A new class of enhanced kinetic sampling methods for building Markov state models

Arti Bhoutekar,[1] Susmita Ghosh,[2] Swati Bhattacharya,[1,2] and Abhijit Chatterjee[1,a)]
[1]*Department of Chemical Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India*
[2]*Department of Physics, Indian Institute of Technology Guwahati, Guwahati 781039, India*

(Received 23 February 2017; accepted 22 May 2017; published online 6 June 2017)

Markov state models (MSMs) and other related kinetic network models are frequently used to study the long-timescale dynamical behavior of biomolecular and materials systems. MSMs are often constructed bottom-up using brute-force molecular dynamics (MD) simulations when the model contains a large number of states and kinetic pathways that are not known *a priori*. However, the resulting network generally encompasses only parts of the configurational space, and regardless of any additional MD performed, several states and pathways will still remain missing. This implies that the duration for which the MSM can faithfully capture the true dynamics, which we term as the validity time for the MSM, is always finite and unfortunately much shorter than the MD time invested to construct the model. A general framework that relates the kinetic uncertainty in the model to the validity time, missing states and pathways, network topology, and statistical sampling is presented. Performing additional calculations for frequently-sampled states/pathways may not alter the MSM validity time. A new class of enhanced kinetic sampling techniques is introduced that aims at targeting rare states/pathways that contribute most to the uncertainty so that the validity time is boosted in an effective manner. Examples including straightforward 1D energy landscapes, lattice models, and biomolecular systems are provided to illustrate the application of the method. Developments presented here will be of interest to the kinetic Monte Carlo community as well. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4984932]

I. INTRODUCTION

Recent years have witnessed the widespread use of Markov state models (MSMs)[1–5] within the biophysics community. MSMs are detailed kinetic network models wherein the configurational space of a biomolecule under study is partitioned into states. The dynamical evolution of the system is approximated in terms of state-to-state transitions. The number of states can range between tens to thousands depending on the complexity and level of coarse-graining. Each node in the network denotes a metastable state of the system while the connections between the nodes provide rates of interconversion between the states. MSMs have become useful tools for probing the dynamics of nucleic acids and proteins,[6–9] e.g., folding and unfolding events at long time scales. Though we restrict ourselves to biomolecular systems, MSMs are closely related to kinetic Monte Carlo (KMC) models[10–12] used in the materials and reactions areas for studying catalysis,[13] crystal growth,[14] material processing,[15] and adsorption phenomena[16] to name a few. Both approaches solve a master equation and have benefitted from the exchange of ideas between the respective communities. For instance, knowledge of the network structure can be exploited to accelerate the KMC dynamics by eliminating fast degrees of freedom.[17–20] Despite their widespread usage, some aspects of MSM construction are still poorly understood.

A key step in the MSM construction entails determining states and kinetic pathways to be included in the model. The availability of a large number of parallel processors has enabled rapid construction of high fidelity MSMs using brute-force molecular dynamics (MD) calculations.[21–23] Herein states and kinetic pathways are identified via coarse-graining several independent MD trajectories. The MD trajectories can be seeded from different starting configurations, which allows for better sampling of the configurational space. Other simulation techniques offering resolution greater than the MSMs can also be employed.[24–26] Enhanced thermodynamic-sampling techniques that can sample rare events with large energy barriers can aid in the efficient construction of the model.[27–31] However, often overlooked is an additional challenge associated with building MSMs (and indeed with KMC models as well[32]); namely, a fundamental limitation remains that the entire configurational space cannot be sampled by a finite number of MD trajectories, i.e., a MSM is never complete. Even when the MD trajectories used for network-building collectively exceed microsecond time scales, there are bound to be rare states and pathways missing from the MD data. When *relevant* states and pathways are missing from the constructed MSM, thermodynamic/kinetic quantities being sought can be inaccurate. The main purpose of this article (as part of the *Special Topic Issue on Reaction Pathways*) is to highlight the danger arising from missing relevant states and pathways in a network, develop a strategy to quantify the *completeness* of a kinetic network model, and identify regions of configuration

[a)] Electronic addresses: swaticb@che.iitb.ac.in and abhijit@che.iitb.ac.in space relevant to the dynamical evolution, which can guide further network construction.

Many network-building procedures entail pruning/lumping of states and kinetic pathways to enforce detailed balance and avoid absorbing states. Although the length of the MD trajectory used to build a network is generally reported, it is not enough to establish the maximum duration for which the dynamics is faithfully predicted by the network model. In the worst case, missing states can be important to the ensemble-average quantities calculated from an "incomplete" network model. Given that network models are nowadays generated by seeding the MD calculations starting from different states while using a variety of computational tricks, comparing the dynamics from models for the same system is subject to error/uncertainty resulting from the missing kinetic information. A conceptual framework that accounts for missing states/pathways will bolster endeavors to generate reliable network models.

Estimators for missing rates from a state first developed in Refs. 33 and 34 have been applied to different material systems.[32,35–37] However, more than the missing rates, conceptually, it is the time scale where the missing pathways become relevant to the dynamics that is of interest. Extending our previous work, the largest time scale where the network model continues to yield the correct dynamics, termed as the validity time for the model, is introduced here. The validity time allows one to systematically compare the behavior of two models of the same system while accounting for known kinetic rates, topology, and relaxation times of the network, as well as missing pathways and states that have not been included in the model. The main idea is contingent on identifying states that may have a large probability flux into configurational space that is not part of the existing model. One can then compute the validity time where the error in the dynamics is small, i.e., the existing network model can be regarded as complete till its validity time and is safe to use. The theoretical underpinning of the validity time is discussed in Sec. II.

An enhanced kinetic-sampling technique called programmed state-constrained calculation is presented that guides selection of states where additional MD must be performed to extend the validity time. The usefulness of the validity time is illustrated in Sec. III with the help of prototype network models that are completely known to us at the outset. We demonstrate application of MSMs of a desired validity time to the study of stretched deca-alanine under tension in Sec. IV. Using the validity time, we conclude that a large number of rarely visited states are also relevant to the dynamics. The MSMs provide new physical insights into state-specific properties that are crucial towards understanding the folding/unfolding mechanism and the forces required for pulling the molecule. Finally, conclusions are presented in Sec. V.

II. VALIDITY TIME FOR A MARKOV STATE MODEL

Consider a MD trajectory of a duration τ_{MD}. Analysis of the trajectory using a combination of distance metrics[38–43] and clustering methods,[4,44] and tests for the Markovian approximation (e.g., implied time scales for discrete-time MSMs[23] and tests for first-order behavior for continuous time MSMs[36,45]) can yield information about the states of the system and the associated kinetic rates. Although states are randomly visited in the trajectory, the occupation $\pi_S(t)$ for a Markov state S at time t is deterministic and is given by the master equation

$$\frac{d}{dt}\pi_S(t) = \sum_{S'\neq S} k_{S'\to S}\pi_{S'}(t) - \sum_{S'\neq S} k_{S\to S'}\pi_S(t). \tag{1}$$

Here, $k_{S\to S'}$ is the kinetic rate from state S to state S', and $k_{S'\to S}\pi_{S'}$ and $k_{S\to S'}\pi_S$ denote the inflow and outflow probability flux for S. Kinetic rates can be obtained using a statistical approach, such as maximum likelihood estimation (MLE).[46] While Eq. (1) forms the basis of a MSM, the MSM constructed using finite MD trajectories is approximate because of various errors,[47,48] statistical uncertainty,[49,50] as well as missing kinetic information.[33,34] Here, we shall focus on error from the missing kinetic information. Equation (1) can be written as

$$\frac{d\pi}{dt} = T\pi, \tag{2}$$

the continuous-time MSM, where T is the rate matrix. Equations (1) and (2) are solved with a specified initial distribution and rate matrix.

The number of states and pathways in the MSM can increase when additional MD data are made available. Pathways with large (small) probability flux are more (less) likely to be selected in the dynamics. Due to a number of factors including randomness inherent in sampling, time scales accessed, the network topology, and the starting conformation, it is possible that certain pathways that have a reasonably large probability flux are still missing in the MD trajectory. As a consequence, states that can be reached only via the missing pathways are also missing in the MSM. It is well known that topologically different MSMs are often generated for the same system when MD calculations are seeded from different starting conformations that are separated by large free energy barriers (even when enhanced conformation sampling techniques are used[27,51–56]). The MD time spent in each state is a key parameter that affects the MSM accuracy.

It is convenient to define a time τ_V termed as the MSM validity time such that all kinetic pathways that are likely to be selected within τ_V are present in the existing MSM. Pathways that are less relevant can be missing from the MSM without affecting the accuracy of the kinetic model. MD data pertaining to less relevant pathways do not result in appreciable increase in τ_V. Next, we relate τ_V to the MD time.

A. Core, periphery, and missing states

States in Eq. (1) can be partitioned into three types: core, periphery, and missing states. A state where the system has resided for a significant time in the MD calculation is termed a core state. The probability flux out of the core states can be estimated from the MD data, i.e., they constitute the *source* terms in Eq. (1). For the remainder of our discussion, a MSM is the core network model, and we use these terms interchangeably. The MSM is ergodic since it is possible to reach a core state from any part of the network. States that are visited for a short time in the MD calculation preventing estimation of kinetic rates from such states with reasonable confidence are termed as periphery states. Note that the rates from core states to periphery states might be available. The need for periphery states will become clear in Sec. II D. These states are redundant at the MSM validity times. Periphery states correspond to *absorbing* or *sink* states in Eq. (1). The dynamics of periphery states can be quite different from the one predicted by the MSM at longer times. We term a network model comprising of core and periphery states as a full network model. The validity time of a MSM can be increased by performing additional MD in the core and periphery states. A periphery state becomes a core state when sufficient time has been spent in the state so that one/more kinetic rates can be estimated. States that have never been visited in the MD trajectory are termed as missing states. Some missing states later become periphery or core states as additional MD is performed. Estimates for the rates from the core states to the missing states are given in Sec. II B.

Figure 1 shows the structure of the rate matrix in Eq. (2) constructed with MD where core, periphery, and missing states are considered. Each off-diagonal term in the row (column) j in the matrix denotes the kinetic rates into (from) the state j. Core (periphery and missing) states are placed at the top (bottom) of the occupation vector on the right-hand side of Eq. (2). The matrix can be divided into three parts. The top-left (green) corner involves the core states, i.e., it forms the rate matrix for the MSM. We lump the periphery and missing states together as absorbing states. The last column of the matrix contains the rates from the absorbing states. The rates are set to zero. The bottom row of the matrix contains estimates of rates (termed *leakage* rates) from the core states to the absorbing states. As we shall show next, the validity time of the MSM is determined by the leakage rates, which in turn depends on the time spent in the core states. The validity time for the core network can be made large when all leakage rates are kept small.

B. Upper bound for the missing rate from a core state

Consider a collection of pathways from a state S with total rate k. Assuming first order kinetics, the probability of not selecting these paths in time t_S spent in S is $\exp(-kt_S)$. The likelihood that the rate equals $-\ln\delta/t_S$ given that these paths are not selected in MD is δ. All values of k that result in likelihood greater than δ satisfy $k < -\ln\delta/t_S$. Thus, $1-\delta$ can be regarded as a confidence associated with the estimate for k. An upper bound for the missing rates for a core state S is given by

$$k_S^{\max} = \frac{\ln(1/\delta)}{t_S}. \tag{3}$$

Note that the entire MD duration $\tau_{MD} = \sum_S t_S$. Similarly, an upper bound for the missing flux consistent with the MD data is $F_S^{\max}(t) = k_S^{\max}\pi_S(t)$.

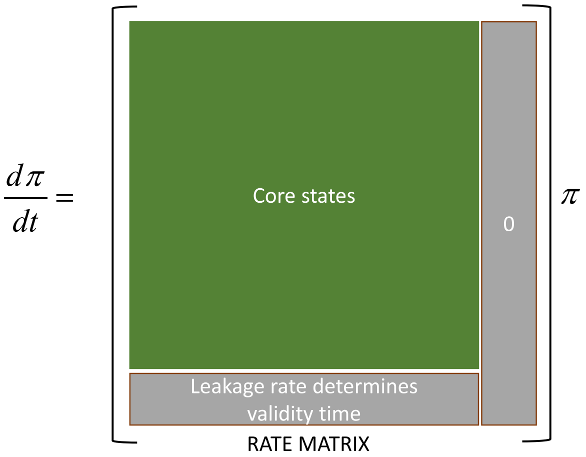

FIG. 1. Structure of a rate matrix in the master equation [Eq. (2)] when independent MD trajectories are used to build a Markov state model (MSM). $\pi(t)$ denotes the occupation column vector.

C. Leakage flux from core network

For a core state S, Eq. (1) can be rewritten in terms of the core (C), periphery (P), and missing (M) states as

$$\frac{d\pi_S}{dt} = \sum_{S'\in C}(k_{S'\to S}\pi_{S'} - k_{S\to S'}\pi_S) - F_S, \tag{4}$$

where the leakage flux

$$F_S = \sum_{S'\in P\cup M}(k_{S\to S'}\pi_S - k_{S'\to S}\pi_{S'}). \tag{5}$$

In the worst-case scenario, using Eq. (3) to replace F_S with an upper bound, we define the maximum leakage flux from S as

$$F_S^{leak} = k_S^{leak}\hat{\pi}_S = \hat{\pi}_S\left(\frac{\ln(1/\delta)}{t_S} + \sum_{S'\in P} k_{S\to S'}\right). \tag{6}$$

The caret denotes maximum leakage into the periphery/missing states. State occupations are obtained by solving

$$\frac{d\hat{\pi}_S}{dt} = \sum_{S'\in C}(k_{S'\to S}\hat{\pi}_{S'} - k_{S\to S'}\hat{\pi}_S) - F_S^{leak}. \tag{7}$$

The leakage flux can be ignored when $\left|\sum_{S'\in C}(k_{S'\to S}\hat{\pi}_{S'}(t) - k_{S\to S'}\hat{\pi}_S(t))\right| \gg F_S^{leak}(t)$. Once a state with large leakage is detected, one can analyze the source of leakage. Large leakage due to $k_S^{\max}$ implies additional MD in state S would be beneficial. The other possibility is that one or more periphery states have become important to the dynamics. By performing additional MD, a periphery state S' is converted into a core state and its contribution to the leakage flux is eliminated. The flux from the core to periphery states is determined by the network topology. Topologies where the core states are connected to a large number of periphery states generally have a short validity time. The missing flux from state S can be made small by performing MD calculations in S, which extends the time t_S.

The occupation $\tilde{\pi}_S(t)$ is obtained by solving the MSM,

$$\frac{d\tilde{\pi}_S}{dt} = \sum_{S'\in C}(k_{S'\to S}\tilde{\pi}_{S'} - k_{S\to S'}\tilde{\pi}_S). \tag{8}$$

Detailed balance is not assumed, i.e., presence of *both* forward and backward pathways is not required. Dynamics from the core network [Eq. (8)] and full network [Eq. (7)] models diverge beyond the validity time. Consider a special case where the stationary distribution $\tilde{\pi}_S^{st}$ is attained and leakage in the full network model becomes significant at time scale τ beyond the relaxation time scales. We write

$$\frac{d\tilde{\pi}_S}{d\tau} = -k_S^{leak}\tilde{\pi}_S;\ \tilde{\pi}_S(0) = \tilde{\pi}_S^{st}. \tag{9}$$

Since the core network dynamics is fast, the ratio of occupations for two core states is fixed, i.e., $\tilde{\pi}_S(\tau) = \tilde{\pi}_S(0)f(\tau)$ with $f(\tau)$ denoting the probability of residing in the core states at time τ. Equation (9) is rewritten as

$$\frac{df(\tau)}{d\tau} = -k^{leak}f(\tau);\ f(0) = 1, \tag{10}$$

and

$$k^{leak} = \sum\nolimits_S k_S^{leak}\tilde{\pi}_S^{st}. \tag{11}$$

Here, k^{leak} denotes the total leakage rate for the network. The occupation for a core state S is

$$\tilde{\pi}_S(t) = \tilde{\pi}_S^{st}\exp(-k^{leak}t). \tag{12}$$

In general, Eqs. (7) and (8) need to be solved simultaneously to determine the validity time for the core network. For convenience, we can approximate the validity time scale τ_V in terms of the time constant for leakage, namely,

$$\tau_V = 1/k^{leak}. \tag{13}$$

According to Eq. (12), the core state occupations decrease by a factor of exp(−1) at these time scales. The advantage of Eq. (13) is that only the core network model needs to be solved to calculate the validity time using Eq. (11). The assumption in Eqs. (11)–(13) that the relaxation time scale within the core network is smaller than the time scale associated with probability leakage is violated when the MD time accumulated in core state(s) is shorter than the relaxation time scale. We shall consider this case in Sec. III B. Note that in this work, a version of Eq. (13) using the time-dependent occupations is employed for calculating the validity time.

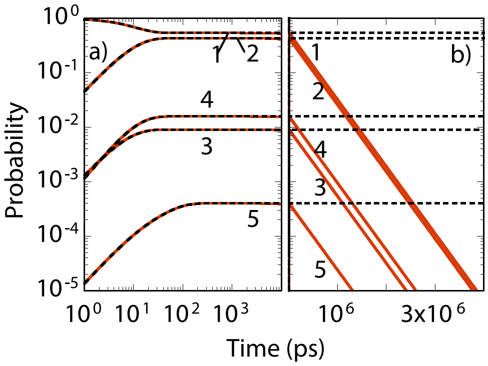

FIG. 3. Occupation for core states (states 1-5) of Fig. 2 found by solving the core network [Eq. (8), dashed black line] and full network [Eq. (7), red line] models at (a) short (log-log) and (b) long (semilog) time scales. Both models were constructed using MD trajectories. The full network model denotes the worst-case scenario where probability leakage into periphery/missing states occurs because of which the core state occupations decay exponentially at long time scales. The core network model is a compact MSM that does not contain any periphery/missing states.

As an illustration, a MSM constructed for solvated alanine dipeptide at 300 K is shown in Fig. 2. States and kinetic pathways were found by analyzing *on-the-fly* several thousand MD trajectories as they were being generated in parallel. See Sec. S3 of the supplementary material for details regarding setup and comparison to literature results. States 1-5 form the core network. Detection of states 1-4 is possible within 0.1 μs MD, while detection of state 5 can require longer trajectories. States 1 and 3 closely represent the α_R conformation of alanine dipeptide. States 2 and 4 are located in the β/PII/C7$_{eq}$ region in (ϕ, ψ) space. Occasionally, the system will visit 11 other states (see Fig. 2) only to quickly return to the core states. We consider these 11 states as periphery states. Kinetic rates were estimated using MLE when a pathway is sighted 10 times or more. Standard error in rates shown in Fig. 2 is computed using the Bootstrap method. Figure 3 shows the core state occupation in dashed lines obtained by solving Eq. (8). Although the MD trajectory exceeds 0.3 μs, it is conceivable that in the

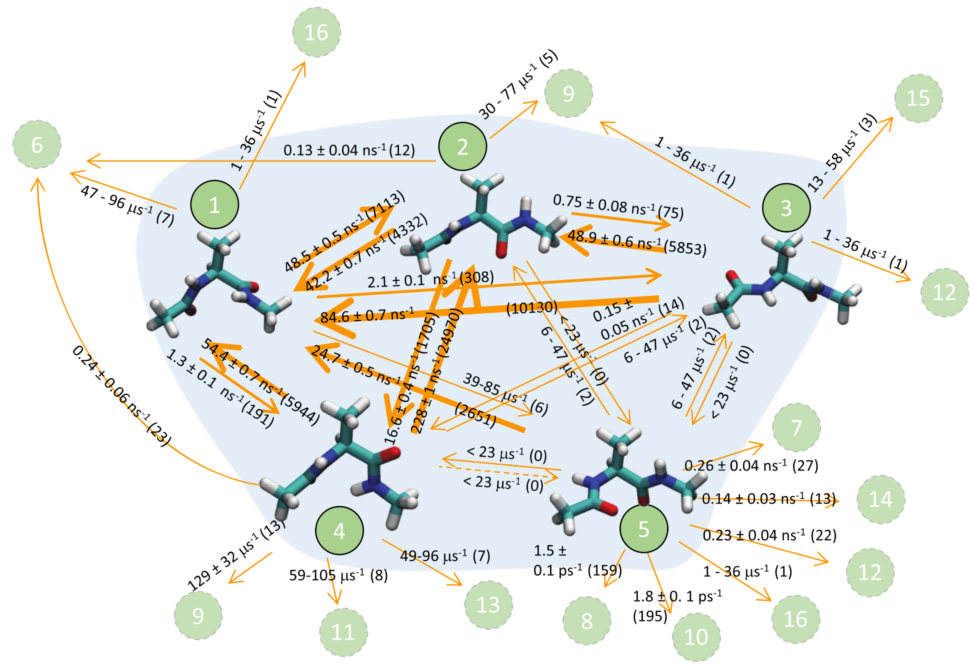

FIG. 2. Markov state model for solvated alanine dipeptide at 300 K. States are numbered in the order they were discovered with MD. States 1-5 are core states and 6-16 are periphery states. Kinetic rates are shown along with the number of sightings for the event in parentheses.

worst-case scenario absorbing-states will be visited via pathways that are missing in the core network model. The leakage flux, calculated based on the rates from core states to periphery states and the missing rate from each core state using $\delta = 0.1$, is included in the full network model of Eq. (7). Equations (7) and (8) agree at short times [Fig. 3(a)]; however, they diverge at longer time scales [Fig. 3(b)]. In the worst case, all core state occupations will decay exponentially with the same rate consistent with Eq. (9).

Suppose the system is trapped/equilibrated in N core states corresponding to deep basins in the energy landscape and periphery states are absent, the MD time t_S for a core state S is proportional to $\tilde{\pi}_S^{st}$, i.e., $t_S = \tilde{\pi}_S^{st}\tau_{MD}$. Equation (11) simplifies to

$$k^{leak} = N\frac{\ln(1/\delta)}{\tau_{MD}}. \tag{14}$$

Subsequent coarse-graining of states into superstates can result in a more compact MSM, but the validity time is still determined by the total time spent in the superstate.

D. Programmed state constrained MD

Adaptive sampling methods[50,57] that seek rare-configurations so that new MD trajectories can be seeded from such configurations are generally used for calculating thermodynamic properties. Relevant kinetic information can be gained by efficiently extending the MSM validity time. If state S is poorly sampled in a dynamical trajectory even though it is relevant, one will require a longer trajectory with the hope that at some point enough transitions from the state will be sampled. Such situations can be tackled using state-constrained MD calculations. In state-constrained MD calculations, one performs MD in state S while checking for a transition at regular intervals. Once a transition is detected, the MD calculation is stopped and the waiting time and final state are noted. A fresh independent MD calculation is seeded from S such that the system is in thermal equilibrium (see Sec. S1 of the supplementary material for flowchart). More transitions from S are sought. This prevents the system from freely diffusing over the potential energy landscape and confines it to a particular state for the purpose of detecting kinetic pathways from the state, calculating the rates, and lowering leakage flux of state S. Core- and full-network models can be constructed efficiently with *programmed* state-constrained MD (PSC-MD) by automatically targeting states with the largest leakage flux and performing state-constrained MD in those states (see Sec. S2 of the supplementary material for flowchart).

The PSC-MD scheme usually introduces many periphery states in the network model as it is used to build the MSM in patches. Occasionally, states are chosen from one part of the network for state-constrained MD calculations, and later, another part may be selected based on the calculated leakage flux. When the largest leakage is from core state S to a periphery state S', state-constrained MD is performed in S' until S' becomes a core state. To achieve this goal with regular MD, a trajectory of duration $m/(k_f\pi_{S'})$ is required, where m denotes the number of transitions needed for estimating the rate with reasonable statistical accuracy and k_f is the fastest rate from S'. When the occupation $\pi_{S'}$ is small, the time taken using state-constrained MD calculations given by m/k_f is orders of magnitude smaller than the one using regular MD.

III. PROTOTYPE EXAMPLES

That the validity time is a vital parameter for quantifying the MSM accuracy becomes evident by examining simple networks that are fully known to us from the outset. We consider (i) a landscape containing trapping states, which is representative of deep basins of protein systems, and (ii) random walk in a corrugated landscape containing shallow basins reminiscent of a disordered system. This understanding will be useful later when we study stretched deca-alanine in Sec. IV.

A. Network with trapping states

Energy landscapes of biomolecular systems often contain low lying minima separated by large barriers. The inset of Fig. 4 shows a one-dimensional network with 5 states. Initially the system is in state 2. The rate is calculated as $k = 10^5 \exp(-\Delta F/k_B T)\,\mathrm{ps}^{-1}$, where ΔF is the energy barrier given in the caption of Fig. 4, k_B is the Boltzmann constant, and $T = 300$ K is the temperature. Since the barrier for the move from state 2 to 3 is large (0.35 eV), the system remains trapped in states 1-2 at short times. States 4 and 5, which are also kinetic traps, are accessed at longer times.

A dynamical trajectory is generated using a kinetic Monte Carlo procedure wherein a move is selected randomly from the current state with a probability proportional to its rate and the time is advanced by $-\ln(\xi)/k_S$. Here, ξ is a uniform random deviate and k_S is the sum of rates from the current state S. A MSM is constructed with the help of the dynamical trajectory. Kinetic rates are estimated using MLE when a pathway is sighted 10 times or more.

The fraction of time spent in a state in the dynamical trajectory is plotted in Fig. 4. After the initial transient, the occupations for states 1-2 plateau. Based on the short dynamical trajectory, one may correctly conclude that the MSM containing only states 1-2 will suffice at short time scales. The picture changes at the longer time scales. After 50 ns states 4-5 are accessed and the fraction of time spent in states 1-2 decays.

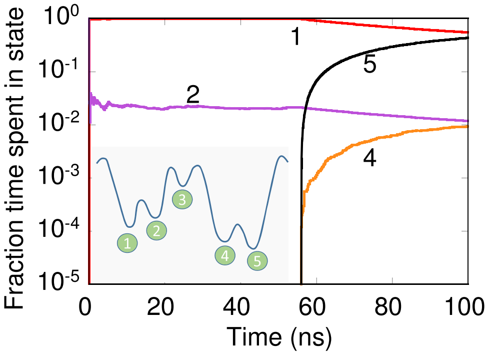

FIG. 4. Fraction of time spent in states belonging to the 1-D network shown in the inset over the course of the simulation. The x-axis denotes the time elapsed in a dynamical trajectory. Energy barrier for forward moves from left to right are 0.3, 0.35, 0.1, and 0.2 eV. Barriers for backward moves from left to right are 0.2, 0.1, 0.45, and 0.3 eV. The system resides in state 2 at time t = 0 ps.

State 3 is not shown in Fig. 4 for convenience. Similar behavior is expected from other dynamical trajectories started from state 2, although the time required to access states 4-5 will vary. The relevance of state 3 cannot be ignored as it provides access to states 4-5. Thus, the MSM should include states 1-5 at longer time scales. So when should one include states 3-5 to ensure that the MSM remains accurate?

PSC calculations were performed with state 2 as the starting state. The validity time is calculated using $\delta = 0.1$. Initially, the MSM contains only two states and the MSM validity is extended by performing state-constrained calculations in states 1-2. The leakage rate from Eq. (14), namely, $2\ln(1/\delta)/\tau_{MD}$, is in close agreement with PSC calculations [Fig. 5(a)]. The validity time keeps increasing with the MD time except for an abrupt decrease when the periphery states 3-5 are detected [Fig. 5(b)]. All states and pathways were available in the MSM at 553 ns. The corresponding validity time was 20.1 ns. Although we are aware of the total number of states and pathways in this toy model, in general for complex systems such as proteins whether the network model is complete will remain unknown to us. Therefore, one would continue to search for newer states and pathways. At longer times, the leakage rate is given by $5\ln(1/\delta)/\tau_{MD}$.

It can be shown that the fraction of time spent in states 1-2 is identical for the dynamical trajectory of Fig. 4 and the state-constrained calculations of Fig. 5 at short times. Similar behavior is true for states 4-5 at longer times. Therefore, conclusions from Fig. 5(b) can be extended to Fig. 4. From Fig. 5(b), a MSM generated from a 40 ns long dynamical trajectory is valid only till 8 ns. The MSM of Fig. 4 only contains the pathways between states 1-2 at 40 ns, i.e., the pathways missing in the MSM are less relevant to the dynamics till approximately 8 ns. This is confirmed in Fig. 6, where the time-dependent state occupations are plotted. States 3-5 are visited before 60 ns in Fig. 4, which corresponds to a validity time of approximately 10 ns. This implies states 1-5 should be present in a MSM when 10 ns time scales are accessed.

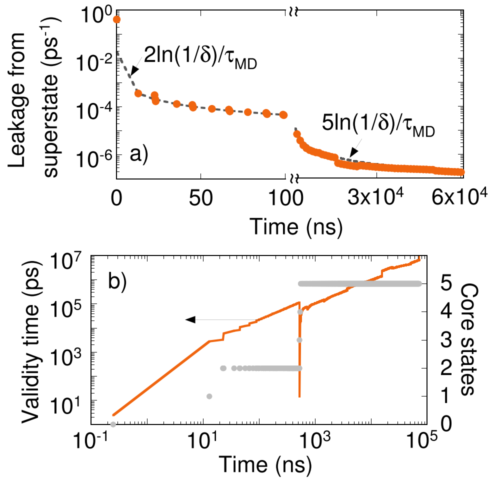

FIG. 5. (a) Leakage flux calculated using Eq. (14) (dashed line) and state-constrained calculations (filled circles). (b) Validity time calculated for the MSM constructed for the network in Fig. 4 using state-constrained calculations (orange line). The number of core states shown in filled grey circles increases as the trajectory grows longer. The x-axis in both panels denotes time elapsed in the dynamical trajectory.

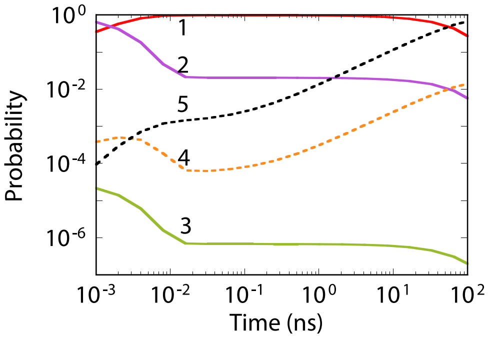

FIG. 6. State occupation for the network shown in Fig. 3 obtained by solving the MSM with validity time of 10^4 ns. Initial state is 2.

Figure 6 shows that roughly in 1 in 10 dynamical trajectories the system will be in state 5 at 10 ns. When 1/10 is set as the tolerance limit, the dynamical behavior can no longer be predicted using a two-state MSM, highlighting the importance of missing pathways.

B. Network with diffusive behavior

Dynamics in a large number of biomolecular and disordered systems can be described in terms of random walk in shallow energy basins. The inset of Fig. 7(a) shows a periodic 4 × 4 lattice where each site is connected to 4 neighboring sites, each pathway having a rate of 1 ns^{-1}. A particle, initially placed at the orange-colored site, can randomly hop to any of its nearest-neighbor sites. The exact kinetic model contains 16 states and 64 pathways. The state occupation is obtained by solving the master equation analytically (solid lines in Fig. 7). Along the lines of Sec. III A, a MSM is constructed using PSC

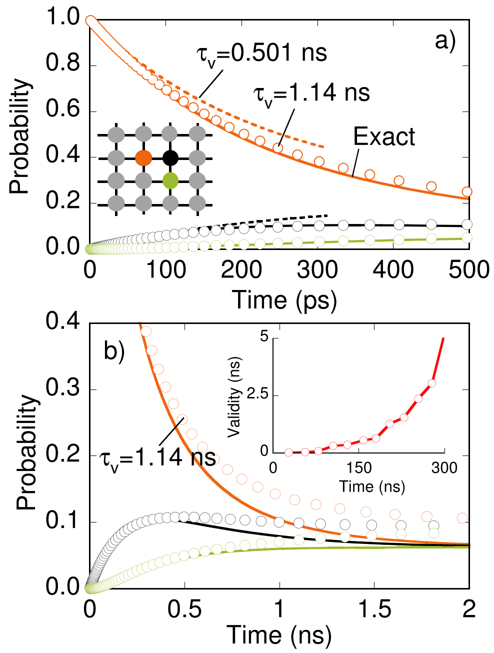

FIG. 7. State occupation calculated for MSMs of validity time 501 ps (dashed lines) and 1140 ps (open circles) at (a) short and (b) longer times. The MSM is constructed for the network in the inset of panel (a). Occupation for initial (orange), a nearest neighbor (black), and a next nearest neighbor (green) site is shown in their respective colors. Dependence of the validity time on the duration of the state constrained trajectory is shown in the inset of panel (b).

calculations of the total duration of 105 ns. Kinetic rates were estimated using MLE when a pathway is sighted 10 times or more. The MSM contains 4 core states, and the validity time calculated using Eq. (13) is 0.501 ns. Dashed lines in Fig. 7(a) denote the occupation obtained by solving this MSM. While the MSM predictions are reasonable at short time scales, deviations are observed later. The large leakage from core states to periphery states results in a small validity time in diffusing systems as shown in the inset of Fig. 7(b). A MSM constructed with 278 ns PSC calculations remains accurate for a longer period of time (see open-circles). The MSM contains 11 core states and has a validity time of 1.14 ns. All states and pathways are available in the core network model once PSC calculations have accrued 1045 ns.

IV. MARKOV STATE MODEL OF STRETCHED DECA-ALANINE

Sections I–III lead to the conclusion that the MSM validity time is always shorter than the duration of MD used to construct the MSM. This has serious implications on long-time studies using MSMs. Despite this, crucial insights can be obtained with MSMs. We demonstrate this aspect by building a MSM with a desirable validity time to study the kinetics of a deca-alanine molecule in vacuum under tension. A capped deca-alanine (Ala10) with acetylated N-terminus and amidated C-terminus was selected for the study. The initial configuration was obtained from the 104-atom helical model of Ref. 58. The C_α atoms at the two ends (residues 1 and 10) are tethered to two anchor points by harmonic restraints with a spring constant of 0.86 (kcal/mol)/Å^2 (Fig. 8 inset). The anchor separation d is kept fixed during the construction of a MSM and fluctuations in molecular extension, and forces are measured.

Past studies of deca-alanine demonstrate that unravelling of the helical structure results in higher free-energy configurations, although the presence of metastable configurations of stretched deca-alanine has not been reported. Hence, specific 3D metastable structures encountered in the dynamics are determined as d is varied between 16 and 26 Å. Questions related to appropriateness of employing the same Markov state definitions and pathways across different anchor separations as well as changes in the kinetic rates are examined. Analysis of dominant configurations helps us probe the importance of multiple pathways for unravelling of the helical structure as the molecule is stretched. Since helix winding/unwinding is a reversible process, helical and stretched configurations co-exist, which causes the force experienced by the AFM tip to fluctuate. Ensemble-average forces from MSMs and MD are compared. Through this study, we conclude that forces are inaccurately predicted using MSMs with short validity time, thus highlighting the important role of missing states/pathways in ensemble-averaged quantities.

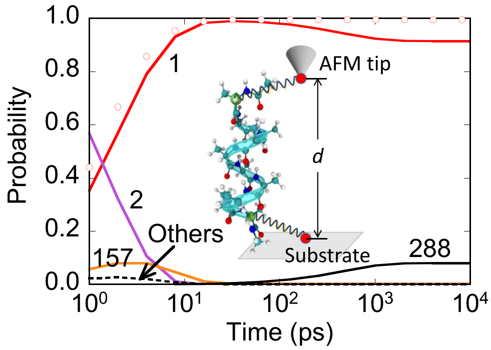

FIG. 8. Occupation for the top-four states using a MSM with 65 ns validity time when the anchor separation d = 16 Å. The force-spectroscopy setup for deca-alanine in vacuum at 300 K is shown in the inset. Harmonic restraint is applied to the light-green colored C_α atoms. Open circles show state 1 occupation from a 5-state MSM with 2 ns validity time (constructed from a 0.57 μs long MD trajectory).

MSMs are constructed on-the-fly with 90% confidence ($\delta = 0.1$) using thousands of short-MD calculations that are run in parallel as part of the PSC procedure. All MD simulations were performed with NAMD 2.9[59] with the CHARMM36 force fields.[60] Temperature was held at 300 K using a Langevin thermostat. Bonds involving hydrogen atoms were constrained to their equilibrium values using RATTLE.[61] An integration time step of 2 fs was used. States were determined by comparing the backbone atoms after aligning the molecule using the Kabsch algorithm. A tolerance of 3 Å was found to be suitable for identifying the states. MD snapshots were collected every 0.2 ps because of rapid interconversion between the states. A transition was said to have successfully occurred when the system continues to reside in the new state for at least 1.2 ps after the transition was detected in the MD trajectory. This prevents recrossing events to be counted as transitions.

Steered MD simulations were performed with the deca-alanine for several nanoseconds to obtain a preliminary collection of unfolded structures. These configurations were provided as inputs to several nanosecond-long regular MD trajectories with chosen anchor separations. A preliminary catalog of states was constructed that could be employed with different anchor separations. States were indexed in the order they are found. State-constrained MD calculations, which are more efficient than regular MD (see Sec. II D), were used to confirm Markovian behavior (see Sec. S5 of the supplementary material) and that the kinetic pathways can be described as a first-order process. Poor MSM validity was achieved when states or MD duration were selected ad-hoc in the state-constrained MD calculations. For example, in a preliminary MSM-building attempt we found a 4 μs long MD trajectory resulted in a validity time of 0.064 ns, which is the reason why only PSC-MD calculations are performed.

The full network model consisted of 810 states. Most states are periphery states. The number of core states and their relevance varies with anchor separation. State occupation at $d = 16$ Å obtained by solving a MSM of 65 ns validity is shown in Fig. 8. The MSM constructed using a 9 μs trajectory contains 25 core states and 78 pathways. The system was initially present in state 2 shown in Fig. 9. Rapid conversion to α-helical configuration (state 1) with a large rate of 0.44 ps^{-1} is observed consistent with previous studies. The average distance between the terminal C_α atoms is 16.4 Å for state 1. Another dominant configuration, namely state 288, is selected beyond 100 ps with nearly 10% probability. State 288 is accessible from state 1 with a small rate of 1.4 ns^{-1}. As a

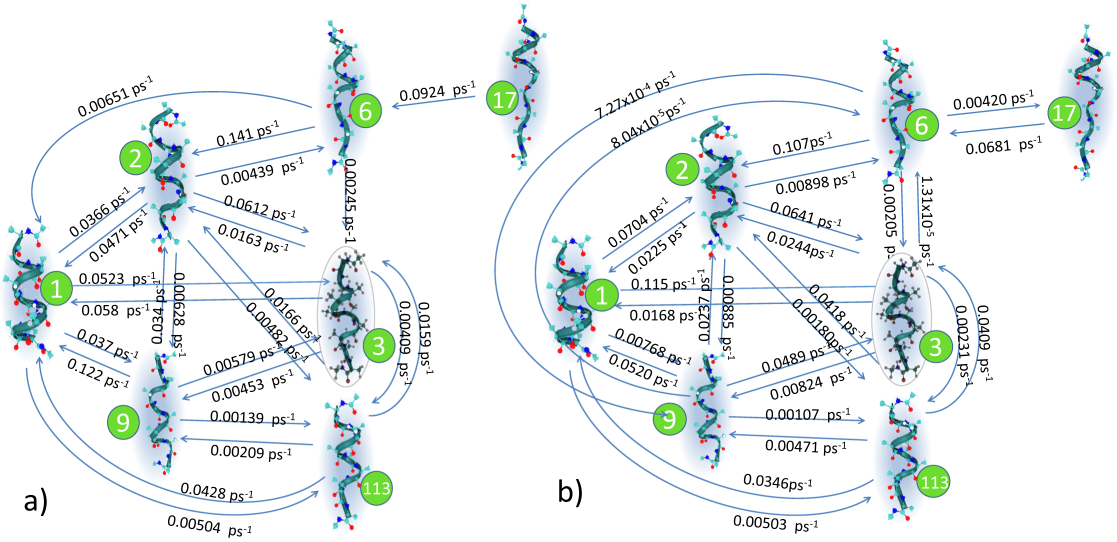

FIG. 9. Network model obtained when the anchor separation is (a) 22 and (b) 23 Å. Only frequently visited states are shown. State 3 (encircled) is the starting configuration for the subsequent figures.

consequence, state 288 is absent in a MSM with a shorter validity time and the time-dependent occupation for state 1 from such a MSM is incorrectly predicted (see open circles in Fig. 8). This behavior is analogous to the one observed in Fig. 4. State 288 is preferred over state 2 for two reasons. First, it has an end-to-end distance smaller than that of state 2, which is favored at compressive conditions. Second, the average α-helicity for states 1, 2, and 288 are 0.8, 0.2, and 0.74, respectively. Multiple backbone hydrogen bonds impart more stability to state 228 than state 2. It appears that a two-state model (states 1 and 288) might suffice for the calculation of average force at $d = 16$ Å.

The abundance of energetically-favorable stretched-out configurations causes state 288 to lose its relevance at higher anchor separations, but state 1 still continues to be relevant. Figure 9 shows the dominant core states for anchor separations 22 and 23 Å. The unravelling of a helical structure to an elongated one proceeds via multiple intermediate states. For instance, one pathway for visiting state 17 from state 1 involves only "local" readjustments. First, a partial opening of lower coils (C-terminal) is observed (state 1 to 6 via state 2) followed by subsequent stretching of the lower coils (state 6 to 17). Alternatively, fluctuations in the middle residues can frequently cause deformation in the helix (state 1 to 3) that may sometimes lead to an elongated configuration (state 3 to 6). States 2, 3, and 6 have a non-negligible 3_{10}-helicity (see Sec. S4B of the supplementary material). State 6, which is an essential intermediate for both pathways, can be reached faster from states 2 and 3 as the anchor separation is increased. The preferred winding/unwinding mechanism proceeds predominantly at the C-terminal, i.e., the former pathway.

In the past, the end-to-end distance has been employed as a reaction coordinate for stretched deca-alanine. End-to-end distance distribution for a state tends to be sharply-peaked with a standard deviation of nearly 1-2 Å; however, the distribution is a function of the anchor separation. Large overlap in end-to-end distribution for the core states makes it practically impossible to distinguish states when only end-to-end distances are employed. Inclusion of the 3D structure, which is implicit in our state description, helps resolve intermediate states and state-specific properties. In particular, we are interested in the stiffness of deca-alanine, which determines the force on the AFM tip. The stiffness, calculated using state-constrained MD calculations, is found to vary from one state to another depending on the intramolecular interactions. Presence of strong hydrogen bonds in state 1 results in a large spring constant of 39.44 pN/Å. On the other hand, state 3 has a smaller spring constant of 25.38 pN/Å. A natural consequence is that the average force experienced by the AFM tip can be altered by as much as 100 pN in either direction during a state-to-state transition because of the differences in the state-specific spring constants. The average force is given by the sum of force experienced for each state times the state occupations.

MSMs are sensitive to small changes in the anchor separation [see Figs. 9(a) and 9(b)]. Pathways between states 6 and 9 are absent in the MSM for d = 22 Å, but they are dynamically more relevant at validity times of 8 ns when d = 23 Å. Forward and backward rates are found for many pairs of states. Exceptions in Fig. 9(a) include the move from state 1 to 6, although this is not an issue since a stationary solution is obtained without the requirement of detailed balance. The rate constants involving the core states vary over several orders of magnitude between 10^{-1} and 10^{-5} ps^{-1}. Exponential increase/decrease in rates is witnessed between 16 and 26 Å (see Sec. S4B of the supplementary material). While a handful states can describe the dynamics for small anchor separations, additional states should be included in the MSM when the molecule is stretched extensively. An explosion in the number of core states is witnessed from 36 to 78 states between 22 and 25 Å. Correspondingly, the validity time plummets by almost 10 times from 12 ns at 22 Å separation to 1.5 ns at 25 Å separation for a 0.5 μs long MD trajectory. Note the length of (PSC-)MD trajectory required to reach the nanosecond-long validity time.

Figure 10 shows the evolution obtained with different MSMs for anchor separations between 22 and 25 Å. The rise and fall in the relevance of states and kinetic pathways is witnessed. The initial state of the system is state 3. One might expect that the helical structures (state 1) would not be selected at high separations; however, even at 24 Å separation there is a 1% chance of finding state 1. This is attributed to the lower energy of state 1 and the small stiffness of the tethers. In other words, it is possible for the tether to stretch to an extent where state 1 can still be visited in the dynamics at 24 Å separation. State 3 is an important configuration between 22 and 25 Å separation. The maximum state 3 occupation is witnessed around 23 Å. States 9, 6, and 17 compete with each other at large anchor separations. As the number of core states increases, there is an increased chance that the system will visit these states. The occupation in all other core states combined is shown in the dashed lines in Fig. 10. The time spent in the top-five states decreases from 96% at 22 Å separation to 30% at 26 Å separation. Based on Fig. 10, the stationary solution for deca-alanine is reached at nearly 100 ps, i.e., time-dependent forces can be resolved only at sub-100 ps time scales.

The free energy difference for a state A with respect to state B is calculated from the stationary distribution as $\Delta F_{A-B} = -k_B T \ln(p_A/p_B)$, where p denotes the stationary occupation of a state. Since the stationary solution depends on the relevant pathways, absence of one or more pathways can introduce errors in the free energies. Occupations in Fig. 10 were used to calculate the free-energy difference between the states. Insights into the separation-dependence of the kinetic rates can be obtained from the Bell-Evans-Polyani principle.[62,63] Consider states 1 and 2. Since the free energy of state 1 increases with reference to state 2 as deca-alanine is stretched, the free barrier for the move from 1 to 2 (2 to 1) decreases (increases), which explains the corresponding shift in the rate constants (see Sec. S4B of the supplementary material). Similarly, the free energy of state 2 decreases with respect to state 3, causing the rate constant from state 2 to 3 to increase. As shown in Sec. S4A of the supplementary material, the work done while stretching deca-alanine can be calculated using the stationary occupations at different anchor separations. The calculated work is in agreement with previous values in the literature.

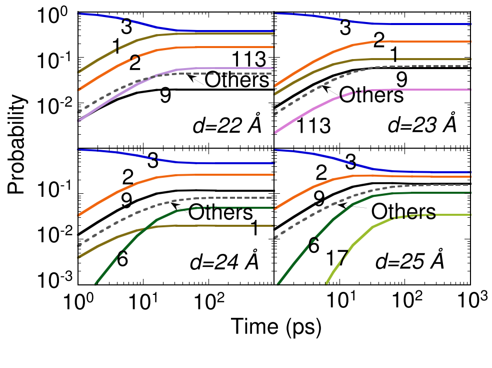

FIG. 10. Probability evolution for different anchor separations. Numbers denote the state index in Fig. 9. The initial state of the system was state 3. The MSM validity time exceeded 1 ns in all cases.

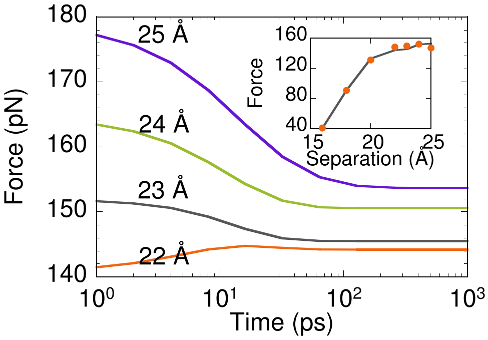

FIG. 11. Solid lines show the average force acting on the AFM tip when the deca-alanine is stretched (using MSMs with validity time exceeding 4 ns). Behavior for anchor separation 22-25 Å is shown. Steady state forces calculated from the MSM (line) and MD (filled-circles) shown in inset are in good agreement.

Figure 11 shows the time-dependent ensemble-averaged force experienced by the AFM tip corresponding to the evolution shown in Fig. 10 for different anchor separation. Average forces were computed for the individual core states using state-constrained MD. Since the force experienced with state 3 is smaller than with state 1, the average force increases with time at 22 Å separation as state 1 occupation increases (see Fig. 10). Although state 3 plays a minor role in the winding/unwinding of deca-alanine, it is important for calculation of average forces due to its large occupation. Beyond 23 Å separation, one finds that state 2 is preferred over state 3. The average force decreases with time since the force experienced with state 2 is smaller than with state 3. The average force at steady-state from the MSM and MD are in good agreement (Fig. 11 inset) validating our MSM. Such an agreement between MD and the MSM is not witnessed when the validity time is small as many relevant states are missing. The sudden drop in the slope of the force-separation curve is attributed to the smaller stiffness of states encountered at higher separations.

Weaker electrostatic interactions are expected when deca-alanine is present in water. To mimic the effect of water, we construct MSMs at anchor separations 16 and 24 Å with a dielectric constant of 80. In order to find common features in the dynamical behavior for dielectric constants 1 and 80, the list of 810 states obtained previously was used as the starting known structures for our new calculations. Deca-alanine was initially kept in state 3. The dominant states for anchor separation of 16 Å include states 1, 20, 113, and 288. Figure 12(a) shows that the occupation for state 3 decreases continually in time. As in Fig. 8, a maximum value of the occupation for state 1 is witnessed. However, the steady state occupation for state 1, namely, 0.401, is much smaller than the one observed in Fig. 8. The occupation for other states combined increases to a significant value. We find that 11 states possess steady state occupations greater than 0.01. By considering the full network model, we conclude that the MSM obtained with 10.8 ns long MD calculations has a validity time of nearly 30 ps. At 30 ps, the difference in the state 1 occupations in the MSM and the full network model has exceeded 0.1. More states are observed with the anchor separation of 24 Å. In all 25 states possessed a steady state occupation in excess of 0.01 with state 3 being the only state common to both Figs. 10 and 12 with

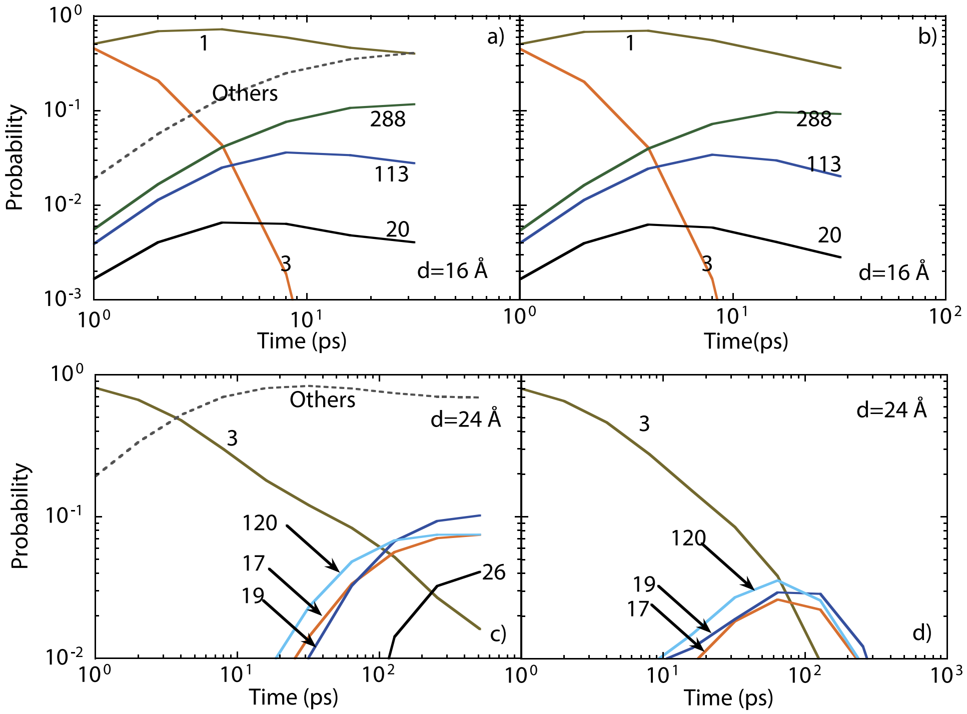

FIG. 12. Probability evolution for anchor separations 16 Å [panels (a) and (b)] and 24 Å [panels (c) and (d)] using the dielectric constant of 80 to mimic deca-alanine in water. Numbers denote state index. Initial state of the system was state 3. Left panels [(a) and (c)] show results from the MSM while right panels [(b) and (d)] show results obtained with the full network model.

d = 24 Å. The total number of core states was found to be 137 using 53 ns long PSC-MD. Figures 12(c) and 12(d) show state occupations as a function of time from the MSM and the full network model. Predictions from the two models diverge at nearly 0.1 ns. One can determine whether states and kinetic pathways in the present MSM will continue to remain relevant at longer time scales only by extending the validity time of the MSM.

V. CONCLUSIONS

MSMs have the potential to become even more powerful computational tools in the future for studies of biomolecular, materials, and reacting systems as new advances emerge that enable one to accurately encode the kinetic and thermodynamic information on the multidimensional landscape in terms of state-to-state transitions. Unfortunately, MSMs constructed bottom-up from finite MD trajectories are rarely complete, which has a direct implication on its accuracy. Since the (kinetic) information content in a single trajectory can be different from that of an ensemble at the same length of time, questions related to validity of the MSM due to missing information arise. We introduce the fundamental concept of validity time of a MSM to quantify its completeness. The concept guarantees that all states and kinetic pathways that are relevant to the dynamics will be present in the MSM provided the time scales accessed by the model are smaller than its validity time. Put differently, this concept helps us understand the time scales where states and pathways missing in the available MD data or the MSM can become relevant to experimentally measurable quantities, and differences between the experimental quantities and MSM predictions might originate from the missing information in the MSM. Our methodology is flexible in terms of its ability to handle a wide-range of kinetic rates, number of states, relaxation times in the network, topology of the network, as well as missing pathways and states that have not been found with MD. Development of programmed state-constrained MD calculations in this work provides an efficient means for extending validity time of a MSM.

MSMs with sufficiently large validity time provide following key insights into the stretching of deca-alanine. Unwinding of deca-alanine proceeds mainly via breaking of hydrogen bonds at the C-terminal. Characterization of states might not be possible only using simple reaction coordinates such as end-to-end distance. Each state possesses its own mechanical characteristics, e.g., spring constants, that ultimately determine the force experienced by the AFM tip in the force spectroscopy (FS) setup. To calculate the force, one also needs to accurately estimate the state occupations. It is not straightforward to guess the relevant states/pathways at different anchor separations. The list of relevant states and kinetic pathways, and associated rates, i.e., the network topology, can change dramatically with the anchor separation. The size of the MSM increases for large anchor separations, which has a direct bearing on the amount of MD required to know the state occupations accurately. Our studies demonstrate that the absence of relevant states/pathways in the MSM can lead to incorrect prediction of the kinetic and thermodynamic quantities being sought, which is the main reason why studies employing MSMs and other related kinetic network models should state the validity time of the model.

SUPPLEMENTARY MATERIAL

See supplementary material for the flow chart for state constrained and programmed state constrained MD, simulation protocol used for solvated alanine dipeptide, additional figures for the deca-alanine system, and discussion on test for Markovian behavior.

ACKNOWLEDGMENTS

A.C. acknowledges support from Science and Engineering Research Board, Department of Science and Technology Grant No. SB/S3/CE/022/2014 and Indian National Science Academy Grant No. SP/YSP/120/2015/307. S.B. acknowledges support from Department of Biotechnology Grant No. BT/PR8311/BID/7/451/2013.

[1] C. R. Schwantes, R. T. McGibbon, and V. S. Pande, J. Chem. Phys. **141**, 090901 (2014).
[2] V. S. Pande, K. Beauchamp, and G. R. Bowman, Methods **52**, 99 (2010).
[3] M. Senne, B. Trendelkamp-Schroer, A. S. J. S. Mey, C. Schütte, and F. Noé, J. Chem. Theory Comput. **8**, 2223 (2012).
[4] N.-V. Buchete and G. Hummer, J. Phys. Chem. B **112**, 6057 (2008).
[5] C. Schütte, F. Noé, J. Lu, M. Sarich, and E. Vanden-Eijnden, J. Chem. Phys. **134**, 204105 (2011).
[6] S. Gnanakaran, H. Nymeyer, and J. Portman, Curr. Opin. Struct. Biol. **13**, 168 (2003).
[7] G. G. Dodson, D. P. Lane, and C. S. Verma, EMBO Rep. **9**, 144 (2008).
[8] K. Klenin, B. Strodel, D. J. Wales, and W. Wenzel, Biochim. Biophys. Acta **1814**, 977 (2011).
[9] M. T. Woodside and S. M. Block, Annu. Rev. Biophys. **43**, 19 (2014).
[10] A. Chatterjee and D. G. Vlachos, J. Comput.-Aided Mater. Des. **14**, 253 (2007).
[11] A. F. Voter, in *Radiation Effects in Solids*, edited by K. E. Sickafus, E. A. Kotomin, and B. P. Uberuaga (Springer, NATO Publishing Unit, Dordrecht, 2006).
[12] D. J. Wales, Int. Rev. Phys. Chem. **25**, 237 (2006).
[13] R. M. Ziff, E. Gulari, and Y. Barshad, Phys. Rev. Lett. **56**, 2553 (1986).
[14] G. H. Gilmer, H. C. Huang, T. D. de la Rubia, J. Dalla Torre, and F. Baumann, Thin Solid Films **365**, 189 (2000).
[15] P. Haldar and A. Chatterjee, Acta Mater. **127**, 379 (2017).
[16] A. Chatterjee, M. A. Katsoulakis, and D. G. Vlachos, Phys. Rev. E **71**, 026702 (2005).
[17] A. Chatterjee and A. F. Voter, J. Chem. Phys. **132**, 194101 (2010).
[18] A. Chatterjee and D. G. Vlachos, J. Chem. Phys. **124**, 64110 (2006).
[19] E. Weinan, B. Engquist, and Z. Y. Huang, Phys. Rev. B **67**, 092101 (2003).
[20] M. A. Snyder, A. Chatterjee, and D. G. Vlachos, Comput. Chem. Eng. **29**, 701 (2004).
[21] G. R. Bowman, X. Huang, and V. S. Pande, Methods **49**, 197 (2009).
[22] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, J. Chem. Phys. **126**, 155101 (2007).
[23] 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).
[24] L. Xu and G. Henkelman, J. Chem. Phys. **129**, 114104 (2008).
[25] D. Konwar, V. J. Bhute, and A. Chatterjee, J. Chem. Phys. **135**, 174103 (2011).
[26] A. F. Voter, J. Chem. Phys. **106**, 4665 (1997).
[27] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **99**, 12562 (2002).
[28] B. Ensing, M. De Vivo, Z. Liu, P. Moore, and M. L. Klein, Acc. Chem. Res. **39**, 73 (2006).
[29] A. F. Voter, F. Montalenti, and T. C. Germann, Annu. Rev. Mater. Res. **32**, 321 (2002).
[30] S. Divi and A. Chatterjee, J. Chem. Phys. **140**, 184115 (2014).
[31] V. Imandi and A. Chatterjee, J. Chem. Phys. **145**, 034104 (2016).
[32] P. Haldar and A. Chatterjee, Modell. Simul. Mater. Sci. Eng. **23**, 025002 (2015).
[33] V. J. Bhute and A. Chatterjee, J. Chem. Phys. **138**, 244112 (2013).
[34] V. J. Bhute and A. Chatterjee, J. Chem. Phys. **138**, 084103 (2013).
[35] S. T. Chill and G. Henkelman, J. Chem. Phys. **140**, 214110 (2014).
[36] A. Chatterjee and S. Bhattacharya, J. Chem. Phys. **143**, 114109 (2015).
[37] A. Chatterjee and S. Bhattacharya, J. Phys.: Conf. Ser. **759**, 012024 (2016).
[38] Y. Mu, P. H. Nguyen, and G. Stock, Proteins **58**, 45 (2005).
[39] E. H. Kellogg, O. F. Lange, and D. Baker, J. Phys. Chem. B **116**, 11405 (2012).
[40] T. Zhou and A. Caflisch, J. Chem. Theory Comput. **8**, 2930 (2012).
[41] R. T. McGibbon and V. S. Pande, J. Chem. Theory Comput. **9**, 2900 (2013).
[42] G. Pérez-Hernández, F. Paul, T. Giorgino, G. de Fabritiis, and F. Noé, J. Chem. Phys. **139**, 015102 (2013).
[43] C. R. Schwantes and V. S. Pande, J. Chem. Theory Comput. **9**, 2000 (2013).
[44] K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, J. Chem. Theory Comput. **7**, 3412 (2011).
[45] S. Ghosh, A. Chatterjee, and S. Bhattacharya, J. Chem. Theory Comput. **13**, 957–962 (2017).
[46] D. L. Ensign and V. S. Pande, J. Phys. Chem. B **113**, 12410 (2009).
[47] M. Sarich, F. Noé, and C. Schütte, Multiscale Model. Simul. **8**, 1154 (2010).
[48] N. Djurdjevac, M. Sarich, and C. Schütte, Multiscale Model. Simul. **10**, 61 (2012).
[49] P. Metzner, F. Noé, and C. Schütte, Phys. Rev. E **80**, 021106 (2009).
[50] G. R. Bowman, J. Chem. Phys. **137**, 134111 (2012).
[51] T. Zhou and A. Caflisch, J. Chem. Theory Comput. **8**, 2134 (2012).
[52] A. K. Faradjian and R. Elber, J. Chem. Phys. **120**, 10880 (2004).
[53] L. Maragliano and E. Vanden-Eijnden, Chem. Phys. Lett. **426**, 168 (2006).
[54] A. Laio and F. Gervasio, Rep. Prog. Phys. **71**, 126601 (2008).
[55] A. C. Pan, D. Sezer, and B. Roux, J. Phys. Chem. B **112**, 3432 (2008).
[56] C. A. F. de Oliveira, D. Hamelberg, and J. A. McCammon, J. Chem. Phys. **127**, 175105 (2007).
[57] X. Huang, G. R. Bowman, S. Bacallado, and V. S. Pande, Proc. Natl. Acad. Sci. U. S. A. **106**, 19765 (2009).
[58] S. Park, F. Khalili-Araghi, E. Tajkhorshid, and K. Schulten, J. Chem. Phys. **119**, 3559 (2003).
[59] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, J. Comput. Chem. **26**, 1781 (2005).
[60] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. Mackerell, J. Chem. Theory Comput. **8**, 3257 (2012).
[61] H. C. Andersen, J. Comput. Phys. **52**, 24 (1983).
[62] M. G. Evans and M. Polanyi, Trans. Faraday Soc. **31**, 875 (1935).
[63] R. A. Marcus, J. Phys. Chem. **72**, 891 (1968).