seeding from GTT(mutation_ref) can predict both thermodynamic and kinetic properties of FiP35(WT)

Husic, B. E., & Pande, V. S. (2018). Markov State Models: From an Art to a Science. *Journal of the American Chemical Society*, *140*(7), 2386–2396.

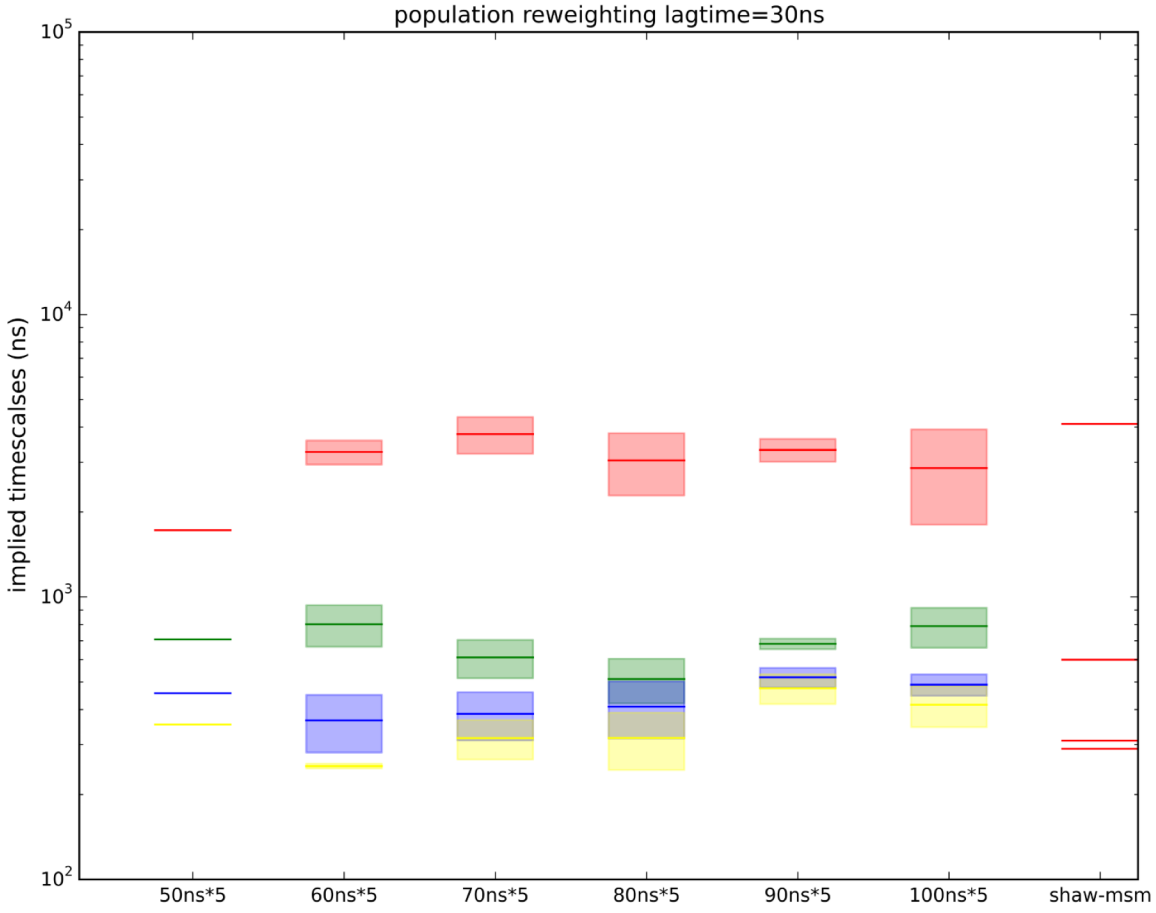
MLE underestimates folding time



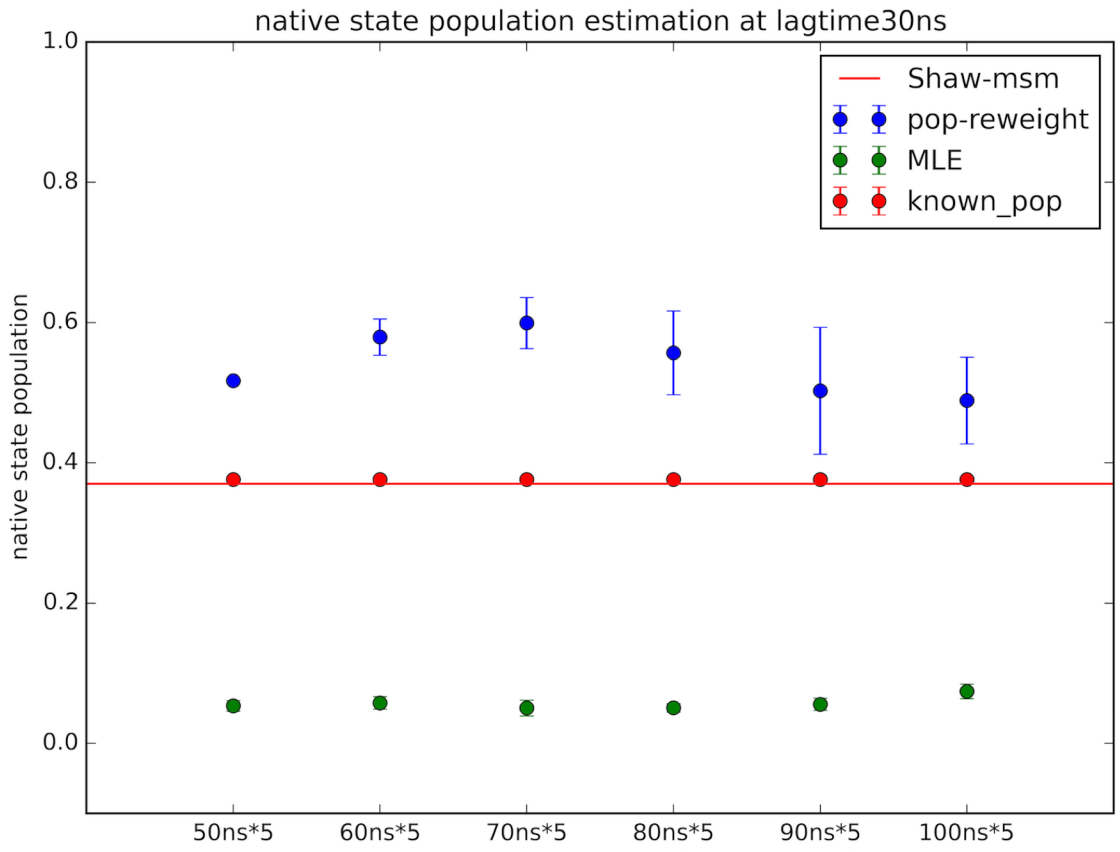
MLE-known-population overestimates folding time



MLE-pop-reweight converges as more sampling is collected



The native population estimation from MLE-pop-reweight(blue) gets better as more sampling is collected



Conclusions:

- Seeding from a reference MSM
  - MLE estimates a faster folding time and a lower native-state population.
  - MLE_known-pop can quickly get the right free energy profile but over-estimates the folding time.
  - MLE_population-reweight can properly estimate the folding time, and the free energy profile converges with more sampling.

Thanks Voelz lab