MRS Communications (2018), *8*, 850–857
© Materials Research Society, 2018
doi:10.1557/mrc.2018.93

MRS Communications

Research Letter

Uncertainty quantification of Kinetic Monte Carlo models constructed on-the-fly using molecular dynamics

Abhijit Chatterjee, Department of Chemical Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India
Address all correspondence to Abhijit Chatterjee at **abhijit@che.iitb.ac.in**

(Received 23 January 2018; accepted 30 April 2018)

Abstract

Kinetic Monte Carlo (KMC) models of complex materials and biomolecules are increasingly being constructed using molecular dynamics (MD). A KMC model contains a catalog of states and kinetic pathways, which enables study of the dynamics. The completeness of the catalog is crucial to the model accuracy and is linked to the quality of the MD data used for model construction. Therefore, quantifying the uncertainty due to missing states and pathways is important. A review on computational procedures available for on-the-fly KMC model construction using MD, uncertainty measurement, and algorithms for guiding further MD sampling in an accelerated manner is presented.

Introduction

Kinetic Monte Carlo (KMC) is often the method of choice while studying the long-time dynamical evolution of molecular- and meso-scale materials and biomolecular systems.[1–5] To generate dynamical trajectories of the system, a KMC model requires a catalog of metastable states, i.e., long-lived configurations of the system, and the kinetic pathways that specify how fast the system can interconvert between the states. In KMC, pathways are defined in terms of the starting and end states and their associated rate constants. The continuous time trajectories generated using a KMC model are available as "discrete" state-to-state transitions. Achieving dimensionality reduction from the large number of atomic degrees of freedom to a significantly smaller "network" of states constitutes a major advantage of the KMC method. However, the same property also introduces a challenge since the catalog of states and pathways is generally not known at the outset for most complex materials and biomolecules.[6] For instance, configurations visited during the unfolding of a protein from its folded state might be unknown but these states are required as an input in the KMC model to accurately describe the dynamical unfolding process.[7,8]

To address this issue, KMC models can be constructed using molecular dynamics (MD).[2,5,9] In MD, one solves the Newton's equations of motion to generate dynamical trajectories.[10] The input provided to an MD calculation includes the initial atomic positions and velocities, the MD time-step (typically in femtoseconds) and an interatomic potential or a force field for the system. MD is computationally more expensive than KMC. One can have best of both worlds by learning about the discrete states and kinetic pathways of the system by post-processing several independent MD trajectories to generate a concise KMC model.[11,12] States and pathways sampled in the MD trajectory are incorporated into the KMC model. Alternatively, we can direct our efforts toward parts of the configurational space that we think are important for the construction of the KMC model. In such a case, transitions are sought only from chosen states using MD to prepare a KMC model "on-the-fly". In either case, even when terabytes of MD data is available, the resulting KMC model will be only as good as the MD data used to construct it. A related issue is the underlying network complexity, namely, states that are relevant to the dynamics can be extremely large in number and hard to guess,[13–15] which precludes the construction of "complete" KMC catalogs.[8,16,17]

It becomes important to ascertain the uncertainty in the KMC model before the model is used to generate long-time dynamical trajectories and gain new physical insights. The KMC model has an uncertainty associated with it because of the states and pathways missing in the MD data. Uncertainty quantification helps us assess how important the missing information could be to the true dynamical evolution of the system.[7,15–17] We answer the following questions:

- How does one determine uncertainty in a KMC model from the MD data used to construct the model?
- If there were provision for additional MD, is there an optimal way to generate high-quality MD data which lower the uncertainty?

Since the true dynamics is not known to us, the model uncertainty is based on probability theory. The strategy used revolves around two key points. First, pathways that are likely to be selected in the dynamics should be present in the KMC model. As long as the missing pathways continue to be selected with an extremely low probability, they are deemed less relevant to

https://doi.org/10.1557/mrc.2018.93 Published online by Cambridge University Press

the dynamics and can be missing from the model without affecting the KMC accuracy. "Excellent sampling" of a state in question can guarantee that (at least up to a certain timescale) most of the relevant pathways from the state would be present in the catalog. Issues related to missing pathways also arise in some accelerated MD techniques.[18] Second, optimal usage of the computational resources requires that states likely to be visited in the dynamics should be studied in detail using MD. As long as other states are visited with a lower probability, they can be ignored while sampling the configurational space. A KMC model that appears to be reasonably accurate at short timescales can become grossly inaccurate at longer timescales where the missing states and pathways matter to the dynamics. Thus, uncertainty in the KMC model is a timescale-dependent property, which is captured by a quantity we term as validity time of the KMC model. A KMC model remains accurate (with a confidence level) till its validity time. A detailed discussion on uncertainty in the KMC model constructed using MD can be found in Ref. 8. Biomolecular simulations often employ Markov State Models (MSMs) that are closely related to the KMC models. MSMs are constructed using post-processing of MD trajectories.[3] Therefore, the ideas presented here are useful to the practitioners of both KMC and MSM approaches.

Questions related to the completeness of the KMC model created using dynamical approaches[8,16,17] require slightly different terminology from non-dynamical approaches.[19–21] Since the KMC model uncertainty is a timescale-dependent property, dynamical approaches are better suited toward uncertainty quantification. In this article, we provide a review on a computational procedure that can (i) construct KMC models on-the-fly using MD, (ii) measure the uncertainty in the current KMC model, and (iii) optimally target uncertain regions of configuration space for further network construction. "Two extreme KMC model construction examples" section provides two examples to motivate the need for on-the-fly KMC model construction using MD. A step-by-step description of the construction procedure is provided. The basis for uncertainty quantification of the KMC model is discussed. It is important to recognize that accessing long timescales with MD can be prohibitively expensive. Typically, standard MD is restricted to the μs timescales, whereas KMC can easily access ms timescales and beyond with significantly smaller computational resources. Two novel enhanced sampling strategies are described next, which can be used to seek pathways rarely selected with standard MD. These methods offer significant advantages over regular MD and help lower the uncertainty in the KMC model in a very efficient manner.

Two extreme KMC model construction examples

A KMC model can be described in terms of a directed network with the nodes representing the states and edges/connections representing the pathways [Figs. 1(a) and 1(b)]. The potential energy surface (PES) is a useful concept to visualize the dynamics.[22] The landscape contains several minima. A state of the system can contain one or more minima, such that it is a long-lived configuration. The number of connections between the states, i.e., the network complexity, depends on the underlying PES. It is essential to distinguish between the two extreme forms of landscapes. The first one involves a "diffuse" energy landscape characterized by a large number of states [Fig. 1(a)]. The probability of the system visiting these states is low as the system can select one of many possible pathways during the evolution. The other extreme involves a funnel-shaped landscape, which generally would contain a small number of states (<1000) and the system will remain trapped in these states [Fig. 1(b)]. Two examples of these landscapes are provided next to highlight similarities and differences in the KMC model development and uncertainty quantification. In general, landscapes can contain features corresponding to both Figs. 1(a) and 1(b).

Figure 1(c) shows structural rearrangements observed in 26-atom Ag cluster on Ag(001) surface studied using KMC. The Ag cluster evolves on a diffuse landscape from an initially less-compact structure and becomes more compact as time progresses. The Ag cluster can evolve from its initial state in many different ways but the pathways are not known initially. A state is described in terms of its all atom coordinates. Kinetic pathways from states were sought using MD calculations. Two hundred fifty-five MD calculations were performed in the current state at 300 K. Each MD calculation was independent by virtue of the random seeds used to initialize the trajectories in the state. An MD calculation was stopped once a transition from the state was detected. Thus, the 255 MD calculations enabled construction of a catalog of kinetic pathways from the current state, and new states are now known to us. Additional MD calculations performed in the state could lead to further discovery of connected states. The completeness of the catalog is a function of the total MD simulation time t_S elapsed in a state S.

An extensive search over the configurational space using 255 processors over 18 days yielded 9964 states and 16,475 kinetic pathways. The model is more accurate than standard lattice KMC models used in the literature. Standard models incorrectly assume that the cluster evolves only via hopping and exchange mechanisms. MD calculations clearly demonstrate that several important concerted atom moves, which are not accounted in the standard models, play an important role in the dynamics (see Supplementary Material). As a consequence, the timescales of evolution from the lattice-based KMC models would significantly differ from an MD-based KMC model. In addition, many off-lattice states need to be included in the KMC model (see Supplementary Material). Determining all possible states and mapping the connectivity between the states is an impossible task. Use of pattern-recognition schemes that reuse previously discovered kinetic pathways for similar looking local atomic environments can aid in the construction of the KMC model.[23] In addition, rates for certain types of mechanisms found that using MD calculations can be extended to new atomic environments, which remain unexplored using cluster expansion[24–26] and neural-network techniques.

The second example involves stretching of a deca-alanine molecule by maintaining the distance d between two anchor

https://doi.org/10.1557/mrc.2018.93 Published online by Cambridge University Press

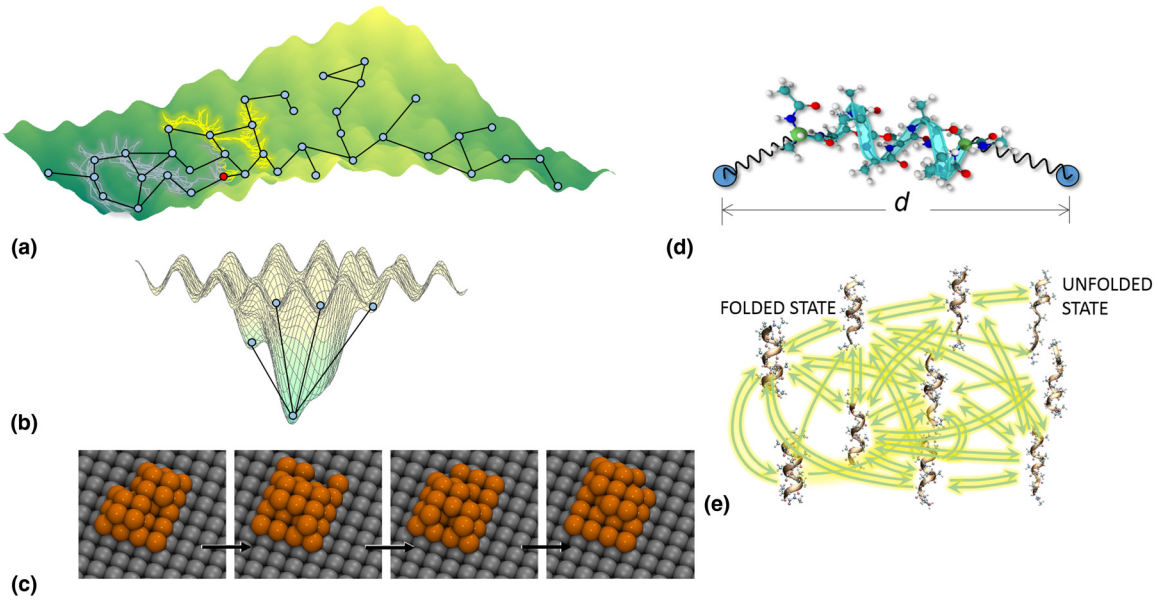


Figure 1. Schematic representation showing two extreme examples of potential energy surfaces: (a) diffuse and (b) funnel-shaped landscape. Circles denote long-lived metastable states of the system. Straight lines denote kinetic pathways between states. Random gray and yellow lines in panel a denote MD trajectories initiated from the red-colored state. (c) Evolution of an Ag nanocluster on Ag(100) at 300 K. (d) A deca-alanine molecule with force applied at its two ends. The anchor separation is kept at *d*. (e) Network model constructed for stretched deca-alanine molecule. Adapted with permission from Ref. 27, American Chemical Society, 2017.

points (blue circles) shown in Fig. 1(d). The ends of a molecule are attached to the anchor points mimicking a force spectroscopy setup. The initial state of system corresponding to a folded structure is specified [Fig. 1(e)]. Guessing the states and pathways is impractical. States of deca-alanine and their connectivity are discovered using 300 independent MD calculations at a time for a state in question. After 3 days of computational time, 810 states and more than 1000 kinetic pathways are known. The state recognition scheme involved comparing the backbone atoms after aligning the molecule using the Kabsch algorithm.[8] The PES for the stretched deca-alanine is funnel-shaped. As a result, the system remains trapped mostly in the ten states in Fig. 1(e) for 95% of the time when d <24 Å. The KMC model in the form of a network is shown in Fig. 1(e). However, this behavior changes for higher extensions and knowledge of frequently visited states and their conformational properties becomes essential toward the calculation of the average forces experienced by the anchors. It is always possible that some states and pathways that are accessed at much longer MD timescales remain to be discovered. One would like to calculate the uncertainty in the two models despite differences in the network topology.

On-the-fly construction of a KMC model and its uncertainty

A step-by-step procedure to generate a KMC model using MD is described next. Consider the initial state S shown in red in Fig. 2(a). An MD calculation is performed in state S [red circle in Fig. 2(a)] in the canonical ensemble till a transition from the state is detected. The new state of the system and the waiting time are noted. As mentioned earlier, kinetic pathways are defined only in terms of their initial and final states. Additional waiting times are collected by seeding independent MD trajectories from S. The advantage of constraining the system to state S is that it prevents the system from performing a random walk on the energy landscape and a large number of transitions from S can be sought. We term this procedure as the state-constrained MD method. Next, the rate constant k_p for a pathway p is estimated using a statistical technique. For instance, the estimate using the maximum-likelihood analysis is given by

$$k_p = n_p/\tau_{S,p}, \tag{1}$$

where n_p is the number of transitions involving p and $\tau_{S,p}$ is the time elapsed in state S for the n_p transitions. Independent MD trajectories can be generated in parallel when multiple processors are available. The total MD time invested into the construction of the model [with two pathways from state S as shown in Fig. 2(a)] is t_1.

The kinetic information content, e.g., states visited, waiting times, in a single trajectory is limited and varies from one trajectory to another. It becomes important to compile the behavior of a large collection of trajectories and to assess the missing kinetic information in terms of the ensemble. The probability of residing in state S at time t is given by the state occupation π_S. The dynamics of first-order processes is described by the

https://doi.org/10.1557/mrc.2018.93 Published online by Cambridge University Press

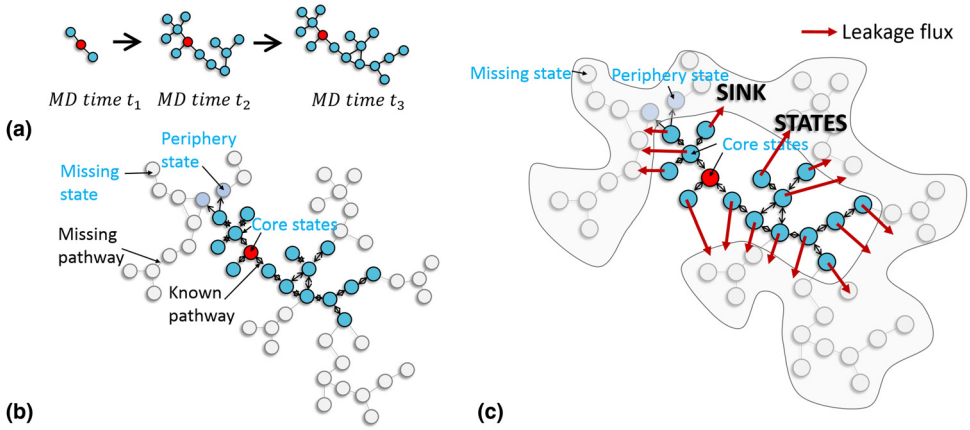


Figure 2. (a) Growing kinetic network model as a function of the duration of the molecular dynamics (MD) calculations. (b) Types of states in the network model: core (blue and red circles), periphery (gray-filled circles), and missing states (white circles). (c) Schematic representation illustrating the concept of leakage flux from the core states.

master equation:

$$\frac{d}{dt}\pi_S(t) = \sum_{S' \neq S} k_{S' \to S}\pi_{S'} - \sum_{S' \neq S} k_{S \to S'}\pi_S. \qquad (2)$$

The rate of change in π_S is given by the probability flux into S minus the probability flux out of state S. Writing Eq. (2) requires as an input the catalog of states and pathways discovered in our MD trajectories. Therefore, Eq. (2) captures our current understanding of the dynamics. Since in our present KMC model (MD time t_1) only pathways from state S have been sampled [Fig. 2(a)], the rates constants $k_{S' \to S} = 0, \ \forall\, S' \neq S$. Equation (2) reduces to $d[\pi_S(t)]/dt = -\sum_{S' \neq S} k_{S \to S'}\pi_S$. We term state S as a core state as the state has been sampled and one or more pathways from the state are known. States $S' \neq S$ are termed as the periphery states of the model since they have not been adequately sampled and none of the pathways from S' are known. States that have not been visited so far in the MD calculations are called missing states. For the periphery states, $d[\pi_{S'}(t)]/dt = k_{S \to S'}\pi_S$ highlights that the core states are source and periphery states are sink states. It is shown later that missing states are also sink states. The periphery states can be converted one-by-one into core states by performing MD in these states. For instance, MD calculations performed in a periphery state S' may reveal a pathway from S' to S'', which was not present in the KMC model earlier. In the process, the network model grows and many new periphery states can emerge. Figure 2(a) shows a schematic of the growing network model as a function of the MD time. In Fig. 2(a), the MD time invested t_1, t_2, and t_3 are well separated so that significant changes in the model topology is observed. Blue circles in Fig. 2(a) are states discovered using MD. We shall later discuss a systematic strategy to determine which periphery or core states need to be optimally explored.

Defining the KMC model as one consisting only of the core (C) states [see red and blue states in Fig. 2(b)] ensures that the system reaches a steady state. Thus, the KMC model is given by

$$\frac{d}{dt}\tilde{\pi}_S(t) = \sum_{\substack{S' \in C \\ S' \neq S}} (k_{S' \to S}\tilde{\pi}_{S'} - k_{S \to S'}\tilde{\pi}_S), \quad S \in C. \qquad (3)$$

The tilde in Eq. (3) emphasizes that the system of equations is different from the one in Eq. (2). White circles in Fig. 2(b) denote the states that are missing (M) from the model. A more complete model would have included some of the missing states. Since $k_{S' \to S}$, $S' \in M$ is not known, we assume these rates to be zero. Estimates for $k_{S \to S'}$, $S' \in M$ are obtained later. The missing and periphery states are lumped together into the sink states, which allow us to write the full network model

$$\frac{d}{dt}\hat{\pi}_S(t) = \sum_{\substack{S' \in C \\ S' \neq S}} (k_{S' \to S}\hat{\pi}_{S'} - k_{S \to S'}\hat{\pi}_S) - \sum_{S' \in P \cup M} k_{S \to S'}\hat{\pi}_S, \quad S \in C. \qquad (4)$$

In comparison with Eq. (3), Eq. (4) additionally contains the leakage flux to the missing and periphery states. The caret indicates that the solution to Eq. (4) can be different to the one from Eq. (3). A similar approach has been used in Ref. 20 without accounting for the missing states.

Figure 2(c) explains the full network model using a schematic representation. Probability from core states (blue or red circles) can leak (red arrows) into the periphery and missing states (shown in gray and gray-blue circles, respectively, within the gray-shaded region). The system cannot return to the core states as pathways from the sink to core states are unknown.

https://doi.org/10.1557/mrc.2018.93 Published online by Cambridge University Press

The KMC model behaves identically to the full network model as long as the leakage to sink states remains small. An upper bound for the probability flux from a core state S to the sink states is obtained. This bound is termed as the leakage flux F_S [see Fig. 2(c)]. The leakage flux is given by

$$F_S = \hat{\pi}_S \left(\sum_{S' \in P} k_{S \to S'} + \sum_{S' \in M} k_{S \to S'} \right). \quad (5)$$

The leakage flux is small as long as $\hat{\pi}_S$ or $\sum_{S' \in P} k_{S \to S'} + \sum_{S' \in M} k_{S \to S'}$ or both are small. The leakage rate to the periphery states $\sum_{S' \in P} k_{S \to S'}$ in Eq. (5) is already known from MD calculations. The leakage rate to missing states is obtained from the available MD data. Estimates for the missing rates of the form

$$\sum_{S' \in M} k_{S \to S'} = \frac{\ln(1/\delta)}{t_S} + k_b (N_b - n_b). \quad (6)$$

were obtained in Ref. 16. When the number of pathways from a state is small, the missing rate is bounded by $\ln(1/\delta)/t_S$. Here t_S is the MD time spent in the state S and $1 - \delta$ is the confidence, such that $0 < \delta < 1$. Non-dynamical sampling techniques[19–21] cannot provide estimates for the missing rate from a state since t_S is not available. Low values of δ (typically $\delta < 1$) are used. Additional corrections given by the second term in Eq. (6) are needed when a large number of pathways, N_b in number, from S possess similar values of rate constants. These pathways occur at the current timescales t_S. Suppose only N_b of the N_b pathways from S have been sampled. Letting k_b denote the average rate for the frequency band b, the goal is to determine N_b. The number of transitions involving the pathways p_1, p_2, …, p_{N_b} is distributed according to the multinomial distribution

$$p(n_{p1}, n_{p2}, \ldots, n_{pN_b}) = \frac{n_S!}{\prod_{p=1,2,\ldots,N_b} (n_p!)} \prod_{p=1,2,\ldots,N_b} \left(\frac{1}{N_b} \right)^{n_p}. \quad (7)$$

The correction, which involves estimating N_b from the waiting times sampled for N_b pathways, can be ignored when $N_b < 10$.

Three questions can be addressed as this point:

I. How to quantify the error in Eq. (3)—the KMC model.
II. Find how long Eq. (3) can faithfully predict the dynamics by estimating the uncertainty in the model.
III. How to perform the MD sampling to reduce the uncertainty.

The difference between the two models $\hat{\pi}_S(t) - \tilde{\pi}_S(t)$ remains small as long as the total leakage flux from the core states is negligible. The timescales where the two models begin to diverge determines the validity time for the KMC model. The validity time allows systematic comparison of the two models of the same system constructed with different amounts of MD time in each state, while accounting for known kinetic rates, topology, and relaxation times of the network as well as pathways and states missing from the model. The missing and periphery states may become important to the dynamics beyond the validity time of the model. The KMC model can be regarded as complete with confidence $(1 - \delta)$ till the validity time. We illustrate some of these points next.

Figure 3(a) shows an example with a single alanine dipeptide molecule in solvent. Initially only state 1 is known to us. State-constrained simulations were performed in state 1, which resulted in the discovery of states 2–4 with relatively few MD calculations. The transition to state 5 is rare and therefore it took a longer time to discover the pathway from state 1 to 5. Rates from state 1 calculated using Eq. (1) for the different pathways are shown in the figure next to the arrows. The number in parentheses next to the rate constant gives the number of MD calculations used to estimate the rate. Similarly, we performed several thousand MD calculations in states 2–5 to learn about the pathways on-the-fly as the trajectories are generated in parallel. Note that pathway from state 3 to 5 was not sampled but missing rate from state 3 can be estimated using Eq. (6). The total MD time invested in building the five core state (states 1–5) KMC model exceeded 0.3μs. Eleven periphery states (states 6–16) are found. These states are shown in the dotted circles in Fig. 3(a). Solving the KMC model [i.e., Eq. (3) with only core states] shows that the system largely resides in states 1 and 2. The full model [Eq. (4)] contains the five core and sink states. Since the flux from states 1–5 to states 6–16 are small, MD calculations were not performed in states 6–16. This aspect can be easily checked by comparing the time-dependent occupations from the KMC model [black lines in Fig. 3(b)] and the full model [red lines in Fig. 3(b)]. The models agree till about 0.1μs beyond which the leakage to the periphery and missing states becomes significant for us to lose confidence in the KMC model. The validity time for the KMC model is 0.1μs. Missing states and pathways in the KMC model can become important to the dynamics after the time simulated with the KMC model exceeds its validity time.

Post-processing of MD trajectories can be used to create a fresh KMC model or update a pre-existing KMC model. The validity time for the resulting model is obtained as follows:

Step 1: Identify the states and pathways of the material/biomolecular system from the collection of MD trajectories. Obtain rate constants for pathways using a statistical approach, e.g., using Eq. (1), provided the number of transitions for the pathway n_p is large enough for us to calculate the rate with a high confidence.

Step 2: Determine the core and periphery states in the updated KMC model.

Step 3: Use Eq. (6) to calculate the missing rate from each core state (in case the core state has been updated).

Step 4: Construct the KMC [Eq. (3)] and full network model [Eq. (4)].

Step 5: Solve the system of equations in Eqs. (3) and (4) to obtain the dynamical evolution $\tilde{\pi}_S(t)$ and $\hat{\pi}_S(t)$, respectively.

Step 6: The validity time is determined as the time where $\tilde{\pi}_S(t)$ and $\hat{\pi}_S(t)$ begin diverge.

https://doi.org/10.1557/mrc.2018.93 Published online by Cambridge University Press

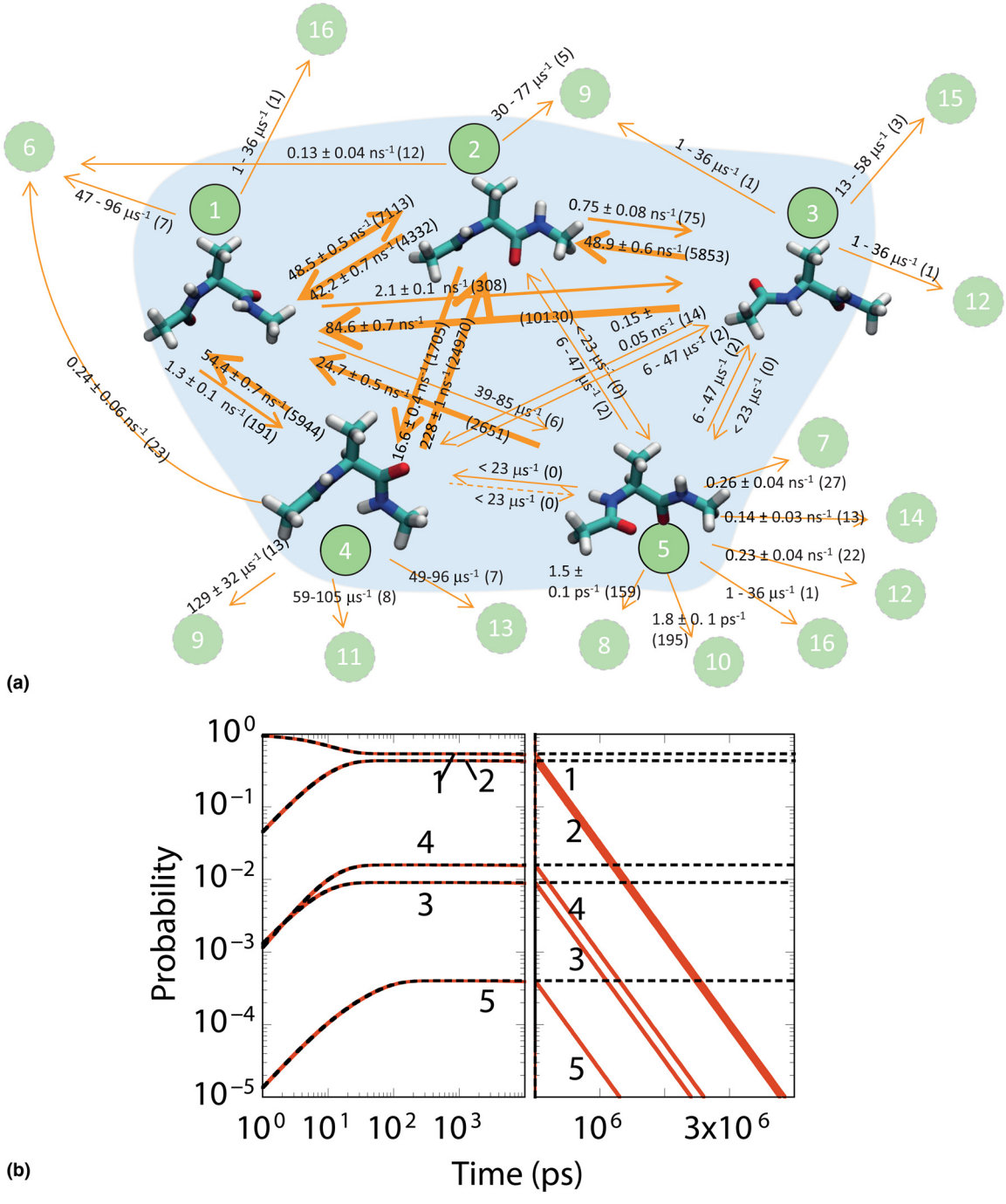


Figure 3. (a) Kinetic network model for solvated alanine dipeptide containing 16 states. States 1–5 form the core states while 6–16 are periphery states. (b) Calculation of the validity time for the kinetic network model. Adapted with permission from Ref. 8, AIP Publishing, 2017.

An automated procedure called the programmed state-constrained MD to efficiently build the KMC model was reported in Ref. 8. The procedure involves systematically finding the largest leakage rates and performing additional MD to plug the probability leakage from the core states. When the leakage to a specific periphery state S' is the largest source of leakage in Eq. (4), MD calculations are performed in the periphery state S' so that pathways from S' can be discovered and S' converted to a core state. On the other hand, when the largest leakage results from a core state S to missing states performing MD in S is recommended in order to lower the missing flux. The programmed state-constrained MD enables automatic construction of the KMC model in the most efficient manner. The main steps involved are a continuation of steps 1–6 described earlier:

https://doi.org/10.1557/mrc.2018.93 Published online by Cambridge University Press

Step 7: Calculate the leakage flux from each core states using Eq. (5) at the point where $\tilde{\pi}_S(t)$ and $\hat{\pi}_S(t)$ begin diverge, i.e., at the validity time.

Step 8: Identify sources of large leakage. When a significant leakage occurs from a core state S to a periphery state S' additional MD should be performed in the periphery state S'. On the other hand when the missing flux from a core state S is large, additional MD is required in the core state S.

Step 9: Perform additional MD in states that associated with large leakage. For instance, significant improvement can be made to the network model by generating 100 or more short MD trajectories. These could be state-constrained MD trajectories ending with a transition. Figure 4 shows the flow chart for the state-constrained MD calculations.

Step 10: Go to step 1 and continue as long as a target validity time is achieved or until more computer time can be afforded.

Accelerated construction of KMC models

When the system is equilibrated in N trapping core states corresponding to deep basins in the funnel-shaped energy landscape, the total leakage rate for the network is $k^{\text{leak}} = \sum_S F_S$. It can be shown that $k^{\text{leak}} = N\tau_{\text{MD}}^{-1} \ln(1/\delta)$, where τ_{MD} is the total MD time. The validity time is approximately given by $(k^{\text{leak}})^{-1}$, i.e., validity time scales as τ_{MD}. Accelerated techniques can help increase the validity time and reduce the KMC model uncertainty by revealing missing pathways from core and periphery states. A number of (dynamical) enhanced sampling techniques[28–31] are available that can readily access orders-of-magnitude longer times t_S or τ_{MD} than possible with regular MD. Recent examples of exploiting dynamical-enhanced sampling techniques include Ref. 7, 15, 32 and 33. Two methods developed by the author's research group are described next.

The temperature-programmed molecular dynamics (TPMD) method[34,35] is a versatile technique that can be used with a variety of materials with thermally activated events provided the system permits the use of high-temperature MD calculations. TPMD has been employed with metals, semiconductor, ionic materials, and solvated systems. In TPMD, one performs state-constrained MD calculations with a temperature program. The temperature program specifies the set-point temperature to be used with short MD calculations. Typically, the temperature is raised in stages. The activated events are triggered from the state in question when higher temperatures are accessed. The main feature of the TPMD method is that rate constants can be estimated over a range of temperatures from the temperature-dependent waiting times sampled for pathways from the state. Pathways selected at ms and longer timescales are readily accessed with short (ps–ns long) TPMD calculations.

Another powerful technique called the force spectroscopy method has been applied to study the rare events in single biomolecules.[27] Typically, the native structure of biomolecules tends to be a folded state of the molecule. Unfolded conformations are rarely accessed. Application of high temperatures is generally not advisable in these systems. The main idea is to stretch the molecule similar to the ones shown in Fig. 1(d), which allows it to access unfolded conformations more readily.

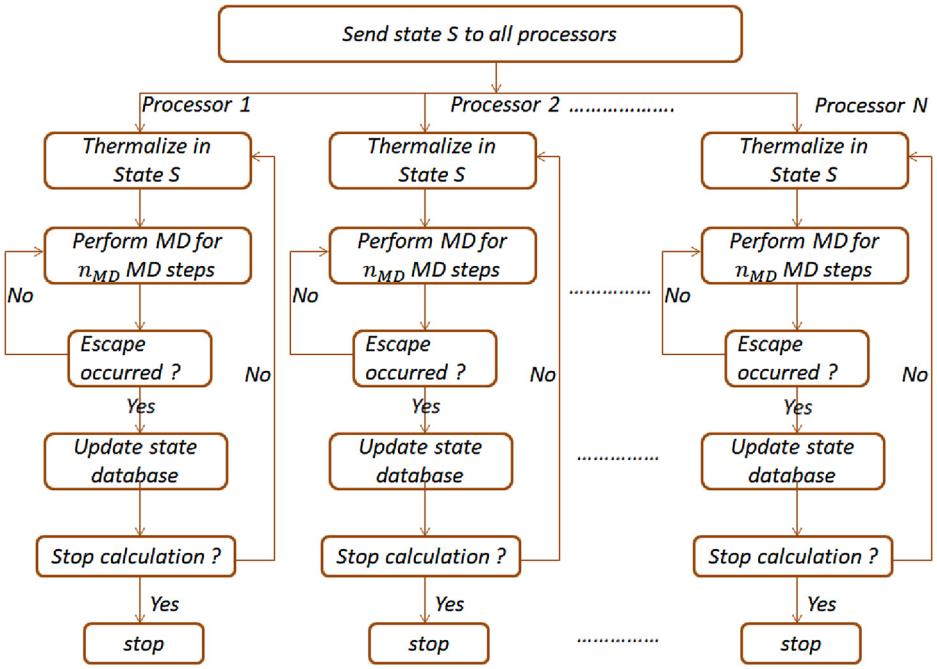


Figure 4. Flowchart for the state-constrained MD calculation. Depending on the material/biomolecular system typically n_{MD} is chosen such that the corresponding time is smaller than the average waiting time for a transition from state S.

https://doi.org/10.1557/mrc.2018.93 Published online by Cambridge University Press

The rates measured at various stretched conditions can be used to estimate the rates for zero-force conditions.

Conclusions

KMC models constructed bottom-up from finite MD trajectories MSMs have the potential to become powerful computational tools for studies of biomolecular, materials, and reacting systems. Unfortunately, such models are rarely complete as many states and pathways that could be important to the dynamical evolution are missing from the model. This has a direct implication on our confidence in the model. The concept of validity time for the KMC model reviewed in this article provides a fundamental basis of measuring the model uncertainty. The validity time is generally smaller than the total amount of MD time invested toward construction of the model. The situation is far worse for diffuse energy landscapes where a large number of states can be visited in the dynamics and validity times can be extremely low. Future algorithmic developments will need to focus on improving the model confidence by constructing the networks in an accelerated manner. Such algorithms could allow for validity times that are greater than the total MD time invested.

Supplementary material

The supplementary material for this article can be found at https://doi.org/10.1557/mrc.2018.93

Acknowledgments

The author acknowledges the 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.

References

1. A. Chatterjee and D.G. Vlachos: An overview of spatial microscopic and accelerated Kinetic Monte Carlo methods. *J. Comput. Aided Mater. Des.* **14**(2), 253 (2007).
2. A.F. Voter: Introduction to the kinetic Monte Carlo method. In *Radiat. Eff. Solids*, edited by K.E. Sickafus, E.A. Kotomin, and B.P. Uberuaga (Springer, NATO Publishing unit, Dordrecht, 2006).
3. C.R. Schwantes, R.T. McGibbon, and V.S. Pande: Perspective: Markov models for long-timescale biomolecular dynamics. *J. Chem. Phys.* **141** (9), 90901 (2014).
4. R.M. Ziff, E. Gulari, and Y. Barshad: Kinetic phase transitions in an irreversible surface-reaction model. *Phys. Rev. Lett.* **56**(24), 2553 (1986).
5. G.H. Gilmer, H.C. Huang, T.D. de la Rubia, J. Dalla Torre, and F. Baumann: Lattice Monte Carlo models of thin film deposition. *Thin Solid Films* **365** (2), 189 (2000).
6. G. Henkelman and H. Jonsson: Long time scale kinetic Monte Carlo simulations without lattice approximation and predefined event table. *J. Chem. Phys.* **115**, 9657 (2001).
7. A. Chatterjee and S. Bhattacharya: Uncertainty in a Markov state model with missing states and rates: application to a room temperature kinetic model obtained using high temperature molecular. *J. Chem. Phys.* **143** (11), 114109 (2015).
8. A. Bhoutekar, S. Ghosh, S. Bhattacharya, and A. Chatterjee: A new class of enhanced kinetic sampling methods for building Markov state models. *J. Chem. Phys.* **147**(15), 152702 (2017).
9. A. Chatterjee: Accelerating rare events and building kinetic Monte Carlo models using temperature programmed molecular dynamics. *J. Mater. Res.* **33**(7), 835 (2017).
10. D. Frenkel and B. Smit: *Understanding Molecular Simulation: From Algorithms to Applications* (Academic Press, New York, 1996).
11. V.S. Pande, K. Beauchamp, and G.R. Bowman: Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **52**(1), 99 (2010).
12. J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J.D. Chodera, C. Schütte, and F. Noé: Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* **134**(17), 174105 (2011).
13. P. Haldar and A. Chatterjee: Seeking kinetic pathways relevant to the structural evolution of metal nanoparticles. *Model. Simul. Mater. Sci. Eng.* **23**(2), 25002 (2015).
14. P. Haldar and A. Chatterjee: Connectivity-list based characterization of 3D nanoporous structures formed via selective dissolution. *Acta Mater.* **127**, 379 (2017).
15. S.T. Chill and G. Henkelman: Molecular dynamics saddle search adaptive kinetic Monte Carlo. *J. Chem. Phys.* **140**(21), 214110 (2014).
16. V.J. Bhute and A. Chatterjee: Building a kinetic Monte Carlo model with a chosen accuracy. *J. Chem. Phys.* **138**, 244112 (2013).
17. V.J. Bhute and A. Chatterjee: Accuracy of a Markov state model generated by searching for basin escape pathways. *J. Chem. Phys.* **138**, 84103 (2013).
18. M.R. Sorenson and A.F. Voter: Temperature-accelerated dynamics for simulation of infrequent events. *J. Chem. Phys.* **112**(21), 9599 (2000).
19. G.C. Boulougouris and D. Frenkel: Monte Carlo sampling of a Markov web. *J. Chem. Theory Comput.* **1**(3), 389 (2005).
20. G.C. Boulougouris and D.N. Theodorou: Dynamical integration of a Markovian web: a first passage time approach. *J. Chem. Phys.* **127**, 84903 (2007).
21. L. Xu and G. Henkelman: Adaptive kinetic Monte Carlo for first-principles accelerated dynamics. *J. Chem. Phys.* **129**, 114104 (2008).
22. D.J. Wales: Energy landscapes: calculating pathways and rates. *Int. Rev. Phys. Chem.* **25**(1–2), 237 (2006).
23. D. Konwar, V.J. Bhute, and A. Chatterjee: An off-lattice, self-learning kinetic Monte Carlo method using local environments. *J. Chem. Phys.* **135**, 174103 (2011).
24. S. Verma, T. Rehman, and A. Chatterjee: A cluster expansion model for rate constants of surface diffusion processes on Ag, Al, Cu, Ni, Pd and Pt (100) surfaces. *Surf. Sci.* **613**, 114 (2013).
25. T. Rehman, M. Jaipal, and A. Chatterjee: A cluster expansion model for predicting the activation barrier of atomic processes. *J. Comput. Phys.* **243**, 244 (2013).
26. N. Kulkarni and A. Chatterjee: Capturing local atomic environment dependence of activation barriers in metals using cluster expansion models. *J. Phys. Conf. Ser.* **759**(1), 12041 (2016).
27. S. Ghosh, A. Chatterjee, and S. Bhattacharya: Time-dependent Markov State Models for single molecule force spectroscopy. *J. Chem. Theory Comput.* **13**(3), 957 (2017).
28. A.F. Voter, F. Montalenti, and T.C. Germann: Extending the time scales in atomistic simulation of materials. *Annu. Rev. Mater. Res.* **32**, 321 (2002).
29. L.K. Béland, P. Brommer, F. El-Mellouhi, J.-F. Joly, and N. Mousseau: Kinetic activation-relaxation technique. *Phys. Rev. E* **84**(4), 46704 (2011).
30. R. Miron and K.A. Fichthorn: Accelerated molecular dynamics with the bond-boost method. *J. Chem. Phys.* **119**(12), 6210 (2003).
31. A. Laio and M. Parrinello: Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA* **99**(20), 12562 (2002).
32. A. Chatterjee and S. Bhattacharya: Probing the energy landscape of alanine dipeptide and deca-alanine using temperature as a tunable parameter in molecular dynamics. *J. Phys. Conf. Ser.* **759**(1), 12024 (2016).
33. T.D. Swinburne and D. Perez: Self-optimized construction of transition rate matrices from accelerated atomistic simulations with Bayesian uncertainty quantification. arXiv:1803.05273v1 (2018).
34. S. Divi and A. Chatterjee: Accelerating rare events while overcoming the low-barrier problem using a temperature program. *J. Chem. Phys.* **140** (18), 184115 (2014).
35. V. Imandi and A. Chatterjee: Estimating Arrhenius parameters using temperature programmed molecular dynamics. *J. Chem. Phys.* **145**(3), 34104 (2016).

https://doi.org/10.1557/mrc.2018.93 Published online by Cambridge University Press