Identifying mechanistically distinct pathways in kinetic transition networks

Daniel J. Sharpe[1] and David J. Wales[1, a)]
Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

(Dated: 24 May 2019)

We present an implementation of a scalable path deviation algorithm to find the k most kinetically relevant paths in a transition network, where each path is distinguished on the basis of having a distinct rate-limiting edge. The potential of the algorithm to identify distinct pathways that exist in separate regions of the configuration space is demonstrated for two benchmark systems with double-funnel energy landscapes, namely a model 'three-hole' network embedded on a 2D potential energy surface, and the cluster of 38 Lennard-Jones atoms (LJ_{38}). The path cost profiles for the interbasin transitions of the two systems reflect the contrasting nature of the landscapes. There are multiple well-defined pathway ensembles for the three-hole system, whereas the transition in LJ_{38} effectively involves a single ensemble of pathways *via* disordered structures. A by-product of the algorithm is a set of edges that constitute a cut of the network, which is related to the discrete analogue of a transition dividing surface. The algorithm ought to be useful for determining the existence, or otherwise, of competing mechanisms in large stochastic network models of dynamical processes, and for assessing the kinetic relevance of distinguishable ensembles of pathways. This capability will provide insight into conformational transitions in biomolecules and other complex slow processes.

Keywords: energy landscape, kinetic transition network, discrete path sampling, stochastic network model, master equation, reaction paths, graph theory, shortest paths

I. INTRODUCTION

Kinetic transition networks constructed by discrete path sampling[1–4] (DPS) or alternative[5–13] methods provide a powerful framework for modeling many physical systems. The DPS methodology, where the N-dimensional potential energy function $V(\mathbf{r}^N)$ is mapped to a network of V nodes and E weighted and bidirectional edges by determination of the transition states connecting pairs of minima,[14–18] avoids explicit simulation of the dynamics. The framework is therefore particularly useful for modeling systems that feature broken ergodicity.[19] Kinetic transition networks determined by DPS are an attractive coarse-grained representation of the energy landscape, since they preserve the high dimensionality of the surface.[20] The DPS framework has provided insight into the thermodynamics and dynamics of many systems, including atomic and molecular clusters,[21–24] biomolecules,[4,25] and glasses.[26]

A simple way to gain mechanistic and kinetic information from a transition network is to use Dijkstra's algorithm[27,28] to determine the shortest path between any pair of nodes that belong to two defined states A and B, $a \in A$ and $b \in B$, and calculate the contribution of this discrete path to the steady state rate constant.[28] However, in many physical systems there are liable to be a multitude of kinetically relevant pathways,[29] and competing paths may exist in separate regions of the network. A particularly pertinent question, and one that is the subject of current debate,[30–32] is the longstanding problem of understanding whether protein folding takes place *via* multiple parallel pathways, or else by a single and well-defined dominant pathway. There is therefore a need for algorithms that are capable of identifying mechanistically distinct pathways covering the whole of a transition network.

The problem of finding the k shortest paths between source and sink nodes in a network is a fundamental problem in computer science, and many algorithms to solve the problem are known.[33,34] Which algorithm provides the optimal solution depends on the properties of the network being analysed, such as its sparsity. Kinetic transition networks from DPS have weighted and bidirectional edges, and are normally sparse, due to both the connectivity of minima on a potential energy surface being relatively low, even in high dimensions,[35] and also due to inexhaustive sampling. The networks typically range in size from tens to hundreds of thousands of nodes. The less general problem of determining the k shortest *loopless* paths between source and sink nodes in a network[36,37] is a significantly harder one.[38] Some classical algorithms for this problem scale poorly, such as the algorithm of Yen,[39] which has time complexity $\mathcal{O}(kV(E + V \log V))$.

In previous work, the recursive enumeration algorithm (REA) of Jiménez and Marzal[40] was implemented to solve the general k shortest paths problem for kinetic transition networks.[41] As for most general k shortest path algorithms, the required input to the REA is an initial shortest path tree. The REA can then find the set of k shortest paths efficiently, specifically in $\mathcal{O}(E + kV \log(E/V))$ time. If the edge weights of the transition network are chosen appropriately (Section II.A), then these are the paths that give the k largest contributions to the steady state rate constant for the transition between the sets of endpoint nodes A and B. However, since

a)Electronic mail: dw34@cam.ac.uk

the REA allows loops, the discrete paths may differ only trivially from one another. In particular, typical reactive trajectories on a transition network are liable to exhibit 'flickering' between nodes for which the transition rates are fast compared to the rare barrier-crossing events between metastable states.[42–44] Thus, practically, one finds that in order to explore pathways existing in separate regions of a transition network, and therefore obtain a complete picture of the possible mechanisms for interstate transitions, one must run the algorithm for a large number of paths k.[45] The same problem affects other fast algorithms for the general k shortest paths problem, such as the algorithms of Martins and Santos,[34,46–49] (time complexity $\mathcal{O}(kV \log V)$), Eppstein[50,51] (time complexity $\mathcal{O}(E + V + k \log k)$) and Azevedo *et al.*[52,53] (time complexity $\mathcal{O}(kE)$). No known algorithms for the k shortest loopless paths problem achieve such favourable asymptotic time complexity. In any case, high-performance algorithms for this less general problem, such as the algorithm of Martins, Pascoal and Santos,[34,37,38,54] may still return successive paths that differ only trivially from one another, *i.e.* by minor variations away from the dynamical bottleneck region. In the context of transition networks, such pathways belong to the same ensemble of trajectories.[29]

In the present work, we adopt an alternative algorithm that is more appropriate for identifying mechanistically distinct paths in a transition network. Hence we can harvest representative pathways to assess the kinetic relevance of, and competition between, distinguishable processes. This capability will allow detailed insight into the slow dynamics of complex systems, from the perspective of the underlying energy landscape. Specifically, we utilise the algorithm of Frigioni, Marchetti-Spaccamela and Nanni for the dynamic updating of shortest path trees[55] to find the k shortest loopless paths in a transition network. Following determination of the initial shortest path tree, the algorithm runs in worst case time $\mathcal{O}(k\sqrt{E} \log V)$ and memory costs scale linearly. Path deviation algorithms for solving the k shortest paths problem allow one to select a criterion for distinguishing successive paths, which we here choose to be that each path has a distinct rate-limiting edge. We subsequently refer to the algorithm as the "k distinct paths" (kDP) algorithm. We show that the use of this algorithm also allows approximate characterisation of the transition state ensembles for interstate transitions, by determination of the complete set of rate-limiting edges, which induces an A-B cut in the network. The dynamical bottleneck region of the energy landscape has a dominant effect on the nature of slow transitions between metastable states,[56,57] and obtaining an accurate description of this region presents a challenging problem.[58,59]

The algorithm described in Ref. 55 for the dynamic updating of shortest path trees was first applied to analyse transition networks by Noé *et al.*[60,61] for a set of random networks, and for conformational switching in polypeptides modeled with a coarse-grained potential. The transition networks were constructed from a mapping of the stationary points on the potential energy landscape. Due to computational limitations, the number of minima and transition states in the networks were relatively small. Initially the transition state energies were bracketed by upper and lower bounds,[61,62] only being determined absolutely if the corresponding edges appeared in either of the shortest paths given by Dijkstra's algorithm based on the upper or lower bounds. Hence the true Dijkstra shortest path was not necessarily located. Furthermore, the minima were determined by local minimisation after uniform sampling, and edge connectivity was simply inferred. This methodology is significantly different from the DPS framework employed here. A further crucial difference between our approach and that of the previous studies is in the definition of the edge weights. In the original work, the edges were weighted according to the inverse Boltzmann factors of the transition states, neglecting dynamical prefactors. Determining the shortest path in this way, with the total cost of a path being a sum over transition states, is then conceptually similar to finding the path of 'maximum flux' on the continuous energy surface.[63] This definition of the edge weights is limited since the weights do not directly relate to transition rates, and the bidirectional nature of the edges of a transition network is neglected. In the present work, the edge weights are chosen so that their sum along a discrete path is related to the negative logarithm of the path contribution to the overall interstate rate constant under a steady state approximation,[28] as described in Section II.A. The practical limitations of the methodology described in Refs. 60 and 61 allowed for calculation of only a small number of the most kinetically relevant paths. The transition state ensemble was characterised by an alternative greedy method. The implementation used in the present contribution, where all edge weights are known, is found to run to completion rapidly, even for large transition networks. Indeed, for the benchmark systems considered herein, the bottleneck of the computation is the determination of the initial shortest path tree by Dijkstra's algorithm.

We illustrate our results with two benchmark kinetic transition networks, for a roughened model three-hole potential with a Poissonian node degree distribution, and for the cluster of 38 Lennard-Jones atoms (LJ_{38}).[21–24,64,65] The three-hole potential has been studied as a standard test system for methods to determine reaction paths[63,66] and for transition path theory.[67,68] The LJ_{38} system has been investigated extensively in many theoretical studies relating to flows on stochastic networks,[69–72] as well as in simulation studies, for example by parallel tempering,[73,74] direct transition current sampling,[75] Monte Carlo[76] and other[5,77,78] methods. Both systems can be described in terms of double-funnel energy landscapes.[5,65]

II. METHODOLOGY

A. The energy landscape framework

Kinetic transition networks are constructed by the discrete path sampling (DPS) method as described below. Firstly, low-energy minima on the landscape, including the global minimum, are located by basin-hopping.[64,79,80] A discrete path (minimum-transition state-minimum sequence) connecting two chosen endpoint minima is then determined as follows. A doubly-nudged[81,82] elastic band[83–85] (DNEB) calculation is performed, and images along the minimum energy path that are local maxima are selected as transition state candidates. An attempt is made to converge each transition state candidate tightly by hybrid eigenvector-following[86–88] (HEF). The pair of minima connected by each transition state are located by limited-memory Broyden-Fletcher-Goldfarb-Shanno[89,90] (L-BFGS) local minimisation, following small displacements along directions parallel and antiparallel to the transition vector of the saddle point. Double-ended searches are continued until a complete connected path is found, with a Dijkstra-based missing connection algorithm used to prioritise pairwise connection attempts.[91] The minima and transition states are mapped to the nodes and edges of a network, respectively. Given a network constructed from one or more initial paths, further sampling is achieved using many parallel interpolation calculations. The priority values of connection attempts for pairs of endpoint minima are based on one of a number of distance and barrier metrics.[92,93]

In the Dijkstra algorithm and in the k distinct paths algorithm described herein, the edge weights $M_{\alpha\beta}$ of the transition network representing the $A \leftarrow B$ interstate transition are chosen to be the negative logarithms of branching probabilities

$$M_{\alpha\beta} = -\ln P_{\alpha\beta} = -\ln \frac{k_{\alpha\beta}}{\sum_\gamma k_{\gamma\beta}} \quad \forall\ \alpha, \beta. \tag{1}$$

Here, $P_{\alpha\beta}$ and $k_{\alpha\beta}$ are the branching probability and the transition rate for the $\alpha \leftarrow \beta$ internode transition, respectively, and the sum is over all neighbouring nodes γ. In the context of transition networks constructed by DPS, the elements $k_{\alpha\beta}$ are the $\alpha \leftarrow \beta$ minimum-to-minimum rate constants. The contribution of a single discrete path $a \leftarrow b$, connecting endpoint nodes $a \in A$ and $b \in B$, to the overall steady state rate constant for the $A \leftarrow B$ transition is a product of branching probabilities for all nodes along the path, weighted by the inverse of the waiting time $\tau_b = 1/\sum_\gamma k_{\gamma b}$ for node b, and by the ratio π_b/π_B. Here, π denotes an equilibrium occupation probability, and $\pi_B = \sum_{b\in B} \pi_b$. The overall steady state rate constant k_{AB}^{SS} is equal to a sum of individual contributions $k_{a\leftarrow b}^{\mathrm{SS}}$ from all possible $a \leftarrow b$ paths

$$k_{AB}^{\mathrm{SS}} = \frac{1}{\pi_B} \sum_{a\leftarrow b} P_{ai_1} P_{i_1 i_2} \ldots P_{i_n b} \tau_b^{-1} \pi_b, \tag{2}$$

or, equivalently,

$$k_{AB}^{\mathrm{SS}} = \frac{1}{\pi_B} \sum_{b\in B} \frac{q_b^+ \pi_b}{\tau_b}. \tag{3}$$

Here, q_b^+ is the forward committor probability for node b,[56] *i.e.* the probability that the system, initially at node b, will visit the state A before returning to B. i is used to denote intermediate nodes, $i \notin A \cup B$. The Dijkstra and k shortest paths algorithms require positive edge weights. If the edge weights are given by Eq. 1, and the set B contains a single node b, then by comparison with Eq. 2, the shortest $a \leftarrow b$ path in the network, where the total path cost is a sum over edge weights, is that for which $-\ln k_{a\leftarrow b}^{\mathrm{SS}}$ is minimal. That is, the shortest discrete path in the network is that which makes the maximum contribution to the $A \leftarrow B$ steady state rate constant. Furthermore, the relative contribution of two paths to the overall steady state rate constant is given directly by the ratio of exponentials of path costs. If the set B contains multiple nodes, then the additional factors of $\tau_b^{-1}\pi_b$ must be accounted for. Estimates for the minimum-to-minimum rate constants $k_{\alpha\beta}$ are obtained here from harmonic transition state theory.[94] However, any appropriate unimolecular rate theory could be used, including methods based on explicit dynamics.

Transition networks can be visualised as disconnectivity graphs,[65,95,96] which preserve the full dimensionality of the system, and therefore provide a faithful representation of the barriers and basins on the corresponding energy landscape. To construct a disconnectivity graph, all nodes are initially considered to belong to the same set, which is then cut at incremental threshold energies. The cuts partition groups of nodes into disjoint sets, termed superbasins. Nodes in the same superbasin at a given threshold energy are mutually accessible, whereas a transition between different superbasins must proceed *via* an edge with energy exceeding the threshold. The leaves of the graph terminate at the energies of the corresponding nodes.

B. Finding distinct pathways

We briefly describe the algorithm of Frigioni, Marchetti-Spaccamela and Nanni[55] in the form employed here. That is, to iteratively find k pathways that are distinct, in the sense that they are disjoint with respect to the identity of their rate-limiting edges. This statement provides an intuitive and physical working definition for what constitutes separate pathways that can be considered to differ non-trivially. The pseudocode for this "k distinct paths" (kDP) algorithm is given as Algorithm 1. Full details of the strategy for updating shortest path trees, and proof of correctness, can be found in Ref. 55. A step-by-step illustration of a single iteration of the algorithm for a toy network is shown in Figs. S1-S10 of the Supplementary Information. The main loop of Algorithm 1 traces the shortest path tree, determines and

blocks the rate-limiting edge (u, v), directed $u \leftarrow v$, of the shortest path for the current iteration, and marks u and all of the descendants of u in the transition network as 'red', where u is a child of v in the tree. The remainder of the operations in the main loop constitute two sequential inner loops. The first inner loop of the algorithm iterates over all red nodes, z, searching for the best alternative route to z *via* a neighbour t of z that is not red. If such a node exists, then z is added to a queue with priority equal to the cost of the new route, and t is set as the parent of z. The second inner loop of the algorithm iterates over the red nodes z that have been queued, searching for alternative paths to neighbouring red nodes h *via* z that improve the cost of the path to h. If a shorter path to h exists, then the tree structure is updated accordingly, and h is queued with a new priority value. In the present work, the input shortest path tree that contains the first shortest path between two defined states A and B is given by Dijkstra's algorithm.

In Algorithm 1, the user defines a single node $b \in B$ that is the source node of the shortest path tree at every iteration, along with a set of sink nodes $\{a\} \in A$. If the set A has more than one member, then the shortest $A \leftarrow B$ path at the current iteration is that with the lowest cost considering all nodes in A. Contributions of individual paths to the steady state rate constant are directly related to the path costs, by taking explicit account of the waiting time and occupation probability for the node $b \in B$ (Eq. 2). The definition of multiple sink nodes is especially useful if the rate-limiting edges of determined pathways tend to occur late along the pathway in the $A \leftarrow B$ direction, such that a cut separating the source and sink nodes is rapidly induced in the network as edges are blocked.

A second important consideration is deciding the criterion for what constitutes a rate-limiting edge. Three different definitions for the rate-limiting edge are considered; based on maximum edge weight, and based on the energy barrier height or absolute transition state energy associated with the edge. The first two definitions are conceptually similar, but the former accounts for the branching probability, as well as for the explicit form of the transition rates $k_{\alpha\beta}$. Within the DPS framework, the expression for the minimum-to-minimum rate constants $k_{\alpha\beta}$ depends on the harmonic vibrational frequency of the transition state. Blocking the edge corresponding to the highest absolute transition state energy along a path is the most useful way to explore mechanistically distinct pathways if the sets of endpoint nodes are low-energy states separated by a high energy barrier. The set of rate-limiting edges then characterises the transition dividing surface separating the two basins corresponding to states A and B,[56] as discussed below. Blocking the edge corresponding to the largest edge weight or barrier height provides a more appropriate definition of paths that can be considered to be kinetically distinct.

A by-product of Algorithm 1 is a set C of rate-limiting edges. The distribution of transition state energies corresponding to edges in the set C gives an idea of the landscape entropy contribution to A-B pathways, and inspection of the configurations of the transition states allows the identification of distinct transition state ensembles in the configuration space. The local densities of states that appear in the transition state theory expression for the transition rates within the DPS methodology account for local vibrational entropy. If there is a single collection of dominant (*i.e.* low-energy) transition states corresponding to the edges in C, then there is little competition from alternative pathways. The landscape entropy contribution of pathways to the overall interstate rate constant is then small, and the dynamical bottleneck of the transition is well-defined.[61] Conversely, if there is more than one collection of transition states in the set C with similar energies, which can be identified as belonging to distinct regions in configuration space, then the reverse is true. The separation of transition state and pathway ensembles may be quantified by order parameters, or else by analysing the corresponding pathways for the similarity of configurational changes along a particular pair of pathways,[97] or for the cost of transforming one path into another.

If the algorithm runs to completion, then the edges in the set C constitute a *cut* in the network that partitions the A and B sets into two disconnected components. The A-B cut in the transition network induced by the cut set C is related to the concept of a transition dividing surface,[98,99] in a discretised space. A transition dividing surface is a deformed plane in $N-1$ dimensions that partitions the N-dimensional potential energy surface into reactant and product states, such that the flux between reactants and products is maximal. However, the edge weights of the kDP algorithm (Eq. 1) do not constitute a flow, and C is not a minimum cut, and thus does not represent maximum 'flow' between the A and B sets. Nonetheless, the set C conveys similar information to the *rate-limiting cut* discussed by Noé *et al.* in Ref. 61. In the original work of Noé *et al.*, practical limitations prohibited the determination of more than a small number of distinct paths, and hence of the complete cut set. The notion of a 'topographical ridge cut', an approximation to the transition dividing surface determined by a greedy method, was introduced as an alternative. The present formulation enabled us to rapidly determine the complete cut set, even for the largest transition networks considered.

The i-th distinct path is the member of the infinitely large set of discrete paths associated with the i-th rate-limiting edge that makes the largest contribution to the steady state rate constant k_{AB}^{SS}. For a typical reactive path, there are liable to be a number of unproductive loops[71] and, in particular, a large number of recrossings of the transition dividing surface.[98] Therefore the relative contributions of two individual distinct paths to k_{AB}^{SS}, determined *via* the ratio of path costs (*cf.* Eqs. 1 and 2), may not necessarily be representative of the relative contributions of the corresponding rate-limiting edges or sets

of paths to the overall reactive A-B flux. To assess the contributions of the rate-limiting edges contained in the cut set C to the reactive flux, we could calculate the steady state rate constant for the residual network at each iteration of the kDP algorithm, using the graph transformation (GT) method.[42,100,101] The GT method avoids evaluating the infinite sum over pathways in Eq. 2, and instead iteratively removes nodes from the network in a deterministic manner, and preserves the $A \leftarrow B$ mean first passage time by renormalisation of the branching probabilities and waiting times of nodes. We here denote the $A \leftarrow B$ steady state rate constant calculated using the GT method by k_{AB}^{NGT}, after Ref. 42. Significant decreases in k_{AB}^{NGT} as specific edges are blocked would suggest that these edges are well-defined reaction bottlenecks. In contrast, if k_{AB}^{NGT} decreases steadily as successive rate-limiting edges are blocked, then this would suggest that the set of reactive pathways for the $A \leftarrow B$ transition has significant entropic character. When all edges of the A-B cut set C are blocked, then $k_{AB}^{\mathrm{NGT}} = 0$. This analysis is possible for small networks, but it is not feasible to repeat the GT calculation multiple times for large networks. An alternative and more scalable analysis is to calculate k_{AB}^{NGT} at each iteration i of the kDP algorithm in a cumulative manner. That is, to calculate k_{AB}^{NGT} for the network formed of the nodes and edges that appear along the i distinct paths determined by the kDP algorithm thus far. If the reaction bottleneck is well-defined, then k_{AB}^{NGT} for the accumulated network will rapidly converge to a value close to the value calculated for the complete network when a small number of specific nodes and edges are included. If, however, the transition path ensemble has significant entropic character, then k_{AB}^{NGT} will increase steadily as successive distinct paths are added to the network.

C. Constructing model kinetic transition networks

Low-dimensional potential energy surfaces, where the number of distinguishable configurational ensembles is apparent, and the characterisation of dynamical bottlenecks and pathway ensembles for interstate transitions is tractable, can provide useful benchmarks. We can embed a kinetic transition network onto a simple potential by assigning the location of stationary points. This procedure essentially corresponds to a roughening of the potential by superimposing small barriers on the smooth energy surface. Furthermore, an artificial kinetic transition network ought to have an 'ideal' node degree distribution that follows a given statistical law commonly observed for real-world networks. We present an algorithm for constructing such networks randomly, given as pseudocode in Algorithm 2. Similar to the procedure described in Ref. 60, the algorithm iteratively updates the positions of nodes selected at random, one at a time, and also periodically updates the Euclidean cutoff distance d, below which two nodes are considered to be connected by an edge. Updates are accepted if they lead to a decrease in the error between the observed node degree distribution and the target distribution specified by the user. The update loop terminates if the maximum number of iterations n_{it} is reached, or if the error value ϵ_{obs} associated with the degree distribution decreases below the tolerance ϵ_{tol}. Then the energies E_u of nodes u are assigned according to a potential energy function provided by the user, and energies E_{uv} of transition states corresponding to the edges (u, v) are assigned according to a specified mean barrier height, with noise incorporated into both. The target node degree distribution can be Poissonian, power-law or Gaussian, all of which have been reported for complex networks covering a variety of real-world systems.[102]

III. RESULTS

A. Model network

A model kinetic transition network was constructed according to Algorithm 2, with the node degree distribution fitted to a Poisson form with $\lambda = 8$, and node energies assigned according to a shifted two-dimensional three-hole potential[67,68] $V(x, y)$ with domain $x, y \in [-2, 2]$. Other parameters of the algorithm were $n_{\mathrm{V}} = 1000$, $d = 0.15$, $\sigma_{\mathrm{E}} = 0.04$, $\mu_b = 0.2$ and $\epsilon_{\mathrm{tol}} = 0.05$. The final network consisted of 998 nodes and 3981 bidirectional edges. The elements of the weighted adjacency matrix (Eq. 1) were calculated according to a reduced temperature of $T = 0.6$. The disconnectivity graph[65,95,96] for the resulting network is shown in Fig. 1a, and exhibits a clear double-funnel topology, with a third, more shallow funnel corresponding to values of x close to zero. The 2D potential energy surface on which the network is embedded is shown in Fig. 2a.

The kDP algorithm found 150 distinct paths in the kinetic transition network before the pair of endpoint nodes, chosen as the lowest-energy nodes of each of the two major funnels, became disconnected. The $A \leftarrow B$ transition is in the direction of increasing x (see Figs. 1 and 2). The criterion for determining the rate-limiting edge of a given discrete path was based on the corresponding transition state energies. The 150 distinct paths of the network, and the 150 edges that are the rate-limiting edges of each of the distinct paths, are shown on the 2D potential energy surface in Figs. 2b and 2c, respectively. The energy profiles and the costs of each of the 150 distinct paths are shown in Figs. 3a and 3b, respectively.

Immediately one can identify three ensembles of trajectories on the underlying potential energy surface (Fig. 2b). This result illustrates the ability of the kDP algorithm to explore distinct regions of a transition network, and hence separated regions of the underlying configuration space. The 15 shortest distinct paths all follow a similar route in the 2D space, with y initially increas-

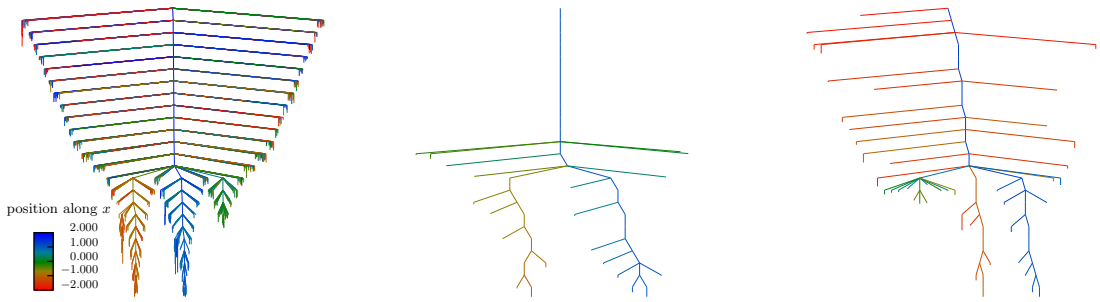
ing, starting from either endpoint minimum, but not entering the third, more shallow basin centered at $x = 0$. The next, approximately 60, distinct paths are mostly of a second distinguishable pathway ensemble, with y initially decreasing, although some paths also belong to the first ensemble. The energies of the transition states corresponding to the rate-limiting edges of the paths comprising this second ensemble are only slightly greater than for the first ensemble, and the paths are slightly longer. The profiles of path costs and of the steady state rate constant for the residual network at each iteration of the kDP algorithm (Fig. 3b) show that, although the first pathway ensemble makes a dominant contribution to the rate constant, the second ensemble is competitive, at the relatively high reduced temperature of $T = 0.6$. Interestingly, the second ensemble of pathways, *i.e.* those proceeding *via* the lower channel, becomes dominant at low reduced temperatures ($T \approx 0.05$), an effect which has previously been termed 'entropic switching'.[67] Approximately 75 more distinct paths are determined before an A-B cut is induced in the network and the kDP algorithm terminates. These distinct paths mostly correspond to x initially decreasing and y increasing, starting from the B endpoint (with $x \approx -1$), and the transition states associated with the the rate-limiting edges are located at $x \approx -1.8$, $y \approx 0.8$. This third ensemble of pathways has highest-energy transition states with significantly greater energies than for the first two ensembles, and the total path costs are much greater. Analysis of the steady state rate constant for the residual network at each iteration of the kDP algorithm demonstrates that this third ensemble of pathways makes a negligible contribution to the rate constant at $T = 0.6$ (Fig. 3b).

There are three distinguishable transition state ensembles in the 2D space, corresponding to each of the three principal pathway ensembles described (Fig. 2c). The existence of three distinguishable mechanisms is also apparent from the energy profiles of the 150 distinct paths, shown in Fig. 3a. The numbers of stationary points along the distinct paths of the first pathway ensemble are small, and the highest-energy transition states have relatively low potential energies, $V \approx -1$. The distinct paths of the second ensemble include a slightly greater number of stationary points than those of the first, and the maximum energy along the pathways varies in the range $V \approx -1$ to $V \approx 2.5$. The distinct paths of the third ensemble have a high maximum energy, $V \approx 2.5$, and are much longer than the paths of the first and second ensembles, proceeding *via* formation of a metastable intermediate before traversing the high barrier. Visualisation of the disconnectivity graphs including only stationary points found along one discrete path, or a set of related paths, provides a convenient means to compare pathway ensembles, without requiring a low-dimensional projection of the network. From the disconnectivity graphs for the networks composed of stationary points along the first (Fig. 1b) and 150$^{\text{th}}$ (Fig. 1c) distinct paths, the differences between the first and third pathway ensembles are clear, including the fact that pathways of the third ensemble enter the shallow basin at $x \approx 0$. The REA,[40,41] for the general k shortest paths problem, is very inefficient in exploring the multiple pathway ensembles (see the Supplementary Information). The 50000 shortest paths determined by the REA roughly correspond to the first 5 distinct paths determined by the kDP algorithm, even though analysis of the steady state rate constant for the residual network suggests that approximately 100 shortest distinct paths are kinetically relevant (Fig. 3b).

There is evident correlation between the profiles of path costs and of steady state rate constants for the residual (Fig. 3b) and accumulated (Fig. S12b of the Supplementary Information) networks with increasing number of distinct paths, at a reduced temperature of $T = 0.6$. The network formed of the stationary points present within the set of the 100 shortest distinct paths, which comprises 182 nodes and 282 bidirectional edges, captures almost half of the rate constant of the complete network ($k_{AB}^{\text{NGT}} = 1.48 \times 10^{-3}$ compared to $k_{AB}^{\text{NGT}} = 3.83 \times 10^{-3}$, respectively). Taken together, these observations suggest that, although for typical reactive A-B trajectories there are significant fluctuations from the set of distinct paths at a reduced temperature of $T = 0.6$, the complete set of distinct paths appropriately characterises all possible reactive paths on the network. That is, although loops and deviations from the set of distinct paths make non-negligible contributions to the reactive flux, this mechanistic information is essentially redundant, and hence the relative costs of the distinct paths reflect the kinetic relevance of the corresponding rate-limiting edges and sets of paths. It is in this sense that the complete set of distinct paths is representative of all reactive paths on the network, including those that are rare. The landscape entropy contribution to the steady state rate constant at a reduced temperature of $T = 0.6$ is large. This effect is reflected in the steady decrease of the rate constant for the residual network, and in the steady increase in the rate constant for the accumulated network, with increasing number of distinct paths. Hence there is no well-defined reaction bottleneck. The multiple jumps in these profiles, and in the profile of path costs, indicates that there are multiple pathway ensembles that exist in separate regions of the pathway space, separated by high energy barriers.

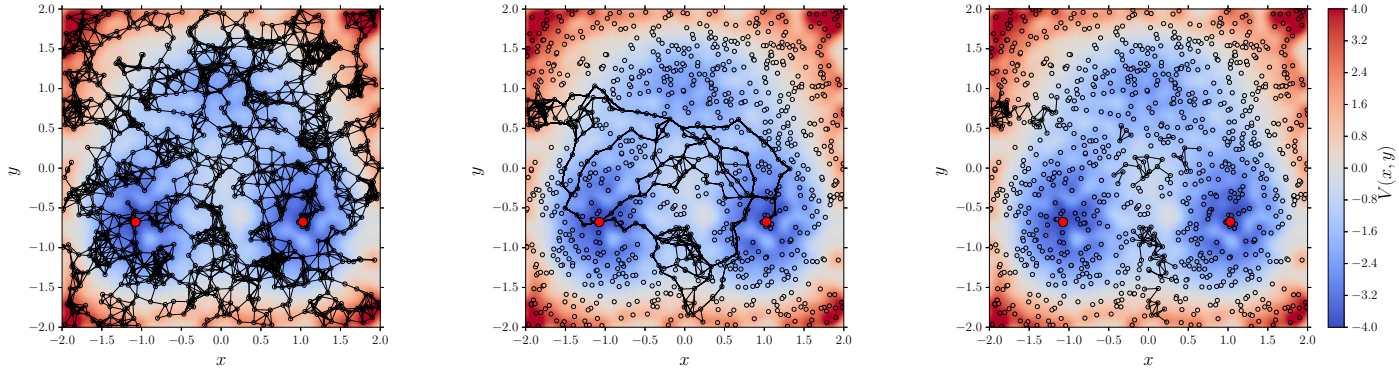
B. LJ_{38}

Having established the ability of the kDP algorithm to determine distinguishable ensembles of pathways in configuration space, we now move on to a much larger benchmark system, namely a kinetic transition network for the cluster of 38 Lennard-Jones atoms (LJ_{38}), consisting of 63706 nodes and 203624 bidirectional edges. This network was constructed by the discrete path sampling method,[1–4] as described in Section II.A, and the elements of the weighted adjacency matrix (Eq. 1) were calculated



(a) (b) (c)

FIG. 1: (a) Disconnectivity graph for the model kinetic transition network based on the three-hole potential, at an energy threshold increment of $\Delta V = 0.25$. The minima (that is, the leaves of the tree) are coloured according to their position in the x direction (see Fig. 2a). (b) Disconnectivity graph including only the stationary points along the shortest distinct path between the lowest-energy minima of the two major funnels. (c) Disconnectivity graph including only the stationary points along the 150$^{\text{th}}$ shortest distinct path.



(a) (b) (c)

FIG. 2: (a) The model Poissonian kinetic transition network embedded on the 2D three-hole potential $V(x, y)$. The pair of endpoint minima are marked in red. (b) The set of 150 distinct paths determined by the kDP algorithm for the $A \leftarrow B$ interbasin transition before an A-B cut is induced in the network. (c) The set of 150 edges that are the rate-limiting edges of each of the distinct paths.

for a reduced temperature of $T = 0.05$. Similar to the model three-hole system discussed in Section III.A, the potential energy landscape for LJ_{38} has a double-funnel topology.[22,65] We find that the physical features of the set of distinct paths for the transition between the two major basins in the network for LJ_{38} are significantly different from the three-hole system.

The competing low-energy morphologies in the LJ_{38} kinetic transition network are an incomplete Mackay icosahedron[64] (denoted I_h) and a face-centered cubic (F) structure. The latter structure is the global potential energy minimum, and corresponds to a much smaller region of configuration space than icosahedral structures.[14] The $F \leftarrow I_h$ transition takes place *via* initial formation of a high-energy hexagonal close-packed (H) metastable intermediate, which subsequently forms a low-energy decahedral state that converts to the F structure by one or more diamond-square-diamond rearrangements associated with moderate barrier heights.[103]

We calculated the complete set of distinct paths for the $F \leftarrow I_h$ interbasin transition, associating the rate-limiting edge with the transition state of highest energy. We also calculated the complete set of distinct paths for a transition from a high-energy H-type state to the F state, which we denote the $F \leftarrow H$ transition, defining the rate-limiting edge as that with the greatest weight. This transition is an example of a transition that is downhill in energy. That is, there is a low energy barrier to escape the slightly defective H state, and the F state is much more thermodynamically stable than the metastable H state. The kDP algorithm found 23460 and 11904 distinct paths for the respective transitions, before an A-B cut was induced in the network. In both cases, the values for the costs of successive paths converged after approx-

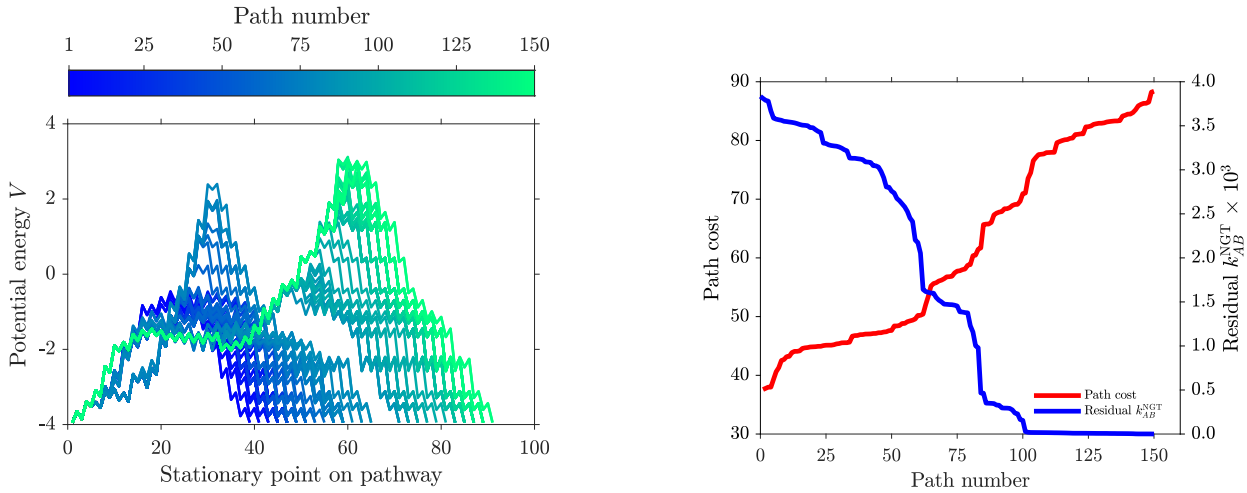


FIG. 3: Results of the kDP algorithm applied to the interbasin transition of the model transition network based on the three-hole potential, at a reduced temperature of $T = 0.6$. (a) Energy profiles of each of the 150 distinct discrete paths. The paths are coloured from blue (shortest distinct path) to green (150^{th} shortest distinct path). (b) (Red) Costs of the 150 distinct paths. (Blue) Steady state rate constant, calculated by the graph transformation method, for the residual network at each iteration of the kDP algorithm.

imately 5000 iterations (Fig. S13 of the Supplementary Information). Energy profiles of representative distinct paths for these transitions are shown in Fig. 4. Profiles of the path costs, and of the steady state rate constants for the accumulated networks, for the 1000 shortest distinct paths of the two transitions are shown in Fig. 5.

For the $F \leftarrow I_h$ transition (Fig. 4a), the low-symmetry, high-energy intermediate states formed by melting of the initial I_h state are liquid-like, *i.e.* structurally disordered, of similar energy and separated by low energy barriers. The transition state ensemble for the $F \leftarrow H$ transition is similarly disordered (Fig. 4b). Therefore there is little energetic preference for the transitions to proceed *via* any particular set of intermediate structures. Furthermore, since the energy barriers for the interconversions of these disordered structures are small, there is only a slightly greater path cost associated with transitions *via* a larger number of stationary points. It is not possible to distinguish any well-defined alternative mechanisms for either transition. That is, for both transitions there is effectively only a single pathway ensemble, where individual paths are separated by low energy barriers in pathway space. We therefore anticipate that, for higher reduced temperatures, there are likely to be a large number of paths that have similar costs. However, the contributions of individual discrete paths to the steady state rate constant are exponentially sensitive to the heights of the energy barriers along the paths (Eq. 2). Consequently, as the reduced temperature is decreased, the number of individual members of the single pathway ensemble that are kinetically relevant decreases, until eventually the dynamical bottleneck of the reactive transition becomes well-defined.

Indeed, we find that for both the $F \leftarrow I_h$ and $F \leftarrow H$ transitions in LJ_{38}, the landscape entropy contribution to the steady state rate constant is small at a reduced temperature of $T = 0.05$ (Fig. 5). The path cost profiles exhibit rapid decay of the relative contributions of successive distinct paths to the steady state rate constant. The rate constants for the accumulated networks at each iteration of the kDP algorithm rapidly converge to values close to those for the complete network within a small number of distinct paths. For the $F \leftarrow I_h$ transition, the majority of the rate constant is achieved for the network formed from the 116 shortest distinct paths, comprising 358 nodes and 472 bidirectional edges ($k_{AB}^{\text{NGT}} = 2.37 \times 10^{-27}$ for the reduced network, compared to $k_{AB}^{\text{NGT}} = 2.77 \times 10^{-27}$ for the complete network). The remaining contribution to the reactive flux is due to small contributions from a large number of many alternative pathways (Fig. S13a). Similarly, for the $F \leftarrow H$ transition, the majority of the rate constant is achieved for the network formed from the 307 shortest distinct paths, comprising 876 nodes and 1188 bidirectional edges ($k_{AB}^{\text{NGT}} = 4.50 \times 10^{-11}$ for the reduced network, compared to $k_{AB}^{\text{NGT}} = 5.23 \times 10^{-11}$ for the complete network). Again, the remaining contribution to the rate is a sum of small contributions from many alternative pathways (Fig. S13b). For both of these transitions, there are a small number of significant jumps in the rate constant profiles for the accumulated networks, indicating that the dynamical bottlenecks of the respective reactive transitions are well-defined at $T = 0.05$.

The landscape entropy contribution to the steady state rate constant for the $F \leftarrow I_h$ transition is significantly greater at a higher reduced temperature of $T = 0.15$

(Fig. S14). This effect is evident from the profile of distinct path costs, which is much flatter than for $T = 0.05$ (Fig. 5a), indicating that there are many kinetically relevant pathways, with each individual pathway making a comparable small contribution to the rate constant. Likewise, the profile of the steady state rate constant for the accumulated network at each iteration of the kDP algorithm is much smoother than for $T = 0.05$. The network formed from the 1000 shortest distinct paths, comprising 1765 nodes and 2784 bidirectional edges, captures only around two-thirds of the reactive flux ($k_{AB}^{\mathrm{NGT}} = 1.55 \times 10^{-6}$ for the reduced network, compared to $k_{AB}^{\mathrm{NGT}} = 2.21 \times 10^{-6}$ for the complete network). These observations demonstrate that the dynamical bottleneck of the $F \leftarrow I_h$ transition is less well-defined at higher temperature.

At the low reduced temperature of $T = 0.05$, where the dynamical bottlenecks of the $F \leftarrow I_h$ and $F \leftarrow H$ transitions are well-defined, the profiles of the steady state rate constants for the accumulated networks are characteristically step-like. Small plateau regions in the profiles may arise if subsequent distinct paths involve only a small number of nodes and edges that are not present in the set of previously determined distinct paths. Furthermore, the inclusion of additional nodes and edges may accommodate new paths that are similar to paths already existing in the network, and which therefore serve merely to split the probabilities of existing paths, yielding a negligible change in the rate constant. The steps in the rate constant profiles correspond to the addition of key edges to the network. With decreasing temperature, a small number of critical transition states have an increasingly dominant effect. For instance, the second jump in the profile of the rate constant of the accumulated network for the $F \leftarrow I_h$ transition at $T = 0.05$, occurring with the inclusion of the 15^{th} distinct path, arises due to the addition of a single edge corresponding to a transition state with a defective F-type structure. This low-energy transition state connects the relatively small volume of configuration space corresponding to F-type structures with the configurational ensemble of disordered structures. Compared to alternative 'late' transition states for the $F \leftarrow I_h$ transition, this key transition state separates the F-type and disordered configurational ensembles by a relatively high energy barrier. It can therefore be inferred that this transition state facilitates pathways for which the probability of recrossings between the F-type and disordered states is small, and hence the contribution to the reactive flux is substantial. A similar effect is observed for the $F \leftarrow H$ transition at $T = 0.05$, where the network formed from the first 35 distinct paths makes a negligible contribution to the steady state rate constant. It is only with the inclusion of the next distinct path, which proceeds *via* the formation of a more regular H-type intermediate structure, that the accumulated network captures an appreciable fraction of the rate constant. The preceding distinct paths feature low energy barriers, and therefore the corresponding sets of pathways are associated with a large number of recrossings and unproductive loops. After the formation of a more regular H-type structure, however, it is improbable to return to the endpoint H state due to an asymmetrical energy barrier that is large in the reverse ($F \rightarrow H$) direction. That is, compared to the more diffusive dynamics[104] for the previous distinct paths, where all energy barriers are small, there is a sharp increase in the committor function associated with the formation of a more regular H-type intermediate. It is for this reason that the reactive flux associated with pathways proceeding *via* such a metastable state is high. These observations demonstrate that caution must be exercised in the interpretation of the distinct path costs, which may not reflect the contributions of the corresponding infinitely large sets of paths, represented by individual distinct paths, to the reactive flux.

Differences between the profiles of distinct path costs for the interbasin transitions of the three-hole potential (Fig. 3b) and of LJ_{38} (Figs. 5a, S13a and S14) are indicative of the contrasting characteristics of the underlying energy landscapes, and hence of the transition path ensembles, for the two systems. For the interbasin transition of LJ_{38}, all reactive trajectories follow a broadly similar mechanism. Effectively, there is only a single pathway ensemble, and the transition state ensemble is a set of disordered structures with similar energies. Conversely, for the interbasin transition of the three-hole potential, there are several distinguishable ensembles of pathways. That is, pathways of high probability are contained within localised regions of pathway space that are separated by high energy barriers. This leads to jumps in the cost profile of successive distinct paths, as the kDP algorithm explores pathways corresponding to separated regions of configuration space. These separate pathway ensembles are characterised by distinguishable transition state ensembles. Conversely, the path cost profiles for the interbasin transition of LJ_{38} at both low and high reduced temperatures tail off smoothly, without significant jumps. The disordered nature of the transition state ensemble for the interbasin transition in LJ_{38} is apparent from the disconnectivity graph,[65] which, in contrast to that for the network based on the three-hole potential (Fig. 1a), does not have a funnelled structure in the high-energy region. The existence of a large family of kinetically relevant paths for both the $F \leftarrow I_h$ and $F \leftarrow H$ transitions at a high reduced temperature, where each successive path comprises a marginally greater number of stationary points than the preceding path, may be a consequence of the small-world character of the kinetic transition network.[105]

IV. CONCLUSIONS

We have implemented an algorithm that is a member of the path deviation family of methods for solving the k shortest loopless paths problem, which allows efficient

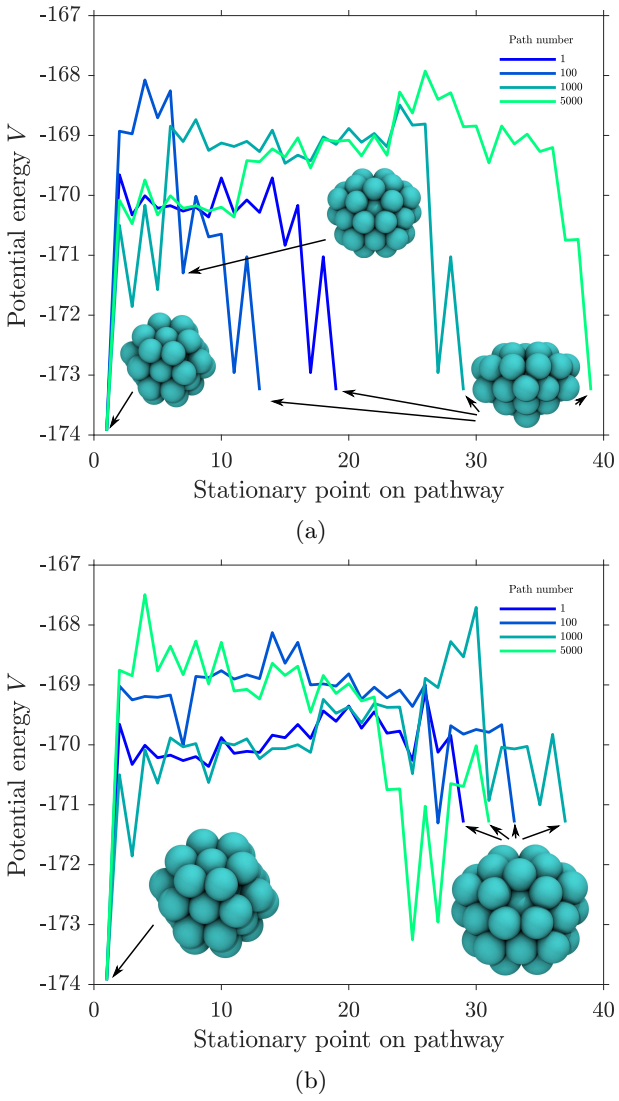


FIG. 4: Energy profiles of the first, hundredth, thousandth and five-thousandth shortest distinct discrete paths for transitions in LJ_{38}, at a reduced temperature of $T = 0.05$. (a) The $F \leftarrow I_h$ transition. (b) The $F \leftarrow H$ transition.

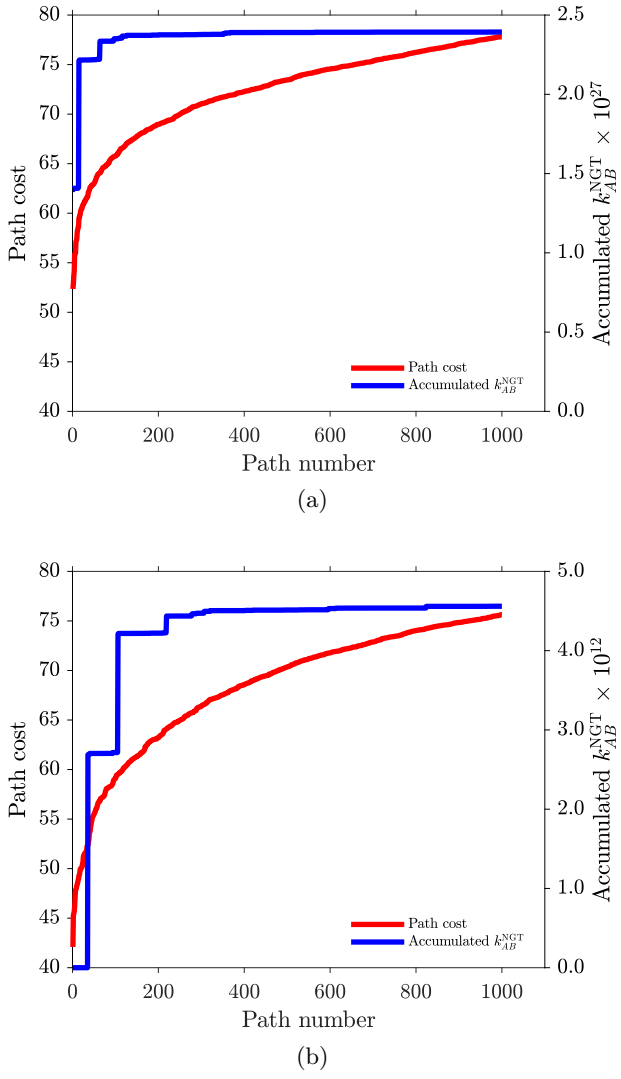


FIG. 5: Results of the kDP algorithm applied to transitions in LJ_{38}, at a reduced temperature of $T = 0.05$. (a) The $F \leftarrow I_h$ transition. (b) The $F \leftarrow H$ transition. (Red) Costs of the 1000 shortest distinct paths. (Blue) Steady state rate constant, calculated by the graph transformation method, for the accumulated network composed of the distinct paths determined by the kDP algorithm.

identification of mechanistically distinct pathways in a transition network. The algorithm is ideal for our purpose, as it allows one to define a criterion for distinguishing successive pathways. Here we impose the condition that each path has a distinct rate-limiting edge compared to all other paths. This constraint provides an intuitive definition for paths that can be considered to differ non-trivially from one another, and encourages the algorithm to rapidly progress to separate regions of the network. A by-product of the algorithm is a set of rate-limiting edges that approximates the transition state ensemble. The algorithm has favourable time and space complexity, an essential consideration given the large size of kinetic transition networks constructed by discrete path sampling for real systems of interest.

We have tested the method on two benchmark systems that demonstrate the capability of the algorithm to explore pathway ensembles existing in separate regions of a transition network, and therefore of the underlying configuration space. Inspection of the cost profiles for the distinct paths found by the algorithm, and analysis of the steady state rate constants for the accumulated networks composed of the determined distinct paths, shows that multiple kinetically relevant pathways are present for both systems. Hence Dijkstra's algorithm alone does not give a complete picture of the dynamics, even in

such simple cases. The contrasting path cost profiles for the interbasin transitions of networks for the model 'three-hole' potential and for the LJ_{38} cluster reflect the fact that the former system features multiple well-defined pathway ensembles separated by high energy barriers in the configuration space. This leads to jumps in the profile of path costs as well-defined transition state ensembles become blocked and the algorithm explores separate regions of the network. Similarly, there are jumps in the profile of the rate constant for the accumulated network, as distinct paths belonging to separate pathway ensembles are incorporated in the network. For the interbasin transition in LJ_{38} there is effectively a single pathway ensemble with many possible alternative disordered intermediate states following initial melting, and low energy barriers between these different structures. This is evident from the distribution of energies for the transition state ensemble, which is relatively homogeneous. This characteristic of the transition path ensemble is indicated by the smooth tailing off of the profiles of the path costs and of the rate constant for the accumulated network. At lower temperatures, the reaction bottleneck is more well-defined, and hence the contributions of successive distinct paths to the steady state rate constant decay more rapidly, and the rate constant for the accumulated network also converges more rapidly.

The full set of rate-limiting edges that induces an A-B cut in the network may be interpreted similarly to a transition dividing surface of the corresponding continuous configuration space.[60,61] Thus, as a by-product of the algorithm, we obtain a partition of the high-dimensional configuration space, which characterises the dynamical bottleneck of the $A \leftarrow B$ interstate transition. This region has a critical role in determining the dynamical features of the rare barrier-crossing events between metastable states.[56,57]

The k distinct paths algorithm is applicable to any network where edge weights correspond to transition rates, probabilites, or probability fluxes, such that suitable edge weights for use with a shortest path algorithm can be assigned (*cf.* Eq. 1). A notable class of transition networks is Markov state models[12,13,106,107] (MSMs), networks parameterised by a transition matrix determined from many short simulation trajectories. MSMs have been used extensively to study biomolecular conformational transitions,[13,107–109] and are amenable to analysis by transition path theory[110–112] to calculate important dynamical quantities such as reactive fluxes. Stochastic network models are also of fundamental importance in many other domains, such as in systems biology,[113,114] and in studies of epidemic spread[115] and finance.[116]

Further work will focus on the analysis of existing kinetic transition networks, for which evidence for multiple competing pathway ensembles, or otherwise, and assessment of the kinetic relevance of alternative pathway ensembles, yields valuable insight into the underlying dynamics. Variations with temperature in the profiles of distinct path costs and of steady state rate constants for the accumulated networks formed from the distinct paths reflect the entropic character of an interstate transition. The kDP algorithm therefore provides a novel framework for understanding the fundamental features of reactive transitions on energy landscapes. A particularly interesting application is to conformational transitions of biomolecules, for which patterns of dynamical behaviour are closely related to biological function.[4,25]

V. ACKNOWLEDGEMENTS

DJS gratefully acknowledges the Cambridge Commonwealth, European and International Trust for a PhD scholarship. DJW gratefully acknowledges support from the EPSRC. We are grateful to Dr Sam P. Niblett for assistance with constructing stationary point databases for testing the algorithm.

VI. SUPPLEMENTARY INFORMATION

The algorithm detailed in this paper for determining k distinct paths in a transition network is available in the Fortran 90 language as the KDISTINCTPATHS subroutine of the PATHSAMPLE program (http://www-wales.ch.cam.ac.uk/PATHSAMPLE/), which is freely available software under the GNU General Public License. A Python implementation of the algorithm and a C++ script to construct model kinetic transition networks are publically available online at https://github.com/danieljsharpe.

[1] D. J. Wales, Mol. Phys. **100**, 3285–3305 (2002).
[2] D. J. Wales, Mol. Phys. **102**, 891–908 (2004).
[3] D. J. Wales, Phil. Trans. Roy. Soc. A **370**, 2877–2899 (2012).
[4] J. A. Joseph, K. Röder, D. Chakraborty, R. G. Mantell, and D. J. Wales, Chem. Commun. **53**, 6974–6988 (2017).
[5] S. V. Krivov and M. Karplus, J. Phys. Chem. B **110**, 12689–12698 (2006).
[6] F. Rao and M. Karplus, Proc. Natl. Acad. Sci. USA **107**, 9152–9157 (2010).
[7] N. M. Amato, K. A. Dill, and G. Song, J. Comput. Biol. **10**, 239–255 (2003).
[8] L. Gong and X. Zhou, J. Phys. Chem. B **114**, 10266–10276 (2010).
[9] F. Marinelli, F. Pietrucci, A. Laio, and S. Piana, PLoS Comput. Biol. **5**, e1000452 (2009).
[10] N. Singhal, C. D. Snow, and V. S. Pande, J. Chem. Phys. **121**, 415–425 (2004).
[11] N.-V. Buchete and G. Hummer, J. Phys. Chem. B **112**, 6057–6069 (2008).
[12] G. R. Bowman, V. S. Pande, and F. Noé (Eds.), *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, 1st ed. (Springer, Netherlands, 2014).
[13] J. D. Chodera and F. Noé, Curr. Op. Struct. Biol. **25**, 135–144 (2014).
[14] D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, 2003).
[15] F. Noé and J. C. Smith, "Transition networks: A unifying theme for molecular simulation and computer science," in *Mathematical Modeling of Biological Systems, Volume I.*, edited by

A. Deutsch, L. Brusch, J. Byrne, G. de Vries, and H.-P. Herzel (Birkhäuser, Boston, 2007) pp. 125–144.
[16]F. Noé and S. Fischer, Curr. Op. Struct. Biol. **18**, 154–162 (2008).
[17]F. Noé, I. Horenko, C. Schütte, and J. C. Smith, J. Chem. Phys. **126**, 155102 (2007).
[18]J. D. Stevenson and D. J. Wales, J. Chem. Phys. **141**, 041104 (2014).
[19]D. J. Wales and P. Salamon, Proc. Natl. Acad. Sci. USA **111**, 617–622 (2014).
[20]D. J. Wales, Annu. Rev. Phys. Chem. **69**, 401–425 (2018).
[21]J. P. K. Doye, D. J. Wales, and M. A. Miller, J. Chem. Phys. **109**, 8143–8153 (1998).
[22]J. P. K. Doye, M. A. Miller, and D. J. Wales, J. Chem. Phys. **110**, 6896–6906 (1999).
[23]J. P. K. Doye, M. A. Miller, and D. J. Wales, J. Chem. Phys. **111**, 8417–8428 (1999).
[24]M. A. Miller, J. P. K. Doye, and D. J. Wales, Phys. Rev. E **60**, 3701–3718 (1999).
[25]K. Röder, J. A. Joseph, B. E. Husic, and D. J. Wales, Adv. Theory Simul. **2**, 1800175 (2019).
[26]S. P. Niblett, M. Biedermann, D. J. Wales, and V. K. de Souza, J. Chem. Phys. **147**, 152726 (2017).
[27]E. W. Dijkstra, Numer. Math. **1**, 269–271 (1959).
[28]D. A. Evans and D. J. Wales, J. Chem. Phys. **121**, 1080–1090 (2004).
[29]T. J. H. Vlugt and B. Smit, PhysChemComm **4**, 11–17 (2001).
[30]S. W. Englander and L. Mayne, Proc. Natl. Acad. Sci. USA **111**, 15873–15880 (2014).
[31]S. W. Englander and L. Mayne, Proc. Natl. Acad. Sci. USA **114**, 8253–8258 (2017).
[32]W. A. Eaton and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **114**, E9759–E9760 (2017).
[33]D. R. Shier, Networks **9**, 195–214 (1979).
[34]E. Q. V. Martins, M. M. B. Pascoal, and J. L. E. Santos, International Journal of Foundations of Computer Science **10**, 247–261 (1999).
[35]D. J. Wales and J. P. K. Doye, J. Chem. Phys. **119**, 12409–12416 (2003).
[36]A. Perko, Networks **16**, 149–160 (1986).
[37]E. Q. V. Martins, M. M. B. Pascoal, and J. L. E. Santos, "The k shortest loopless paths problem," Tech. Rep. (Universidade de Coimbra, 1998).
[38]E. Q. V. Martins and M. M. B. Pascoal, 4OR - Quarterly Journal of the Belgian, French and Italian Operations Research Societies **1**, 121–134 (2003).
[39]J. Y. Yen, Management Science **17**, 712–716 (1971).
[40]V. M. Jiménez and A. Marzal, "Computing the k shortest paths: a new algorithm and experimental comparison," in *Algorithm Engineering: 3rd International Workshop, WAE'99, London, UK*, edited by J. S. Vitter and C. D. Zaroliagis (Springer Berlin, Heidelberg, 1999) pp. 15–29.
[41]J. M. Carr and D. J. Wales, "The energy landscape as a computational tool," in *Latest Advances in Atomic Cluster Collisions: Structure and Dynamics from the Nuclear to the Biological Scale*, edited by A. Solov'yov and J.-P. Connerade (Imperial College Press, London, 2008) pp. 321–330.
[42]D. J. Wales, J. Chem. Phys. **130**, 204111 (2009).
[43]A. Chatterjee and A. F. Voter, J. Chem. Phys. **132**, 194101 (2010).
[44]M. Athènes and V. V. Bulatov, Phys. Rev. Lett. **113**, 230601 (2014).
[45]J. M. Carr and D. J. Wales, Phys. Chem. Chem. Phys. **11**, 3341–3354 (2009).
[46]E. Q. V. Martins, J. Op. Res. **18**, 123–130 (1984).
[47]E. Q. V. Martins and J. L. E. Santos, "A new shortest paths ranking algorithm," Tech. Rep. (Universidade de Coimbra, 1996).
[48]E. Q. V. Martins, M. M. B. Pascoal, and J. L. E. Santos, Investigação Operacional **21**, 47–60 (2001).
[49]E. Q. V. Martins and J. L. E. Santos, Investigação Operacional **20**, 47–62 (2000).
[50]D. Eppstein, "Finding the k shortest paths," Proc. 35th IEEE Symp. FOCS , 154–165 (1994).
[51]D. Eppstein, SIAM J. Comput. **28**, 652–673 (1999).
[52]J. A. Azevedo, M. E. O. S. Costa, J. J. R. E. S. Madeira, and E. Q. V. Martins, European J. of Op. Res. **69**, 97–106 (1993).
[53]J. A. Azevedo, J. J. R. E. S. Madeira, E. Q. V. Martins, and F. M. A. Pires, European J. Op. Res. **73**, 188–191 (1994).
[54]E. Q. V. Martins, M. M. B. Pascoal, and J. L. E. Santos, "A new algorithm for ranking loopless paths," Tech. Rep. (Universidade de Coimbra, 1997).
[55]D. Frigioni, A. Marchetti-Spaccamela, and U. Nanni, J. Algorithms **34**, 251–281 (2000).
[56]C. Dellago, P. G. Bolhuis, and P. L. Geisller, Adv. Chem. Phys. **123**, 1–78 (2002).
[57]J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, J. Chem. Phys. **134**, 174105 (2011).
[58]A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769–6779 (2005).
[59]R. Elber, J. M. Bello-Rivas, P. Ma, A. E. Cardenas, and A. Fathizadeh, Entropy **19**, 219 (2017).
[60]F. Noé, M. Oswald, G. Reinelt, S. Fischer, and J. C. Smith, Multiscale Model. Simul. **5**, 393–419 (2006).
[61]F. Noé, D. Krachtus, J. C. Smith, and S. Fischer, J. Chem. Theory Comput. **2**, 840–857 (2006).
[62]F. Noé, F. Ille, J. C. Smith, and S. Fischer, Proteins **59**, 534–544 (2005).
[63]S. Huo and J. E. Straub, J. Chem. Phys. **107**, 5000–5006 (1997).
[64]D. J. Wales and J. P. K. Doye, J. Phys. Chem. A **101**, 5111–5116 (1997).
[65]D. J. Wales, M. A. Miller, and T. R. Walsh, Nature **394**, 758–760 (1998).
[66]S. Park, M. K. Sener, D. Lu, and K. Schulten, J. Chem. Phys. **119**, 1313–1319 (2003).
[67]P. Metzner, C. Schütte, and E. Vanden-Eijnden, J. Chem. Phys. **125**, 084110 (2006).
[68]P. Metzner, C. Schütte, and E. Vanden-Eijnden, Multiscale Model. Simul. **7**, 1192–1219 (2009).
[69]M. K. Cameron, J. Stat. Phys. **152**, 493–518 (2013).
[70]M. K. Cameron, J. Chem. Phys. **141**, 184113 (2014).
[71]M. K. Cameron and E. Vanden-Eijnden, J. Stat. Phys. **156**, 427–454 (2014).
[72]M. K. Cameron, Networks Heterogen. Media **9**, 383–416 (2014).
[73]J. P. Neirotti, F. Calvo, D. L. Freeman, and J. D. Doll, J. Chem. Phys. **112**, 10340 (2000).
[74]F. Calvo, J. P. Neirotti, D. L. Freeman, and J. D. Doll, J. Chem. Phys. **112**, 10350 (2000).
[75]M. Picciani, M. Athenes, J. Kurchan, and J. Taileur, J. Chem. Phys. **135**, 034108 (2011).
[76]W. Polak and A. Patrykiejew, Phys. Rev. B **67**, 115402 (2003).
[77]F. Calvo, J. P. K. Doye, and D. J. Wales, J. Chem. Phys. **114**, 7312–7329 (2001).
[78]J. P. K. Doye and F. Calvo, J. Chem. Phys. **116**, 8307–8317 (2002).
[79]Z. Li and H. A. Scheraga, Proc. Natl. Acad. Sci. USA **84**, 6611–6615 (1987).
[80]Z. Li and H. A. Scheraga, J. Mol. Struct. **179**, 333–352 (1988).
[81]S. A. Trygubenko and D. J. Wales, J. Chem. Phys. **120**, 2082–2094 (2004).
[82]S. A. Trygubenko and D. J. Wales, J. Chem. Phys. **120**, 7820–7820 (2004).
[83]G. Henkelman and H. Jònsson, J. Chem. Phys **111**, 7010–7222 (1999).
[84]G. Henkelman, B. P. Uberuaga, and H. Jònsson, J. Chem. Phys. **113**, 9901–9904 (2000).
[85]G. Henkelman and H. Jònsson, J. Chem. Phys. **113**, 9978–9985 (2000).

[86]L. J. Munro and D. J. Wales, Phys. Rev. B **59**, 3969–3980 (1999).
[87]Y. Zeng, P. Xiao, and J. Henkelman, J. Chem. Phys. **140**, 044115 (2014).
[88]R. G. Mantell, C. E. Pitt, and D. J. Wales, J. Chem. Theory Comput. **12**, 6182–6191 (2016).
[89]D. Liu and J. Nocedal, Math. Program. **45**, 503–528 (1989).
[90]J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. (Springer-Verlag, Berlin, 2006).
[91]J. M. Carr, S. A. Trygubenko, and D. J. Wales, J. Chem. Phys. **122**, 234903 (2005).
[92]J. M. Carr and D. J. Wales, J. Chem. Phys. **123**, 234901 (2005).
[93]B. Strodel, C. S. Whittleston, and D. J. Wales, J. Am. Chem. Soc. **129**, 16005–16014 (2007).
[94]B. Peters, *Reaction Rate Theory and Rare Events* (Elsevier, Oxford, UK, 2017).
[95]O. M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495–1517 (1997).
[96]S. V. Krivov and M. Karplus, J. Chem. Phys. **117**, 10894–10903 (2002).
[97]S. L. Seyler, A. Kumar, M. F. Thorpe, and O. Beckstein, PLoS Comput. Biol. **11**, e1004568 (2015).
[98]A. Berezhkovskii, G. Hummer, and A. Szabo, J. Chem. Phys. **130**, 205102 (2009).
[99]H. Jung, K. Okazaki, and G. Hummer, J. Chem. Phys. **147**, 152716 (2017).
[100]S. A. Trygubenko and D. J. Wales, Mol. Phys. **104**, 1497–1507 (2006).
[101]S. A. Trygubenko and D. J. Wales, J. Chem. Phys. **124**, 234110 (2006).
[102]A.-L. Barabási and M. Pósfai, *Network Science*, 1st ed. (Cambridge University Press, Cambridge, 2016).
[103]J. P. K. Doye, *The Structure, Thermodynamics and Dynamics of Atomic Clusters*, Ph.D. thesis, University of Cambridge (1996).
[104]R. J. Allen, P. B. Warren, and P. R. ten Wolde, Phys. Rev. Lett. **94**, 018104 (2005).
[105]J. W. R. Morgan, D. Mehta, and D. J. Wales, Phys. Chem. Chem. Phys. **19**, 25498–25508 (2017).
[106]V. S. Pande, K. Beauchamp, and G. R. Bowman, Methods **52**, 99–105 (2010).
[107]B. E. Husic and V. S. Pande, J. Am. Chem. Soc. **140**, 2386–2896 (2018).
[108]D. Shukla, C. X. Hernández, J. K. Weber, and V. S. Pande, Acc. Chem. Res. **48**, 414–422 (2015).
[109]W. Wang, S. Cao, L. Zhu, and X. Huang, WIREs Comput. Mol. Sci. **8**, e1343 (2018).
[110]E. Vanden-Eijnden, "Transition path theory," in *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1*, edited by M. Ferrario, G. Ciccotti, and K. Binder (Springer Berlin, Heidelberg, 2006) pp. 453–493.
[111]W. E and E. Vanden-Eijnden, J. Stat. Phys. **123**, 503–523 (2006).
[112]W. E and E. Vanden-Eijnden, Annu. Rev. Phys. Chem. **61**, 391–420 (2010).
[113]D. Schultz, A. M. Walczak, J. N. Onuchic, and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **105**, 19165–19170 (2008).
[114]M. J. Tse, B. K. Chu, C. P. Gallivan, and E. L. Read, PLoS Comput. Biol. **14**, e1006336 (2018).
[115]C. Eskin, J. S. Shamma, and J. S. Weitz, Sci. Rep. **7**, 44122 (2017).
[116]R. Gong and P. Frank, "Systemic risk and the dynamics of temporary financial networks," Tech. Rep. (Systemic Risk Centre, The London School of Economics and Political Science, 2016).

input : a kinetic transition network G with weights M_{uv} for edges (u, v), directed $u \leftarrow v$
shortest path tree T with costs $D(u)$ for nodes u
source node $b \in B$
set of sink nodes $\{a\} \in A$
number of paths to compute k
output: set of k pathways $P = \{P_1, P_2, ..., P_k\}$
set of k rate-limiting edges $C = \{e_{\mathrm{RLE},1}, e_{\mathrm{RLE},2}, ..., e_{\mathrm{RLE},k}\}$. $e_{\mathrm{RLE},i} = (u, v)$,
where u is the parent node of v in T, and the weight of the edge is $M_{\mathrm{RLE},i}$

$i = 1$, $Q \leftarrow \emptyset$; `// Q is a minimum-priority queue of nodes`
while $i \leq k$ **do**
  $P_i \leftarrow$ `GetShortestPath(`T,A,B`)`;
  $e_{\mathrm{RLE},i} = (u, v) \leftarrow$ `GetRateLimEdge(`P_i`)`. $e'_{\mathrm{RLE},i} = (v, u)$;
  $M_{\mathrm{RLE},i} \leftarrow$ inf, $M'_{\mathrm{RLE},i} \leftarrow$ inf;
  if A and B are **not** connected **then**
    break;
  u and all nodes in the set `Descendants(`u,T`)` are coloured red;
  foreach node $z \in T$ that is red **do**
    $t \leftarrow$ the neighbouring node t of $z \in G$ for which $D(t) + M_{zt}$ is minimal **and** such that t is **not** red;
    if t == null **then**
      `UpdateParent(`z,null,T`)`, `UpdateCost(`z,inf`)`;
    else
      `UpdateParent(`z,t,T`)`, `UpdateCost(`z,$D(t) + M_{zt}$`)`, `Push(`Q,z,$D(t) + M_{zt}$`)`;
  while $Q \neq \emptyset$ **do**
    $z \leftarrow$ `Pop(`Q`)`;
    foreach node $h \in G$ that is a neighbour of z **and** is red **do**
      if $D(z) + M_{hz} < D(h)$ **then**
        `UpdateParent(`h,z,T`)`, `UpdateCost(`h,$D(z) + M_{hz}$`)`, `UpdatePriority(`Q,h,$D(z) + M_{hz}$`)`;
  set all nodes to **not** red;
  $i \leftarrow i + 1$;
if A **and** B are connected **then**
  $P_k \leftarrow$ `GetShortestPath(`T,A,B`)`;
  $e_{\mathrm{RLE},k} \leftarrow$ `GetRateLimEdge(`P_k`)`;
return P, C;

Algorithm 1: Outline of the "k distinct paths" algorithm to find the k shortest paths in a transition network that are distinct, in the sense that each has its own unique identifying rate-limiting edge, given an initial shortest path tree.

input : Target node degree distribution $P(k)$
N-dimensional potential function $V(\mathbf{r}^N)$
function domain $r_1, r_2, ..., r_N \in [-r_{\max}, r_{\max}]$
number of nodes n_{V}, the n-th node has position vector $\mathbf{r}_n^N$
distance threshold for determining node connectivity d
standard deviation of minima and transition state energies σ_{E}
mean energy barrier height μ_b
error tolerance for current $P_{\mathrm{obs}}(k)$ and target $P(k)$ degree distributions ϵ_{tol}
maximum number of iterations n_{it}
interval for attempting changes in distance threshold n_{intvl}
output: A kinetic transition network G with fitted degree distribution and minima and transition state energies $\{E_u\}$ and $\{E_{uv}\}$, respectively.

for $n \leftarrow 1$ **to** n_{V} **do**
  $\mathbf{r}_n^N = (r_{n1}, r_{n2}, ..., r_{nN}), r_{ni} \leftarrow$ `RandUnifFloat(`$-r_{\max}$`,`$r_{\max}$`)`;
$G \leftarrow$ `GetEdges(`$\mathbf{R}^N = (\mathbf{r}_1^N, \mathbf{r}_2^N, ..., \mathbf{r}_N^N)$`,`$d$`)`;
$n_{\mathrm{step}} \leftarrow 0$, $d' \leftarrow d$, $\epsilon_{\mathrm{obs}} \leftarrow$ inf, $\epsilon'_{\mathrm{obs}} \leftarrow$ inf;
while $n_{\mathrm{step}} \leq n_{\mathrm{it}}$ **and** $\epsilon_{\mathrm{obs}} > \epsilon_{\mathrm{tol}}$ **do**
  if $n_{\mathrm{step}} \,\%\, n_{\mathrm{intvl}} == 0$ **then**
    $d' \leftarrow d \cdot$ `RandNormal(`$\mu = 1$`,`$\sigma = 0.1$`)`, $\mathbf{R'}^N \leftarrow \mathbf{R}^N$;
  else
    $a \leftarrow$ `RandUnifInt(1,`n_{V}`)`;
    $\mathbf{r}_a^N = (r_{a1}, r_{a2}, ..., r_{aN}), r_{ai} \leftarrow$ `RandUnifFloat(`$-r_{\max}$`,`$r_{\max}$`)`;
    $\mathbf{R'}^N \leftarrow$ `UpdateNodePosn(`$\mathbf{R}^N$`,`a`,`$\mathbf{r}_a^N$`)`;
  $G' \leftarrow$ `GetEdges(`$\mathbf{R'}^N$`,`d`)`;
  $P_{\mathrm{obs}}(k) \leftarrow$ `GetDegDistrib(`G'`)`;
  $\epsilon'_{\mathrm{obs}} \leftarrow$ `GetDegDistribErr(`$P_{\mathrm{obs}}(k)$`,`$P(k)$`)`;
  $n_{\mathrm{step}} \leftarrow n_{\mathrm{step}} + 1$;
  if $\epsilon_{\mathrm{obs}} < \epsilon'_{\mathrm{obs}}$ **then**
    $\epsilon_{\mathrm{obs}} \leftarrow \epsilon'_{\mathrm{obs}}$, $\mathbf{R}^N \leftarrow \mathbf{R'}^N$, $G \leftarrow G'$, $d \leftarrow d'$;
for node u **in** G **do**
  $E_u \leftarrow V(\mathbf{r}_u^N)+$ `RandNormal(0,`σ_{E}`)`;
for edge (u, v) **in** G **do**
  $\mathbf{r}_{uv}^N \leftarrow (\mathbf{r}_u^N - \mathbf{r}_v^N)/2$;
  $t \leftarrow$ `FindHigherEnergyNode(`u`,`v`)`;
  $E_{uv} \leftarrow E_t + \mu_b +$ `RandNormal(0,`σ_{E}`)`;
return $G \leftarrow$ `GetLargestConnectedComponent(`G`)`;

Algorithm 2: to construct a kinetic transition network corresponding to a roughened potential provided by the user and with a node degree distribution fitted to a specified distribution. μ and σ denote the mean and standard deviation of a Gaussian, respectively.