On the use of short reseeding trajectories to sample Markov state models

Hongbin Wan[1] and Vincent A. Voelz[1, a)]
Department of Chemistry, Temple University, Philadelphia, PA 19122, USA

(Dated: 21 November 2018)

In the last decade, advances in molecular dynamics (MD) and Markov State Model (MSM) methodologies have made possible accurate and efficient estimation of kinetic rates and reactive pathways for complex biomolecular dynamics occurring on slow timescales. A promising approach to enhanced sampling of MSMs is to use so-called adaptive methods, in which new MD trajectories are seeded preferentially from previously identified states. Here, we investigate the performance of various MSM estimators on reseeding trajectory, for both a simple 1D free energy landscape, and for mini-protein folding MSMs of WW domain and NTL9(1-39). Our results reveal the practical challenges of reseeding simulations, and suggest a simple way to better estimate both thermodynamic and kinetic information.

PACS numbers: Valid PACS appear here
Keywords: kinetics, molecular simulation

I. INTRODUCTION

In the last decade, Markov State Model (MSM) methodologies have made possible accurate and efficient estimation of kinetic rates and reactive pathways for slow and complex biomolecular dynamics.[1–5] One of the key advantages touted by MSM methods is the ability to use large ensembles of short-timescale trajectories for sampling events that occur on slow timescales. The main idea is that sufficient sampling using many short trajectories can circumvent the need to sample long trajectories.

With this in mind, many "adaptive" methods have been developed for the purpose of accelerating sampling of MSMs. The simplest of these can be called *adaptive seeding*, where one or more new rounds of unbiased simulation are performed by "seeding" swarms of trajectories throughout the landscape.[6] The choice of seeds are based on some initial approximation of the free energy landscape, possibly from non-equilibrium or enhanced-sampling methods. Adaptive seeding can be performed by first identifying a set of metastable states, then initiating simulations from each state. If the seeding trajectories provide sufficient connectivity and statistical sampling of transition rates, an MSM can be constructed to accurately estimate both kinetics and thermodynamics.

Similarly, so-called *adaptive sampling* algorithms have been developed for MSMs in which successive rounds of targeted seeding are performed, updating the MSM after each round[7,8]. A simple adaptive sampling strategy is to start successive rounds of simulations from under-sampled states, for instance, from the state with the least number of transition counts.[9] A more sophisticated approach is the FAST algorithm, which is designed to discover states and reactive pathways of interest by choosing new states based on an objective function that balances under-sampling with a reward for sampling desired structural observables.[10?] The REAP algorithm efficiently explores folding landscapes by using reinforcement learning to choose new states.[11]

A key problem with adaptive sampling of MSMs arises because we are often interested in *equilibrium* properties, while trajectory seeding is decidedly *non-equilibrium*. This may seem like a subtle point, because the dynamical trajectories themselves are unbiased, but of course, the ensemble of starting points for each trajectory are almost always *statistically* biased, i.e. the seeds are not drawn from the true equilibrium distribution. This can problematic because most MSMs are constructed from transition rate estimators that enforce detailed balance and assume trajectory data is obtained at equilibrium. The distribution of sampled transitions, however, will only reflect equilibrium conditions in the limit of long trajectory length.

One way around this is to focus mostly on the kinetic information obtained by adaptive sampling. A recent study of the ability of FAST to accurately describe reactive pathways concluded that the most reliable MSM estimator to use with adaptive sampling data is a row-normalized transition count matrix.[?] Indeed, weighted-ensemble path sampling algorithms focus solely on sampling the kinetics of reactive pathways, information which can be used to recover global thermodynamic properties.[13–17] A major disadvantage of this approach is that it ignores potentially valuable equilibrium information. As shown by Trendelkamp-Schroer and Noé,[18] detailed balance is a powerful constraint to infer rare-event transition rates from equilibrium populations. Specifically, when faced with limited sampling, dedicating half of one's simulation samples toward enhanced thermodynamic sampling (e.g. umbrella sampling) can result in a significant reduction in the uncertainty of estimated rates, simply because the improved estimates of equilibrium state populations inform the rate estimates through detailed balance.

Another way around this problem, recently described by Nüske et al., is to use an estimator based on observer operator model (OOM) theory which utilizes information from transitions observed at lag times τ and 2τ to obtain estimates unbiased by the initial distribution of seeding trajectories.[?] Although the OOM estimator is able to make better MSM estimates at shorter lag times, it requires the storage of a transition count ar-

a)Electronic mail: vvoelz@temple.edu

ray that scales as $\propto N^3$ for a MSM with M states, and a dense-matrix singular-value decomposition step, which may make it impractical for MSMs with large numbers of states. Nüske et al. derive an expression quantifying the error incurred by non-equilibrium seeding, from which they conclude that such bias is difficult to remove without either increasing the lag time or improving the state discretization. Both strategies are antithetical to adaptive seeding, the purpose of which is to utilize swarms of very short trajectories from previously-defined MSM states.

Here, we explore an alternative way to recover accurate MSM estimates from biased seeding trajectories, by reweighting sampled transitions counts to better approximate counts that would be observed at equilibrium. Like the Trendelkamp-Schroer and Noé method, this requires some initial estimate of state populations, perhaps obtained from previous rounds of adaptive sampling.

Moreoever, we are particularly interested in examining how our reweighting method performs in cases where it is impractical to generate long trajectories and instead are forced to rely on ensembles of short seeding trajectories. An example of this is adaptive seeding of protein folding MSMs built from ultra-long trajectories simulated on the Anton supercomputer[?] Because such computation is not a widely available resource, adaptive seeding using more conventional computers may be one of the only ways to leverage these MSMs to predict the effect of mutations, for example.

In this manuscript, we first prepare an adaptive seeding scenario on a 1D-potential energy model, test a number of estimator

make the case that even the most basic kind of adaptive sampling, i.e. single-round *adaptive seeding*, benefits from MSM estimates that include equilibrium population estimates. We first show this in a simple 1-D model, and then show its application in a practical scenario: adaptive seeding of protein folding MSMs using large-scale distributed computing. Our results suggest a simple method to make reseeding simulations more accurate. Finally we apply the method to predict mutation induced kinetics. The overall results indicate the wide ranging applicability from doing proper estimations in seeding simulations to predict perturbed kinetics.

A. Adaptive seeding of folding landscapes

To character of the protein folding landscapes remains the challenging in molecular biology even using GPU-acceleration or simulation on the *Anton*, special-purpose supercomputer) due to the rough and large energy barriers. On the other hand, there are great interests in using molecular dynamics simulations along with MSM methodologies, which can offer a seemingly simple way to improve the conformational sampling efficiency for obtaining the full folding/unfolding behaviors.[7?,8] In concept, MSMs should guide the simulations to explore conformational spaces. Based on this idea that motivates the recent development of multiensemble Markov model **MEMM**, which applies transition-based reweighting method to properly estimate MSMs simulated with bias/un-bias.[?] It should be noted that it is unlike Weighted-ensemble (WE) path sampling algorithms, which can also be classified as adaptive sampling algorithms. In WE approaches such as WExplore[?] and WESTPA[14], successive rounds of new trajectories are spawned to better sample a quantity of interest (reactive flux, for example), while the statistical weights of each trajectory are carefully managed so that no bias is introduced. The above adaptive methods are unbiased, where the bias here are in the sense that each trajectory is simulated with changed potential/pre-defined reaction coordinates. This statistical bias has interesting consequences in estimating various quantities, consequences which arise from the trade-off between sampling of transition rates versus equilibrium probabilities. Thus, the adaptive seeding simulations need an proper estimator instead of the maximum-likelihood estimator (MLE) that is default in packages like MSMBuilder.[3,19–23] This is because the MLE enforces detailed balance; i.e. MLE assumes that the observed counts are sampled from the equilibrium distribution, an assumption which is (purposely) violated by adaptive sampling. One might expect, then, that while quantities like rates and pathways are accurately estimated by adaptive methods, quantities like equilibrium populations may have more uncertainties and/or bias. Indeed, while the weighted ensemble algorithms have recently been used to efficiently sample very slow folding rates[15] ad unbinding rates (residence times > 10 s), these same algorithms(.....???? need to edit)

In this paper, we consider a specific kind of adaptive method for sampling MSMs: adaptive seeding of perturbed folding landscapes. It is often the case that large numbers of expensive simulations in a particular force field are utilized to model the folding landscape of a particular protein sequence. We would like to model the folding of a sequence variant, or perhaps use a different force field potential, without having to perform a heroic amount of simulation. Since there is much prior information from the *wild-type* MSM, it is reasonable to think that adaptive seeding could provide a good picture of how folding rates and populations change with the perturbation. Here, we explore the accuracy of several estimators for obtaining folding rates and populations from adaptive seeding simulations. Using a 1-D two-state potential as simple model, we explore different estimators and find interesting differences in their relative accuracies in estimating rates versus equilibrium populations from adaptive seeding trajectory data. We also explore the effects of using different trajectory lengths and number of seeds. In general, we find that rates and free energies are more accurately estimated by estimators that incorporate some prior knowledge of the equilibrium populations. We then show how adaptive seeding can be used to model changes in folding rates and populations for GTT WW domain and K12M mutant of NTL9, based on MSMs built from ultra-long simulation trajectories.

B. Estimators

We explored the accuracy and efficiency of several different transition probability estimators using the adaptive seeding trajectory data as input: (1) a maximum-likelihood estimator (MLE), (2) a MLE estimator where the equilibrium populations π_i of each state are known *a priori*, (3) the MLE estimator where each input trajectory is weighted by the *a priori* equilibrium population of its starting point,(4) row-normalized transition counts and (5) Observable operator model estimators.

1. Maximum-likelihood estimator (MLE).

The MLE for a reversible MSM assumes that observed transition counts are independent, and drawn from the equilibrium distribution, so that reversibility (i.e. detailed balance) can be used a constraint. The likelihood of observing a set of given transition counts, $L = \prod_i \prod_j p_{ij}^{c_{ij}}$, when minimized under the constraint that $\pi_i p_{ij} = \pi_j p_{ji}$, for all i, j, yields a self-consistent expression that can be iterated to find the equilibrium populations,[3,26,27]

$$\pi_i = \sum_j \frac{c_{ij} + c_{ji}}{\frac{N_j}{\pi_j} + \frac{N_i}{\pi_i}} \tag{1}$$

where $N_i = \sum_j c_{ij}$. The transition probabilities p_{ij} are given by

$$p_{ij} = \frac{(c_{ij} + c_{ji})\pi_j}{N_j \pi_i + N_i \pi_j} \tag{2}$$

2. Maximum-likelihood estimator (MLE) with known populations π_i

Minimization of the likelihood function above, with the additional constraint of fixed populations π_i, yields a similar self-consistent equation that can be used to determine a set of Lagrange multipliers,[28]

$$\lambda_i = \sum_j \frac{(c_{ij} + c_{ji})\pi_j \lambda_i}{\lambda_j \pi_i + \lambda_i \pi_j}, \tag{3}$$

from which the transition probabilities p_{ij} can be obtained as

$$p_{ij} = \frac{(c_{ij} + c_{ji})\pi_j}{\lambda_j \pi_i + \lambda_i \pi_j}. \tag{4}$$

3. Maximum-likelihood estimator (MLE) with population-weighted trajectory counts.

For this estimator, first a modified count matrix c'_{ij} is calculated,

$$c'_{ij} = \sum_k w^{(k)} c_{ji}^{(k)}, \tag{5}$$

where transition counts $c_{ji}^{(k)}$ from trajectory k are weighted in proportion to $w^{(k)} = \pi^{(k)}$, the equilibrium population of the initial state of the trajectory. The idea behind this approach is to counteract the statistical bias from adaptive seeding by scaling the observed transition counts proportional to their equilibrium fluxes.The modified counts are then used as input to the MLE.

4. Row-normalized counts.

For this estimator, the transition probabilities are approximated as

$$p_{ij} = \frac{c_{ij}}{\sum_j c_{ij}}. \tag{6}$$

This approach does not guarantee reversible transition probabilities, which only occurs in the limit of large numbers of reversible transition counts. In practice, however, the largest eigenvectors of the transition probability matrix have very nearly real eigenvalues, such that we can report relevant relaxation timescales and equilibrium populations.

5. Observable operator model(OOM) theory

For this estimator, it requires count matrix C_{Eq}^{τ} and two-step count matrix $C_{Eq}^{2\tau}$ to obtain the unbiased equilibrium correction matrix and unbiased equilibrium distribution.[?] The non-reversible and reversible transition probabilities are approximated as

$$T_{Eq}^{\tau}(i,j) = \frac{C_{Eq}^{\tau}(i,j)}{\pi_i} \tag{7}$$

$$T_{Eq}^{\tau}(i,j) = \frac{T_{Eq}^{\tau}(i,j) + T_{Eq}^{\tau}(j,i)}{\sum_{j=1}^{N} T_{Eq}^{\tau}(i,j) + \sum_{j=1}^{N} T_{Eq}^{\tau}(j,i)} \tag{8}$$

This approach does not only correct MSM estimates with a shorter lag times but also can play an indicator of the equality from of such MSMs. However, the computational cost of this estimator is heavy, which is proportional to N^3 (N is the number of MSM states) for example.

II. RESULTS

A. Adaptive seeding of a 1-D potential energy surface

We consider the following potential energy surface, as used by Stelzl et al.[24]: $U(x) = -\frac{2k_BT}{0.596} \ln[e^{-2(x-2)^2-2} + e^{-2(x-5)^2}]$ for $x \in [1.5, 5.5]$, and $k_BT = 0.596$ kcal mol^{-1}. The state space is uniformly divided into 20 bins to calculate discrete-state quantities. Diffusion on the 1-D landscape is approximated by a Markov Chain Monte Carlo (MCMC) procedure in which new moves are translations randomly chosen from $\delta \in [-0.05, +0.05]$ and accepted

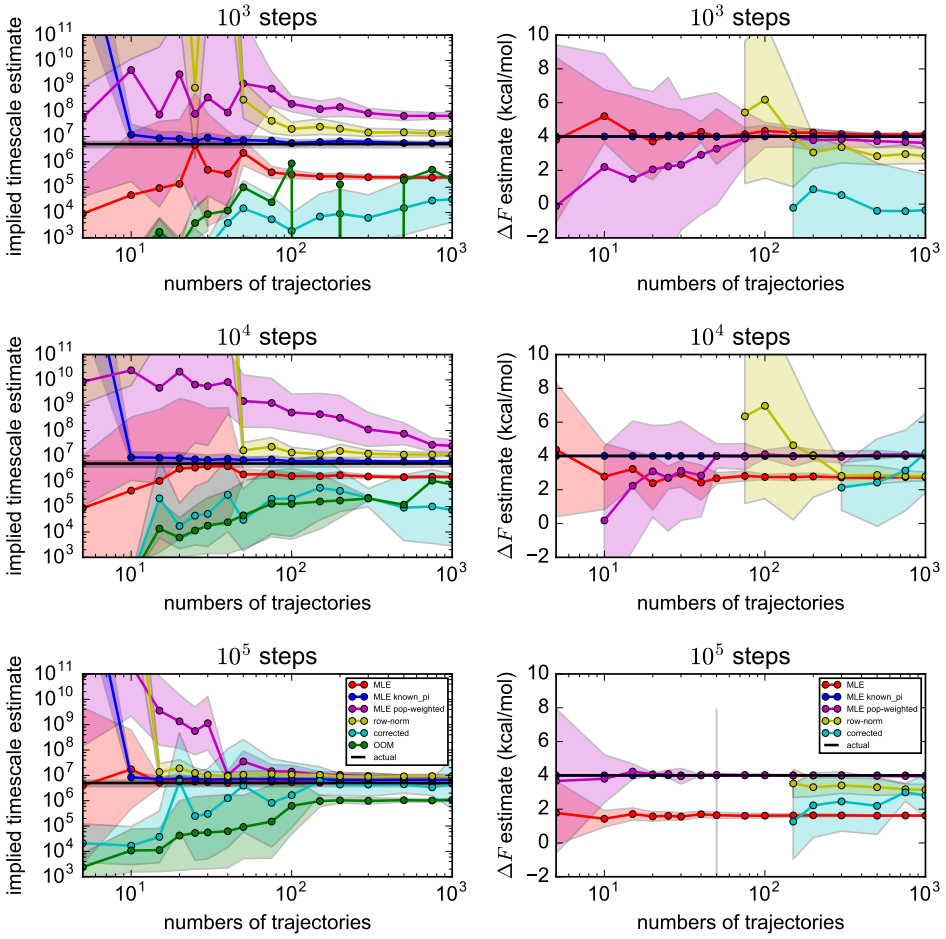

FIG. 1. The performances on predictions of implied timescales and free energy from different estimators

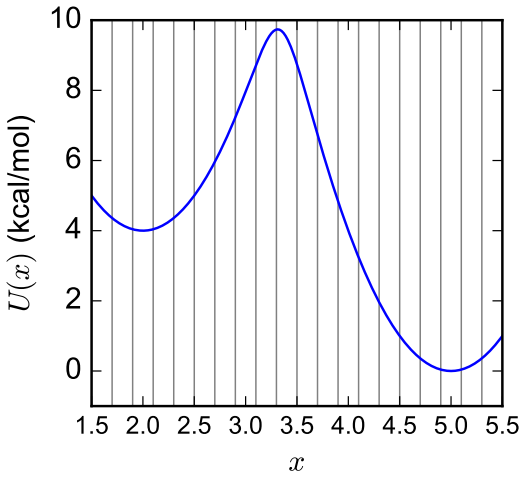

FIG. 2. A 1-D two-state potential

.

with probability $\min(1, \exp(\beta[U(x+\delta) - U(x)]))$, i.e. the Metropolis criterion.

For all tests with this toy model, we limit the lag time to $\tau = 100$ steps. To estimate the "true" relaxation timescale of the two-state model, we generated long MCMC trajectories of 10^9 steps, sampling from a series of scaled potentials $U^{(\lambda)}(x) = \lambda U(x)$ for $\lambda \in [0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$. For each λ value, 20 trajectories were generated, with half of them starting from $x = 2.0$ and the other half starting from $x = 5.0$, resulting in a total of 120 trajectories. The DTRAM estimator of Wu et al.[25] was used to estimate the slowest relaxation timescale as 9.66 ($\pm$ 1.37) $\times 10^6$ steps, using a lag time of 1000 steps.

To emulate adaptive seeding trajectory data, various numbers of trajectories s, each of length $n\tau$ were initiated from the center positions of all twenty bins. The resulting data consists of $20 \times s \times n$ transition counts between states i and j in lag time τ, stored in a 20×20 count matrix of entries c_{ij}. From these counts, estimates of the transition probabilities, p_{ij} can be made. The MLE with known populations undoubtedly gets converged on the predictions of both implied timescales and free energy profiles due to only 1 reaction coordinate defined in the simulations. The Maximum Likelihood estimator(MLE) requires longer simulations to get accurate implied timescale estimations and OOM estimator always underestimates the implied timescale. As expected, the row normalization estimator

is hard to work with large lagtime in these short seeding simulations. With population-weighted trajectory counts,Maximum-likelihood estimator(MLE) reaches the right implied timescale very fast if the trajectory length is long or more seeding trajectories ware simulated. Interestingly, MLE successfully captures the right free energy profile with short simulations but starts to lose performances in longer seeding simulations. The free energy profile estimated from OOM starts to become better if more trajectories data can be collected. The population-weighted MLE gets converged very fast on predicting the free energy profile even with short seeding simulation length or smaller number of simulations trajectories.

B. Seeding of folding landscapes for WW Domain and NTL9(1-39)

The WW domain, one of the fast-folding protein domains, is 35-residue proline-rich protein that consists of three β-sheet structure, whose kinetics has been extensively studied by experiments[29–33]. And it also has been characterized to be a benchmark system in molecular dynamics to explore two-state fast-folding behaviors and downhill folding manners[34–36]. A large number of mutation experiments have been done to investigate its folding mechanisms. Among the variants of its families,a WW domain of the FIP mutant(FiP35) of protein human pin1 has a folding time of 13μs[37]. Another optimization near loop2 where native sequence Asn-Ala-Ser (NAS) is replaced with Gly-Thr-Thr(GTT) tri-mutations was discovered by Piana et al. that has a fast relaxation time 4 μs[36]. Two independent ultra-long trajectories(651 and 486 μs) provided by Shaw et al.[38] were performed at 360K using the CHARMM22* force field.

NTL9(1-39) domain is the N-terminal domain of the ribosomal protein L9, and its folded state consists of one α-helix and three-strand β-sheet. NTL9 domain is widely used as a classic system to study two-state folders in both experimental and computational studies[39–43]. The K12M mutant in the hydrophobic core destabilizes the non-native states and has a faster folding rate[44]. The K12M mutant of NTL9(1-39) trajectories provided by Shaw et al. were four MD simulations of 1052μs, 990μs, 389μs, and 377μs at 355K with CHARMM22* force field[38].

1. Markov State Model construction

The MSM of the GTT WW domain at 360K was previously descried by Wan[45], which used eight tICA components and a lag time of 10ns to find the best low dimensional subspace. A 1000 micro-state MSM with a lag time 100ns using k-center clustering,which was determined by GMRQ method,[?] to best capture its folding behaviors. The relaxation time scales of 10.2μs was estimated from MSM, which was well matched to a folding time 21 6μs predicted by Shaw[38] and 8μs for a three-state model of GTT folding by Beauchamp[46]. As a better test of different estimating approaches due to large timing used in OOM analysis, we built a 200 micro-state MSM which showed 2.3μs folding time.

The MSM of the K12M mutant of NTL9(1-39) was previously described by Schwantes and Lin[43,46,47]. A folding time 18μs was observed from the MSMs and the 6 components was chosen as input for tICA metric to best project trajectories on low-dimensional subspace. The optimal lagtime 200ns and 1000 micro-states, determined by GMRQ[48], were used to accurately capturing folding dynamics of NTL9(1-39). A relaxation time 18μs was observed, compares well to 29μs predicted by Shaw[38] and to the folding rate of 26.4 μs (at 80 °C), obtained from the T-jump experiments[47]. Similarly to GTT system, we built a 200 micro-state MSM for a better comparison using didfferent estimators, which showed 5.2μs folding time.

Reseeding simulations trajectories starting from each micro-state of both GTT and NTL9 system were randomly picked from these MSMs built from these two ultra-large data-sets($\sim$1.1ms for GTT and $\sim$ 2.8ms for NTL9). Totally, 12 and 14 reversible folding behaviors were observed for GTT and NTL9 in these two data-sets respectively[38]. We believe that random seeded simulations trajectories from these two data-sets are amenable to explore the effects of reseeding methods. To test the effects from reseeding numbers of trajectories and lengths of trajectories, several rounds of analysis were managed to construct the reseeding simulation trajectories. We collected 5 and 10 reseeding trajectories starting from all micro states with different lengths from both 200 micro-state and 1000 micro-state models and explored the performances of different estimators. The sizes of cumulative seeding trajectories for each MSM were chosen to be less than original ultra-large data-sets.

The equilibrium populations used to do population re-weighting estimation and known population estimation were predicted from MSMs. And notably, Row-normalized estimator could not work with large lagtime in short seeding trajectories, thus, we added the minimum counts to each element in count matrices to avoid getting un-representable values in division calculations. We are expecting a faster folding rate estimation and a more evenly distributed population estimation.

2. Predictions of Folding Rates from reseeding trajectories with different estimators

Given that our references 200 and 1000 micro-state MSMs predicted the GTT WW domain and NTL9 have folding time at 2.3, 10.2[45] and 5.2, 18 μs[43,47] respectively. We then applied different estimator approaches to our reseeding trajectories to predict both kinetics and thermodynamics of each system. The computed folding relaxation time scales from MLE estimator for these 200 micro-state MSMs built from 5 independent seeding trajectories starting from each micro state are within tolerance but way off for 1000 micro-state MSMs, even we extended all the trajectories length from 50ns to 250ns in 1000 micro-state MSMs, which still failed to improve the time scales estimations. We then examined the performance of MLE estimator with 10 independent seeding trajectories form each state, and again MLE inaccurately

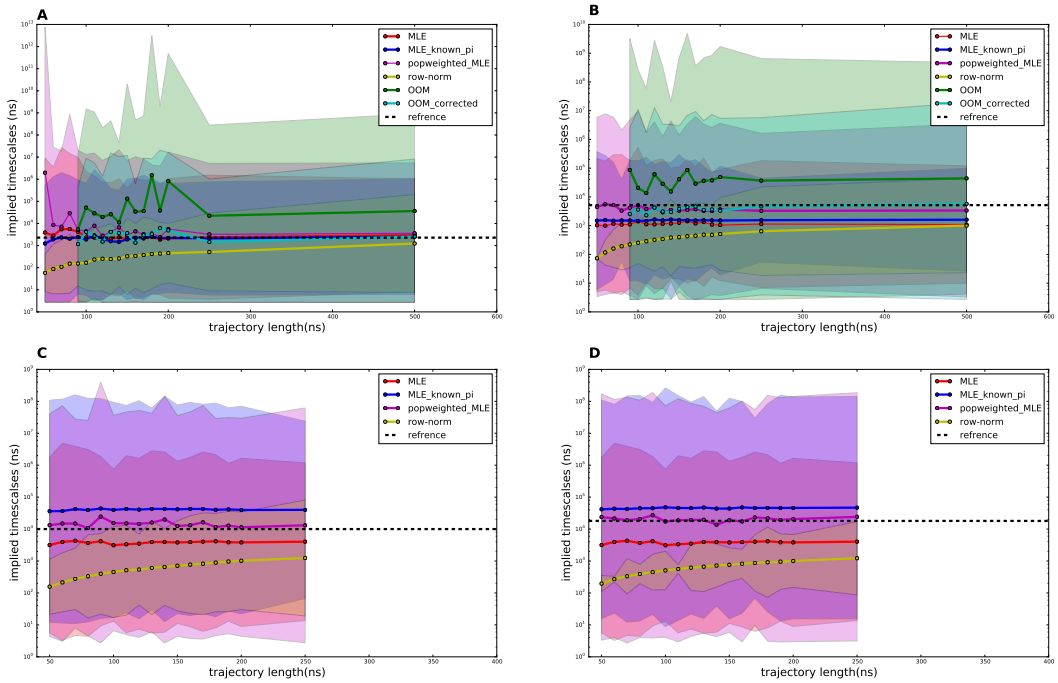

FIG. 3. ((a)and(b)) are the estimates for the slowest relaxation time scales from different estimators on the 200-microstate MSMs of GTT(left) and NTL9(right).((c)and(d)) are the estimates for the slowest relaxation time scales from different estimators on the 1000 micro-state MSMs of GTT(left) and NTL9(right).

predicted all relaxation timescales(SI).

The folding time scales of GTT and NTL9 predicted with Maximum Likelihood estimator(MLE) with known populations successfully rescued the poor performances with MLE estimator, which yield really closed folding rates comparing to the references. Importantly, we note that the rate estimations do not rely on the length of trajectories.This is of course a huge advantage in doing seeding simulations which would reduce the computational time. We also find that 5 seeding trajectories suffering minimally comparing to more seeding trajectories(SI). However, only native state population was known at most time ahead of simulations,which limited the usages with this estimator in common model-builds.

With population-weighted trajectory counts, Maximum-likelihood estimator (MLE) correctly rescued the relaxation time estimations in both 200 and 1000 micro-state MSMs. Amazingly, we note that this estimator does not depend on the length of trajectories. The data size needed to accurately capture the right folding time is much smaller than the original data-sets, which reduces the computational time a lot. More importantly, this estimator is more applicable on predicting the effects from perturbations including mutations, temperatures and different force-fields if a well-defined MSM was provided ahead.

Row-normalized counts estimator can only work if the connectivity between states is not zero, however it is tough to be approached with a larger lag time in short seeding simulations. The performances in both GTT and NTL9 systems were so poor because of the poor connectivity, even with the pseudo counts added to purposely make this estimator work. However, it is a mathematical solution in a long seeding simulations, which is against the purpose of doing short seeding simulations.

The OOM framework with bootstrapping count metric only works with longer seeding simulations, which ensures the connectivity between the pair-wised states. And the OOM overestimates the relaxation timescales in both GTT and NTL9 systems, while the corrected MSMs estimate time scales are one magnitude smaller and closely capture the referenced range. Furthermore, the practical application of OOM methods are limited by the number of states, simply due to the (N^3) scaling of the algorithm limits the calculation for the all-atom models to 200 states.

3. Predictions of native-state population from reseeding trajectories with different estimators

Both GTT and NTL9 were two classical two-state folder systems from extensively studied. The accuracies of native state population estimations are expected to closely match to the predictions on the kinetics because of the expected single energy barrier in these two systems. The analysis above showed MLE with known populations estimator could accurately capture the slow-

est relaxation time,thus the native state population was expected to be well estimated, which is always true in this case. The row-normalized counts estimator all had a fast folding rate for both GTT and NTL9 systems, which indicated these estimators inaccurately predicted a flatter energy barriers for these two-state folders. Notably, row-normalized counts estimator wrongly predicted a small native-state population due to the pseudo counts, which made the folding landscape even flatter in this 1000 micro-state model. The MLE estimator underestimated the native state population in both 200 and 1000 micro-state models. The MLE with population-weighted trajectory counts estimator successfully capture the the native-state population in 200 micro-state model but a little over estimated the native state population in 1000 micro-state models, however, the estimation got better with longer seeding simulations. Furthermore, The dataset needed to get a converged population estimation with population-weighted MLE is much smaller than the other estimators. The OOM poorly estimated the population of the native state in both GTT and NTL9 systems and the estimates were independent on the the trajectory length, where it may be resulted from bad state discretization or much longer seeding simulations trajectories are needed for the application of this method.

4. Predictions of mutation effect from reseeding trajectories with different estimators

As a further test of the Maximum-likelihood estimator (MLE) with population-weighted trajectory counts estimator in predicting perturbation effects, several more MSMs were built from seeding trajectories of GTT and FiP35 from four additional GTT simulations trajectories(83, 118, 124, and 272 μs)[36,37] with and six 100 μs FiP35 simulations trajectories with AMBER ff99SB-ildn force field at 395K[36,49].As a fast folding sequence, FiP35 has a folding time of 13μs which is slower than GTT mutant[?] .

The results in 200 micro-state model for both GTT and FiP35, OOM overestimated the folding times and all the other estimators undereatimated the folding time. However, it is amazingly that population-weighted trajectory counts estimator closely captured the folding time for both GTT and FiP35. And population-weighted estimator were the best to predict the native-state populations always. Notably, population-weighted estimator did not successfully predict the folding time and notabilities resulted in mutation effects, but we assume it may be caused by finite sampling errors, however, the population-weighted estimator was independent on seeding trajectory length in both systems, which raised our interests about how robust that this estimator can work in predicting the larger perturbations of landscape from mutations, force fields, temperature etc.

III. DISCUSSION

The estimators approaches turned out that the folding time estimation from common MLE suffers greatly from doing seeding simulations; MLE with known populations incredibly rescued the relaxation time, but limited to use due to unknown equilibrium populations ahead of analysis; Row-normalized estimator could not always work with large lagtime in short seeding simulations, which obey the purposes of doing seeding simulations; MLE with population-weighted trajectory counts estimator also successfully improved the kinetics estimators and independent on the lengths of seeding trajectories. The potential applications of population-reweighted MLE on Markov State models are exciting. One application can be used to predict larger perturbation effects from different temperatures, force fields without running ultralong simulations. However, the success of population-reweighted MLE is highly dependent on that perturbations wont vary too much on the metastable state definitions. Our method is similar to the weighted ensemble (WE), which each trajectory is assigned with a weight. However, WE mainly focus on pathway sampling where MSM can predict both kinetics and thermodynamics at the same time. Moreover, our approach is from a well-defined MSM, which all seeding simulations optimally can get converged faster.

This idea–that both kinetic and thermodynamic information are necessary to obtain good MSMs–has motivated the development of multiensemble MSM estimators like DTRAM, TRAM, DHAM and DHAMed. These estimators combine both free energy estimation and MSM transition rate estimation, leveraging both thermodynamic and kinetic information from sampled trajectories to estimate an MSM.

All AIP journals require that the initial citation of figures or tables be in numerical order. LaTeX's automatic numbering of floats is your friend here: just put each `figure` environment immediately following its first reference (`\ref`), as we have done in this example file.

ACKNOWLEDGMENTS

The authors thank the participants of Folding@home, without whom this work would not be possible. We graciously acknowledge D. E. Shaw Research for providing access to the WW domain folding trajectory data. This research was supported in part by the National Science Foundation through major research instrumentation grant number CNS-09-58854 and National Institutes of Health grants 1R01GM123296-01 and NIH Research Resource Computer Cluster Grant S10-OD020095.

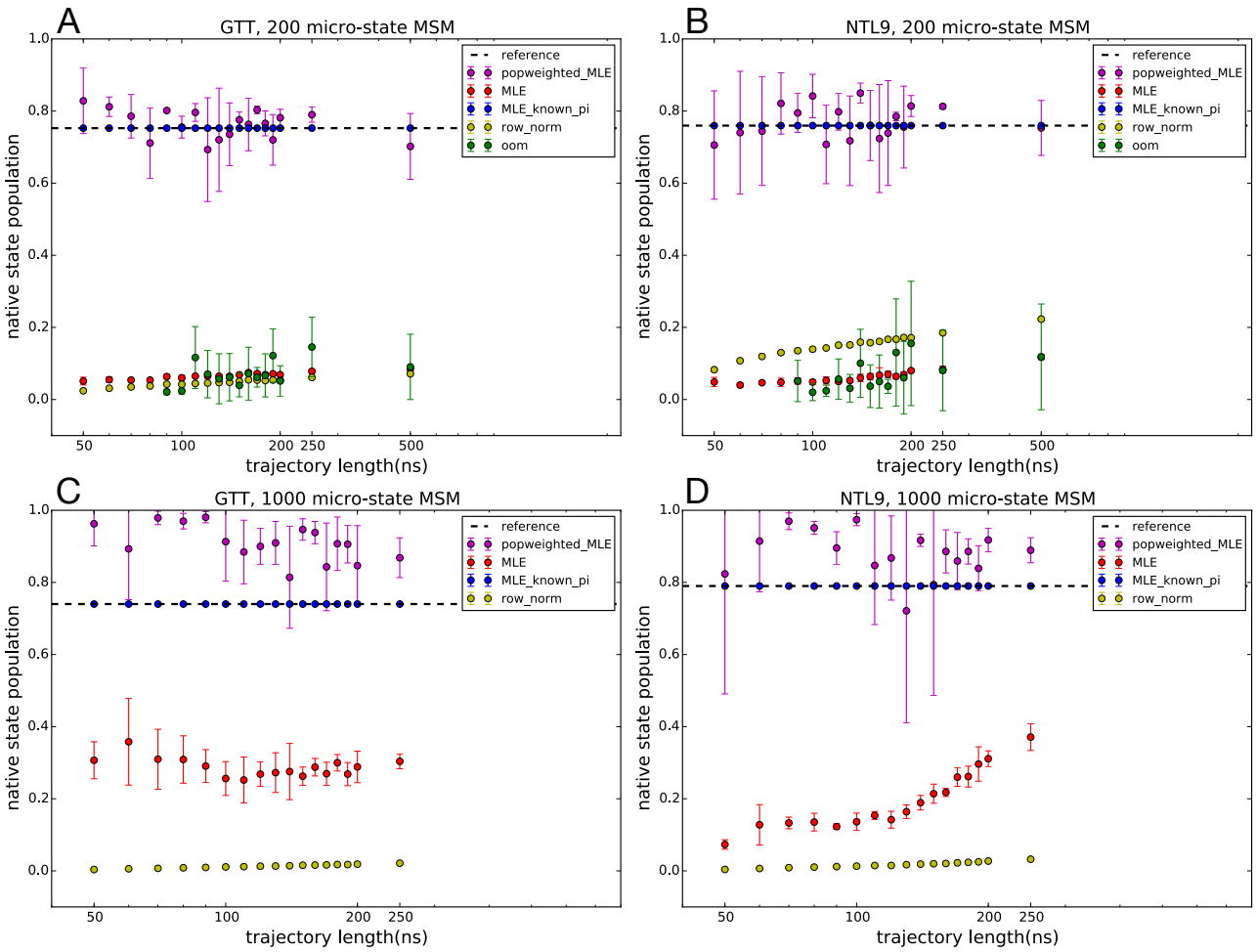

FIG. 4. The native state population estimations of both GTT((a) and (c)) and NTL9((b) and (d)) with different estimators from 200 and 1000 micro-state MSMs.

IV. BULLET POINTS

A. The accuracy of different estimators on seeding data from GTT and NTL9

In general, for this kind of seeding data, row-normalized counts work poorly in estimating both timescales and native populations. MLE does better at estimating timescales, but not native populations. The two methods that "rescue" the native populations (by properly weighting the contributions of counts) are MLE-known-pi and population reweighting. By definition. the known-pi constraints the populations, while pop-reweighted has a lot of potential errors, but estimate values in the right range.

What about OOM and corrected timescales? Our results suggest that these methods don't work particularly well with short re-seedin data. The OOM timescales are pretty good, but the native populations are very bad. Furthermore, the practical application of OOM methods are limited by the number of states, simply due to the $O(N^3)$ scaling of the algorithm limits the calculation for the all-atom models to 200 states.

When we are able to increase the number of states in the all-atom model to 1000 state,both the timescales and the populations are much better, and furthermore, we can get good estimates form much shorted trajectory lengths (smaller than the complete 1 ms data set)

Lastly, we can use the populations of GTT to reweight a mutant srquence, Fip35, and get pretty good est,ates of the slower rate of this folder.

[1]F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. U. S. A. **106**, 19011 (2009).
[2]V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, J. Am. Chem. Soc. **132**, 1526 (2010).
[3]J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, J. Chem. Phys. **134**, 174105 (2011).
[4]J. D. Chodera and F. Noé, Curr. Opin. Struct. Biol. **25**, 135 (2014).
[5]F. Noé, J. Chodera, G. Bowman, V. Pande, and F. Noé, "An introduction to markov state models and their application to long timescale molecular simulation, vol. 797 of advances in experimental medicine and biology," (2014).
[6]X. Huang, G. R. Bowman, S. Bacallado, and V. S. Pande, Proceedings of the National Academy of Sciences **106**, 19765 (2009).
[7]V. A. Voelz, B. Elman, A. M. Razavi, and G. Zhou, J. Chem. Theory Comput. **10**, 5716 (2014).
[8]Z. Shamsi, A. S. Moffett, and D. Shukla, Nature Publishing Group **7**, 12700 (2017).
[9]S. Doerr and G. De Fabritiis, Journal of chemical theory and computation **10**, 2064 (2014).

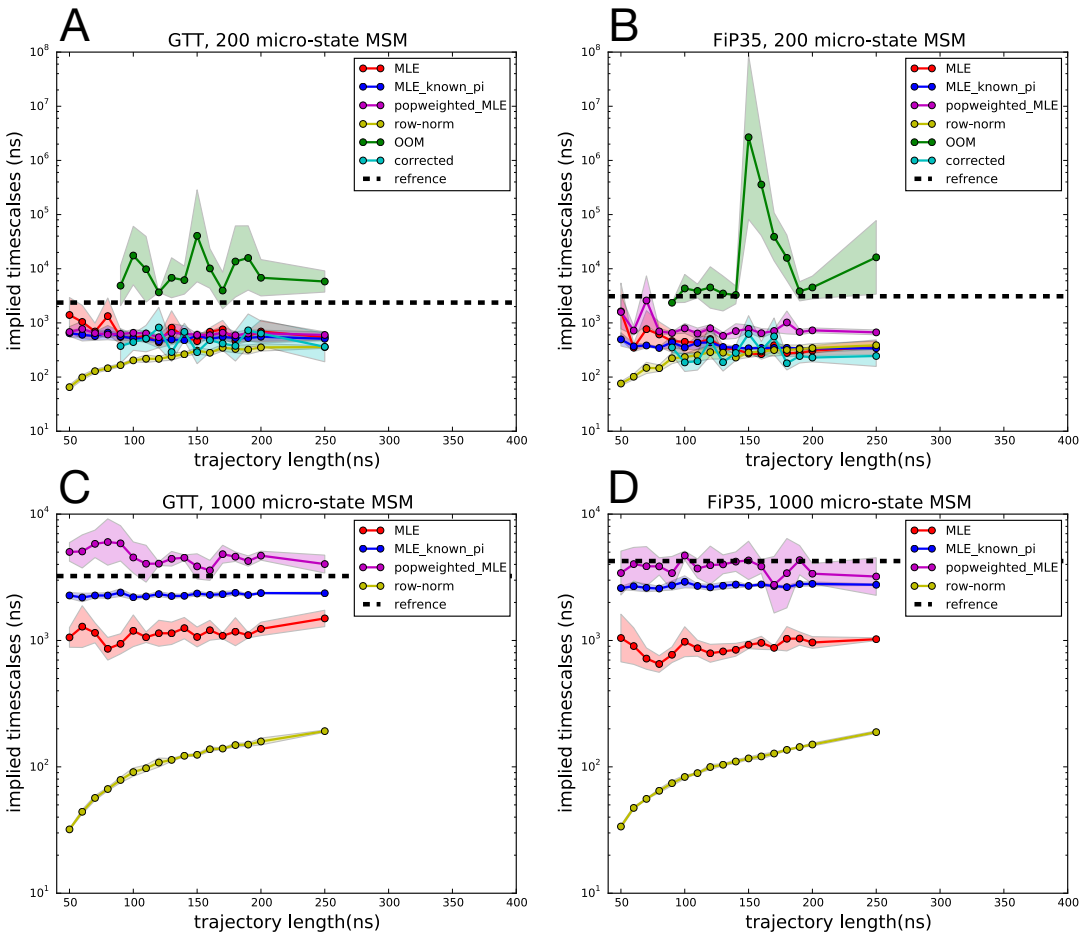

FIG. 5. Maximum-likelihood estimator (MLE) with population-weighted trajectory counts estimator for FiP35(right) and MLE estimator for GTT (left) show FiP35 folds slower than GTT

[10] M. I. Zimmerman and G. R. Bowman, J. Chem. Theory Comput. **11**, 5747 (2015).
[11] Z. Shamsi, K. J. Cheng, and D. Shukla, arXiv preprint arXiv:1710.00495 (2017).
[12] M. I. Zimmerman, J. R. Porter, X. Sun, R. R. Silva, and G. R. Bowman, arXiv preprint arXiv:1805.04616 (2018).
[13] B. W. Zhang, D. Jasnow, and D. M. Zuckerman, The Journal of Chemical Physics **132**, 054107 (2010).
[14] M. C. Zwier, J. L. Adelman, J. W. Kaus, A. J. Pratt, K. F. Wong, N. B. Rego, E. Suárez, S. Lettieri, D. W. Wang, M. Grabe, D. M. Zuckerman, and L. T. Chong, Journal of chemical theory and computation **11**, 800 (2015).
[15] A. Dickson and S. D. Lotz, The Journal of Physical Chemistry B **120**, 5377 (2016).
[16] S. D. Lotz and A. Dickson, Journal of the American Chemical Society , jacs.7b08572 (2018).
[17] T. Dixon, S. D. Lotz, and A. Dickson, Journal of Computer-Aided Molecular Design **15**, 547 (2018).
[18] B. Trendelkamp-Schroer and F. Noe, Physical Review X **6**, 011009 (2016).
[19] V. S. Pande, K. Beauchamp, and G. R. Bowman, Methods (San Diego, Calif.) **52**, 99 (2010).
[20] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, Biophysj **109**, 1528 (2015).
[21] E. H. Kellogg, O. F. Lange, and D. Baker, The Journal of Physical Chemistry B **116**, 11405 (2012).
[22] P. Metzner, F. Noé, and C. Schütte, Physical Review E **80**, 021106 (2009).
[23] F. Nüske, B. G. Keller, G. Perez-Hernandez, A. S. J. S. Mey, and F. Noé, J. Chem. Theory Comput. **10**, 1739 (2014).
[24] L. S. Stelzl, A. Kells, E. Rosta, and G. Hummer, Journal of chemical theory and computation **13**, 6328 (2017).
[25] H. Wu, A. S. J. S. Mey, E. Rosta, and F. Noé, The Journal of Chemical Physics **141**, 214106 (2014).
[26] H. Wu, A. S. J. S. Mey, E. Rosta, and F. Noé, The Journal of Chemical Physics **141**, 214106 (2014).
[27] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, J. Chem. Phys. **131**, 124101 (2009).
[28] B. Trendelkamp-Schroer and F. Noé, Physical Review X **6**, 011009 (2016).
[29] H. Nguyen, M. Jäger, A. Moretto, M. Gruebele, and J. W. Kelly, Proc. Natl. Acad. Sci. U. S. A. **100**, 3948 (2003).
[30] M. Jäger, Y. Zhang, J. Bieschke, H. Nguyen, M. Dendle, M. E. Bowman, J. P. Noel, M. Gruebele, and J. W. Kelly, Proc. Natl. Acad. Sci. U. S. A. **103**, 10648 (2006).
[31] F. Liu and M. Gruebele, Chem. Phys. Lett. **461**, 1 (2008).
[32] K. Dave, M. Jäger, H. Nguyen, J. W. Kelly, and M. Gruebele, Journal of Molecular Biology **428**, 1617 (2016).
[33] F. Liu, D. Du, A. A. Fuller, J. E. Davoren, P. Wipf, J. W. Kelly, and M. Gruebele, Proc. Natl. Acad. Sci. U. S. A. **105**, 2369 (2008).
[34] Y. Gao, C. Zhang, X. Wang, and T. Zhu, Chemical Physics Letters **679**, 112 (2017).
[35] R. B. Best, G. Hummer, and W. A. Eaton, Proceedings of the National Academy of Sciences **110**, 201311599 (2013).
[36] S. Piana, K. Sarkar, K. Lindorff-Larsen, M. Guo, M. Gruebele, and D. E. Shaw, Journal of Molecular Biology **405**, 43 (2011).
[37] Houbi Nguyen, Marcus Jäger, J. W. Kelly, , and Martin Gruebele, *Engineering a β-Sheet Protein toward the Folding Speed*

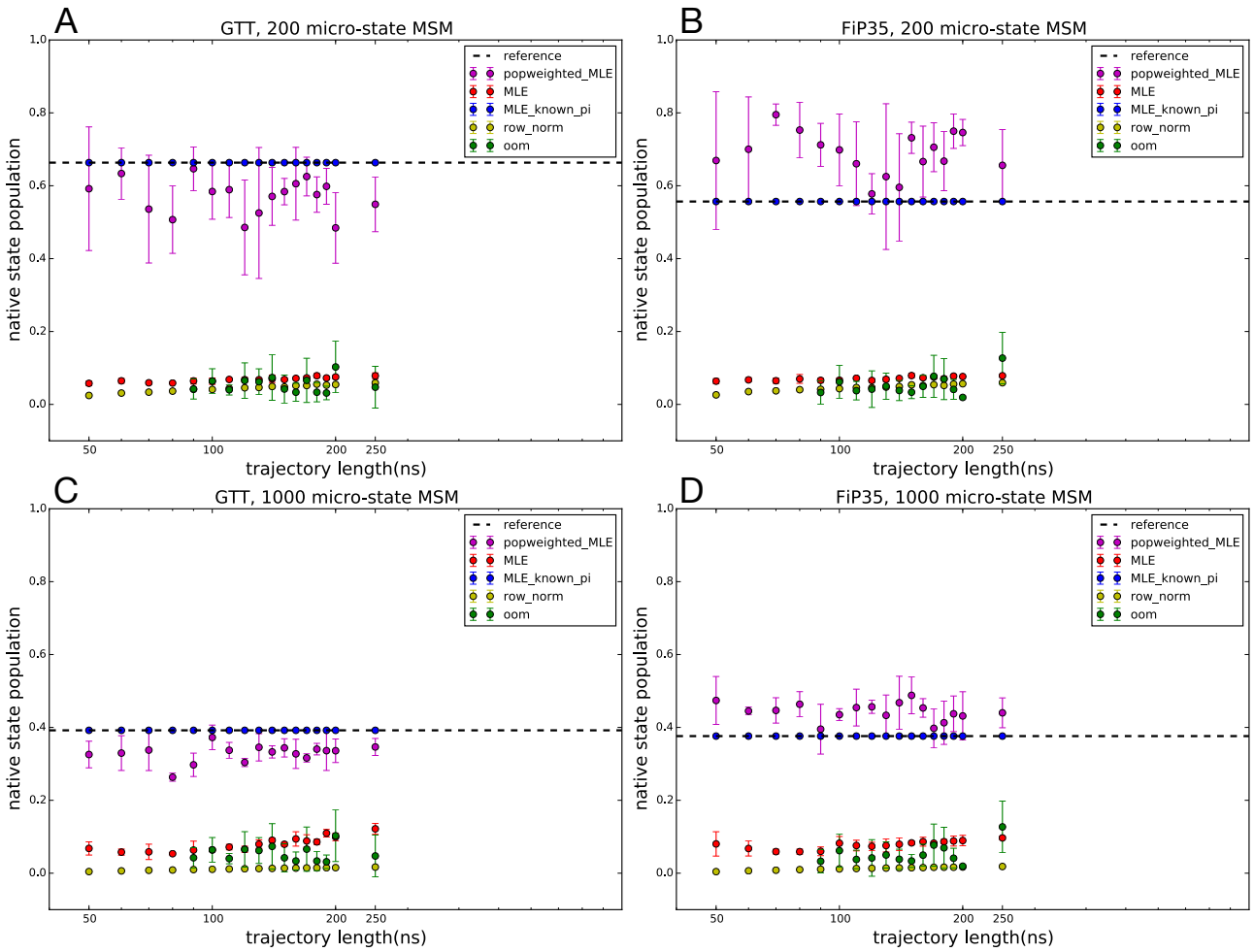

FIG. 6. The native state population estimations of both GTT((a) and (c)) and NTL9((b) and (d)) with different estimators from 200 and 1000 micro-state MSMs.

Limit, Vol. 109 (American Chemical Society, 2005).
[38] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, Science (New York, N.Y.) **334**, 517 (2011).
[39] J.-H. Cho, S. Sato, and D. P. Raleigh, Journal of Molecular Biology **338**, 827 (2004).
[40] J.-H. Cho and D. P. Raleigh, Journal of Molecular Biology **353**, 174 (2005).
[41] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, Journal of the American Chemical Society **132**, 1526 (2010).
[42] D. W. Hoffman, C. Davies, S. E. Gerchman, J. H. Kycia, S. J. Porter, S. W. White, and V. Ramakrishnan, The EMBO . . . **13**, 205 (1994).
[43] C. R. Schwantes and V. S. Pande, Journal of chemical theory and computation **9**, 2000 (2013).
[44] J.-H. Cho, W. Meng, S. Sato, E. Y. Kim, H. Schindelin, and D. P. Raleigh, Proceedings of the National Academy of Sciences **111**, 12079 (2014).
[45] H. WAN, G. Zhou, and V. A. Voelz, Journal of chemical theory and computation , acs.jctc.6b00938 (2016).
[46] K. A. Beauchamp, R. McGibbon, Y.-S. Lin, and V. S. Pande, Proc. Natl. Acad. Sci. U. S. A. **109**, 17807 (2012).
[47] C. R. Baiz, Y.-S. Lin, C. S. Peng, K. A. Beauchamp, V. A. Voelz, V. S. Pande, and A. Tokmakoff, Biophysj **106**, 1359 (2014).
[48] R. T. McGibbon and V. S. Pande, The Journal of Chemical Physics **142**, 124105 (2015).
[49] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, Science **330**, 341 (2010).

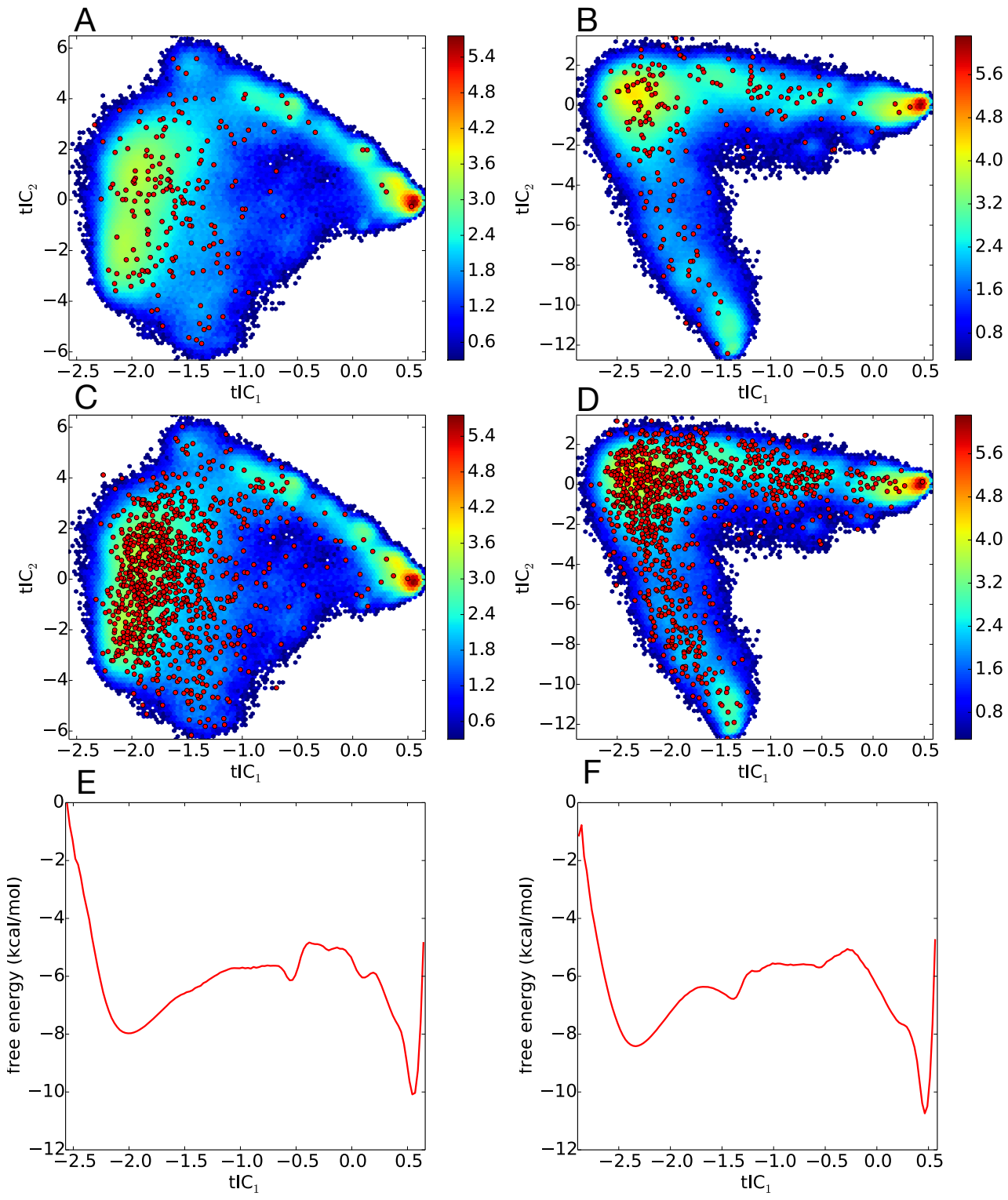

FIG. 7. Projection of GTT ((a) and (c)) and NTL9 ((b) and (d)) trajectory data onto the two largest tICA components, overlaid with the locations of the 200 and 1000 microstates respectively (red dots), and their free energy profiles calculated from the first tICA subspace with bin-width=0.025 unit.

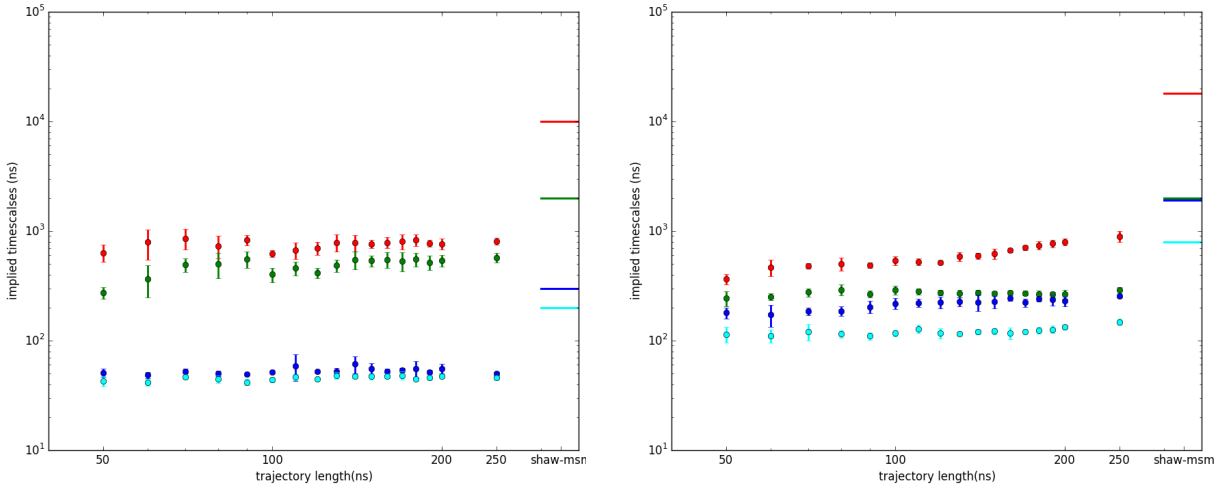

FIG. 8. Maximum-likelihood estimator (MLE) for both GTT(left) and NTL9(right) under-estimates implied timescales.

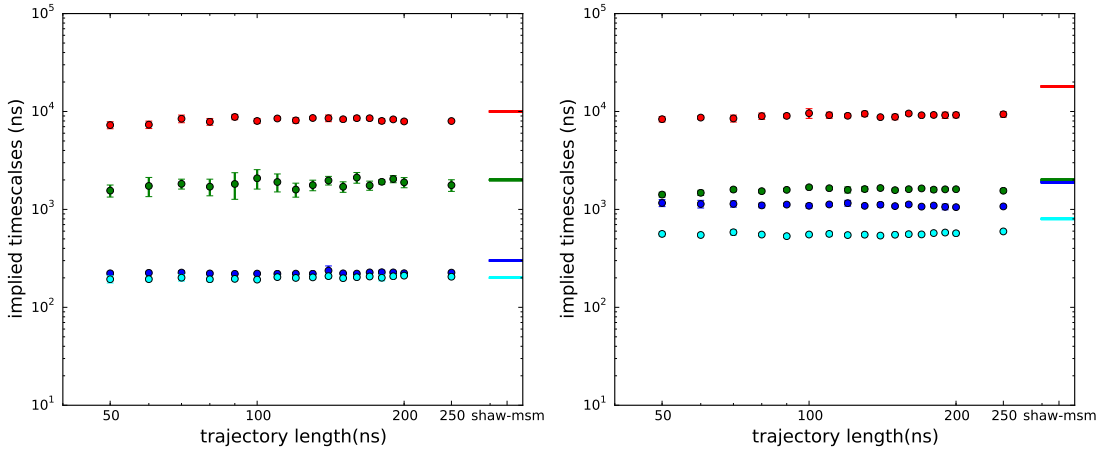

FIG. 9. Maximum-likelihood estimator (MLE) with known populations estimator (MLE) for both GTT(left) and NTL9(right) well predicted implied timescales.

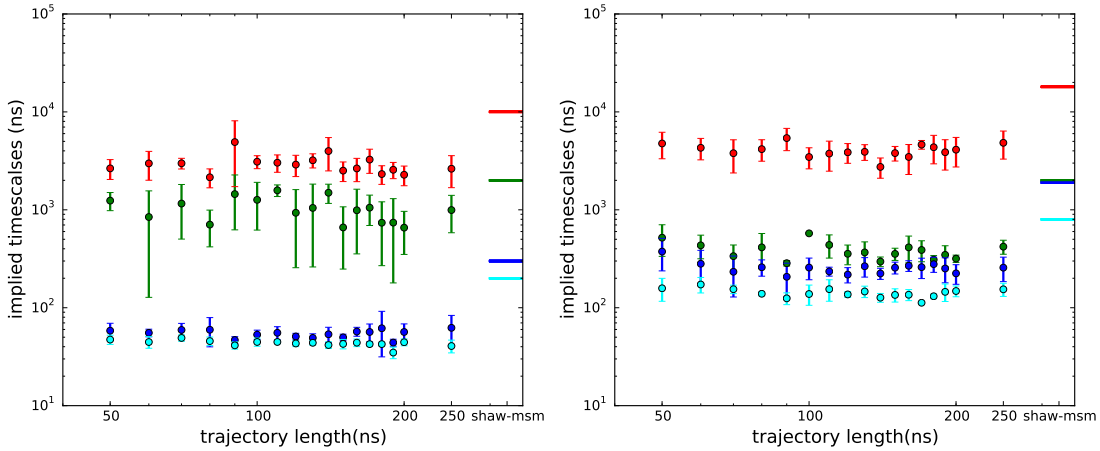

FIG. 10. Maximum-likelihood estimator (MLE) with population-weighted trajectory counts estimator (MLE) for both GTT(left) and NTL9(right) rescued implied timescales.

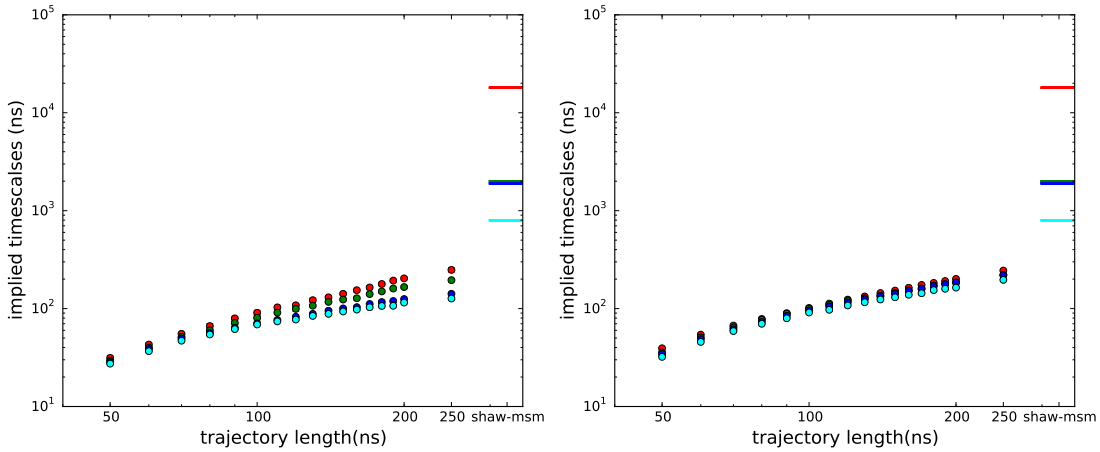

FIG. 11. Row-normalized counts estimator for both GTT(left) and NTL9(right) badly predicted implied timescales.

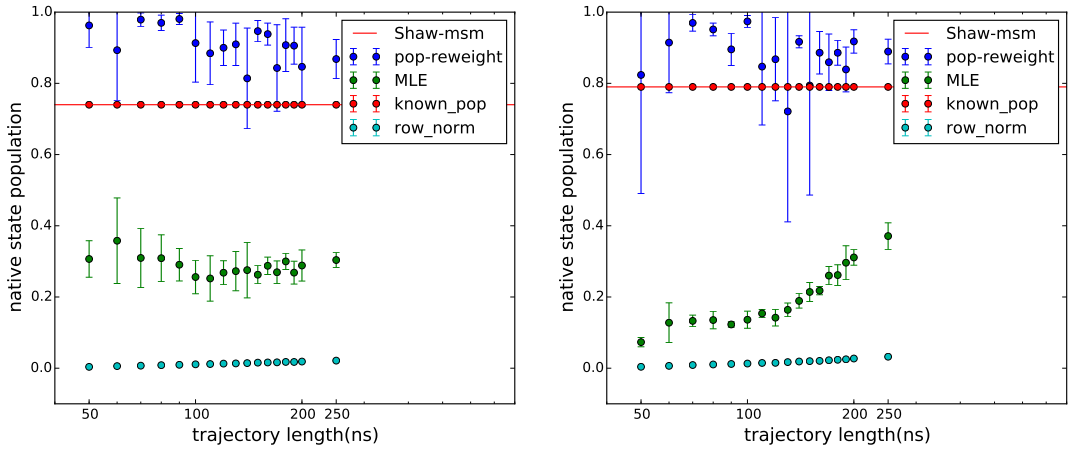

FIG. 12. The native state population estimations of both GTT(left) and NTL9(right) with different estimators.