Efficient and exact sampling of transition path ensembles on Markovian networks

Cite as: J. Chem. Phys. 153, 024121 (2020); https://doi.org/10.1063/5.0012128
Submitted: 30 April 2020 . Accepted: 18 June 2020 . Published Online: 13 July 2020

Daniel J. Sharpe, and David J. Wales

COLLECTIONS

This paper was selected as Featured

ARTICLES YOU MAY BE INTERESTED IN

A well-behaved theoretical framework for ReaxFF reactive force fields
The Journal of Chemical Physics 153, 021102 (2020); https://doi.org/10.1063/5.0013906

Rare events and first passage time statistics from the energy landscape
The Journal of Chemical Physics 153, 134115 (2020); https://doi.org/10.1063/5.0016244

Optimal estimates of self-diffusion coefficients from molecular dynamics simulations
The Journal of Chemical Physics 153, 024116 (2020); https://doi.org/10.1063/5.0008312
Efficient and exact sampling of transition path ensembles on Markovian networks

Cite as: J. Chem. Phys. 153, 024121 (2020); doi: 10.1063/5.0012128
Submitted: 30 April 2020 • Accepted: 18 June 2020 • Published Online: 13 July 2020

Daniel J. Sharpe and David J. Wales

AFFILIATIONS
Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

Author to whom correspondence should be addressed: dw34@cam.ac.uk

ABSTRACT
The problem of flickering trajectories in standard kinetic Monte Carlo (kMC) simulations prohibits sampling of the transition path ensembles (TPEs) on Markovian networks representing many slow dynamical processes of interest. In the present contribution, we overcome this problem using knowledge of the metastable macrostates, determined by an unsupervised community detection algorithm, to perform enhanced sampling kMC simulations. We implement two accelerated kMC methods to simulate the nonequilibrium stochastic dynamics on arbitrary Markovian networks, namely, weighted ensemble (WE) sampling and kinetic path sampling (kPS). WE-kMC utilizes resampling in pathway space to maintain an ensemble of representative trajectories covering the state space, and kPS utilizes graph transformation to simplify the description of an escape trajectory from a trapping energy basin. Both methods sample individual trajectories governed by the linear master equation with the correct statistical frequency. We demonstrate that they allow for efficient estimation of the time-dependent occupation probability distributions for the metastable macrostates, and of TPE statistics, such as committor functions and first passage time distributions. kPS is particularly attractive, since its efficiency is essentially independent of the degree of metastability, and we suggest how the algorithm could be coupled with other enhanced sampling methodologies. We illustrate our approach with results for a network representing the folding transition of a tryptophan zipper peptide, which exhibits a separation of characteristic timescales. We highlight some salient features of the dynamics, most notably, strong deviations from two-state behavior, and the existence of multiple competing mechanisms.

Published under license by AIP Publishing. https://doi.org/10.1063/5.0012128

I. INTRODUCTION
The stochastic dynamics of many complex systems can be formulated as a continuous-time Markov chain (CTMC), with dynamics governed by the master equation. In molecular and condensed matter systems, the construction of transition networks circumvents the timescale problem, which precludes the use of unbiased molecular dynamics (MD) simulations for systems exhibiting rare event dynamics. One approach to constructing a master equation representation is to map the stationary points of the potential energy landscape onto the nodes and edges of a network, as in discrete path sampling (DPS). This strategy avoids explicit simulation of the dynamics and is therefore especially useful for modeling systems featuring broken ergodicity. Moreover, the resulting network representation of the underlying energy landscape preserves the full dimensionality of the configuration space. Both of these considerations are often important for biomolecular systems, since a suitable low-dimensional projection of the energy landscape does not necessarily exist. Projection of the energy landscape onto inappropriate reaction coordinates is liable to misrepresent the transition state ensemble (TSE) region that is crucial to the description of the rare events.

Many alternative methods for constructing transition networks representing the dynamics of continuous-state systems have been developed, and we do not aim to review them all here. Transition networks described by a linear master equation also appear in social and economic models. Transition networks for which the dynamics are governed by a nonlinear master equation, where higher-order terms arise due to interactions of species with discrete populations, are common in systems biology, for example, in the modeling of gene regulatory networks, as well as in epidemiology and ecology, and can be mapped to linear transition networks.
In the present work, we describe efficient and exact methods for explicit simulation of the nonequilibrium stochastic dynamics for arbitrary CTMCs. In particular, we are interested in performing a detailed analysis of the transition path ensemble (TPE), the set of \( A \leftarrow B \) transition paths from initial to absorbing macrostates, denoted as \( B \) and \( A \), respectively. Key properties characterizing the global \( A \leftarrow B \) dynamics, such as mean first passage times (MFPTs) and committor probabilities, can be calculated robustly by the graph transformation method. In the present contribution, our aim is to describe methods that can elucidate how these quantities are encoded in the ensembles of pathways from which they are computed. The MFPT alone is not particularly informative, and the full FPT distribution may have a complex form. The moments of the probability distributions for path properties in the first passage can be calculated numerically by exploiting recursion relations that iteratively perform the infinite sum over paths of increasing length. However, this approach is not scalable. Often, information on the sequence of events for a typical \( A \leftarrow B \) transition is of interest. For complex systems, the dominant transition paths, which can be determined by \( k \) shortest paths algorithms with appropriate edge weights, account for only a small fraction of the \( A \leftarrow B \) flux, and therefore, analysis of the shortest paths alone may be misleading. For these reasons, to gain rigorous, detailed, and quantitative insight into an \( A \leftarrow B \) dynamical process, it is necessary to sample the TPE explicitly.

There are various numerical methods for obtaining detailed trajectory information. Direct solution of the master equation by linear algebra methods rapidly becomes intractable with increasing system size. In strongly metastable stochastic networks, the global dynamical behavior is dominated by the rare fluctuations across the boundaries between the long-lived states. Separation of characteristic timescales is a ubiquitous feature of realistic models for dynamical processes, including biophysical and biochemical systems. In this regime, linear algebra methods are numerically unstable, and mean-field methods based on deterministic ordinary differential equations (ODEs) may severely misrepresent the dynamics. Therefore, the most generally applicable approach to analyze the time-dependent occupation probability distribution on transition networks is to use kinetic Monte Carlo (kMC) simulation to sample the solution to the master equation, either exactly or approximately, by the generation of individual realizations of trajectories.

The problem of “flickering” trajectories within metastable macrostates seriously limits the efficiency of standard rejection-free exact kMC algorithms, such as the Bortz–Kalos–Lebowitz (BKL) algorithm extended herein, and the equivalent Gillespie algorithm for stochastic reaction networks. Hence, there is a need to employ some enhanced sampling methodology to accelerate the observation of rare events in kMC simulations. Many solutions have been proposed to the timescale problem associated with standard kMC methods. Strategies to ensure that the entire state space is representatively sampled include biasing the simulations and reweighting trajectories, perturbing existing transition paths, and repeatedly simulating portions of transition paths in parallel based on a division of the state space. One class of accelerated kMC methods is based on the formulation of the escape of a trajectory from a metastable trapping basin as an absorbing Markov chain, as in the Monte Carlo with absorbing Markov chains (MCAMC) algorithm. The master equation of the absorbing Markov chain can be solved exactly by first passage time analysis (FPTA). Alternatively, the master equation can be solved approximately by the mean rate method or by assuming a local equilibrium within the active basin. The metastable macrostates may be specified in advance or can be determined on-the-fly. A basin that is being actively sampled may be built up as a Markovian web of explored microstates. Methods based on absorbing Markov chains, especially those utilizing FPTA, incur a significant computational overhead that severely limits the feasible size of the trapping basins and hence the potential computational gains achievable by such methods. Employing the approximate alternatives to FPTA, or solving the dynamics within metastable sets of microstates using ODEs, forfeits a statistically exact description of the trajectories within the metastable basins. The graph transformation method can be leveraged to keep the number of microstates in a trapping basin small, by iteratively eliminating nodes from the transition network, and renormalizing the transition probabilities and waiting times in the reduced network to preserve the mean of the escape time distribution.

An efficient approximate approach to facilitate the escape of trajectories from metastable macrostates is provided by accelerated superbasin kMC (AS-kMC). In AS-kMC, the repeated observation of a transition between a pair of microstates triggers a search to determine the complete metastable basin to which the pair belongs. This neighbor search determines a subnetwork of microstates that are internally connected by fast transition rates, according to a given threshold. Then, the rates of all inter-microstate transitions within the basin are raised by a scale factor, in a way that maintains the accuracy of the kMC trajectory within a specified error tolerance. This biasing of the individual transition rates eventually encourages escape from a metastable basin. Many other approaches to accelerating kMC simulations exist, including strategies based on an importance function, "leapfrog" moves, tau-leaping, multi-level algorithms, sliding windows, uniformization, stochastic complements, waste recycling, and more.

In the present work, we analyze two complementary enhanced sampling methods, which facilitate efficient kMC simulations for CTMCs of varying dimensionality and metastability. The first method that we consider is weighted ensemble (WE) sampling, originally proposed in Ref. 134 and pioneered by Zuckerman and co-workers. The WE methodology has been discussed primarily in the context of stochastic MD simulations, including applications to biomolecular conformational transitions but has also been applied to kMC simulations of stochastic network models, specifically to the solution of nonlinear master equations in systems biology. WE sampling belongs to a family of enhanced sampling methods where the state space is partitioned into nonoverlapping bins and an ensemble of trajectories are simulated in parallel. Each trajectory is associated with a statistical weight, and the ensemble of trajectories is maintained by resampling in the pathway space. The simulation of trajectory segments for transitions between the bins facilitates the simulation of complete reactive trajectories between two defined endpoint macrostates of interest. We describe WE-kMC in more detail in Sec. 2 D. The WE method is exact in sampling the path probability distribution for Markovian dynamics and can be highly
efficiency in accelerating the observation of rare events.\cite{148} Since the WE method utilizes multiple trajectories, it provides a natural approach for identifying multiple pathway ensembles.\cite{151,152} This capability is desirable in the present context, since Markovian networks representing realistic dynamical processes, such as biomolecular conformational transitions,\cite{149,154} frequently contain competing sets of pathways.\cite{155}

The second approach to enhanced sampling of the dynamics on arbitrary CTMCs that we consider is kinetic path sampling (kPS),\cite{156,157} which provides a powerful alternative to methods based on the explicit kMC simulation of trajectories. kPS leverages graph transformation\cite{158} to generate a stochastic escape path from a defined trapping basin to an absorbing boundary, along with an associated waiting time, which is exactly consistent with the master equation. The order of inter-microstate transitions in the escape path is not computed, but microstates at the absorbing boundary are sampled with the correct probability distribution. The kPS method, which is close in spirit to the MCAMC algorithm,\cite{97,98} is described in more detail in Sec. II E.

The WE-kMC and kPS methods have some complementary desirable features. The time complexity of the kPS algorithm is nearly independent of the metastability of the transition network, and therefore, the method is essentially immune to kinetic trapping. However, the cost of generating an escape path from a trapping basin is $O(N_{E}^{6})$, where $N_{E}$ microstates that are eliminated in the graph transformation stage of the algorithm, scales roughly as $O(N_{E}^{6})$. Hence, there is a significant computational overhead associated with the method if the trapping basins are large.\cite{152} The memory requirements of a single iteration of the kPS algorithm likewise scales strongly with the size of the community. In its simplest form, the kPS algorithm does not simulate ordered trajectories on the transition network but instead computes a non-Markovian trajectory on the coarse-grained network defined by the partitioning of the state space. Although kPS can be extended to compute the order of transition events along a detailed escape trajectory on the network, the time complexity is then adversely affected by metastability.\cite{151}

In WE-kMC, there is no such restriction on the size of communities, but it is essential that the communities reflect all uncoupled slow dynamical modes of the system, since dynamical modes orthogonal to the bin coordinates must be sampled by standard rejection-free kMC.\cite{149} For this reason, WE-kMC is less efficient for more strongly metastable Markov chains.\cite{151} However, the method is highly parallelizable,\cite{149} and in the context of simulating the dynamics on transition networks, where the bins to which microstates belong are identified simply by labels, the computational overhead associated with the trajectory resampling procedure is negligible. Moreover, by utilizing a protocol to accelerate the establishment of a steady state, WE-kMC can also be used to conduct equilibrium simulations.\cite{154}

Both WE-kMC and kPS are based on a partitioning of the state space into disjoint sets of microstates, although this division need not be known a priori. Here, we propose the use of an unsupervised stochastic community detection algorithm, namely, multi-level regularized\cite{154,155} Markov clustering (MLR-MCL), to identify the metastable sets of nodes in a transition network,\cite{158} which are then used as the fixed bins for the WE-kMC and kPS simulations. Strategies for the choice of metastable macrostates, which can also be performed adaptively, are discussed in more detail in Sec. II C.

II. METHODOLOGY

A. Master equation dynamics

The simulation methodology and the theory we present are applicable to an arbitrary CTMC. We illustrate our approach with a transition network constructed using geometry optimization methods\cite{159} to locate the stationary points on a potential energy landscape. Here, local minima and transition states are mapped to the $V$ nodes and $E$ weighted and bidirectional edges, respectively, of a transition network.\cite{11,12,13} The edge weights are minimum-to-minimum rate constants, usually estimated by harmonic transition state theory,\cite{14,15} although any unimolecular rate theory, including methods based on explicit dynamics, may be used. Formally, the partition function can be written as a sum of contributions from local minima,\cite{16,163} and so all thermodynamic properties can be extracted from the discretized representation of the energy landscape if the local densities of states are known. In practice, the equilibrium occupation probabilities associated with the nodes are usually approximated assuming a locally harmonic density of states.\cite{16,163} Transition networks constructed by geometry optimization typically contain tens to hundreds of thousands of nodes and are sparse.\cite{163} Further details of the discrete path sampling (DPS) methodology for the construction and analysis of transition networks can be found in recent reviews.\cite{162,164,165}

Markovian dynamics on a transition network are described by the linear master equation,\cite{151}

$$\frac{dp(t)}{dt} = \mathbf{K}p(t).$$

which can be written in matrix notation,

$$\frac{dp(t)}{dt} = \mathbf{K}p(t).$$

Here, $p(t) = (p_1(t), p_2(t), \ldots, p_V(t))^T$ is the time-dependent occupation probability vector for the nodes of the transition network, and $\mathbf{K}$ is the transition rate matrix. The off-diagonal elements $k_{ij}$ of $\mathbf{K}$ are the inter-microstate rates for the $i \leftrightarrow j$ transitions. The diagonal elements are chosen so that the columns of the matrix sum to zero, $k_{ii} = -\sum_{j \neq i} k_{ij}$. At equilibrium, the occupation probability vector is
equal to the stationary probability vector \( \pi \), and the detailed balance condition is satisfied, \( k_{ij} \pi_j = k_{ji} \pi_i, \forall i \neq j \).

Typical numerical methods for the linear algebra solution of the master equation, \( \mathbf{p}(t) = \exp(\mathbf{K}t)\mathbf{p}(0) \), have time complexity \( O(V^3) \), and this direct approach is therefore intractable, except for transition networks of relatively low dimensionality.\(^{145}\) This problem is exacerbated for rare event dynamics, where numerical instability is a pervasive problem affecting linear algebra methods to calculate the MFPT between two endpoint macrostates\(^{166}\) or the time-dependent occupation probability distribution via solution of Eq. (2).\(^{144}\) It is the aim of the present work to sample the exact solution to the linear master equation [Eqs. (1) and (2)], for transitions between two endpoint macrostates of interest, \( A \) and \( B \), by the explicit simulation of trajectories using accelerated kMC.

Let an \( a \leftrightarrow b \) trajectory on a transition network (i.e., a discrete path\(^{141}\)) connecting microstates \( a \in A \) and \( b \in B \), where \( B \) is the initial macrostate and \( A \) is the absorbing macrostate, respectively, be denoted by the ordered sequence of visited microstates \( \xi^{a\rightarrow b} = (a \leftarrow i_0 \leftarrow i_{b-1} \leftarrow \cdots \leftarrow i_1 \leftarrow b) \). Here, \( i \) is used to denote intervening nodes separating the endpoint macrostates, \( i \in (A \cup B)^c \). The probability \( P[\xi^{a\rightarrow b}] \) of this reactive \( A \leftrightarrow B \) path is simply a product of branching probabilities \( p_{ij} = k_{ij}/\sum_{j \neq i} k_{ij} \) for the inter-microstate transitions along the path, weighted by the probability of a reactive path starting from the initial microstate \( b, \pi_b/\sum_{b \neq a} \pi_b \). The contribution of a transition path to the \( A \leftrightarrow B \) steady state rate constant\(^{146}\) is the path probability \( P[\xi^{a\rightarrow b}] \) weighted by the inverse of the mean waiting time \( \tau_b = 1/\sum_{j \neq b} k_{bj} \) for the initial microstate \( b \). Therefore, if the weights associated with \( i \leftrightarrow j \) edges are chosen to be \( -\ln p_{ij} \), the set of transition paths that make the dominant contributions to the steady state rate constant for a transition network with a single source node can be extracted using \( k \) shortest path algorithms.\(^{56,166}\) The contribution of an individual path to the \( A \leftrightarrow B \) MFPT is given by the product of the path probability \( P[\xi^{a\rightarrow b}] \) and the sum of mean waiting times for nodes along the path, excluding the absorbing node \( a \). The \( A \leftrightarrow B \) MFPT, and hence the steady state and non-steady state phenomenological rate constants,\(^{145,168}\) can be calculated robustly by the graph transformation method.\(^{56,166}\) The kMC methods described below sample paths of the \( A \leftrightarrow B \) TPE in proportion to their probabilities \( P[\xi^{a\rightarrow b}] \) and therefore yield an unbiased estimate for the \( A \leftrightarrow B \) MFPT.

### B. Rejection-free kinetic Monte Carlo

The BKL (or \textit{n-fold way}) algorithm\(^{77,84,167}\) is a rejection-free\(^{77,168}\) kMC simulation method, frequently employed in condensed matter\(^{6,7,8,78}\) and surface\(^{8,7,8,79}\) physics. In this section, we briefly outline the algorithm in the context of simulating the solution to the master equation for a transition network described by the rate matrix \( \mathbf{K} \) [Eq. (2)]. Inter-microstate transitions \( i \leftrightarrow j \), associated with rates \( k_{ij} \), which, in general, could be time-dependent, are assumed to be independent Poisson processes associated with average transition times equal to the waiting time for the \( j \)th node, \( \tau_j \). For each microstate \( j \), a list of possible transitions is constructed, and at each iteration of the algorithm, a transition event is selected randomly with probability equal to the branching probability \( p_{ij} = k_{ij}/\sum_{j \neq i} k_{ij} \). The move is always accepted, and the simulation clock is incremented by drawing a random value from the exponential distribution of waiting times between transitions, with rate parameter \( \tau_j^{-1} \), \( p(\Delta t) = \tau_j^{-1} \exp(-\tau_j^{-1}\Delta t) \).\(^{341}\) The exponential form for the distribution of waiting times follows from the fact that the competing independent Poisson processes for the \( i \leftarrow j \) transitions together generate a new Poisson distribution. In practice, advancing the system clock by \( \Delta t \) is achieved by drawing a uniform random number \( r \in (0, 1) \) and setting \( \Delta t = -\tau_i \ln r \); a formal derivation of this algorithm is presented in Refs. 73 and 86.

Since the BKL algorithm preserves the branching probabilities for transition events, and with the assumption that inter-microstate transitions are uncorrelated processes with a Poissonian distribution of transition times, the algorithm produces realizations of trajectories governed by the linear master equation [Eq. (2)] with the exactly correct statistical frequency. Many independent trajectories are required to achieve proper sampling of \( \mathbf{p}(t) \) and hence produce a converged solution of the master equation. Although rejection-free kMC methods can eliminate the possibility of self-transitions, and therefore of becoming trapped in any single microstate associated with a small escape rate, standard simulations can be very inefficient for transition networks featuring a separation of characteristic timescales.\(^{146,147}\) In Secs. II C–II E, we discuss how the standard BKL algorithm can be extended to formulate enhanced sampling methods based on a specification of disjoint sets of microstates, which enables this timescale problem to be overcome.

### C. Identifying metastable states of a transition network

Both WE-kMC (Sec. II D) and kPS (Sec. II E) require a specified partitioning of the state space or a criterion for defining the metastable macrostates on-the-fly. This partitioning is an essential consideration that strongly affects the efficiency of the algorithms.\(^{158}\) This community structure must faithfully represent the coordinates corresponding to the independent slow dynamical modes of the system, which describe the rare transition events between metastable states. Hence, the partitioning should appropriately reflect the progress of the overall reactive transition between the two specified endpoint macrostates of interest \( A \) and \( B \).

Here, we propose the use of multi-level regularized\(^{152,154}\) Markov clustering\(^{55,157}\) (MLR-MCL) to efficiently characterize the metastable sets of nodes in high-dimensional and ill-conditioned transition networks at an adjustable level of timescale resolution. MLR-MCL is a stochastic unsupervised community detection algorithm that uses heuristic operations on a coarse-grained transition probability matrix to obtain a clustering that characterizes the average behavior of random walks on the network. The partitioning determined by MLR-MCL should therefore characterize the metastable sets of nodes and hence provides appropriate predefined and fixed bins for use in WE-kMC and kPS simulations.\(^{169}\) We have found that partitionings of the network obtained using MLR-MCL are strongly metastable,\(^{158}\) as indicated by widely used graph-theoretic metrics, such as the weighted normalized cut and the conductance.\(^{171}\) Furthermore, the partitioning can be subsequently refined by variational optimization of the dominant nonzero eigenvalue of the transition rate matrix. These observations, and a description of appropriate community detection algorithms\(^{172}\) to

---


Published under license by AIP Publishing
determine metastable macrostates in transition networks, will be reported elsewhere.\textsuperscript{134,136} The resolution of the community detection can be directly controlled via the choice of input parameters to MLR-MCL, namely, the granularity parameter, the number of iterations of the multilevel graph coarsening algorithm, and the lag time at which the initial transition probability matrix is estimated.\textsuperscript{132,135} The tuning of these parameters allows flexibility in the timescale at which the determined macrostates appear metastable. This feature is highly desirable, as the efficiency of WE-kMC is strongly affected by the average escape time from the macrostates, since coordinates that are not correlated with the bins must effectively be sampled by brute force.\textsuperscript{137} The efficiency of kPS is also strongly affected by the size of the macrostates. For this reason, it is favorable that MLR-MCL penalizes overly large communities, thereby addressing a deficiency of the original MCL algorithm.\textsuperscript{132,135} Because MLR-MCL is an unsupervised learning algorithm, the overall workflow for the WE-kMC and kPS simulations does not require any prior knowledge of the system. This feature addresses one of the key problems with enhanced sampling methods based on a division of the state space, namely, the determination of an appropriate partitioning is a highly non-trivial problem.\textsuperscript{6,172,173}

Neither WE-kMC nor kPS requires that the metastable macrostates are predefined or fixed throughout the simulation. We therefore also investigate defining the macrostates adaptively, using a search protocol similar to that employed in AS-kMC.\textsuperscript{134} Beginning with the currently occupied microstate, a breadth-first search procedure is used to build up a group of microstates that are mutually interconnected by transition rates that exceed a specified threshold, thereby ensuring that the resulting subnetwork is “well-knit.” The search is terminated when the size of the macrostate exceeds a specified limit or when all transitions to neighboring microstates of the subnetwork are associated with low transition rates.

D. Weighted ensemble kinetic Monte Carlo (WE-kMC)

The weighted ensemble (WE) algorithm\textsuperscript{95} is a method for resampling the path probability distribution in pathway space, which can simulate nonequilibrium\textsuperscript{134} or equilibrium\textsuperscript{136} dynamics for a variety of stochastic processes, including Langevin dynamics\textsuperscript{138} and dynamics of stochastic reaction networks.\textsuperscript{48–51} The WE method employs a partitioning of the state space into bins, which can be performed adaptively,\textsuperscript{135,174} and a set of independent trajectories (“walkers”), each associated with a statistical weight. A stochastic splitting and culling procedure, carried out at regular time intervals $t_k$, maintains a target number of trajectories in each bin throughout the simulation, thus ensuring representative sampling of the entire state space. This resampling procedure is exactly correct for Markovian dynamics and generates an unbiased sample of the TPE.\textsuperscript{134–136} The independence of the simulated trajectories leads to linear parallel scaling.

To employ the WE methodology for nonequilibrium $A \leftrightarrow B$ stochastic dynamics, we define a set of non-overlapping bins, and specify target numbers of walkers, $M_j$, for each bin, which are not necessarily equal and which can be updated on-the-fly. We also specify a time interval $t_k$ for conducting the walker resampling procedure. The WE simulation begins by spawning a specified number of weighted trajectories, where the sum of the weights is unity. The starting trajectories can be set up according to any desired initial condition, including spawning many trajectories in multiple different bins and with non-uniform weights. The stochastic dynamics are propagated independently for each of the trajectories, and when all trajectories have exceeded the time interval for resampling, all populated bins are checked for the total numbers of trajectories that they contain. If the number of trajectories in a given bin is less than the target number, then trajectories occupying the bin are chosen randomly, in proportion to their relative weights, to be split. The newly spawned trajectories each inherit an equal share of the weight of the parent trajectory, and they all share the history of the parent. This procedure is repeated until the number of trajectories in the bin exceeds the target value. If the number of trajectories in a bin is greater than the target number, then trajectories are culled by randomly selecting one of the two trajectories of lowest weight in the bin to survive, in proportion to their relative weights, and breaking ties arbitrarily. The surviving trajectory inherits the weight of the culled trajectory, and this procedure is repeated until the target number of trajectories is met. Then, the walkers of the new set of trajectories are again propagated independently until the time for the next resampling operation is reached. The times and weights of walkers hitting the absorbing macrostate $A$ are recorded, yielding the FPT distribution and an estimate for the MFPT straightforwardly. An overview of the WE-kMC algorithm is illustrated in Fig. 1. Because the WE simulation distributes resources to sampling the entire state space, including the crossing of dynamical bottlenecks,\textsuperscript{139} the simulation can yield the TPE even for transitions associated with timescales that are far too long to be accessible by brute force.\textsuperscript{137} The WE simulation is repeated many times to achieve sufficient sampling.

The legitimacy of resampling the trajectory ensemble is justified by a simple factorization of the path probability.\textsuperscript{135} Let $\xi^{(n\leftarrow m)} = (n \leftarrow n-1 \leftarrow \cdots \leftarrow m+1 \leftarrow m)$ be a trajectory initially at microstate $m$ and terminating at microstate $n$. The path probability $P[\xi^{(n\leftarrow m)}]$ can be factorized as

$$P[\xi^{(n\leftarrow m)}] = p_m(0) \prod_{k=m}^{n-1} P_{k+1,k} = p_m(0) \prod_{k=m}^{n-1} P_{k+1,k} \prod_{k=m'}^{m-1} P_{k+1,k}.$$  \hspace{1cm} (3)

Here, $p(0)$ specifies the initial probability distribution over microstates, and $P_{ij}$ is the branching probability for the $i \leftarrow j$ transition. Consider resampling the path probability distribution at a timestep when a particular trajectory $\xi^{(n\leftarrow m')}$ currently occupies the arbitrary microstate $m'$. From Eq. (3), any resampling procedure that exactly preserves the probability distribution of trajectories at that time, such as the splitting and culling procedure outlined above, exactly yields the correct probability distribution in pathway space at all future times, provided that members of the new set of trajectories inherit the histories of the trajectories from which they were derived. That is, if the path probabilities associated with any daughter trajectories $\xi^{(n\leftarrow m)}$ spawned from $\xi^{(n\leftarrow m')}$ are weighted by an equal...
share of the weight of the parent trajectory, then the correct path probability distribution is preserved. This factorization argument holds for both Markovian and non-Markovian dynamics.

Although the WE method generates an unbiased sample of the TPE, the correlated histories of the trajectories are problematic when attempting to make accurate statistical estimates of dynamical properties from WE simulation data. For instance, although the committer functions can, in principle, be computed by tracing the $A \leftarrow B$ paths, in practice, reliable estimation requires averaging over the results of many independent WE simulations to mitigate the effect of the correlated histories of the trajectories within a given WE run. This feature means that WE sampling and other methods that simulate complete $A \leftarrow B$ trajectories by piecing together trajectory segments, such as forward flux sampling (FFS), are not ideal for the estimation of committer functions.

For completeness, we note that WE-kMC can also be used to sample the equilibrium TPE. A steady state must eventually be reached if, when a walker reaches the absorbing macrostate $A$, the WFPT is computed directly as an average over FPTs for the absorbing macrostate $A$. We only consider nonequilibrium simulations in the present work, and hence, we do not obtain the MFPT via Eqs. (4) and (5). Instead, the MFPT is computed directly as an average over FPTs for the weighted transition paths.

E. Kinetic path sampling (kPS)

Kinetic path sampling (kPS) is a method for sampling the solution to the linear master equation [Eq. (2)] without requiring explicit KMC simulation of trajectories. The kPS algorithm uses graph transformation to reduce the representation of an escape trajectory from the active metastable macrostate. An overview of the stages of the kPS algorithm is illustrated in Fig. 2. To generate an escape path from the currently occupied trapping basin, we first define the sets of microstates that constitute the basin $B$ and the absorbing macrostate $A \equiv B^c$. A subset of $N_B$ microstates of the
trapping basin $E \subseteq B$ are marked for elimination and queued. The remaining microstates of the trapping basin, which constitute the set $T \subseteq B$, are identified. The graph transformation method\textsuperscript{[48–60]} is then used to construct the set of transition probability matrices $\{T^{(n)}(τ)\}, 0 ≤ n ≤ N_E$. That is, the set of matrices $\{T^{(n)}\}$ is formed by the iterative elimination of the $N_E$ microstates in the set $E \subseteq B$ from the subnetwork $B \cup \partial A$, where renormalization of the transition probabilities preserves the individual path probabilities and the MFPT for the set of escape trajectories from the current node to the absorbing boundary $\partial A$.\textsuperscript{[51, 54]} The initial transition probability matrix $T^{(0)}$ may be the linearized transition matrix\textsuperscript{[41]} estimated at a lag time $τ$ or the branching probability matrix. In the latter case, the mean waiting times associated with individual nodes are, in general, non-uniform.

A stochastic path from the currently occupied microstate $ε \in B$ to an absorbing microstate $α \in \partial A$ at the boundary of the metastable basin is randomly generated by repeatedly drawing new microstates according to a probability distribution defined from $T^{(0)}$ and $T^{(N_E)}$. Concomitantly, a count matrix $H^{(n)}(τ)$, containing the number of inter-microstate transitions observed for dynamics based on $T^{(n)}(τ)$, is recorded for $n = N_E$. The elements of the matrices $T^{(n)}$ and $H^{(n)}$, for which the $n$th microstate has been eliminated compared to $T^{(n+1)}$ and $H^{(n-1)}$ subsume all indirect inter-microstate transitions that proceed with $n$ as an intermediate microstate, where $n$ is visited an arbitrary number of times. Modeling the dynamics using one of the reduced transition matrices therefore greatly reduces the complexity of a sampled escape path. An escape trajectory from a microstate $ε \in B$ based on $T^{(N_E)}$ contains only a single step, notwithstanding any transitions involving microstates of the set $T$, if $T \neq \emptyset$. The kPS algorithm then exploits the fact that $H^{(n+1)}$ can be generated stochastically from $H^{(n)}$ given $T^{(n)}$ and $T^{(n+1)}$, without explicit simulation of the dynamics using $T^{(n)}$. Note that the sampling rules do not allow self-loop transitions for microstates of the set $T$, and the kPS algorithm reduces to standard rejection-free kMC\textsuperscript{[55]} in the case where $E = \emptyset$ and therefore $B \equiv T$.

The result of the repeated application of the iterative reverse randomization procedure is the hopping matrix $H^{(0)}$ for which the
elements are the number of inter-microstate kMC moves along a detailed stochastic path within the trapping basin. This matrix can therefore be used to generate a time associated with the trajectory escaping to the sampled absorbing microstate \( \alpha \in \partial \Lambda \) by sampling from a gamma distribution. kPS produces escape paths to absorbing microstates that are exactly consistent with the linear master equation [Eq. (1)] and does not necessarily require \textit{a priori} knowledge of the metastable basins. Since the number of kMC moves along the sampled escape path is calculated within the kPS algorithm, the simulation can always revert to the standard BKL algorithm on-the-fly when it is favorable to do so.

**III. RESULTS**

**A. Simulation setup and performance**

We illustrate the sampling methods described above with results for a transition network representing the folding of the tryptophan zipper peptide TZ1,\textsuperscript{159} constructed by discrete path sampling (DPS).\textsuperscript{14,15} The system was modeled using an atomistic potential and implicit solvent. Further details of the force field and the DPS procedures employed, and some preliminary analysis of the dynamics, for instance, using Dijkstra’s algorithm with appropriate edge weights\textsuperscript{13–15} to determine the transition path that makes the dominant contribution to the steady state rate constant (Sec. II A), can be found in Ref. 159. The stationary point database for TZ1 contains 69 780 minima and 99 935 transition states, and the corresponding network constitutes a single fully connected component.

In the present work, we simulate the nonequilibrium dynamics when the probability density is initially localized in an unfolded microstate for which the peptide chain is extended (with no native or non-native contacts), and the native fold is treated as an absorbing macrostate. Besides these unfolded (U) and folded (F) states, we identify two intermediate states of interest, I1 and I2. The I1 macrostate comprises partially folded conformations, which may include on- or off-pathway intermediate states with non-native contacts. The I2 macrostate comprises low-energy conformations for which the backbone is ordered similarly to the native state, but rearrangements of the side chains are required to transition to the native fold. Note also that the macrostate of unfolded structures (U) includes not only high-energy extended structures, such as the initial microstate, but also disordered structures that are collapsed (U\textsuperscript{n}). The U, I1, and I2+F macrostates used in the analysis of the time-dependent occupation probability distributions are determined by MLR-MCL at a low resolution, and the folded macrostate F comprises only a small number of microstates that are lowest in potential energy and are interconnected by fast transition rates, identified manually. The WE-kMC and kPS simulations are based on 390 communities determined by MLR-MCL at a higher resolution, with the manually chosen microstates of the set F designated as the absorbing macrostate. Further information on the community structure detection is included in the supplementary material. We compare the results from low (300 K) and high (330 K) temperature simulations. The transition network models in each case are obtained by calculating the transition rates and stationary probabilities at the chosen temperature, assuming locally harmonic densities of states for the minima and transition states on the potential energy landscape (Sec. II A).

The results from the kPS and WE-kMC simulations, each obtained from 20 000 transition paths, are nearly identical, and the MFPTs estimated from the simulation data are consistent with the exact values calculated using graph transformation 56–57 (Table I). Note that the FPTs reported herein are not directly comparable with experiment; to analyze folding rates would require extended definitions of the endpoint sets of nodes to reflect the experimental states. Our principal interest here is in the diagnosis of alternative pathways and convergence of sampling algorithms. We do not provide a detailed performance comparison of kPS and WE-kMC, since their efficiency is strongly affected by the simulation setup, which is highly flexible, and the optimal parameter choices are system-dependent and can only be discovered empirically through extensive testing. We simply note that it is essential to incorporate appropriate considerations into the design of the simulation protocol and that both accelerated kMC methods considered herein are many orders of magnitude faster than brute-force kMC. In kPS simulations, it is essential that the communities are not too large, since the graph transformation procedure then incurs a significant computational overhead, and that the simulation reverts to the standard rejection-free kMC algorithm when it is favorable to do so. In the present work, a fixed number of rejection-free kMC steps are taken after each kPS basin escape, to ensure that the trajectory moves away from the boundaries between communities, thereby avoiding the computation of expensive kPS iterations for trivial recrossings between communities. We use the same communities, determined by MLR-MCL, for the kPS and WE-kMC simulations, but the optimal choices for the communities in each case could be rather different,\textsuperscript{158} owing to the various factors affecting the efficiency of the methods. The simulation parameters are described in the supplementary material.

The kPS simulation data presented below were obtained in \( \sim 200 \) CPU hours using Intel Core i7-5820K 3.30 GHz processors. The CPU time was the same for simulations at temperatures of \( T = 300 \text{ K} \) and \( T = 330 \text{ K} \), since the time complexity of kPS is largely independent of the metastability of the transition network.\textsuperscript{60,149} The WE-kMC simulation data were obtained in \( \sim 1000 \) CPU hours for a temperature of 300 K and in around 840 CPU hours for a temperature of 330 K. In contrast, simulation of a single folding trajectory at 330 K by the BKL algorithm (Sec. II B) requires, on average, around 10 h of CPU time, and low-probability paths at the tail of the FPT distribution require much more CPU time. Brute-force kMC simulation is therefore unfeasible. Both kPS and WE-kMC simulations

<table>
<thead>
<tr>
<th>Method</th>
<th>MFPT / 10\textsuperscript{11} (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( T = 300 \text{ K} )</td>
</tr>
<tr>
<td>Graph transformation</td>
<td>9.1275</td>
</tr>
<tr>
<td>WE-kMC</td>
<td>8.8 ( \pm ) 0.4</td>
</tr>
<tr>
<td>kPS</td>
<td>8.9 ( \pm ) 0.3</td>
</tr>
</tbody>
</table>

TABLE I. MFPTs for the folding transition of TZ1 calculated by various methods. The graph transformation result is exact.\textsuperscript{1} The values from WE-kMC and kPS explicit simulation data were calculated from 20 000 transition paths and are associated with a standard error. The simulations were performed using a predefined and fixed partitioning of the network into communities determined by MLR-MCL.
employing an adaptive definition of the communities were slower than simulations based on the predefined communities determined by MLR-MCL, which suggests that the MLR-MCL communities accurately characterize the metastable macrostates and are appropriately balanced in size. Therefore, the additional computational time associated with the breadth-first search procedure to identify communities on-the-fly is an unnecessary computational expense for this system.

The superior performance of kPS compared to WE-kMC in this particular instance can be ascribed to the presence of strong kinetic traps in the transition network for which a very large number of kMC steps (sometimes more than $10^{15}$) are required to escape the corresponding community. The computational time for a kPS iteration is essentially agnostic to the number of kMC steps for inter-macrostate transitions, which are the elements of the hopping matrices $H^{(n)}$ and are not explicitly simulated. Instead, these values only enter the calculation as parameters in the binomial and negative binomial distributions from which the elements of the next hopping matrix $H^{(n+1)}$ are drawn and in the gamma distribution from which the time associated with the basin escape trajectory is sampled.\cite{Smeulders2019, Periole2020} In WE-kMC, these kMC steps must be explicitly taken in order to escape from the community, and therefore the flickering problem,\cite{Agmon1989} while not as serious as in standard kMC, may still hinder the WE-kMC calculation. Hence, the required CPU time for the WE-kMC simulation is adversely affected by decreasing temperature. We anticipate that with alternative computational resources or refinement of the WE-kMC simulation protocol, for instance, by dividing the state space into more communities, and increasing the target numbers of walkers for particular communities, significant gains in the efficiency for the WE-kMC calculation could be achieved.

B. Folding mechanism for the TZ1 peptide

To characterize the mechanistic features of the folding transition for TZ1, we calculate the vector $p(t)$ containing the time-dependent occupation probabilities for the four states of interest, $U, F, I_1, I_2$, described above (Fig. 3). Representative trajectories from the explicit simulations are shown in Fig. 4, alongside the transition path that makes the single largest contribution to the $F \leftarrow U$ steady state rate constant, where the transition times are chosen to be the mean waiting times associated with the individual nodes along this shortest path. It is immediately apparent from these calculations that the $F \leftarrow U$ transition of TZ1 does not conform to a simple two-state model of the dynamics, since the macrostate $I_1$ may persist on appreciable timescales (Fig. 3).

There is a rapid collapse of the initially occupied microstate of the state $U$, the extended conformation, to a more compact state (denoted as $U'$), which similarly contains no native or non-native contacts. These conformations are therefore grouped into the same macrostate $U$ in the calculation of the occupation probabilities for the key states, and this unfolded macrostate has a lifetime of around $10^6$ ns. Following escape from the unfolded macrostate $U$, around 70% of the paths avoid becoming trapped in the $I_1$ state, and subsequently the $I_2$ state, and therefore, there is a relatively steep increase in the occupation probability of the native folded state $F$, on a timescale of around $10^8$ ns. The remaining ~30% of the transition paths do not follow this simple fast-folding mechanism but instead become trapped in the $I_1$ state, with a lifetime of around $10^{12}$ ns, before rapidly proceeding to the native state $F$ via the $I_2$ state. The latter folding mechanism becomes slightly more favored at the lower temperature (Fig. 3). The kinetic traps in the TZ1 transition network that have a significant effect on the folding dynamics therefore correspond to partially folded conformations where the backbone is not properly arranged ($I_1$). Kinetic traps corresponding to low-energy states with an ordered backbone but improperly positioned side chains ($I_2$) are not as strongly metastable and have a low probability to appear along a transition path. They therefore have only a small effect on the folding dynamics (Fig. 3). Hence, once the peptide backbone adopts a conformation similar to the native state, the peptide almost always proceeds rapidly to the native fold.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig3.png}
\caption{Time-dependent occupation probability distribution for the four key states in the course of the $F \leftarrow U$ transition at (a) $T = 300$ K and (b) $T = 330$ K, obtained from 20,000 transition paths simulated using the kPS algorithm with a priori communities determined by MLR-MCL. The areas under the curves corresponding to the occupation probabilities of initial ($U$) and absorbing ($F$) states are shaded to aid visualization of the progress of the folding transition.}
\end{figure}
The single path that makes the dominant contribution to the steady state rate constant, determined using Dijkstra’s algorithm with appropriate edge weights, is clearly not representative of explicitly simulated trajectories (Fig. 4). While the sequence of conformational change events along this shortest path is consistent with the family of simulated transition paths corresponding to fast downhill folding, the shortest path contains no useful temporal information for this system. That is, it is not possible to identify the states that, in practice, are associated with long lifetimes. For transition networks featuring metastability, realistic transition paths feature a large number of flickers. For the shortest path at 300 K, the first passage time is $t_{\text{FPT}} \approx 10^{3.5}$ ns, but the path probability of folding trajectories with $t_{\text{FPT}} < 10^{5}$ ns is negligible (Fig. 5). Therefore, the set of shortest paths, which can be determined by an appropriate $k$-shortest paths algorithm, make a negligible contribution to the folding flux for TZ1. Moreover, because the number of kMC steps along the second family of folding trajectories, which become trapped in the I1 state, is even larger than for the first family, the path probability for any one member of this family of longer-timescale trajectories is exceedingly small. Therefore, these transition paths cannot feasibly be identified using a $k$-shortest paths algorithm, even though the paths collectively make an important contribution to the MFPT (Fig. 5). These observations demonstrate the value of explicitly simulating trajectories to obtain dynamical information.

Thus, the folding energy landscape for TZ1 clearly does not satisfy the criteria outlined by Zwanzig in Ref. 184 that ought to be satisfied for the folding transition to exhibit simple two-state kinetics. It is true that there is effectively a single well-defined native folded microstate (the global potential energy minimum) and a large number of unfolded microstates such that any one individual unfolded microstate makes only a very small contribution to the partition function for the unfolded macrostate. However, sets of unfolded microstates can be grouped into metastable clusters. The folding landscape therefore violates a vital condition required for the observation of two-state kinetics: relatively high-energy barriers separate metastable unfolded, misfolded, and partially folded states, and so the full ensemble of non-native structures is not in overall local equilibrium. The existence of metastable on-pathway partially folded and off-pathway misfolded states complicates the dynamical behavior of
the peptide by acting as strong kinetic traps, and their effect is clear in the complex form of the FPT distribution (Fig. 5). The effect of some of the features of the folding landscape on the observed dynamics may be exacerbated by finite sampling of the database of stationary points on the potential energy landscape. However, the existence of deviations from a single-funnel energy landscape and the consequent appearance of complex features in TPE statistics contrasting with simple fast-folding behavior expected for single-funnel folding landscapes (Sec. III C) are not a result of sampling error. A two-state kinetic model oversimplifies the folding dynamics of TZ1, even though it is a small peptide that folds rapidly in experiments.\textsuperscript{19,20}

C. Transition path ensemble statistics

The complete FPT distributions for the F $\leftrightarrow$ U transition at temperatures of 300 K and 330 K are shown in Fig. 5. Notably, the FPT distribution for the folding transition of TZ1 does not follow a simple Poissonian form but is instead double-peaked. The complex form of the FPT’s distribution reflects the fact that there exist multiple competing mechanisms that are kinetically relevant. In particular, the existence of two peaks in the FPT distribution, one corresponding to a much longer timescale, suggests that the paths can be broadly classified into two families, as noted in Sec. III B. The first family of paths, which constitute around 70% of the simulated transition paths, correspond to fast “downhill” folding to the native state F. In contrast, members of the second family of paths become trapped in one or more metastable partially folded intermediate states, collectively represented by the macrostate I1. The separation of the TPE into two competing sets of transition paths that each make a substantial contribution to the MFPT, one corresponding to fast downhill transitions and the other corresponding to longer-timescale pathways that become trapped in a metastable intermediate state, has been observed in other simulation studies on the folding of simple peptides\textsuperscript{185-188} and nucleic acid oligomers.\textsuperscript{199,200} Both local maxima of the FPT distribution are shifted to longer timescales with decreasing temperature, and the difference in the MFPTs at the two temperatures is around half an order of magnitude (Table 1). The shorter-timescale peak in the FPT distribution is much sharper at the lower temperature, suggesting that a small number of transition paths in the subensemble of pathways corresponding to the fast-folding mechanism become increasingly dominant with decreasing temperature. Curiously, the longer-timescale peak of the FPT distribution actually becomes slightly broader at the lower temperature, owing to the increased influence of the kinetic traps.

There are a very small number of paths in the tail of the FPT distribution, with $t_{\text{FPT}} \approx 10^{14}$ ns, collectively accounting for around 0.01% of the transition path probability. These paths become trapped in a metastable cluster of nodes corresponding to low-energy misfolded structures. From this kinetic trap, the peptide must largely unfold before transitioning to the native state is possible. It is relatively common for the FPT distributions in realistic kinetic networks to be fat-tailed so that extremal values for the FPT make a non-negligible contribution to the MFPT.\textsuperscript{195,217} For such systems, brute-force simulations are inefficient, since inadequate computational resources are used for representative sampling of the tail region of the FPT distribution, and it may be desirable to employ a trajectory reweighting scheme.\textsuperscript{7}

To characterize the features of the TPE at a microscopic level of detail, we calculate the $A \lforewarrow B$ committor probability $q_{ij}^B$ and the $A \lforewarrow B$ visitation probability $r_{ij}^B$ for the nodes $j$ of the network. The $A \lforewarrow B$ committor function for the $j$th node is defined as the probability that a trajectory initially at node $j$ will reach the absorbing macrostate $A$ before returning to the initial set $B$.\textsuperscript{19,15,194,218} By definition, $q_{i\rightarrow B} = 0$ and $q_{i\rightarrow A} = 1$. The $A \lforewarrow B$ visitation probability for the $j$th node is defined as the conditional probability that a trajectory visits node $j$, given that the trajectory is a direct $A \lforewarrow B$ transition path.\textsuperscript{7} The committor function is an “ideal” one-dimensional reaction coordinate characterizing the progress of the $A \lforewarrow B$ transition and is especially useful for identifying the transition state ensemble (TSE) region, defined by microstates associated with values for the committor function close to 0.5.\textsuperscript{7} The TSE essentially defines the boundary between the effective basins of attraction associated with the endpoint macrostates $A$ and $B$.\textsuperscript{7} The visitation probability is a measure of the extent to which the TPE is localized in the state space and provides a convenient metric for identifying sets of kinetically relevant pathways that are separated in the state space.

The committor and visitation probabilities for nodes $j$, with $r_{ij}^B \geq 0.01$, for the F $\leftrightarrow$ U transition at $T = 330$ K are shown in the form of potential energy disconnectivity graphs\textsuperscript{193,194} in Fig. 6. At a temperature of 330 K, there are 7818 nodes with $r_{ij}^B \geq 0.01$ and 2008 nodes with $r_{ij}^B \geq 0.1$, compared to 68780 nodes in total. Thus, the visitation probability is quite localized in the state space. This is especially true at the lower temperature of 300 K for which there are 4046 nodes with $r_{ij}^B \geq 0.01$ and 1541 nodes with $r_{ij}^B \geq 0.1$. The increased localization of the TPE in pathway space with decreasing temperature has also been observed in kinetic networks for peptide folding transitions constructed from replica exchange MD simulation data.\textsuperscript{185} There are a very small number of microstates of the I2 state for which the values of the visitation probability are close to unity, $r_{ij}^B \approx 1$ [Fig. 6(b)], and hence, there exists a well-defined region of the state space through which the vast majority of folding transition paths are channeled. However, these microstates do not correspond to a dynamical bottleneck (i.e., to the TSE), since they are associated with committor probabilities close to unity, $q_{ij}^B \approx 1$. That is, the transition from these microstates to the native folded state $F$ is largely irreversible. In fact, the folding transition almost always proceeds very rapidly once the I2 state is reached (Fig. 3). Another notable feature of the visitation probability in the state space is the cluster of microstates for which $r_{ij}^B \approx 0.3$, which comprise a subset of the I1 macrostate. This observation is consistent with the simulated time-dependent occupation probabilities (Fig. 3) and the FPT distribution (Fig. 5), which show that around 30% of transition paths become trapped in the I1 macrostate and hence are of a comparatively long timescale. Although these microstates correspond to relatively high-energy partially folded structures, their associated committor probabilities are $q_{ij}^A \approx 1$, and therefore, the peptide is strongly committed to folding at this point.

Inspection of the distribution of committor probabilities demonstrates that the folding transition of TZ1 exhibits multi-state kinetics [Fig. 6(a)]. For an ideal two-state system, the vast majority of microstates are associated with committor probabilities $q_{ij}^A = 0$ or $q_{ij}^A \approx 1$ and can therefore be divided into two well-defined sets. The small number of microstates for which $q_{ij}^A \approx 0.5$ constitute the

Published under license by AIP Publishing

153, 024121-11
FIG. 6. Disconnectivity graph with leaves colored according to (a) $A \leftarrow B$ committor probabilities $q^+$ and (b) $A \leftarrow B$ visitation probabilities $r^+$ for nodes $j$ of the transition network, for the $F \leftarrow U$ transition at a temperature of 330 K. Only nodes of the network for which $r^+_j \geq 0.01$ are included in the tree. The data were obtained from 20,000 transition paths simulated using the kPS algorithm with a priori communities determined by MLR-MCL. The vertical axis corresponds to potential energy with an incremental value of 3 kcal mol$^{-1}$ for the superbasin analysis.

IV. DISCUSSION
A. Features of the methodology

We have implemented two accelerated kinetic Monte Carlo (kMC) algorithms, weighted ensemble (WE-kMC) and kinetic path sampling (kPS), which sample trajectories in exact accordance with the linear master equation [Eq. (1)] that governs the Markovian dynamics on arbitrary transition networks. In particular, we have considered the problem of sampling the nonequilibrium $A \leftrightarrow B$ transition path ensemble (TPE) between two endpoint macrostates of interest, $A$ and $B$. The choice of enhanced sampling kMC methods employed herein is motivated by their desirable complementary features. WE-kMC is highly parallelizable and can be adapted to sample the equilibrium TPE, while the time complexity of kPS is essentially independent of the
metastability of the transition network. Both methods overcome the “flickering” problem that precludes the application of standard kMC to metastable transition networks. They also both require a division of the state space, and their efficiency is affected by the extent to which this partitioning faithfully characterizes the metastable macrostates. The choice of the disjoint sets is therefore a crucial consideration and, indeed, is often the most challenging aspect in the implementation of enhanced sampling methods.

Here, we address this problem by employing a fast and numerically stable stochastic community detection algorithm, namely, multi-level regularized Markov clustering (MLR-MCL), to identify metastable sets of nodes on the transition network. The MLR-MCL algorithm is unsupervised, and therefore, our simulation strategy is fully automated and does not require a low-dimensional projection of the underlying energy landscape, which might obscure features that have a dominant effect on the dynamics.

The $O(N_E^3)$ time complexity for a single kPS iteration to simulate an escape trajectory from a trapping basin $B$, where $N_E$ nodes of the basin are eliminated in the graph transformation stage of the algorithm, leads to problems when a metastable macrostate is naturally large. This issue may be alleviated by allowing for a number of retained transient nodes so that $T \neq \emptyset$, thereby keeping the number of eliminated nodes $N_E$ manageable, and effectively transforming the kPS iteration into a hybrid BKL-kPS scheme. However, depending on the choice of nodes belonging to the set $T$, this approach may reintroduce the problem of flickering trajectories, requiring the explicit simulation of a large number of standard rejection-free kMC steps. Alternatively, the network can be preprocessed by one of a number of methods, although we then forfeit exact sampling of the original transition network. Use of a recursive regrouping scheme to subsume nodes interconnected by fast transition rates, according to a specified threshold, can be highly effective in removing the effects of the groups of nodes that are primarily responsible for the flickering.

Preprocessing of the network by graph transformation, which preserves the path probabilities in their reduced representation, and which introduces renormalized waiting times for nodes to preserve the MFPT from any given node to a set of absorbing nodes, can also be used to avoid any one trapping basin becoming too large. This idea has been used to limit the size of trapping basins when using the FPTA method to solve the master equation for an absorbing Markov chain in Ref. 80 and will be explored further in future work.

B. Comparison to alternative enhanced sampling methods

There are several other exact enhanced sampling methods that are closely related to WE-KMC, in the sense that they employ a division of the state space to simulate complete trajectories between two endpoint macrostates of interest in a piecewise fashion, and maintain a set of “walkers” on the state space that are simulated in parallel. We expect them to perform similarly, but there are some factors that may make a particular enhanced sampling method more favorable for a given system. We conclude by giving a brief overview of popular exact enhanced sampling methods alternative to those implemented in the present work. We highlight relative advantages and disadvantages of the methods and draw attention to how the methods could be adapted and optimized for the problem of sampling the TPE on arbitrary Markovian transition networks, as opposed to the more common problem of performing simulations on a continuous state space. In particular, we suggest how some of these methods may be coupled with the kPS (or, in the same way, the MCAMC) algorithm.

Milestoning utilizes a partitioning of the state space into disjoint sets, each characterized by an “anchor” microstate. Short trajectories are initialized at so-called milestones, which are hypersurfaces at the interfaces between the states, with probabilities reflecting the equilibrium distribution. The flux across a milestone is measured from the first passage time distributions of incident trajectories initialized from neighboring milestones. The MFPTs between all pairs of milestones can be computed from this information. In a transition network, the analog of a hypersurface is a set of boundary nodes separating a pair of communities. The communities could be obtained by any appropriate community detection algorithm.

Since the full trajectory time distribution is an ingredient in the estimation of the coarse-grained MFPT matrix, milestoning provides a natural method for the estimation of reduced non-Markovian networks. Furthermore, milestoning does not depend on a separation of timescales, and therefore, the choice of partitioning of the state space is less crucial, and the method remains effective when applied to highly diffusive dynamical processes. The only restriction in choosing the non-overlapping states when using the approximate form of milestoning is that the trajectories should lose their memory in transitioning between milestones.

Nonequilibrium umbrella sampling (NEUS) and the related tilting algorithm likewise employ a partitioning of the state space into arbitrary non-overlapping sets to achieve distributed sampling and aim to calculate the flux across interfaces. Again, for discrete-state systems, this partitioning can be obtained using MLR-MCL, as in the present work. In these methods, each state is assigned a number of walkers and an initial weight. Every time a walker reaches an interface, an incremental amount of weight is transferred from the state to the neighboring state associated with the interface. The walker is then moved to a new interface of the original state, with probability in proportion to the associated flux. Eventually, the bin weights converge to steady state values, and the flux can be inferred from the number of trajectories crossing individual interfaces per unit time. The tilting algorithm variant of NEUS allows for more rapid convergence to the steady state. Despite its relatively high computational cost, the flexibility in defining the states, and in the distribution of computational resources via specification of the numbers of walkers for each state, makes NEUS a powerful method for sampling TPEs in a steady state.

If it is more natural to divide the state space by nonintersecting interfaces, which for transition networks could be achieved automatically by the repeated application of minimum-cut algorithms, then it may be preferable to employ forward flux sampling (FFS), which can be thought of as a particular case of NEUS. In FFS, trajectory segments are simulated starting from a distribution at the current interface. The trajectory pieces either reach the next interface, in which case the incident points are stored for use in the initial distribution of trajectories starting from this succeeding interface, or return to the initial macrostate of the system. Note that, unlike milestoning, FFS does not employ the equilibrium probability distribution at the interfaces and therefore simulates nonequilibrium, instead of equilibrium, TPEs. Coordinates orthogonal to the
reaction coordinate that defines the nesting of interfaces must effectively be sampled by brute force. Therefore, FFS is most useful for simulating rare event systems that can be projected onto a single dimension without significant loss of information, in which case the comparably small computational overhead of FFS makes the method attractive.\textsuperscript{27} The treatment of the successive trajectory pieces in a serial fashion leads to a propagation of errors,\textsuperscript{26} and significant computational effort is expended simulating trajectories that do not reach the next interface but instead return to the initial macrostate, especially if there are intermediate states acting as strong kinetic traps.

Enhanced sampling methods based on the parallel simulation of many trajectory segments share many of the same shortcomings. In particular, the correlated histories of trajectories necessitate that rigorous statistical tests be employed to evaluate the quality of the simulation data.\textsuperscript{13} In continuous space, a relative advantage of WE sampling is that the set of trajectories is examined at regular time intervals, whereas the other methods require that a trajectory be caught in the act of crossing an interface. This requirement may be computationally expensive in high-dimensional continuous-state spaces, but checking of trajectories is trivial when the state space is discrete, since the communities to which nodes belong are identified simply by labels, with no coordinate information required.

In milestoning and FFS, trajectory segments are simulated from one hypersurface (in discrete state space, a set of boundary nodes) to another, and unlike NEUS (or WE-kMC), there is not continual “feedback” between adjacent hypersurfaces, although FFS must be carried out in a serial fashion. This feature means that milestone-and FFS are well-suited for use in conjunction with kPS (or MCAMC) using only modest computational resources, since the calculation can be run by focusing on individual communities in turn. Therefore, the graph transformation stage of the kPS simulation\textsuperscript{149,150} and the spectral decomposition of the transition rate matrix in the MCAMC algorithm,\textsuperscript{95,98,101} which are the computational bottlenecks of the respective methods, need only be carried out once for each community. After storing the relevant information to undo the graph transformation in kPS, the iterative reverse randomization procedure\textsuperscript{149,150} can be repeated to generate the desired number of sample trajectory segments within the community. Similarly, the eigenspectrum of a community can be used to repeatedly generate sample trajectory segments within a MCAMC simulation. Since kPS and the FPTA\textsuperscript{17–19,95,139} variant of the MCAMC algorithm correctly preserve the FPT distribution, and sample nodes at the absorbing boundary of the currently occupied community with the exactly correct probabilities, the milestoneing and FFS methods used in conjunction with kPS or FPTA will yield unbiased estimates for MFPTs. These hybrid methods will also yield an unbiased sample of the TPE, albeit with reduced resolution of the pathways, since information on the dynamics within communities is lost. This loss is fairly inconsequential, since the communities ought to reflect the metastable states within which the trajectories flicker unproductively.

An alternative approach to those described above is provided by transition path sampling (TPS),\textsuperscript{19} where an initial transition path is repeatedly perturbed, with the aim of sufficiently sampling the critical dynamical bottleneck region. In previous work, TPS has been reformulated for discrete-state, continuous-time stochastic dynamics.\textsuperscript{260–262} TPS is a procedure for Monte Carlo sampling in pathway space, which is usually used to simulate the equilibrium TPE. TPS does not depend on a division of the state space but instead requires an initial $A \leftrightarrow B$ trajectory. The efficiency depends strongly on the design of moves to propose successive paths, by propagating the dynamics from a selected point along an existing trajectory, which lead to appreciable acceptance probabilities. In the context of TPS simulation on a discrete state space, representative $A \leftrightarrow B$ trajectories can be determined efficiently using $k$ shortest paths algorithms,\textsuperscript{14} and trajectories could be propagated efficiently by the kPS algorithm, which yields the exact path probability of the simulated trajectory as a by-product of the iterative reverse randomization procedure. In perturbing the current transition path, it is desirable to select a microstate for which the committor probability is close to $0.5$,\textsuperscript{110} since this transition state ensemble (TSE) region has a dominant effect on the dynamical behavior.\textsuperscript{11} In a transition network, the TSE is a cut set of edges that can be closely approximated by fast stochastic minimum-cut algorithms and by the “$k$ distinct paths” shortest paths algorithm.\textsuperscript{14} TPS may be unable to achieve ergodic sampling for systems where the visitation probability is delocalized in separated regions of pathway space.\textsuperscript{15} Again, for transition networks, this problem is circumvented with relative ease, simply by running many TPS simulations initialized from different $A \leftrightarrow B$ trajectories determined by the $k$ distinct paths algorithm.\textsuperscript{14}

C. Conclusions

The accelerated kMC methods employed in this work, weighted ensemble\textsuperscript{95,134–137,139} kMC (WE-kMC) and kinetic path sampling\textsuperscript{19,150} (kPS), allow for a detailed quantitative analysis of the $A \leftrightarrow B$ TPEs on arbitrary Markovian transition networks and remain efficient even for networks that are strongly metastable and of high dimensionality. We have demonstrated our simulation strategy, in which the MLR-MCL\textsuperscript{152–155} unsupervised community detection algorithm is used to define a priori fixed bins in the accelerated kMC simulations, for a transition network representing the folding of the TZ1 peptide\textsuperscript{159} constructed by discrete path sampling.\textsuperscript{14,15} The folding transition for TZ1 exhibits complex multi-state and multi-pathway kinetics, and simulation of the folding transition by brute-force kMC is unfeasible. The choice of partitioning of the state space is a crucial consideration that strongly affects the efficiency of enhanced sampling algorithms,\textsuperscript{17–19,95,139} and any appropriate community detection\textsuperscript{41} algorithm could be used to obtain this division in a transition network. In the future, we will discuss the choice of community detection algorithm for the purpose of guiding accelerated kMC simulations on high-dimensional transition networks in detail.\textsuperscript{158} We have also developed a framework for the dimensionality reduction of transition networks based on kPS simulations, which we will report elsewhere.

In future work, we intend to implement and benchmark alternative enhanced sampling methods, such as those described in Sec. IV B, for accelerating kMC simulations on arbitrary transition networks. In particular, it is desirable to have access to methods that are well-suited to sampling the equilibrium TPE, a problem to which the protocol for establishing a steady state in WE sampling\textsuperscript{149,150} does not provide an ideal solution. Hybrid algorithms combining kPS or FPTA\textsuperscript{17–19,95,139} with milestoneing\textsuperscript{38} and with forward flux sampling (FFS)\textsuperscript{17,93,118,177–181} are particularly promising approaches
for efficient and exact sampling of equilibrium and nonequilibrium TPEs, respectively. We have developed the DISCOTRESS (Discrete State Continuous Time Random Event Simulation Suite) software to perform enhanced sampling simulations on arbitrary CTMCs. The simulation methods described in the present work, including kPS, can be adapted to discrete-time Markov chains (DTMCs). Application of these advanced KMC methods to a variety of systems will yield insight into how features of the TPE, such as the existence of multiple competing mechanisms, arise from the topology of the transition network. The identification of archetypal classes of stochastic dynamics and the corresponding network topologies will provide fundamental understanding concerning how dynamical observables such as the MFPT arise from microscopic features of the TPE, and therefore of how these macroscopic dynamical properties are encoded in the underlying energy landscape.

SUPPLEMENTARY MATERIAL

Our C++ implementations of the weighted ensemble kinetic Monte Carlo (W-kMC) and kinetic path sampling (kPS) algorithms are available online at https://github.com/danielsharpe/DISCOTRESS under the GNU General Public License. These functions are part of DISCOTRESS, the Discrete State Continuous Time Random Event Simulation Suite, a highly generalized program for simulating the stochastic dynamics on arbitrary Markov chains. A C++ implementation of the multi-level regularized Markov clustering (MLR-MCL) algorithm for community detection in arbitrary transition networks is also publically available online at https://github.com/danielsharpe. Parameters for the simulation setups are given in the supplementary material.

ACKNOWLEDGMENT

D.J.S. gratefully acknowledges the Cambridge Commonwealth, European and International Trust for a Ph.D. scholarship. D.J.W. gratefully acknowledges support from the EPSRC. We thank Dr. Jerelle A. Joseph for providing the stationary point database for the tryptophan zipper folding transition. We also thank Deepi Kan-

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

4. D. T. Gillespie, Markov Processes: An Introduction for Physical Scientists (Aca-