



**HAL**  
open science

## Rare events and first passage time statistics from the energy landscape

T D Swinburne, Thomas Swinburne, Deepti Kannan, Daniel J Sharpe, David J Wales

► **To cite this version:**

T D Swinburne, Thomas Swinburne, Deepti Kannan, Daniel J Sharpe, David J Wales. Rare events and first passage time statistics from the energy landscape. *Journal of Chemical Physics*, 2020, 153 (13), pp.134115. 10.1063/5.0016244 . hal-02958853

**HAL Id: hal-02958853**

**<https://hal.science/hal-02958853>**

Submitted on 8 Oct 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## Rare Events and First Passage Time Statistics From the Energy Landscape

Thomas D Swinburne,<sup>1</sup> Deepti Kannan,<sup>2</sup> Daniel J Sharpe,<sup>2</sup> and David J Wales<sup>2</sup>

<sup>1</sup>*Aix-Marseille Université, CNRS, CINaM UMR 7325, Campus de Luminy, 13288 Marseille, France<sup>a)</sup>*

<sup>2</sup>*Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK*

(Dated: 24 August 2020)

In this contribution we analyse the probability distribution of rare first passage times corresponding to transitions between product and reactant states in a kinetic transition network. The mean first passage times and corresponding rate constants are analysed in detail for two model landscapes and the double funnel landscape corresponding to an atomic cluster. Evaluation schemes based on eigendecomposition and kinetic path sampling, which both allow access to the passage time distribution, are benchmarked against mean first passage times calculated using graph transformation. Numerical precision issues severely limit the useful temperature range for eigendecomposition, but kinetic path sampling is capable of extending the first passage time analysis to lower temperatures, where the kinetics of interest constitute rare events. We then investigate the influence of free energy based state regrouping schemes for the underlying network. Alternative formulations of the effective transition rates for a given regrouping are compared in detail, to determine their numerical stability and capability to reproduce the true kinetics, including recent coarse-graining approaches that preserve occupancy cross correlation functions. We find that appropriate regrouping of states under the simplest local equilibrium approximation can provide reduced transition networks with useful accuracy at somewhat lower temperatures. Finally, a method is provided to systematically interpolate between the local equilibrium approximation and exact intergroup dynamics. Spectral analysis is applied to each grouping of states, employing a moment-based mode selection criteria to produce a reduced state space, which does not require any spectral gap to exist, but reduces to gap-based coarse graining as a special case. Implementations of the developed methods are freely available online.

---

<sup>a)</sup>Electronic mail: [swinburne@cinam.univ-mrs.fr](mailto:swinburne@cinam.univ-mrs.fr)

## I. INTRODUCTION

Kinetic transition networks (KTNs) provide a convenient framework to represent the energy landscape of a molecular or condensed matter system.<sup>1–3</sup> Such networks can be represented as a graph, where the nodes are metastable states composed of potential or free energy minima, and the edges are pathways between them. The task of network construction can be approached in various ways; we use discrete path sampling,<sup>4–6</sup> which exploits geometry optimisation to identify local potential energy minima and the transition states that connect them. The associated rate constants, obtained using unimolecular rate theory, scale as  $e^{-\Delta F/T}$  and are exponentially sensitive to the free energy barriers connecting minima.<sup>7,8</sup> Analogous models, which are also based on a master equation<sup>9,10</sup> approach, can be built from simulations employing explicit dynamics.<sup>11–19</sup> Further details of the computational energy landscape methodology and the associated geometry optimisation tools that we employ can be found in some recent reviews.<sup>20–22</sup> This framework is particularly appealing for the analysis of kinetic transition networks derived from energy landscapes where the states of interest are separated by barriers that are large compared with the available thermal energy, giving rise to broken ergodicity and rare event dynamics.<sup>23–35</sup>

Having obtained a KTN,<sup>1,2,36,37</sup> and calculated approximate local equilibrium properties and transition rates using standard methods of statistical mechanics, we are still presented with the task of extracting global observable properties, such as phenomenological rate constants and mean first passage times (MFPTs) between reactants and products. There are typically significant numerical problems with these calculations, especially at low temperatures when transition rates are small.<sup>38,39</sup> In the present work we compare alternative approaches for computing rates and first passage times, which advance the efficiency and accuracy of these procedures. Reference values for the rate constants are obtained from graph transformation,<sup>37,38,40–42</sup> where nodes are iteratively removed from the transition network, and the transition probabilities are renormalised so as to preserve the exact MFPT. These results are compared with eigendecomposition<sup>19,42</sup> and kinetic path sampling,<sup>43,44</sup> which can both provide additional information via the probability distribution for the first passage time (FPT). As expected, numerical precision problems limit the useful temperature range for eigendecomposition, but kinetic path sampling, which exploits graph transformation, allows us to extend the analysis to significantly lower temperatures. The full FPT distribution provides detailed insights into the underlying energy landscape, which are not available from the phenomenological rate constants. Hence this information is particularly valuable for interpreting experimental results, especially single molecule experiments.<sup>45</sup>

Because the number of nodes scales exponentially with the system size,<sup>46,47</sup> KTNs are often large and sparse, providing additional numerical challenges for the extraction of FPT statistics. Lumping states in the original network together is an attractive approach for reducing dimensionality.<sup>48–57</sup> We consider a lumping scheme that recursively regroups local minima separated by low-lying barriers below a free energy threshold.<sup>58–60</sup> If rapidly interconverting states can establish local relative equilibrium on the time scale of the rate-determining process, the intergroup rates in the coarse-grained network can be simply determined by the local equilibrium approximation (LEA)<sup>60</sup> while preserving the rare event of interest. We compare the LEA to alternative formulations of the intergroup rates for groups determined via the free energy regrouping procedure, examining the accuracy and the temperature range over which numerical precision is retained.

Finally, we introduce an extension to the LEA based on spectral analysis within each group of states, providing a procedure to interpolate from the LEA approximation to exact results. We apply these techniques to compute MFPTs and rate constants in the low temperature regime, where interconversion of competing morphologies for a benchmark atomic cluster (LJ<sub>38</sub>) constitutes a rare event. All coarse graining techniques described in this contribution are contained in the freely available PyGT package<sup>61</sup>, along with introductory tutorials, whilst the kinetic path sampling data was generated using the DISCOTRESS simulation suite<sup>62</sup>.

## II. FIRST PASSAGE TIME DISTRIBUTIONS AND RATES

We consider a kinetic transition network<sup>1,2,36,37</sup> (KTN) consisting of local minima and the transition states that connect them on the potential energy landscape. The minima are partitioned into three groups, namely product and reactant regions  $\mathcal{A}$  and  $\mathcal{B}$ , and the remaining minima, denoted  $\mathcal{I}$  for ‘intervening’. These sets contain  $N_{\mathcal{A}}$ ,  $N_{\mathcal{B}}$ , and  $N_{\mathcal{I}}$  minima, respectively. The mean  $\mathcal{A} \leftarrow \mathcal{B}$  first passage time  $\mathcal{T}_{\mathcal{A}\mathcal{B}}$  can be calculated by considering a system where  $\mathcal{A}$  is an absorbing state, with an analogous setup for  $\mathcal{B} \leftarrow \mathcal{A}$  and  $\mathcal{B}$  absorbing.<sup>19,42</sup>

For specificity we illustrate the theory for  $\mathcal{A} \leftarrow \mathcal{B}$ . All trajectories then end in  $\mathcal{A}$  for a connected network, but the dynamics are unchanged until the absorption event. We can therefore write a master equation for the dynamics in

$\mathcal{I} \cup \mathcal{B}$ :

$$\begin{aligned} \begin{bmatrix} \dot{\mathbf{P}}_{\mathcal{I}}(t) \\ \dot{\mathbf{P}}_{\mathcal{B}}(t) \end{bmatrix} &= \begin{bmatrix} \mathbf{K}_{\mathcal{II}} - \mathbf{D}_{\mathcal{I}} & \mathbf{K}_{\mathcal{IB}} \\ \mathbf{K}_{\mathcal{BI}} & \mathbf{K}_{\mathcal{BB}} - \mathbf{D}_{\mathcal{B}} \end{bmatrix} \begin{bmatrix} \mathbf{P}_{\mathcal{I}}(t) \\ \mathbf{P}_{\mathcal{B}}(t) \end{bmatrix} \\ &\equiv \mathbf{M} \mathbf{P}_{\mathcal{I} \cup \mathcal{B}}(t), \end{aligned} \quad (1)$$

where  $\mathbf{K}_{\mathcal{X}\mathcal{Y}} \in \mathbb{R}^{N_{\mathcal{X}} \times N_{\mathcal{Y}}}$  is the matrix of minimum-to-minimum rate constants  $\mathcal{Y} \rightarrow \mathcal{X}$ , and  $\mathbf{D}_{\mathcal{X}} \in \mathbb{R}^{N_{\mathcal{X}} \times N_{\mathcal{X}}}$  is a diagonal matrix whose entries are the total escape rate from each state in  $\mathcal{X}$ , i.e.  $[\mathbf{D}_{\mathcal{X}}]_{ii} = \sum_j \mathbf{K}_{ji}$ . Because  $\mathcal{A}$  is absorbing,  $\mathbf{K}_{\mathcal{A}\mathcal{A}} = \mathbf{0}$ , and transitions to  $\mathcal{A}$  appear only in the diagonal matrices  $\mathbf{D}_{\mathcal{I}}$  and  $\mathbf{D}_{\mathcal{B}}$ .

The probability that the first passage time  $\mathcal{T}_{\mathcal{A}\mathcal{B}}$  lies between  $t$  and  $t + dt$ ,  $p(t)dt$ , for a starting distribution with  $\mathbf{P}_{\mathcal{A}}(0) = \mathbf{0}_{\mathcal{A}}$  is the probability flux out of  $\mathcal{I} \cup \mathcal{B}$ :

$$p(t)dt = -(\mathbf{1}_{\mathcal{I}} \dot{\mathbf{P}}_{\mathcal{I}} + \mathbf{1}_{\mathcal{B}} \dot{\mathbf{P}}_{\mathcal{B}})dt \equiv -\mathbf{1}_{\mathcal{I} \cup \mathcal{B}} \dot{\mathbf{P}}_{\mathcal{I} \cup \mathcal{B}} dt, \quad (2)$$

where  $\mathbf{1}_{\mathcal{I}}$  is a row vector of ones, dimension  $N_{\mathcal{I}}$ , etc. Writing equation (1) as  $\dot{\mathbf{P}}_{\mathcal{I} \cup \mathcal{B}}(t) = \mathbf{M} \mathbf{P}_{\mathcal{I} \cup \mathcal{B}}(t)$  we obtain a formal solution as  $\mathbf{P}_{\mathcal{I} \cup \mathcal{B}}(t) = \exp(\mathbf{M}t) \mathbf{P}_{\mathcal{I} \cup \mathcal{B}}(0)$ , and hence

$$p(t) = -\mathbf{1}_{\mathcal{I} \cup \mathcal{B}} \mathbf{M} \exp(\mathbf{M}t) \mathbf{P}_{\mathcal{I} \cup \mathcal{B}}(0). \quad (3)$$

Evaluation of  $p(t)$  is typically achieved through eigendecomposition of  $\mathbf{M}$ , whose eigenvalues  $-\nu_l < 0$  are strictly negative for a connected network, giving

$$\mathbf{M} = -\sum_l \nu_l \mathbf{w}_l^R \otimes \mathbf{w}_l^L, \quad (4)$$

where  $\mathbf{w}_l^L$  and  $\mathbf{w}_l^R$  are the left and right eigenvectors and  $\otimes$  is the diadic (outer) product. Using this expansion we obtain  $\exp(\mathbf{M}t) = \sum_l \exp(-\nu_l t) \mathbf{w}_l^R \otimes \mathbf{w}_l^L$  and thus

$$p(t) = \sum_l \nu_l e^{-\nu_l t} \mathbf{1}_{\mathcal{I} \cup \mathcal{B}} [\mathbf{w}_l^R \otimes \mathbf{w}_l^L] \mathbf{P}_{\mathcal{I} \cup \mathcal{B}}(0). \quad (5)$$

where we used the fact that  $\mathbf{w}_l^L \mathbf{w}_l^R = \delta_{lq}$ . We note that (5) is a general expression for the first passage time distribution between two regions  $\mathcal{A}$  and  $\mathcal{B}$ , which is independent of how the remaining state space  $\mathcal{I}$  is partitioned. In addition, we can readily calculate an approximate first passage time distribution using lower rank rate matrices, as shown in section III. The expectation value for any power of the first passage time  $\langle t^n \rangle = \int_0^\infty t^n p(t) dt$  is given by

$$\langle t^n \rangle = \sum_l \frac{n!}{\nu_l^n} \mathbf{1}_{\mathcal{I} \cup \mathcal{B}} [\mathbf{w}_l^R \otimes \mathbf{w}_l^L] \mathbf{P}_{\mathcal{I} \cup \mathcal{B}}(0). \quad (6)$$

As the eigenbasis of  $\mathbf{M}$  is complete,  $\sum_l \mathbf{w}_l^R \otimes \mathbf{w}_l^L = \mathbb{I}$ , the distribution is normalised:

$$\begin{aligned} \int_0^\infty p(t) dt &= \sum_l \mathbf{1}_{\mathcal{I} \cup \mathcal{B}} [\mathbf{w}_l^R \otimes \mathbf{w}_l^L] \mathbf{P}_{\mathcal{I} \cup \mathcal{B}}(0) \\ &= \mathbf{1}_{\mathcal{I} \cup \mathcal{B}} \mathbf{P}_{\mathcal{I} \cup \mathcal{B}}(0) = 1, \end{aligned} \quad (7)$$

If there are no direct connections between the product and reactant regions  $\mathcal{A}$  and  $\mathcal{B}$  then  $p(0) = 0$  because the probability flux out of  $\mathcal{B}$  is zero, and the probability distribution must exhibit a maximum value. Furthermore, in the limit of single exponential decay the mean and standard deviation are the same.<sup>63</sup> Since typical methods for computation of the eigenspectrum of a transition network comprising  $N$  states scale as  $\mathcal{O}(N^3)$ ,<sup>39</sup> calculation of the moments of the FPT distribution via Eq. 6 is likely more efficient than using a transfer matrix algorithm,<sup>63-65</sup> where the time complexity is strongly dependent on the average connectivity of nodes and the maximum path length for which  $p(t)$  converges. Although KTNs are typically sparse,<sup>66</sup> flickering of trajectories within metastable groups of nodes can lead to unfeasibly long path lengths.<sup>67</sup>

The detailed balance condition for the individual minimum-to-minimum rate constants,  $\pi_\beta k_{\beta\alpha} = \pi_\alpha k_{\alpha\beta}$ , enables us to define a symmetrised rate matrix<sup>68</sup>

$$\tilde{M}_{\alpha\beta} = \sqrt{\frac{\pi_\beta}{\pi_\alpha}} M_{\alpha\beta}, \quad (8)$$

which has the same eigenvalues as  $\mathbf{M}$ . Here,  $\pi_\alpha$  is the equilibrium occupation probability of minimum  $\alpha$ . The eigenvectors defined by  $\widetilde{\mathbf{M}}\mathbf{u}_l = \lambda_l\mathbf{u}_l$  are related to the left and right eigenvectors of  $\mathbf{M}$  as  $[\mathbf{w}_l^R]_j = [\mathbf{u}_l]_j\sqrt{\pi_j}$  and  $[\mathbf{w}_l^L]_j = [\mathbf{u}_l]_j/\sqrt{\pi_j}$ . We obtained equivalent numerical results using  $\widetilde{\mathbf{M}}$  and  $\mathbf{M}$ , except that the symmetrised implementation generally allows for precision to be maintained to slightly lower temperatures, as expected.<sup>39</sup> In the present work we employed LAPACK routines<sup>69</sup> DSYEVR and DGEEV to treat the symmetrised and unsymmetrised problems, respectively. We also compared the Implicitly Restarted Lanczos Method implementation in ARPACK<sup>70</sup> for the symmetrised formulation. All three methods agreed very well for higher temperatures, but diverged at lower temperatures where precision is lost because of the long time scales involved. The lowest temperatures that can be treated with GT-based methods lie far beyond the limits of applicability for eigendecomposition, whichever numerical method we employed. We have also investigated the use of higher precision numerical schemes, such as quadruple floating point precision. Whilst it is then possible to reach slightly lower temperatures, the improvement in stability is limited. Exponentially higher accuracy is required as the temperature decreases, which quickly overwhelms the additional number of significant digits that can be retained.

We note that a recent study<sup>44</sup> has shown the maximum size (rank) of symmetrized rate matrices for which eigendecomposition is possible can be extended by using direct or iterative shared memory solvers. Whilst the low temperature spectral properties that lead to numerical instabilities remain, it is possible that such approaches could further extend the range of temperatures for which eigendecomposition can be applied. The eigendecomposition formulation can provide insight into how observable properties are encoded in the underlying energy landscape. We focus on the mean first passage time  $\mathcal{T}_{AB} \equiv \langle t \rangle$  from equation (6), the standard deviation  $\sigma_{AB} = \sqrt{\langle t^2 \rangle - \langle t \rangle^2}$ , and rate formulations based on discrete path sampling<sup>4,5</sup> and the graph transformation approach:<sup>37,40,41,71</sup>

$$\begin{aligned} k_{A\leftarrow B}^{SS} &\equiv \sum_{b \in \mathcal{B}} \frac{C_b^A \widehat{\pi}_b}{\tau_b}, & k_{A\leftarrow B}^{NSS} &\equiv \sum_{b \in \mathcal{B}} \frac{C_b^A \widehat{\pi}_b}{\tau_b^{\mathcal{I}}}, \\ k_{A\leftarrow B}^F &\equiv \sum_{b \in \mathcal{B}} \frac{\widehat{\pi}_b}{\mathcal{T}_{Ab}}, & k_{A\leftarrow B}^* &= \frac{1}{\langle \mathcal{T}_{Ab} \rangle_{\mathcal{B}}} = \frac{1}{\mathcal{T}_{AB}}, \end{aligned} \quad (9)$$

where  $C_b^A$  is the committor probability<sup>72</sup> that a trajectory leaves minimum  $b \in \mathcal{B}$  and reaches  $\mathcal{A}$  before returning to  $\mathcal{B}$ , and  $\tau_b = 1/[\mathbf{D}_{\mathcal{B}}]_{bb}$  is the escape time from  $b$  to any directly connected minimum. The hierarchy with degree of approximation of these rate expressions is  $k_{A\leftarrow B}^{SS}$ ,  $k_{A\leftarrow B}^{NSS}$  then  $k_{A\leftarrow B}^F$ . The  $k_{A\leftarrow B}^{SS}$  rate assumes all minima in the intervening region  $\mathcal{I}$  are treated in a steady state approximation, and we assume a local equilibrium distribution,  $\widehat{\pi}_{\mathcal{B}}$  in the reactant region  $\mathcal{B}$ , where  $\widehat{\pi}_b = \pi_b/\Pi_{\mathcal{B}}$  and  $\Pi_{\mathcal{B}} = \sum_{b \in \mathcal{B}} \pi_b$ .  $k_{A\leftarrow B}^{SS}$  neglects contributions to the first passage time from the  $\mathcal{I}$  region, which can lead to a significant systematic error.  $k_{A\leftarrow B}^{NSS}$  improves on this assumption by accounting for contributions from  $\mathcal{I}$  via graph transformation, with  $\tau_b^{\mathcal{I}}$  the renormalised escape time from  $b$  to any member of  $\mathcal{A} \cup \mathcal{B}$ .  $k_{A\leftarrow B}^F$  better accounts for transitions between  $\mathcal{I}$ ,  $\mathcal{B}$  during a reactive pathway by using graph transformation to obtain the renormalised escape time  $\tau_b^{\text{GT}}$  from  $b$  to  $b \cup \mathcal{A}$ , with branching probabilities  $P_{Ab}^{\text{GT}} + P_{bb}^{\text{GT}} = 1$  and hence a mean first passage time to  $\mathcal{A}$  of  $\mathcal{T}_{Ab} = \tau_b^{\text{GT}}/P_{Ab}^{\text{GT}}$ . A time independent reaction rate is only well defined when decay to the product region is a simple exponential decay, and thus is proportional to the inverse of the mean first passage time.<sup>42</sup> Indeed, the rate  $k_{A\leftarrow B}^*$  is defined as the reciprocal of the mean first passage time. Outside of this regime, the first passage time distribution should really be used to study transition kinetics.

In previous work we have shown that  $k_{A\leftarrow B}^F$ ,  $k_{A\leftarrow B}^*$  and  $k_{A\leftarrow B}^{NSS}$ , but not  $k_{A\leftarrow B}^{SS}$ , all agree in the metastable limit of single exponential kinetics for a suitable observation time scale, which depends on a spectral gap condition for the rate matrix.<sup>42</sup> Comparing these various rate constants thus provides a measure of how well we expect a time independent reaction rate to describe the kinetics.

We will label the MFPT results computed by eigendecomposition and graph transformation as  $\mathcal{T}_{AB}^{\text{ED}}$  and  $\mathcal{T}_{AB}^{\text{GT}}$  when we wish to distinguish them. They are the same for calculations that achieve proper precision, as we confirm below.

The formulation in terms of the first passage time distribution enables us to understand how a finite kinetic transition matrix can provide an accurate description of kinetics. Conventional simulation techniques, such as Monte Carlo and molecular dynamics, explore a tiny fraction of phase space, but they can provide useful estimates of observable properties if they sample probability distributions, such as the energy density of states, correctly.<sup>73</sup> Equation (6) tells us that we will obtain the correct first passage time moments, and hence the correct rate, if the underlying database of minima and transition states reproduces the correct form for  $p(t)$ . To calculate rates, we therefore need to converge  $p(t)$  with respect to the underlying finite network.

For strongly metastable dynamics, typical in the low-temperature regime, calculation of the eigenspectrum becomes numerically unstable.<sup>39</sup> It is possible

A more robust approach to calculating  $p(t)$  is to perform explicit kinetic path sampling (kPS) simulations using graph transformation<sup>37,38,40</sup> to treat escape from a source subnetwork.<sup>43,44</sup> An iterative reverse randomisation procedure is



used to sample the number of moves along each edge of the incrementally reconstructed subnetwork, and hence a time for the escape trajectory can be obtained.<sup>43,44</sup> This approach extends the leapfrog scheme that was introduced to treat multiple transitions between minima connected by low barriers in kinetic Monte Carlo simulations,<sup>71</sup> which led to the graph transformation procedure.<sup>37,40,71</sup> kPS does not produce detailed trajectories on the original network, but does yield non-Markovian trajectories for the coarse-grained network defined by the subnetworks. Furthermore, kPS is exact in sampling the solution to the linear master equation,<sup>43,44</sup> and therefore yields the correct  $\mathcal{A} \leftrightarrow \mathcal{B}$  first passage time distributions. Owing to the graph transformation component, the efficiency of kPS is essentially unaffected by the presence of metastable states, provided that the subnetworks for the kPS simulations appropriately characterise the metastable sets of nodes.<sup>67</sup> Hence, unlike standard kinetic Monte Carlo methods, kPS remains efficient even at low temperatures.<sup>43,71</sup> Moreover, unlike kMC methods based on eigendecomposition, such as the Monte Carlo with Absorbing Markov Chains (MCAMC) algorithm,<sup>74-76</sup> kPS also remains numerically stable in the metastable regime.<sup>38</sup>

In the present work, the kinetic transition networks are of sufficiently low dimensionality that the subnetwork of each kPS iteration can simply be chosen as  $\mathcal{B} \cup \mathcal{I}$ , and the endpoint macrostate  $\mathcal{A}$  then corresponds to the set of absorbing nodes for each kPS iteration.  $\mathcal{A} \leftarrow \mathcal{B}$  transition paths are thus generated directly in a single kPS step. This simulation setup avoids the complicated issue of partitioning the state space in a way that accurately captures the metastable macrostates, which strongly affects the efficiency of the kPS simulation.<sup>67</sup> However, trajectory information, including the calculation of committor probabilities, is forfeited, since the *order* of internode transitions in the network is not computed, only the *number* of kMC moves.<sup>43</sup> The FPT distributions obtained from kPS simulations can be converged to lower temperatures than those obtained from eigendecomposition, described above, because graph transformation maintains numerical precision.

All the results that follow were obtained after removing any reactant minima without connections to products, and normalising the restricted equilibrium distribution  $\hat{\pi}_{\mathcal{B}}$  or  $\hat{\pi}_{\mathcal{A}}$  normalised over the remaining reactant states for  $\mathcal{A} \leftarrow \mathcal{B}$  or  $\mathcal{B} \leftarrow \mathcal{A}$ , respectively.

### III. RESULTS FOR MODEL LANDSCAPES

We first compared first passage time distributions and rates for some model landscapes where the local minima, transition states, and associated densities of states were constructed to test the theoretical framework. All the corresponding results are in a consistent set of arbitrary reduced units. For illustration we will summarise some results for two examples with eight minima that differ only in the connectivity in the  $\mathcal{A}$  region. Each landscape has two  $\mathcal{A}$ , two  $\mathcal{B}$ , and four  $\mathcal{I}$  minima, which is sufficient to illustrate some key features. The disconnectivity graphs, rate constants,  $k_{\mathcal{A} \leftrightarrow \mathcal{B}}$ , and mean first passage times,  $\mathcal{T}_{\mathcal{A}\mathcal{B}}$ , are shown in Figure 1, with first passage time moments and probability distributions in Figure 2. Panels (e) and (f) of Figure 1 confirm that graph transformation produces the same results as eigendecomposition from equation (6) for  $\mathcal{T}_{\mathcal{A}\mathcal{B}}$  in both the  $\mathcal{A} \leftarrow \mathcal{B}$  and  $\mathcal{B} \leftarrow \mathcal{A}$  directions. For the first landscape in Figure 1a, the rate constants calculated from graph transformation,  $k^F$ , and from the eigendecomposition,  $k^* = 1/\mathcal{T}_{\mathcal{A}\mathcal{B}}$  from equation (6), also agree quantitatively. They agree for the  $\mathcal{A} \leftarrow \mathcal{B}$  direction in the second landscape in Figure 1b, but increasingly deviate at low temperature for direction  $\mathcal{B} \leftarrow \mathcal{A}$ , due to the different averaging procedures.

These results show that the alternative rate formulations are equivalent when the reactant region is sufficiently metastable, with a well-defined spectral gap. In this limit, single exponential kinetics are rapidly established so that  $1/\mathcal{T}_{\mathcal{A}\mathcal{B}} = \langle 1/\mathcal{T}_{\mathcal{A}b} \rangle_{\mathcal{B}}$ . The deviation observed for  $k_{\mathcal{B} \leftarrow \mathcal{A}}^F$  and  $k_{\mathcal{B} \leftarrow \mathcal{A}}^*$  arises from the structure in the reactant region  $\mathcal{A}$  for the landscape in Figure 1b. Here there is a significant barrier within the reactant region, and so the kinetics are better described as multi-exponential.

The probability distributions for the first passage time, illustrated for selected high temperatures in panels (aii), (bii), (aiii) and (biii) of Figure 2, are similar for the two landscapes. In each case there is a well-defined maximum in the distribution, and then a slow decay. The mean and standard deviation are shown in Figure 2 (ai) and (bi) as a function of inverse temperature. Here the departure from single exponential kinetics at lower temperatures for the landscape in Figure 2b is reflected in the increasing deviation of  $\sigma$  and  $\langle t \rangle$ .

### IV. RESULTS FOR LJ<sub>38</sub>

The second example application is a standard benchmark case, namely the 38-atom cluster LJ<sub>38</sub> bound by the Lennard-Jones potential:<sup>79</sup>

$$V = 4 \sum_{i < j} \epsilon \left[ \left( \frac{\sigma}{r_{ij}} \right)^{12} - \left( \frac{\sigma}{r_{ij}} \right)^6 \right], \quad (10)$$

This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.  
PLEASE CITE THIS ARTICLE AS DOI:10.1063/1.50016244

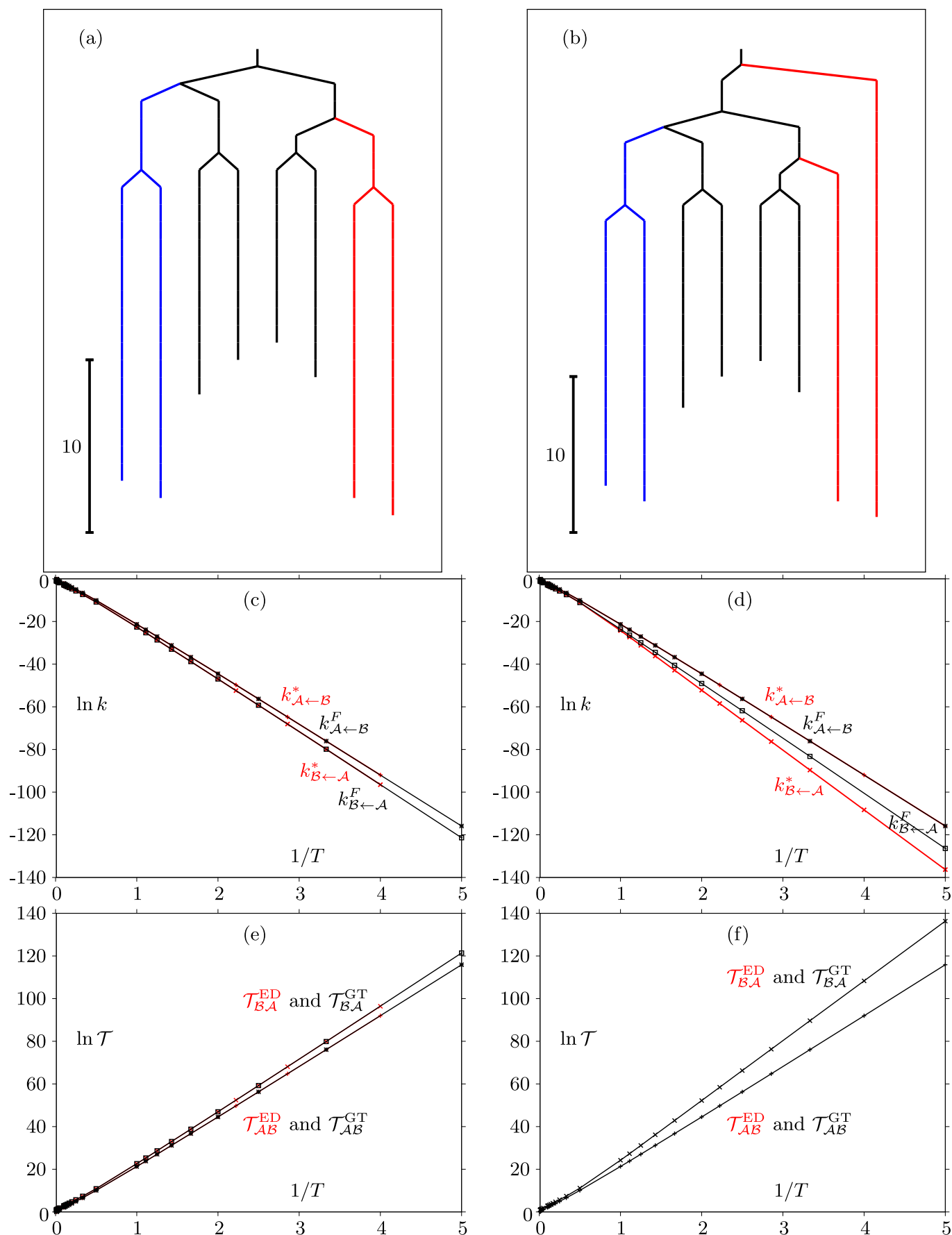


FIG. 1. Results for two model landscapes containing eight local minima. (a) and (b) are the disconnectivity graphs<sup>77,78</sup> with the branches for the two  $\mathcal{A}$  and two  $\mathcal{B}$  minima coloured red and blue, respectively. The scale bars indicate energy in arbitrary units. (c) and (d) show the rate constants obtained from graph transformation,  $k^F$ , and from  $k^*$  via equation (6). (e) and (f) show the corresponding mean first passage times,  $\mathcal{T}_{AB}^{GT}$  and  $\mathcal{T}_{AB}^{ED}$ , which are indistinguishable in the range where eigendecomposition retains precision, up to  $1/T = 4$  in (e) and  $1/T = 5$  in (f).

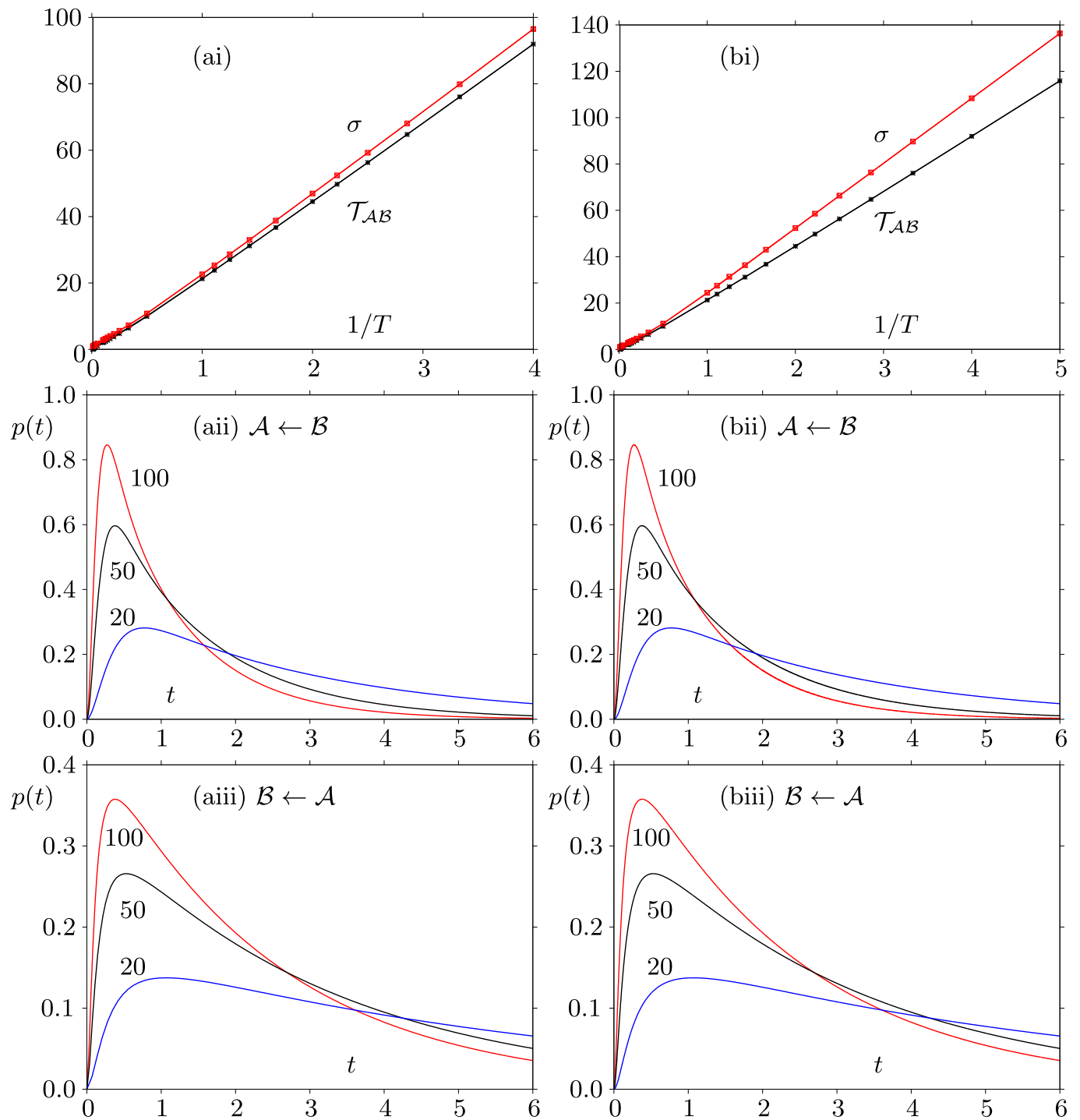


FIG. 2. Statistics of first passage times for the two model landscapes illustrated in Figure 1 panels (a) and (b), which correspond to the results on the left and right in this figure, respectively. (ai) and (bi) show the mean,  $\langle t \rangle$  (black), and standard deviation,  $\sigma$  (red), of the first passage time. The results for directions  $\mathcal{A} \leftarrow \mathcal{B}$  and  $\mathcal{B} \leftarrow \mathcal{A}$  are indistinguishable in these plots. (aii), (bii), (aiii) and (biii) are the probability distribution functions calculated at temperatures of 100, 50, and 20 ( $k_B T$  in energy units, as marked). (aii) and (bii) are for direction  $\mathcal{A} \leftarrow \mathcal{B}$ ; (aiii) and (biii) are for  $\mathcal{B} \leftarrow \mathcal{A}$ . The  $p(t)$  distributions are very similar for the two landscapes, although  $\sigma$  grows significantly faster as the temperature falls in panel (bi).



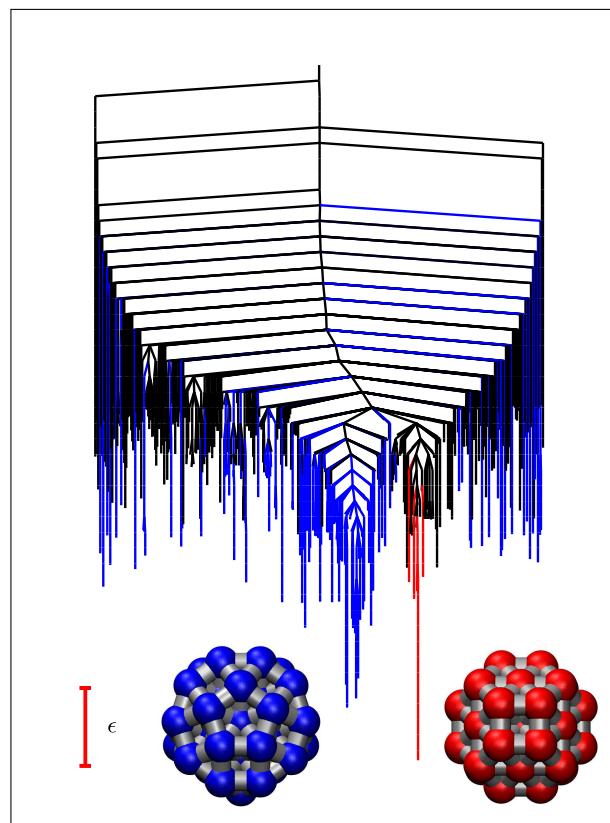


FIG. 3. Disconnectivity graph<sup>77,78</sup> for an LJ<sub>38</sub> database containing 4000 minima with the branches for  $\mathcal{B}$  (icosahedral) and  $\mathcal{A}$  (cuboctahedral) minima coloured blue and red, respectively. The global minimum and the lowest minimum based on icosahedral packing are illustrated close to their corresponding branches.

where  $\epsilon$  and  $2^{1/6}\sigma_{ij}$  are the pair equilibrium well depth and separation for atoms  $i$  and  $j$ , respectively, and  $r_{ij}$  is the corresponding interatomic separation. This system has two competing morphologies that define a double funnel landscape (Figure 3),<sup>78,80–82</sup> which has made it a useful benchmark for global optimisation,<sup>83–86</sup> enhanced thermodynamic sampling to overcome broken ergodicity,<sup>6,81,83,87–95</sup> and rare event dynamics.<sup>4,5,66,81,96–98</sup> This system is amenable to detailed analysis, but it is complex enough to pose challenging problems if we attempt to simplify the configuration space by defining projections such as reaction coordinates.<sup>99</sup>

In treating the rates and thermodynamics of this system we normally consider a set of five minima for the funnel containing the global minimum truncated octahedron, point group  $O_h$ , and 395 minima for the funnel based on an incomplete Mackay icosahedron, which has point group  $C_{5v}$ .<sup>78,80–82</sup> These groups were taken as the  $\mathcal{A}$  and  $\mathcal{B}$  regions for specificity. A large kinetic transition network has been calculated for this system in previous work, and a subset containing 100,000 minima is available from the Cambridge Landscape Database.<sup>100</sup> However, the low temperature interconversion rates of interest converge for a much smaller sample, as shown in Figure 4a and b. Here, all the results correspond to partition functions based on harmonic vibrational densities of states, with consistent minimum-to-minimum rate constants obtained from transition state theory.<sup>7,8</sup>

The rate constants corresponding to  $k^* = 1/\mathcal{T}_{AB}$  from equation (6) agree very well with the  $k^F$  values from graph transformation<sup>37,40,41,71</sup> in the higher temperature range where the eigendecomposition gives meaningful results (Figure 4c and d). For reference, the two morphologies have approximately the same free energy at a reduced temperature around  $k_B T/\epsilon = 0.1$ .  $k^{NSS}$  and  $k^{SS}$  exhibit systematic deviations, especially at high temperature.

The results for the first passage time and standard deviation as a function of temperature are shown in Figure 5 for the database containing 4000 minima and 2677 transition states. The connected component is a subset of 900 minima, and all the subsequent calculations were conducted for this subset. The behaviour is much more complicated than for the simple model landscapes in §III. This structure reflects the detailed organisation of the underlying energy landscape, particularly the presence of internal barriers within the  $\mathcal{A}$  and  $\mathcal{B}$  regions. Although these regions are defined in terms of local minima, based on the expected experimental diagnostics for the two morphologies, the internal dynamics within each set still reflect the structure of the landscape in each region. It is particularly noteworthy that this complexity is not readily apparent in the Arrhenius plots of the rate constants. The first passage

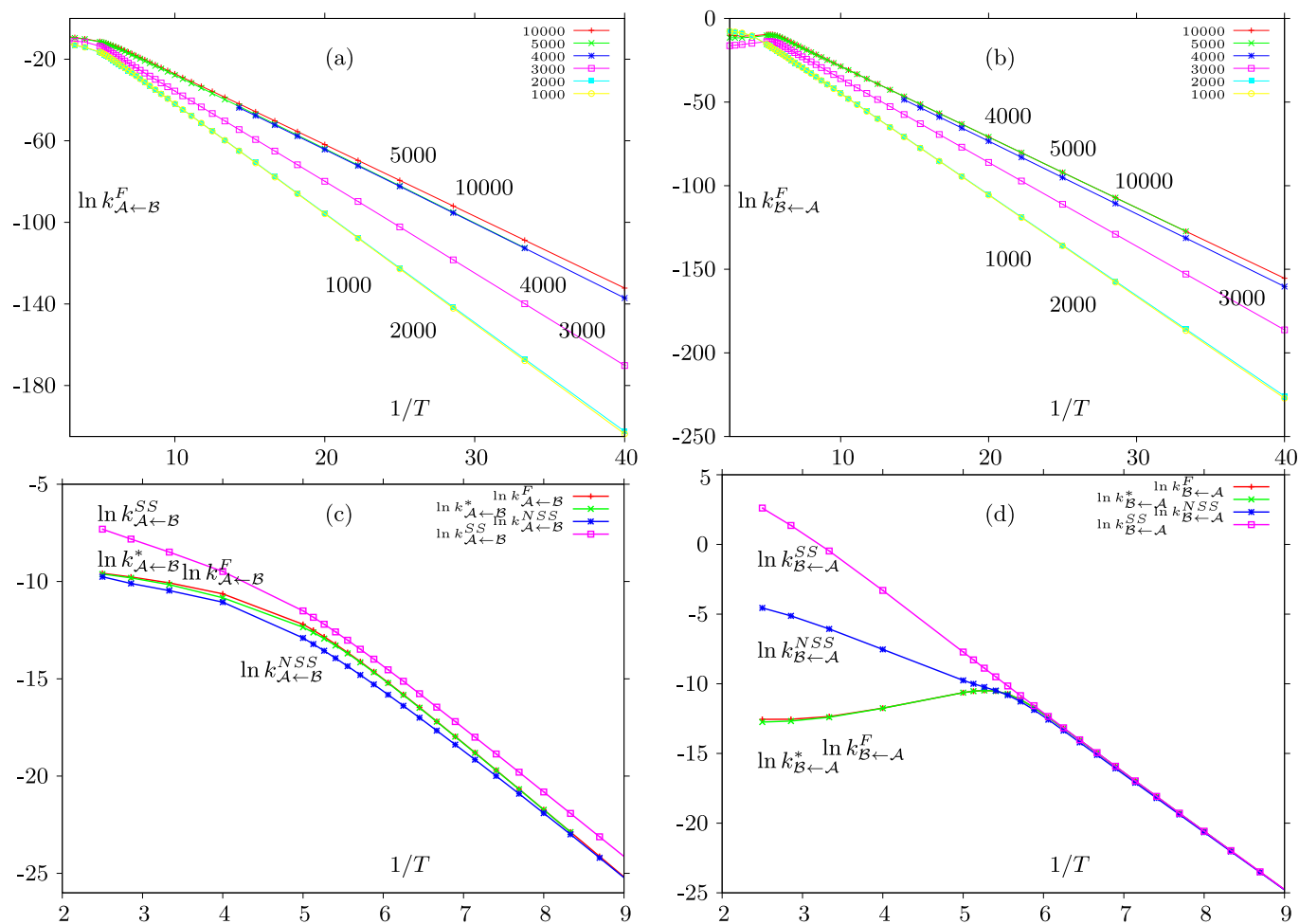


FIG. 4. Results for the LJ<sub>38</sub> cluster. (a) and (b) show the rate constants obtained from graph transformation,  $k_{\mathcal{A} \leftarrow \mathcal{B}}^F$  and  $k_{\mathcal{B} \leftarrow \mathcal{A}}^F$ , for databases with 1000, 2000, 3000, 4000, 5000, and 10,000 minima, as marked, which contain 345, 893, 1481, 2677, 3963, and 10193 transition states, respectively. The results for 1000 and 2000 minima are not distinguishable. There are large increases between 2000 and 3000, and between 3000 and 4000 minima, as more kinetically favourable pathways are progressively included. (c) and (d) compare the  $\mathcal{A} \leftarrow \mathcal{B}$  and  $\mathcal{B} \leftarrow \mathcal{A}$  rate constants for the 4000 minima database.  $k^*$  and  $k^F$  agree closely in each case, while  $k^{NSS}$  and  $k^{SS}$  exhibit systematic deviations, especially at high temperature.

time distribution and its moments are more sensitive to deviations from single exponential behaviour, which are also apparent in projections based on structural order parameters.<sup>99</sup> This behaviour once again illustrates how useful the LJ<sub>38</sub> cluster can be as a benchmark for analysing and understanding the relations between structure, dynamics, and thermodynamics, and how these observables are encoded in the underlying potential energy surface.

The probability distribution for the first passage time at  $k_B T / \epsilon = 0.125$  is shown in Figure 7 for directions  $\mathcal{A} \leftarrow \mathcal{B}$  and  $\mathcal{B} \leftarrow \mathcal{A}$ . The results for two initial probability distributions are compared on a log-log scale. When  $\mathbf{P}_{\mathcal{I} \cup \mathcal{B}}(0)$  is localised in a single high energy minimum we see that  $p(t)$  exhibits two time scales, whereas the time dependence is smooth for an initial restricted equilibrium distribution in the  $\mathcal{B}$  or  $\mathcal{A}$  region. The two time scales observed for relaxation from a high energy minimum correspond to probability that reaches the absorbing product region directly, and a slower component that first relaxes to the reactant, and then escapes to product on a longer time scale. The two relaxation time scales are characteristic of a double funnel landscape.<sup>78,80–82</sup>

Equation (6) can be used to examine the breakdown of contributions to the MFPT from the eigencomponents of the rate matrix. In the low temperature range of interest, we find that one mode, denoted  $d$ , dominates the sum, where the product  $[\mathbf{1}_{\mathcal{I} \cup \mathcal{B}} \mathbf{w}_d^R] [\mathbf{w}_d^L \mathbf{P}_{\mathcal{I} \cup \mathcal{B}}(0)]$  or  $[\mathbf{1}_{\mathcal{I} \cup \mathcal{A}} \mathbf{w}_d^R] [\mathbf{w}_d^L \mathbf{P}_{\mathcal{I} \cup \mathcal{A}}(0)]$  is close to unity. For the LJ<sub>38</sub> test case there are a number of modes corresponding to slower relaxation than mode  $d$ , but their contributions are many orders of magnitude smaller. These modes are all localised on single minima in the intervening region, except for one contribution in the  $\mathcal{A} \leftarrow \mathcal{B}$  direction, which is localised on two higher energy minima in the  $\mathcal{B}$  icosahedral set. We can understand these results by considering how the contributions to the first passage time arise as the initial normalised

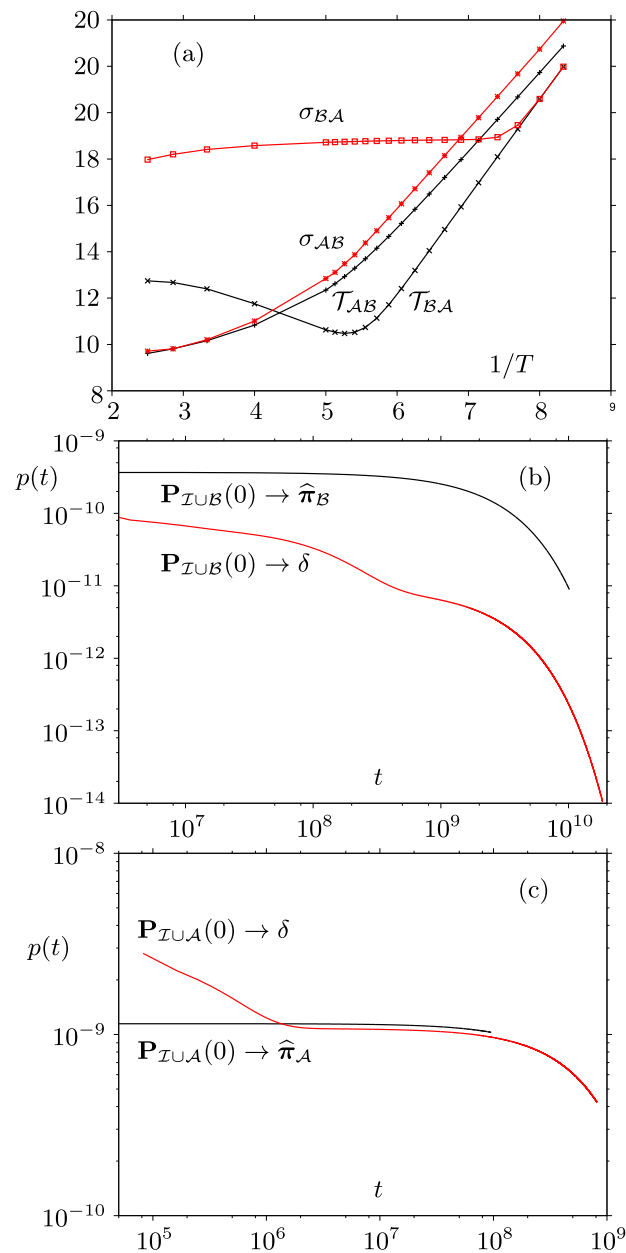


FIG. 5. (a) shows the mean,  $\mathcal{T}^{\text{ED}}$  (black), and standard deviation,  $\sigma$  (red), of the first passage time for directions  $\mathcal{A} \leftarrow \mathcal{B}$  and  $\mathcal{B} \leftarrow \mathcal{A}$ , as marked. (b) and (c) show the probability distribution function for the first passage time,  $p(t)$ , for directions  $\mathcal{A} \leftarrow \mathcal{B}$  and  $\mathcal{B} \leftarrow \mathcal{A}$ , respectively, at a reduced temperature of  $k_B T/\epsilon = 0.125$ . Two initial distributions are considered, one with all the probability localised in the highest energy minimum of the reactant set, denoted by  $\mathbf{P}_{\mathcal{I} \cup \mathcal{B}}(0) \rightarrow \delta$  (red), and the other with a normalised local equilibrium distribution for all the connected  $\mathcal{B}$  or  $\mathcal{A}$  minima, denoted  $\mathbf{P}_{\mathcal{I} \cup \mathcal{B}}(0) \rightarrow \hat{\pi}_{\mathcal{B}}$  and  $\mathbf{P}_{\mathcal{I} \cup \mathcal{A}}(0) \rightarrow \hat{\pi}_{\mathcal{A}}$  (black).

probability distribution drains away to the absorbing boundary. The eigenvectors with slower relaxation than the  $d$  component involve one or more minima separated from the rest by high barriers, where the initial occupation probability at  $t = 0$  is zero or small. The system relaxes without ever populating these states significantly, so that equation (6) corresponds closely to single exponential decay.

As shown in Figure 6, evaluation of  $p(t)$  using eigendecomposition and equation (5) is numerically unstable at low temperatures, because the master equation matrix becomes ill-conditioned.<sup>38,39</sup> To access  $p(t)$  in this regime, we use kinetic path sampling (kPS), a stochastic method based on local graph transformations to sample regional escape time distributions, which are then “undone” by an iterative reverse randomisation procedure, to allow efficient sampling of the exact escape time distribution.<sup>43,44</sup> Numerical instability is avoided through use of the graph transformation

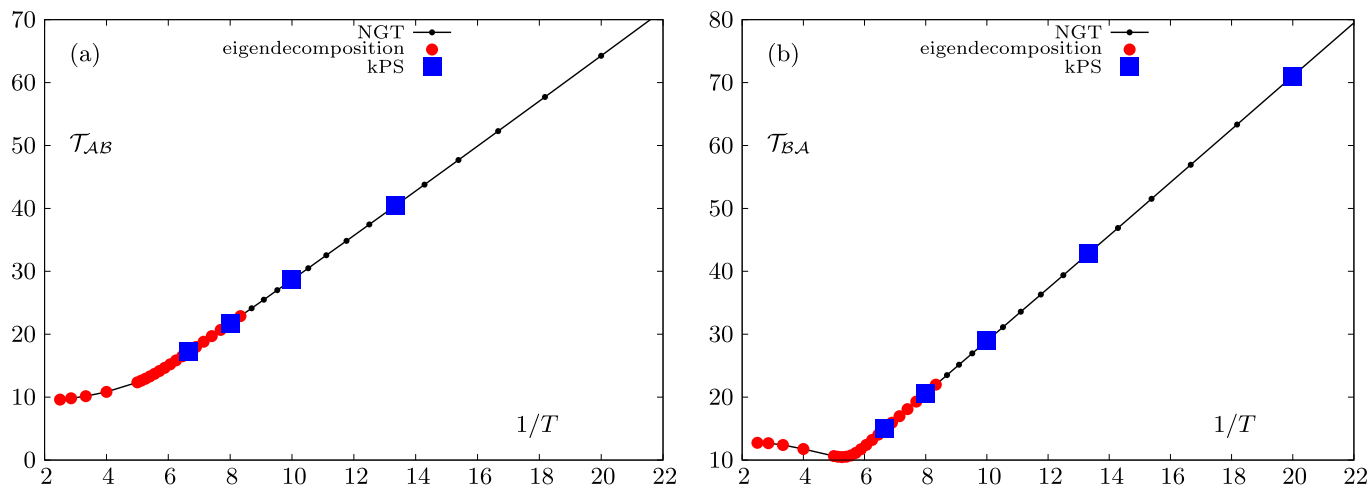


FIG. 6. Mean first passage times for the LJ<sub>38</sub> database containing 4000 minima and 2677 transition states calculated using kinetic path sampling (blue squares) compared to eigendecomposition (red circles), and graph transformation (small black circles). The estimates from kPS simulations were obtained from 20000 transition paths. (a)  $\mathcal{A} \leftarrow \mathcal{B}$ . (b)  $\mathcal{B} \leftarrow \mathcal{A}$ .

method,<sup>38</sup> although the sampled FPT distribution data requires monitoring for convergence.

The mean first passage time estimated from kPS simulations is in excellent agreement with the exact graph transformation result and eigendecomposition in suitable temperature regimes (Figure 6). Figure 7 demonstrates that the sampled escape times agree very well with the exact  $p(t)$ , although there is consistent insufficient sampling of the rare, long escape times. The tails of the FPT distributions are associated with low probability, but they make a significant contribution to the MFPT because the values are extremal.<sup>67</sup> This undersampling leads to significant error when estimating higher moments of the escape time, as  $\langle t^n \rangle = \int t^n p(t) dt$  is clearly more sensitive to large but rare values of  $t$ . These errors can be systematically reduced through increased computational effort. For more complicated systems, however, the ability to converge higher moments of the first passage time in kMC simulations may be limited in practice by the availability of computational resources. The kPS simulation at the lowest temperature considered ( $T = 0.05$ ) fails in one direction, because the number of kMC moves along particular edges in the transition network becomes unfeasibly large.

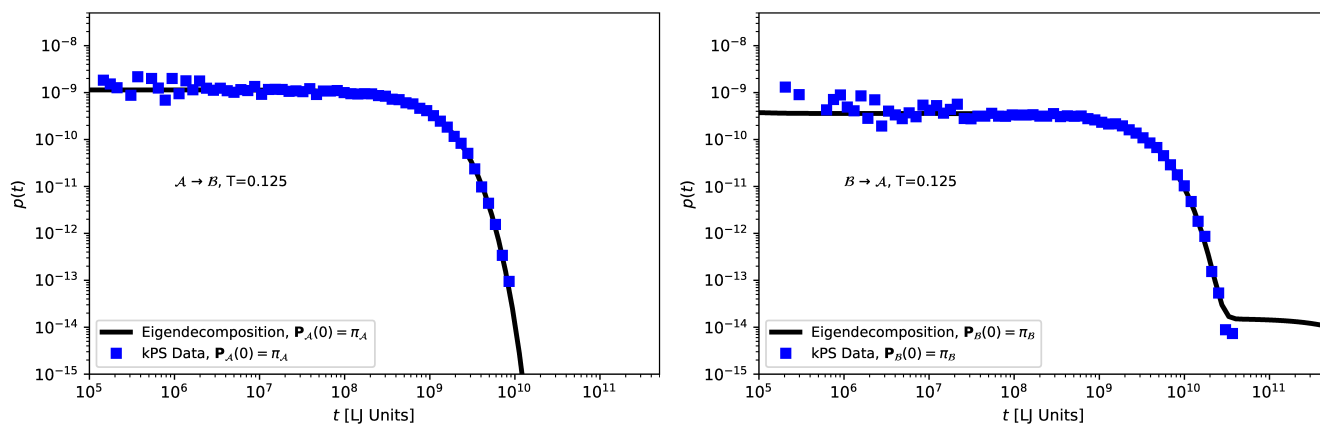


FIG. 7. The exact  $p(t)$  between basins of LJ<sub>38</sub> at  $T = 0.125$ , as determined by eigendecomposition (black line), is compared with a density histogram of data obtained using 20000 transition paths simulated by kinetic path sampling (blue squares). As shown in figure 6, the mean first passage times obtained by kPS are in excellent agreement with eigendecomposition and NGT methods. Undersampling of less probable passage times leads to some error in higher moments, which can be systematically reduced with increased sampling effort.

Once again, the LJ<sub>38</sub> cluster provides us with a tractable benchmark that covers a wide range of behaviour, which can also be found in diverse systems of interest in molecular and condensed matter science. It is particularly noteworthy that the Arrhenius plots for the phenomenological rate constant appear relatively featureless, while detailed analysis

of the first passage time distributions explains how this behaviour is encoded in the underlying potential energy landscape.

These observations are considered again below, where we compare different coarse-graining procedures for the same network, with the objective of pushing the first passage time distribution calculations down to lower temperatures without losing precision. The regrouping scheme yields transition networks that are not only lower dimensional, but are also significantly better conditioned, which therefore (at least partially) alleviates the problem of flickering trajectories in kMC simulations.<sup>43,67</sup>

## V. RESULTS FOR REGROUPED NETWORKS

The eigendecomposition of the rate matrix, which enables us to calculate the probability distribution  $p(t)$  and the moments of the first passage time via equations (5) and (6), provides direct insight into the dynamical properties of a kinetic transition network, and hence allows quantitative assessment of the sampling convergence of the underlying energy landscape. However, in practice, numerical imprecision limits the application of this scheme to high temperatures and small databases.<sup>4,5,38,39</sup> We therefore examined whether the underlying database could be coarse-grained via a suitable regrouping scheme to extend the range of applicability. A number of lumping schemes have been discussed in the literature,<sup>49–57,101</sup> and community detection algorithms<sup>102–110</sup> could also be employed to partition a transition network. However, care must be taken that the resulting groups of nodes accurately reflect the metastable macrostates; otherwise the lumped transition network may not appropriately represent the dynamical properties of the original network. The efficiency of kMC simulation algorithms based on a partitioning of the network, including kPS, is strongly affected by the extent to which the groups of nodes characterise the metastable macrostates.<sup>67</sup>

In this section, we focus on the macrostates, or clusters, determined using a free energy-based regrouping scheme, described in §VA. In §VB we compare the ability of three different intergroup rate expressions, namely the local equilibrium approximation<sup>60</sup> and two versions of the Hummer-Szabo formulation,<sup>101,111</sup> to reproduce the kinetics of interest for the same clustering of the landscape. Then in §VC we introduce a new approximation to intergroup dynamics that exploits a spectral decomposition of each cluster to generalize the local equilibrium approximation. We note that regrouping based on the steady state approximation<sup>112</sup> yields the same form for the resulting rate equations as the steady state assumption for reactant and product regions in the original discrete path sampling formulation.<sup>4,5</sup> A detailed comparison of alternative lumping schemes will be reported elsewhere, including extensive tests of the spectral clustering scheme.

### A. Free Energy Regrouping Using the Local Equilibrium Approximation

We first considered the model landscape shown in Figure 1a and applied a self-consistent regrouping scheme based upon free energy barriers.<sup>60</sup> In this procedure a regrouping threshold  $\Delta G^{\text{RG}}$  is selected and local free energy minima are grouped together if they are separated by free energy barriers below the threshold. The regrouping procedure is applied iteratively to the new groups that result, until no further changes occur, producing occupation probabilities and free energies for a group of minima,  $J$ , as<sup>60</sup>

$$\Pi_J(T) = \sum_{j \in J} \pi_j(T) \quad \text{and} \quad F_J(T) = -k_B T \ln \sum_{j \in J} Z_j(T),$$

where  $Z_j(T)$  is the canonical partition function for local minimum  $j$ . The corresponding free energy and intergroup rate constants for the transition state connecting groups  $J$  and  $L$  in this scheme are simply<sup>60</sup>

$$F_{LJ}(T) = -k_B T \ln \sum_{(lj)^\ddagger} Z_{lj}^\ddagger(T), \quad l \in L, j \in J,$$

$$\text{and} \quad k_{LJ}^\ddagger(T) = \sum_{(lj)^\ddagger} \frac{\pi_j(T)}{\Pi_J(T)} k_{lj}(T), \quad (11)$$

where  $Z_{lj}^\ddagger(T)$  is the partition function for the transition state connecting local minimum  $j$  to  $l$ , and we employ the usual transition state theory formulation<sup>7,8</sup> throughout in the numerical calculations, using harmonic vibrational densities of states.

Although the expression for the intergroup rates in equation (11) assumes the local equilibrium approximation (LEA), the regrouping scheme itself can be used with alternative formulations for the intergroup rates. This lumping

This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.  
PLEASE CITE THIS ARTICLE AS DOI:10.1063/1.50016244

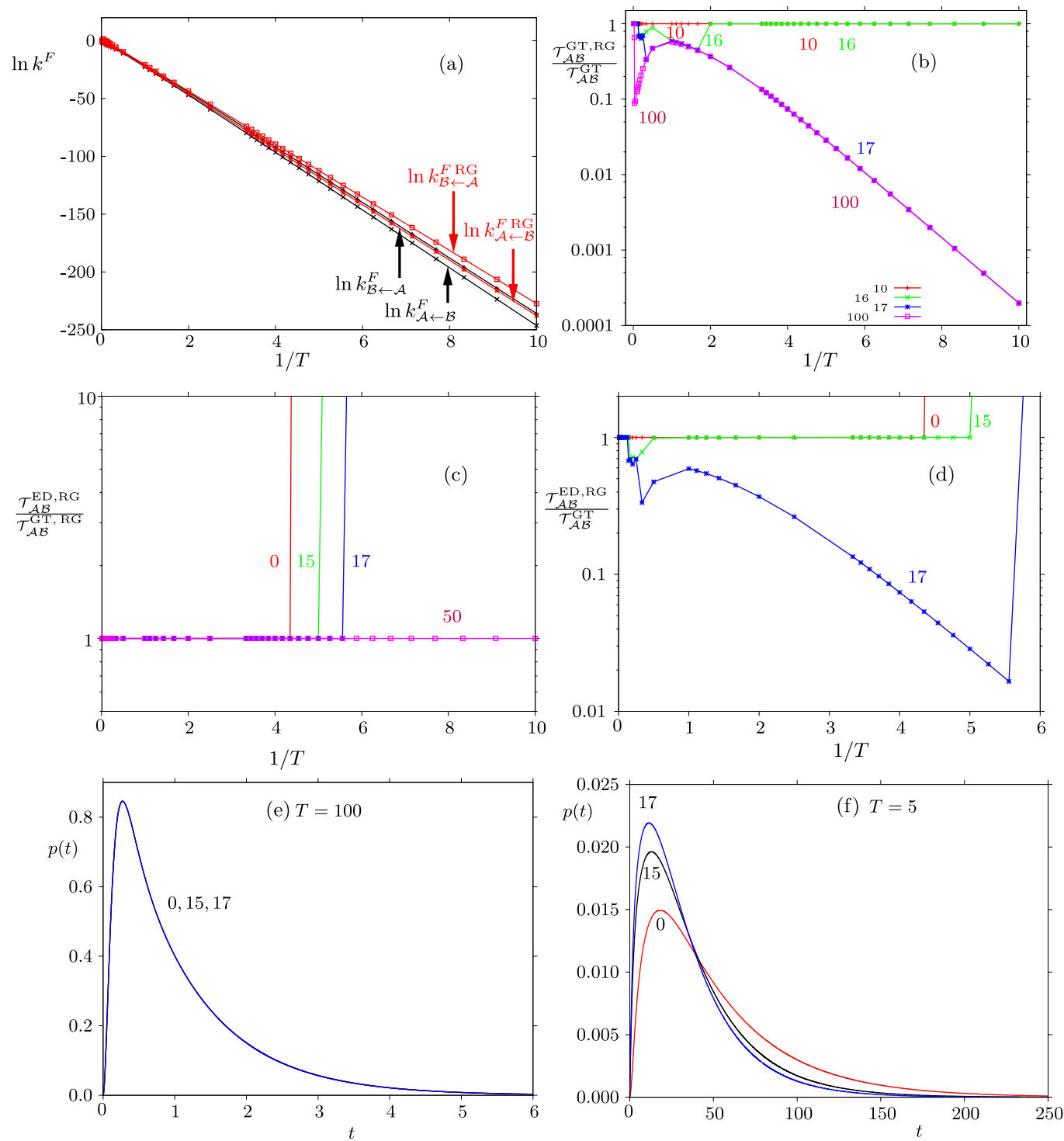


FIG. 8. Comparison of rate constants, MFPT and first passage time distributions calculated using free energy regrouping for the model landscape depicted in Figure 1a. Reduced units are employed throughout. (a) Shows Arrhenius plots of  $k^F$  and  $k^{GT, RG}$  for the largest regrouping threshold  $\Delta G^{RG} = 100$ . The values for  $k^{GT, RG}$  are shifted systematically to higher rates with decreasing temperature, with  $k_{B \leftarrow A}^F$  and  $k_{A \leftarrow B}^{GT, RG}$  lying between the other two plots, as indicated. (b) Ratio of  $\mathcal{T}_{B \leftarrow A}^{GT, RG}$  to  $\mathcal{T}_{B \leftarrow A}^{GT}$  for selected regrouping thresholds. The regrouped rate is correct throughout this temperature range for  $\Delta G^{RG} = 10$ . At  $\Delta G^{RG} = 16$   $\mathcal{T}_{B \leftarrow A}^{GT, RG}$  is correct in the low temperature range, while for thresholds of 17 and 100 the regrouped rates are higher, and the ratio increases systematically in the low temperature range. (c) Ratio of regrouped  $A \leftarrow B$  mean first passage times calculated using eigendecomposition (\*) and graph transformation (F) for the same regrouping thresholds, as marked (zero corresponds to no regrouping). Without regrouping, numerical imprecision leads to inaccurate values for  $\mathcal{T}_{B \leftarrow A}^{ED, RG}$  below  $T = 0.23$ . For a threshold of 50 the regrouped values agree throughout this temperature range. (d) Ratio of the regrouped  $A \leftarrow B$  mean first passage time to the accurate value corresponding to  $\mathcal{T}_{A \leftarrow B}^{GT}$  without regrouping. With a threshold of  $\Delta G^{RG} = 15$  the lowest temperature at which the MFPT is correct is extended down to  $T = 0.2$ . For a threshold of  $\Delta G^{RG} = 17$  the MFPT is within two orders of magnitude of the correct value down to  $T = 0.18$ , before numerical imprecision causes the calculation to fail. (e) The first passage probability distributions for  $T = 100$  are indistinguishable for regrouping thresholds  $\Delta G^{RG}$  of 17, 15 and 0 (no regrouping). (f) For  $T = 5$  the probability distributions are similar, but not identical. At lower temperatures (not shown) the disagreement with increasing  $\Delta G^{RG}$  is even more pronounced.



scheme explicitly groups together states separated by barriers below the threshold, which are expected to interconvert rapidly compared to the rate-determining process of interest. It does not require any structural order parameters, and all degrees of freedom are included, avoiding projection onto a lower dimensional space. Of course, if  $\Delta G^{\text{RG}}$  is chosen too large, then the coarse-grained network that results will no longer reproduce the kinetics properly. Hence, in the present context, we need to assess the effect of regrouping on the rate constants in terms of the true values for the coarse-grained network, and the temperature range where eigendecomposition is numerically correct. We therefore compared the rate constants and mean first passage times calculated for the regrouped network using graph transformation, denoted  $k^{\text{GT, RG}}$  and  $\mathcal{T}^{\text{GT, RG}}$ , and using the eigendecomposition of equation (6), denoted  $k^{*\text{RG}}$  and  $\mathcal{T}^{\text{ED, RG}}$ , to the reference values  $k^F$  and  $\mathcal{T}^{\text{GT}}$  on the full network.

A wide range of regrouping thresholds were applied, and selected results are shown in Figure 8. The largest systematic errors from regrouping with  $\Delta G^{\text{RG}} = 100$  are illustrated in Arrhenius plots for both  $\mathcal{A} \leftrightarrow \mathcal{B}$  rate constants in Figure 8a. Here,  $\mathcal{T}_{\mathcal{A}\mathcal{B}}^{\text{GT, RG}}$  is around three orders of magnitude too small at the lowest temperature considered,  $T = 0.1$ , while thresholds up to  $\Delta G^{\text{RG}} = 16$  provide a faithful representation of the overall kinetics (Figure 8b). For the largest regrouping thresholds the coarse-graining produces two groups, since we do not allow sets containing  $\mathcal{A}$  and  $\mathcal{B}$  to merge.

We find that eigendecomposition and graph transformation produce the same values for the  $\mathcal{A} \leftarrow \mathcal{B}$  MFPT and the corresponding rate constants for the same regrouped databases, as long as sufficient precision is maintained (Figure 8c). The trends for the  $\mathcal{B} \leftarrow \mathcal{A}$  direction are very similar, and are omitted for brevity. For larger  $\Delta G^{\text{RG}}$  thresholds, agreement is maintained to progressively lower temperatures, since numerical precision is improved for the eigendecomposition when the dimension of the rate matrix decreases (fewer free energy groups). The final comparison of  $\mathcal{T}_{\mathcal{A}\mathcal{B}}^{\text{ED, RG}}$  with  $\mathcal{T}_{\mathcal{A}\mathcal{B}}^{\text{GT}}$  in Figure 8d shows that accurate rates can be obtained down to lower temperatures when  $\Delta G^{\text{RG}}$  is chosen judiciously. If an estimate of the rate constant within two orders of magnitude is the objective, then the temperature range can be pushed lower with a larger regrouping threshold. The trends in Figure 8d for  $\mathcal{T}_{\mathcal{A}\leftarrow\mathcal{B}}^{\text{ED, RG}}$  are analogous to Figure 8b for  $\mathcal{T}_{\mathcal{A}\leftarrow\mathcal{B}}^{\text{GT, RG}}$ , except that precision is lost at low temperature.

The probability distributions for the first passage time show how regrouping affects the kinetics. Selected results are illustrated in Figure 8 for two temperatures. The smaller values of  $\mathcal{T}_{\mathcal{A}\mathcal{B}}^{\text{ED, RG}}$  for  $\Delta G^{\text{RG}} = 17$ , visible in Figure 8d, correspond to a first passage time distribution with a higher peak at a lower value of  $t$ , followed by faster decay. The systematic error for  $T = 1$  would be acceptable for an order of magnitude estimate of the rate constant, while the distribution at  $T = 0.25$  deviates significantly from the true form, but still yields a rate within two orders of magnitude of the correct result.

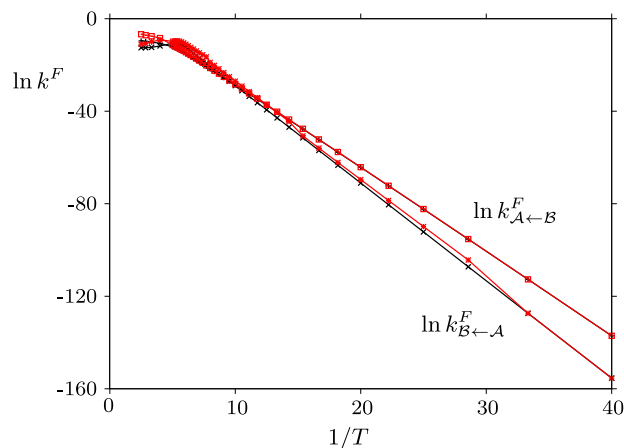


FIG. 9. Arrhenius plots for LJ<sub>38</sub> comparing  $k^F$  (black) and  $k^{F, \text{RG}}$  (red) with a regrouping threshold of  $\Delta G^{\text{RG}} = 0.75 \epsilon$ . Reduced units are employed throughout.

Analogous results for regrouping with LJ<sub>38</sub> are shown in Figures 9 and 10. As for Figure 5 these results are for the database containing 4000 minima and 2677 transition states. Here we introduced a small modification to the regrouping scheme, which produces more accurate rate constants in the reference values  $k^{\text{GT, RG}}$  obtained using graph transformation. The pathway making the largest contribution to  $k^{SS}$  at different temperatures was determined using Dijkstra's shortest path algorithm<sup>113</sup> for edge weights calculated from the corresponding branching probabilities.<sup>4,66,114</sup> Regrouping minima from these paths, especially the highest energy structures, can affect the overall barrier between groups for the most kinetically relevant paths. Excluding selected minima from these paths therefore enables larger regrouping thresholds to be used without affecting the overall barrier for the fastest pathways. Hence the calculated

This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.  
PLEASE CITE THIS ARTICLE AS DOI:10.1063/5.0016244

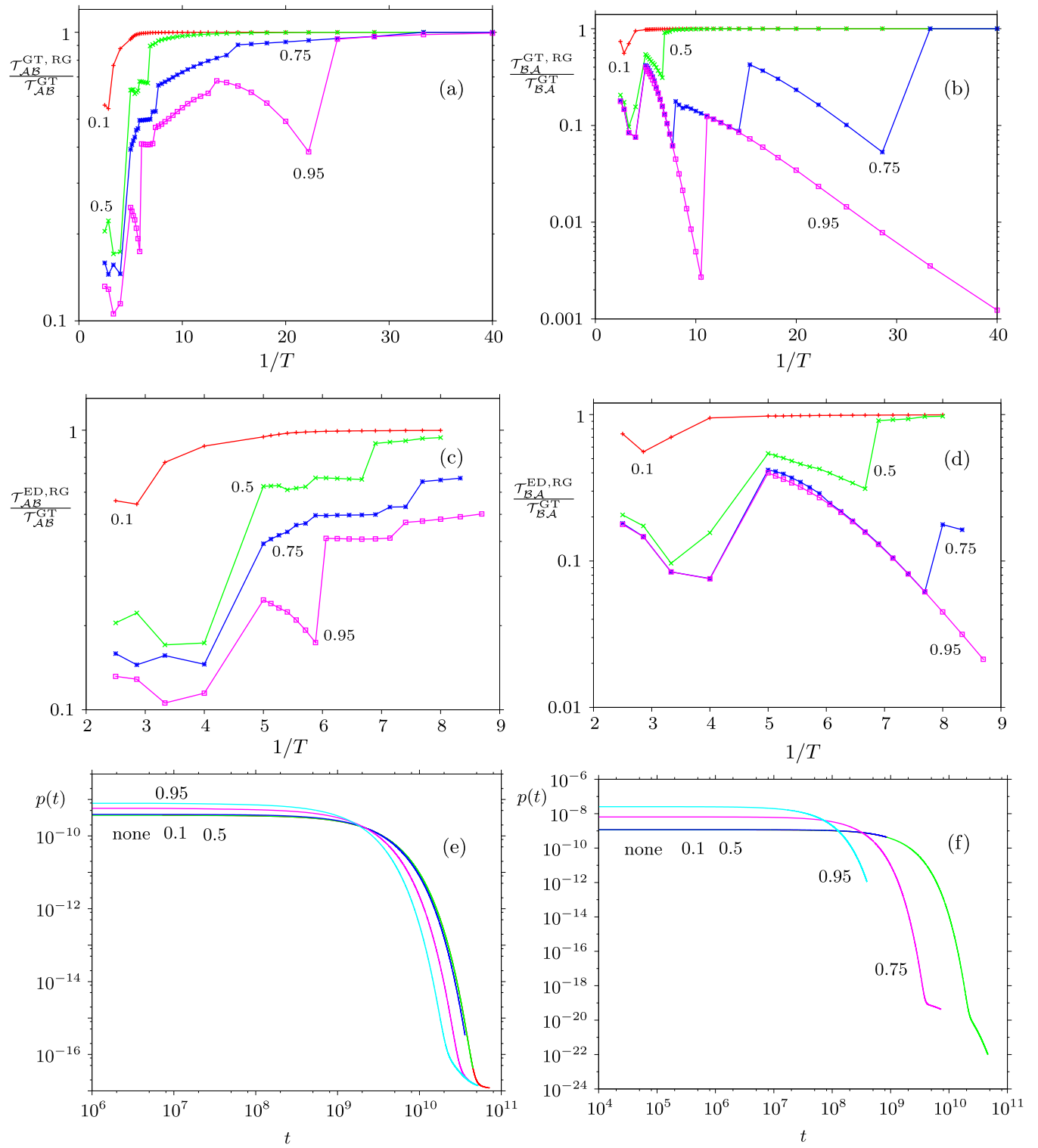


FIG. 10. Comparison of the MFPT and first passage time probability distributions calculated using free energy regrouping for LJ<sub>38</sub> with  $\Delta G^{\text{RG}} = 0.1, 0.5, 0.75,$  and  $0.95$ , as marked. Reduced units are employed throughout. (a)  $\mathcal{T}_{AB}^{\text{GT, RG}}/\mathcal{T}_{AB}^{\text{GT}}$ . (b)  $\mathcal{T}_{BA}^{\text{GT, RG}}/\mathcal{T}_{BA}^{\text{GT}}$ . (c)  $\mathcal{T}_{AB}^{\text{ED, RG}}/\mathcal{T}_{AB}^{\text{ED}}$ . (d)  $\mathcal{T}_{BA}^{\text{ED, RG}}/\mathcal{T}_{BA}^{\text{ED}}$ . (e) Direction  $\mathcal{A} \leftarrow \mathcal{B}$  at  $k_B T/\epsilon = 0.125$  (f) Direction  $\mathcal{B} \leftarrow \mathcal{A}$  at  $k_B T/\epsilon = 0.125$ . The results for thresholds of 0.1 and 0.5 are similar to the reference in each case. As  $\Delta G^{\text{RG}}$  increases the distribution shifts systematically to shorter first passage times.

rates agree better with the true values for larger  $\Delta G^{\text{RG}}$  thresholds, where the coarse-grained database has fewer free energy groups. The reduction in dimensionality results in improved numerical precision in the eigendecomposition. We note that minima could be excluded from regrouping based on more general analysis of distinct pathways classified by the rate-determining step,<sup>66</sup> and we will test this extension in future work.

The effective barriers that result from a best fit of the rate constants to the Arrhenius form are approximately 4.22 and 3.27  $\epsilon$ ,<sup>6</sup> for  $\mathcal{B} \leftarrow \mathcal{A}$  and  $\mathcal{A} \leftarrow \mathcal{B}$ , respectively, for the usual choice of five minima in the octahedral  $\mathcal{A}$  region and 395 associated with  $\mathcal{B}$  region corresponding to the incomplete Mackay icosahedron. Figure 9 shows that both rate constants are reproduced quite accurately for  $\Delta G^{\text{RG}} = 0.75 \epsilon$  with the graph transformation procedure. The effect of regrouping on  $\mathcal{T}^{\text{GT}}$  and  $\mathcal{T}^{\text{ED}}$  is illustrated for selected regrouping thresholds in Figure 10. Here it is the low temperature behaviour that is the focus of interest, and we find that  $\Delta G^{\text{RG}}$  values up to 0.5  $\epsilon$  preserve both  $\mathcal{B} \leftarrow \mathcal{A}$  and  $\mathcal{A} \leftarrow \mathcal{B}$  rate constants very well in this regime.  $\mathcal{T}_{\mathcal{BA}}^{\text{GT, RG}}$  (cuboctahedral to icosahedral) is more sensitive to regrouping, but the deviations do not exceed about an order of magnitude for  $\Delta G^{\text{RG}} = 0.75 \epsilon$  (Figure 10a and b). We found that the regrouped rates agreed well for  $\mathcal{T}^{\text{GT, RG}}$  and  $\mathcal{T}^{\text{ED, RG}}$  up to  $\Delta G^{\text{RG}} = 0.95 \epsilon$  in the temperature range where eigendecomposition retains numerical precision. Panels (c) and (d) demonstrate that regrouping can extend the accessible low temperature range for  $\mathcal{T}^{\text{ED}}$ , but the improvement is relatively small.

Finally, the effect of regrouping on the first passage time distributions at  $k_B T / \epsilon = 0.125$  is shown in Figure 10 panels (e) and (f); the deviations result in a systematic shift to higher probability at shorter time, and an earlier decay. In agreement with the behaviour noted above, the  $\mathcal{B} \leftarrow \mathcal{A}$  first passage time exhibits greater sensitivity to regrouping. It is also important to note that kPS simulation for the  $\mathcal{B} \leftarrow \mathcal{A}$  transition in LJ<sub>38</sub> at the lowest temperature considered ( $T = 0.05$ ), is feasible for the transition network obtained using a regrouping threshold of 2.1  $\epsilon$ . This calculation was intractable for the original network, owing to a small number of edges where the number of kMC moves along a typical transition path is exceptionally large. The error incurred in the resulting estimate for the MFPT is approximately 10% compared to the exact value obtained from graph transformation.

These results show that regrouping has the potential to extend the useful range of the rate constant analysis for the eigendecomposition formulation, and to relieve the problem of flickering trajectories in kMC simulations.<sup>43,67</sup> However, the gains achievable by lumping schemes are limited by the error incurred when nodes separated by higher energy barriers are grouped together. Use of a lumping scheme therefore requires a careful analysis of the dynamics for the resulting coarse-grained Markov chain to ensure consistency with the original transition network.<sup>17,68</sup>

## B. Free Energy Regrouping with Optimal Markovian Coarse-graining of Intergroup Rates

In §V A, the intergroup rates were estimated using the local equilibrium approximation,<sup>60</sup> which assumes that transitions within a free energy group are much faster than transitions between them. This formulation equates the intergroup rate constant from group  $I \leftarrow J$  to a weighted average of the inter-microstate rates between the boundary nodes:  $k_{IJ}^\dagger = \sum_{(ij)^\dagger} k_{ij} \hat{\pi}_j$ , as for equation (11), with  $\hat{\pi}_j = \pi_j / \Pi_J$ . Because this approach assumes that transitions within  $J$  are instantaneous before escape to  $I$ , the local equilibrium approximation is exact for the evolution of the occupation probabilities in the short time limit. This result follows from analysis of the occupancy number cross-correlation functions,

$$C_{ij}(\tau, t) = P(i, t + \tau; j, t) - p_i(t + \tau)p_j(t), \quad (12)$$

where the first term designates the joint probability that the system is in state  $j$  at time  $t$  and in state  $i$  at a lag time  $\tau$  later. The first term can be written in terms of the conditional probability  $P(i, t + \tau | j, t)$ , which is equal to the  $ij^{\text{th}}$  element of the propagator matrix  $e^{\mathbf{K}\tau}$ , where  $\mathbf{K} \in \mathbb{R}^{N \times N}$  is the matrix of minimum-to-minimum rate constants, with diagonal entries corresponding to total escape rate, analogous to equation (1). If the system is in equilibrium, the correlation function becomes a function of the lag time  $\tau$  only:

$$C_{ij}^{\text{eq}}(\tau) = [e^{\mathbf{K}\tau}]_{ij} \pi_j - \pi_i \pi_j, \quad (13)$$

where  $\pi_i$  is the equilibrium occupation probability of state  $i$ , normalised over all minima, as above. It has been shown in Ref.<sup>115</sup> that the local equilibrium approximation is equivalent to

$$\sum_{i \in I} \sum_{j \in J} C_{ij}^{\text{eq}}(0) = C_{IJ}^{\text{eq}}(0), \quad (14)$$

meaning that the number of transitions per unit time occurring at equilibrium is exact in the limit of short lag times. While this approximation works well for clearly-defined metastable basins, it will lead to systematic errors if the assumed time scale separation breaks down; i.e. in flat regions of the landscape.

We now consider an alternative formulation of the intergroup rates due to Hummer and Szabo,<sup>101</sup> who proposed a coarse-graining strategy that minimises the error in the correlation functions under the constraint that the coarse-grained model is Markovian. The area under the correlation functions is then the same for both the coarse-grained network and the full network:

$$\sum_{i \in I} \sum_{j \in J} \int_0^\infty C_{ij}^{\text{eq}}(\tau) d\tau = \int_0^\infty C_{IJ}^{\text{eq}}(\tau) d\tau. \quad (15)$$

Since this relation matches the correlation functions of the coarse-grained and full networks over all time scales, instead of just at short times, it should preserve the kinetics of interest more accurately than the local equilibrium approximation for a given set of clusters.

The Hummer-Szabo formulation, derived below, is ‘optimal’ in the sense that it minimizes the total error in occupation correlation functions, whilst satisfying the Markovian condition and detailed balance,<sup>101,115</sup> balancing accuracy at long and short lag times. We note that some applications prioritize accuracy for a specific timescale, where one instead minimizes correlation error at a set of (typically long) lag times.<sup>116</sup>

Consider a network of  $N$  minima partitioned into  $M$  groups using a clustering matrix  $\mathbf{A} \in \mathbb{R}^{N \times M}$ , where  $A_{iI} = 1$  if  $i \in I$ , and zero otherwise. Then, enforcing (15) results in a coarse-grained rate matrix  $\mathbf{K}^{\text{HS}} \in \mathbb{R}^{M \times M}$  given by<sup>101</sup>

$$\mathbf{K}^{\text{HS}} = \mathbf{\Pi}_M \mathbf{1}_M - \text{diag}(\mathbf{\Pi}_M) (\mathbf{A}^\top [\mathbf{\pi} \mathbf{1}_N - \mathbf{K}]^{-1} \text{diag}(\mathbf{\pi}) \mathbf{A})^{-1}. \quad (16)$$

Here,  $\mathbf{1}_N$  is the  $N$ -dimensional row vector with entries equal to unity, as in §II,  $\mathbf{\pi} \in \mathbb{R}^N$  again contains the normalised equilibrium occupation probabilities of the minima,  $\mathbf{\Pi}_M = \mathbf{A}^\top \mathbf{\pi}$  contains the equilibrium occupation probabilities of the groups, and the diagonal matrices  $\text{diag}(\mathbf{\pi}) \in \mathbb{R}^{N \times N}$  and  $\text{diag}(\mathbf{\Pi}_M) \in \mathbb{R}^{M \times M}$  contain entries corresponding to components of the probability vectors indicated.

The expression in (16) is optimal in that it is the most accurate model for intergroup rates that satisfies the Markovian condition and detailed balance;<sup>101,115</sup> however, it requires inverting an  $N$ -dimensional matrix, which is numerically unstable for large networks and low temperatures. To investigate these issues further, we also implemented an alternative version of the Hummer-Szabo coarse-grained rate matrix written in terms of the pairwise first passage times,  $t_{ij}$ , from states  $i \leftarrow j$ , as derived in Ref.<sup>111</sup>:

$$\begin{aligned} \mathbf{K}^{\text{KKRA}} &= \mathbf{\Pi}_M \mathbf{1}_M - [\mathbf{\Pi}_M \mathbf{1}_M + \mathbf{A}^\top \text{diag}(\mathbf{\pi}) \mathbf{t} \mathbf{\pi} \mathbf{1}_M \\ &\quad - \mathbf{A}^\top \text{diag}(\mathbf{\pi}) \mathbf{t} \text{diag}(\mathbf{\pi}) \mathbf{A} \text{diag}(\mathbf{\Pi}_M)^{-1}]^{-1}. \end{aligned} \quad (17)$$

In principle,  $\mathbf{K}^{\text{KKRA}} = \mathbf{K}^{\text{HS}}$ , but the formulation in equation (17) only requires inverting  $M$ -dimensional matrices, so it can be applied to larger systems than equation (16). However, the calculation of all  $N^2$  first passage times for  $\mathbf{t}$  can also be numerically challenging for large systems and low temperatures.

In Fig. 11, we coarse-grain the model landscape depicted in Fig. 1a using the same grouping scheme as in §V A and compare the intergroup mean first passage times (MFPTs) obtained from the two formulations of the rate matrix  $\mathbf{K}^{\text{HS}}$  and  $\mathbf{K}^{\text{KKRA}}$ . Here we define the MFPTs as

$$\mathcal{T}_{AB} = \sum_{b \in \mathcal{B}} \mathcal{T}_{Ab} \hat{\pi}_b \quad \text{and} \quad \mathcal{T}_{BA} = \sum_{a \in \mathcal{A}} \mathcal{T}_{Ba} \hat{\pi}_a, \quad (18)$$

and we compare the results from graph transformation,  $\mathcal{T}_{AB}^{\text{GT}}$ , obtained from  $\mathcal{T}_{Ab} = \tau_b^{\text{GT}} / P_{Ab}^{\text{GT}}$ , with  $\mathcal{T}_{AB}^{\text{ED}}$ , obtained from the mean first passage time via the eigendecomposition approach in equation (6). To allow direct comparison with the LEA, we assume that microstates in  $\mathcal{A}, \mathcal{B}$  are initially in local equilibrium. We note that this assumption may be inaccurate for weakly metastable or disconnected basins, as discussed below equation (9).

For all regrouping thresholds, both the HS and KKRA schemes lose precision at  $T = 1.0$  and below, but Figure 11 shows that they produce more accurate results than the local equilibrium approximation (LEA) intergroup rates at higher temperatures. For thresholds  $\Delta G^{\text{RG}} \leq 16$ , where the regrouping procedure does not change the number of states, both the HS and KKRA formulations start to become unstable at even higher temperatures. Since the model landscape only has 8 states, we can attribute this instability to a lack of numerical precision, rather than high dimensionality. For  $\Delta G^{\text{RG}} \geq 20$ , where coarse-graining produces a two-state system, the calculations are stable for all temperatures  $T > 1.0$ . In this range of temperatures, the HS rates are more accurate than the LEA in reproducing the kinetics of the full network, as expected. Although all coarse-grained rates are within a factor of two from the true rate in this model system, for more realistic systems like LJ<sub>38</sub>, the LEA rates can deviate systematically at low temperatures from the true values, depending on  $\Delta G^{\text{RG}}$ , as illustrated in §V A. The HS scheme would be more accurate in these cases if it could be applied without losing precision.

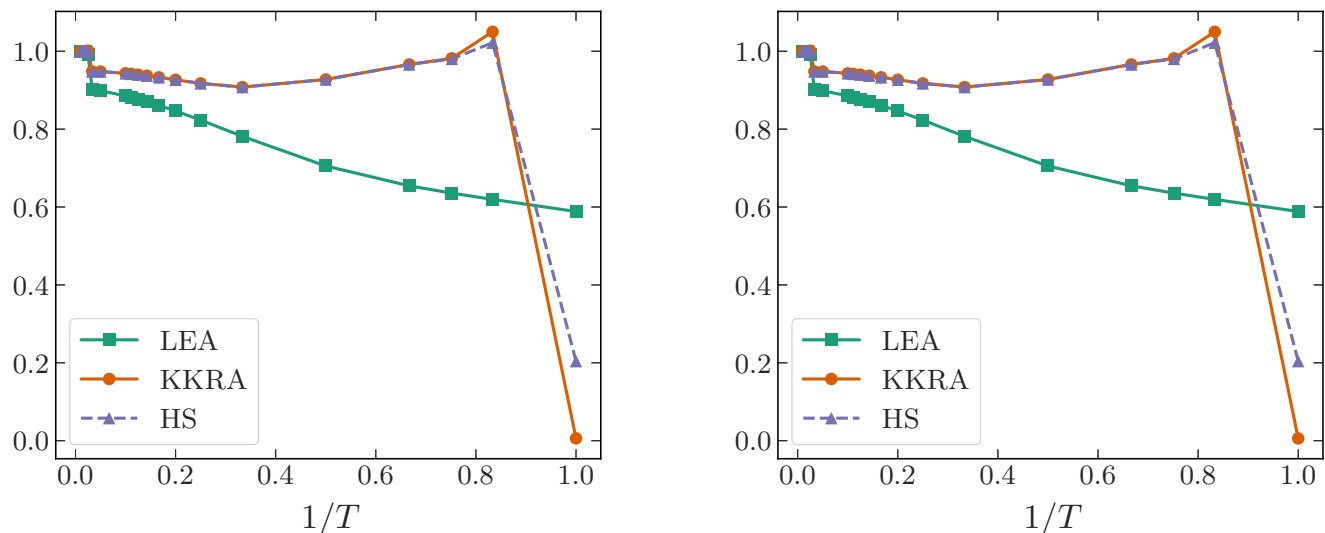


FIG. 11. Comparison of the MFPT for  $\mathcal{B} \leftarrow \mathcal{A}$  calculated for different coarse-grained networks, compared to the exact MFPT ( $\mathcal{T}_{\mathcal{B}\mathcal{A}}^{\text{GT}}$ ) calculated using graph transformation (GT) for the full network for the model landscape depicted in Figure 1a. The coarse-grained network, obtained using a free energy regrouping threshold of  $\Delta G^{\text{RG}} = 100$ , contains just two groups. We compare three different methods for estimating the intergroup rates: the Hummer-Szabo formulation in Eqn. (16) (HS), the same rates calculated instead from the matrix of pairwise mean first passage times, as suggested by Kells et al.<sup>111</sup> (KKRA), and the local equilibrium approximation (LEA).<sup>60</sup> Left: ratio of first passage time calculated using graph transformation (GT) for the coarse-grained network to the GT result for the full network. Right: ratio of first passage time calculated using eigendecomposition (ED) for the coarse network to the GT result for the full network. The HS and KKRA formulations lose precision below  $T = 1.0$ , but are more accurate than LEA for higher temperatures. The MFPTs obtained from eigendecomposition and graph transformation agree to high precision.

### C. Free Energy regrouping using spectral reduction of clusters

We have seen in section V A that the free energy regrouping method, combined with the local equilibrium approximation (LEA), is able to accurately produce the true kinetics providing the regrouping thresholds are chosen judiciously. However, if too high a threshold is chosen, the coarse graining becomes too aggressive and accuracy is lost. In this final section we generalize the LEA procedure for a given state regrouping, providing a method to interpolate between the LEA and the exact dynamics. In particular, our aim is to produce a scheme which receives a given coarse clustering, then proposes additional states to the LEA in order to improve accuracy.

To achieve this, we perform a local eigendecomposition for each cluster, then perform a projection to ensure one mode is equal to the local equilibrium distribution. We then analyse the local spectral properties to decide how many modes per group we retain. If all modes are retained then we recover the full master equation matrix, whilst the retention of one mode per group is equivalent to the LEA.

Although all of the disjoint clusters produced by free energy regrouping can in principle be analyzed simultaneously, for clarity we consider analysing just one cluster  $\mathcal{P}$  with  $N_{\mathcal{P}}$  states. The application to multiple clusters is presented in appendix B. Then, the  $N$ -state master equation matrix can be partitioned into two sets of size  $N_{\mathcal{P}}$  and  $N_{\mathcal{Q}} = N - N_{\mathcal{P}}$  as

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{\mathcal{Q}\mathcal{Q}} & \mathbf{M}_{\mathcal{Q}\mathcal{P}} \\ \mathbf{M}_{\mathcal{P}\mathcal{Q}} & \mathbf{M}_{\mathcal{P}\mathcal{P}} \end{bmatrix}, \quad (19)$$

where the cluster  $\mathcal{P}$  is suitably chosen in rank and conditioning such that we are able to compute a full eigendecomposition for the  $N_{\mathcal{P}}$  relaxation modes, giving the exact relation

$$\mathbf{M}_{\mathcal{P}\mathcal{P}} = - \sum_l^{N_{\mathcal{P}}} \lambda_l^{\mathcal{P}} \mathbf{x}_l^{\mathcal{P}} \otimes \mathbf{y}_l^{\mathcal{P}}, \quad (20)$$

which defines the eigenvalues  $-\lambda_l^{\mathcal{P}} < 0$  and the right and left eigenvectors  $\mathbf{x}_l^{\mathcal{P}}, \mathbf{y}_l^{\mathcal{P}}$ . Although (20) is clearly not a diagonalization of the whole rate matrix, it is an exact representation of the submatrix  $\mathbf{M}_{\mathcal{P}\mathcal{P}}$ . With some initial



condition  $\mathbf{P}_{\mathcal{P}}(0)$  confined to  $\mathcal{P}$ , the dynamics before escape from  $\mathcal{P}$  are given exactly by

$$\mathbf{P}_{\mathcal{P}}(t) = \sum_l^{N_{\mathcal{P}}} [\mathbf{y}_l^{\mathcal{P}} \mathbf{P}_{\mathcal{P}}(0)] \exp(-\lambda_l^{\mathcal{P}} t) \mathbf{x}_l^{\mathcal{P}}, \quad (21)$$

with moments of the escape time from  $\mathcal{P}$

$$\langle t_{\mathcal{Q}\mathcal{P}}^m \rangle = m! \sum_l^{N_{\mathcal{P}}} [\mathbf{y}_l^{\mathcal{P}} \mathbf{P}_{\mathcal{P}}(0)] [\mathbf{1}_{\mathcal{P}} \mathbf{x}_l^{\mathcal{P}}] / [\lambda_l^{\mathcal{P}}]^m, \quad (22)$$

where  $m \in [0, N_{\mathcal{P}}]$  and the relation for  $m = 0$  defines the normalisation  $\mathbf{1}_{\mathcal{P}} \mathbf{P}_{\mathcal{P}}(0) = 1$ , with  $\mathbf{1}_{\mathcal{P}}$  a row vector of ones of dimension  $N_{\mathcal{P}}$ , analogous to equation (7).

To simplify the dynamics in  $\mathcal{P}$ , we reduce the effective rank of  $\mathbf{M}_{\mathcal{P}\mathcal{P}}$  by truncating the sum in the eigendecomposition to include only a subset  $\mathcal{S}_{\mathcal{P}}$  of  $n_1^{\mathcal{P}} < N_{\mathcal{P}}$  modes. Whilst such approaches typically rely on the identification of a spectral gap in the eigenvalues, we define a more general criterion below, which includes this form of selection as a special case.

With a subset  $\mathcal{S}_{\mathcal{P}}$  identified, we project onto rectangular basis matrices  $\mathbf{X}_1^{\mathcal{P}}$  and  $\mathbf{Y}_1^{\mathcal{P}}$  where  $\mathbf{X}_1^{\mathcal{P}} \in \mathbb{R}^{N_{\mathcal{P}} \times n_1^{\mathcal{P}}}$  is a matrix whose  $n_1^{\mathcal{P}}$  columns are the right eigenvectors  $\mathbf{x}_l^{\mathcal{P}}$ ,  $l \in \mathcal{S}_{\mathcal{P}}$  and  $\mathbf{Y}_1^{\mathcal{P}} \in \mathbb{R}^{n_1^{\mathcal{P}} \times N_{\mathcal{P}}}$  is a matrix whose  $n_1^{\mathcal{P}}$  rows are the left eigenvectors  $\mathbf{y}_l^{\mathcal{P}}$ ,  $l \in \mathcal{S}_{\mathcal{P}}$ . More generally, any strategy can be used to form the reduced basis matrices, providing  $\mathbf{X}_1^{\mathcal{P}} \mathbf{Y}_1^{\mathcal{P}} \approx \mathbb{I}_{N_{\mathcal{P}}}$ . The full master equation matrix can then be reconstructed, to within the approximation induced by the truncation, via

$$\mathbf{M} \simeq \begin{bmatrix} \mathbb{I}_{\mathcal{Q}} & \\ & \mathbf{X}_1^{\mathcal{P}} \end{bmatrix} \mathbf{M}_1 \begin{bmatrix} \mathbb{I}_{\mathcal{Q}} & \\ & \mathbf{Y}_1^{\mathcal{P}} \end{bmatrix} \equiv \tilde{\mathbf{X}}_1 \mathbf{M}_1 \tilde{\mathbf{Y}}_1, \quad (23)$$

where the omitted off-diagonal elements are rectangular matrices of zeros, with appropriate dimensions, and the first equation defines the rectangular matrices  $\tilde{\mathbf{X}}_1$  and  $\tilde{\mathbf{Y}}_1$ . The reduced matrix  $\mathbf{M}_1$ , defined in Appendix A, has a reduced rank  $N_{\mathcal{Q}} + n_1^{\mathcal{P}}$ . This procedure can be continued, partitioning and reducing  $\mathbf{M}_1$  in an identical manner to  $\mathbf{M}$ , with corresponding matrices  $\tilde{\mathbf{X}}_2$  and  $\tilde{\mathbf{Y}}_2$ . After  $J$  recursions we have a reduced rate matrix of

$$\mathbf{M} \simeq \left[ \prod_{j=1}^J \tilde{\mathbf{X}}_j \right] \mathbf{M}_J \left[ \prod_{j=1}^J \tilde{\mathbf{Y}}_j \right]. \quad (24)$$

When an  $\mathbf{M}_J$  is found that is suitably well conditioned we then perform a full diagonalization,  $\mathbf{M}_J = -\tilde{\mathbf{X}}_{J+1} \boldsymbol{\lambda}_{J+1} \tilde{\mathbf{Y}}_{J+1}$ , producing an approximate first passage time distribution

$$p(t) \simeq \mathbf{1}_N \left[ \prod_{j=1}^{J+1} \tilde{\mathbf{X}}_j \right] \boldsymbol{\lambda}_{J+1} \exp(-\boldsymbol{\lambda}_{J+1} t) \left[ \prod_{j=1}^{J+1} \tilde{\mathbf{Y}}_j \right] \mathbf{P}(0). \quad (25)$$

In the following analysis we focus on clusters produced using the thresholding criteria and self-consistent regrouping defined in §V A using the LEA approximation, and compare the results of a single iteration ( $J = 1$ ) to LEA. To select the subset of modes  $\mathcal{S}_{\mathcal{P}}$ , the most rigorous approach would be look for a large spectral gap in the eigenvalues, i.e. find a subset  $\mathcal{S}_{\mathcal{P}}$  such that  $\max_{l \in \mathcal{S}_{\mathcal{P}}} \lambda_l^{\mathcal{P}} \ll \min_{l \notin \mathcal{S}_{\mathcal{P}}} \lambda_l^{\mathcal{P}}$ .

If the bulk of the initial probability is associated with the retained modes, i.e.  $\mathbf{1}_{N_{\mathcal{P}}} \mathbf{X}_1^{\mathcal{P}} \mathbf{Y}_1^{\mathcal{P}} \mathbf{P}_{\mathcal{P}}(0) \simeq 1$ , then the system will then have an exponentially small likelihood of leaving  $\mathcal{P}$  at short times where the influence of the fast relaxation modes  $l \notin \mathcal{S}_{\mathcal{P}}$  is greatest. After this short period such modes will have decayed away, meaning that the truncated eigenexpansion  $\mathbf{X}_1^{\mathcal{P}} \exp(-\boldsymbol{\lambda}_1^{\mathcal{P}} t) \mathbf{Y}_1^{\mathcal{P}} \mathbf{P}_{\mathcal{P}}(0)$  will give only exponentially small error in the internal dynamics.

However, a large spectral gap does not always exist, or may require many different regrouping options to be attempted to allow such metastability to be identified.

We propose an alternative criteria, namely selecting  $\mathcal{S}_{\mathcal{P}}$  such that the retained modes approximately account for the local equilibrium distribution  $\hat{\boldsymbol{\pi}}_{\mathcal{P}}$  and some moments of the exit time from  $\mathcal{P}$ , given local equilibrium. For prescribed tolerances  $\epsilon_0, \epsilon_1, \dots > 0$  we require  $\mathbf{1}_{N_{\mathcal{P}}} \mathbf{X}_1^{\mathcal{P}} \mathbf{Y}_1^{\mathcal{P}} \hat{\boldsymbol{\pi}}_{\mathcal{P}} = 1 - \epsilon_0$  and

$$m! (\mathbf{1}_{N_{\mathcal{P}}} \mathbf{X}_1^{\mathcal{P}}) [\boldsymbol{\lambda}_1^{\mathcal{P}}]^{-m} (\mathbf{Y}_1^{\mathcal{P}} \hat{\boldsymbol{\pi}}_{\mathcal{P}}) = (1 - \epsilon_m) \langle t_{\mathcal{Q}\mathcal{P}}^m \rangle, \quad (26)$$

where  $\langle t_{\mathcal{Q}\mathcal{P}}^m \rangle$  is evaluated with  $\mathbf{P}_{\mathcal{P}}(0) = \hat{\boldsymbol{\pi}}_{\mathcal{P}}$ . The error in moment  $m$  from omitting mode  $l$  is  $m! [\mathbf{1}_{N_{\mathcal{P}}} \mathbf{x}_l^{\mathcal{P}}] [\mathbf{y}_l^{\mathcal{P}} \mathbf{P}_{\mathcal{P}}(0)] / (\lambda_l^{\mathcal{P}})^m$ , and the relative error in  $\langle t_{\mathcal{Q}\mathcal{P}}^m \rangle$  will decrease with  $m$ , since it is the largest magnitude eigenvalues that are removed.



As a result, if there is a large spectral gap the criteria in equation (26) will naturally identify the slow ‘below gap’ modes as a special case, but also provide a protocol that can be applied when no clear gap exists, consistent with the conditions for key observables to be reproduced as accurately as possible.

We note that an unsupervised procedure for choosing the subset  $\mathcal{S}_{\mathcal{P}}$  could be devised, where modes  $l \in N_{\mathcal{P}}$  are recursively added to  $\mathcal{S}_{\mathcal{P}}$  until the desired accuracy is reached for a large number of moments. Poor cluster choice would then induce inefficiency rather than inaccuracy. However, in the remainder of this section we only require accuracy for  $m = 0$  and 1. In this context, the free energy regrouping method based upon local equilibrium and associated intergroup rates<sup>60</sup> (LEA) can be viewed as the truncation

$$\begin{aligned} n_1^{\mathcal{P}} &= 1, & \mathbf{Y}_1^{\mathcal{P}} &= \mathbf{y}_0^{\mathcal{P}} \simeq \mathbf{1}_{N_{\mathcal{P}}}, & \mathbf{X}_1^{\mathcal{P}} &= \mathbf{x}_0^{\mathcal{P}} \simeq \hat{\boldsymbol{\pi}}_{\mathcal{P}}, \\ & \Rightarrow \lambda_0^{\mathcal{P}} & \simeq & -\mathbf{1}_{N_{\mathcal{P}}} \mathbf{M}_{\mathcal{P}\mathcal{P}} \hat{\boldsymbol{\pi}}_{\mathcal{P}}, \end{aligned} \quad (27)$$

which should be accurate when there is one eigenvalue (denoted by mode 0) much smaller in magnitude than all the others, i.e. a large spectral gap with one mode below the gap such that  $\mathbf{y}_0^{\mathcal{P}} \simeq \mathbf{1}_{N_{\mathcal{P}}}$  and  $\mathbf{x}_0^{\mathcal{P}} \simeq \hat{\boldsymbol{\pi}}_{\mathcal{P}}$ . Whilst this situation will hold well for highly metastable clusters, with little internal structure, poorly chosen, weakly metastable or complex clusters will not be well represented. However, as application of (27) requires no eigendecomposition it can be rapidly applied for different regrouping schemes.

In Figure 12 we compare the FPT distributions obtained from the LEA truncation in (27) to the moment-based spectral reduction (SR) criteria defined by equation (26), using the clusters produced by the recursive free energy based regrouping scheme<sup>60</sup> described in §V A. The modes in each cluster  $\mathcal{P}$  were ordered via their projection onto the local equilibrium distribution, then added sequentially to  $\mathcal{S}_{\mathcal{P}}$  until the  $m = 0, 1$  moments were within the desired threshold. In the legend of Figure 12  $N_c$  refers to the number of clusters resulting from the free energy regrouping scheme. For the LEA results,  $N_m = N_c$  since only the slowest relaxation mode from each cluster is retained. For the SR results,  $N_m \geq N_c$  is the total number of modes retained summed over all the clusters. In the exact FPT distribution, all modes in all clusters are retained so that  $N_m = N$ .

There are clearly many possible strategies to select  $\mathcal{S}_{\mathcal{P}}$ , which will be explored in future work. Whilst this approach often improved on the LEA approach, in some cases, as can be seen in Figure 12, the spectral reduction scheme produced coarse grained KTNs with higher rank but lower accuracy than the LEA version.

To improve the spectral reduction scheme, we note that the reconstruction (23) only requires  $\mathbf{X}_1^{\mathcal{P}} \mathbf{Y}_1^{\mathcal{P}} \approx \mathbb{I}_{N_{\mathcal{P}}}$ , not that  $\mathbf{X}_1^{\mathcal{P}}$  and  $\mathbf{Y}_1^{\mathcal{P}}$  are formed from eigenvectors of  $\mathbf{M}_{\mathcal{P}\mathcal{P}}$ . As a result, we modified the basis matrices produced by the local eigendecomposition to better interpolate between the LEA approximation and the full dynamics.

The eigenmodes corresponding to  $\mathbf{x}_0^{\mathcal{P}}$  and  $\mathbf{y}_0^{\mathcal{P}}$  with the largest projection onto the vectors corresponding to local equilibrium,  $\hat{\boldsymbol{\pi}}_{\mathcal{P}}$  and  $\mathbf{1}_{N_{\mathcal{P}}}$  are ‘rotated’ to be precisely these vectors, i.e. those used in the LEA approximation (27), whilst all other modes are projected to maintain orthogonality to  $\hat{\boldsymbol{\pi}}_{\mathcal{P}}$  and  $\mathbf{1}_{N_{\mathcal{P}}}$ , ensuring a complete basis if all modes are retained, i.e. as  $\epsilon \rightarrow 0$  the exact result is recovered.

In practice,  $\mathcal{S}_{\mathcal{P}}$  is determined exactly as before but the local basis that forms  $\mathbf{X}_1^{\mathcal{P}}, \mathbf{Y}_1^{\mathcal{P}}$  is modified by the ‘equilibration projection’ (denoted SR-EP)

$$\begin{aligned} \mathbf{x}_0^{\mathcal{P}} &\rightarrow \hat{\boldsymbol{\pi}}^{\mathcal{P}}, & \mathbf{x}_l^{\mathcal{P}} &\rightarrow [\mathbb{I}_{N_{\mathcal{P}}} - \hat{\boldsymbol{\pi}}^{\mathcal{P}} \otimes \mathbf{1}_{N_{\mathcal{P}}}] \mathbf{x}_l^{\mathcal{P}}, \\ \mathbf{y}_0^{\mathcal{P}} &\rightarrow \mathbf{1}_{N_{\mathcal{P}}}, & \mathbf{y}_l^{\mathcal{P}} &\rightarrow \mathbf{y}_l^{\mathcal{P}} [\mathbb{I}_{N_{\mathcal{P}}} - \hat{\boldsymbol{\pi}}^{\mathcal{P}} \otimes \mathbf{1}_{N_{\mathcal{P}}}] . \end{aligned} \quad (28)$$

This new basis retains completeness as  $\epsilon \rightarrow 0$ , but also reduces to the LEA approximation (27) when  $\epsilon$  is large, therefore naturally interpolating between LEA and the exact formulation. We also investigated a modified version of (28), (denoted by SR-EPO) which retains the pairwise orthonormality  $\mathbf{y}_l^{\mathcal{P}} \mathbf{x}_m^{\mathcal{P}} = \delta_{lm}$ , and can thus be considered a rotation of the original rate matrix followed by a coarse-graining. Figure 12 shows that the modified basis vectors (28) are always superior to the simpler LEA formulation, allowing for further systematic improvement in reproducing the KTN observables. The pairwise orthonormality modification gave very slightly worse performance than (28). These approximations will be tested systematically with a wider variety of clustering schemes in future work.

## VI. CONCLUSIONS

We have explored the first passage time distributions for kinetic transition networks representing various synthetic potential energy landscapes, along with the LJ<sub>38</sub> atomic cluster, a realistic benchmark system with competing morphologies separated by a high barrier. In particular, we have examined the mean first passage time (MFPT) and the standard deviation of the first passage time, computed using alternative approaches based on eigendecomposition and kinetic path sampling.<sup>43,44</sup> For the MFPT, accurate reference values are provided by the graph transformation procedure.<sup>37,40</sup> This approach enables us to analyse how observable dynamical properties, such as first passage

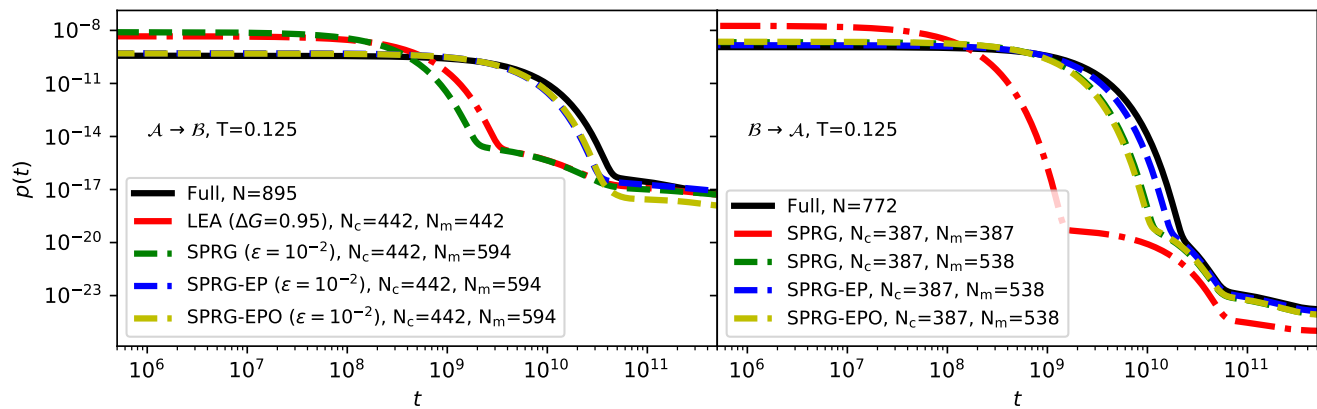


FIG. 12. Exact (black line) and approximate  $p(t)$  between basins of  $LJ_{38}$  at  $T = 0.125$ . Multiple ( $N_c$ ) disjoint clusters are formed using the thresholding technique described above with  $\Delta G = 0.95$ . The LEA approximation (red line) is equivalent to (27), with one mode per cluster ( $N_c = N_m$ ). The spectral reduction (SR) method (green line) had a threshold of  $\epsilon = 10^{-2}$ , giving  $N_m > N_c$  modes. Modifying the SR method (SR-EP) via (28) with the same threshold (blue line) gave better agreement with the exact result and naturally interpolates from the LEA. A further modification (SR-EPO), which enforces pairwise orthogonality for the retained eigenvectors, having set  $\mathbf{x}_0^p \rightarrow \hat{\pi}^p$  and  $\mathbf{y}_0^p \rightarrow \mathbf{1}_{N_p}$  has a slight negative effect on accuracy.

times and rate constants, are encoded in the underlying energy landscape, and provides a guide to converging the corresponding kinetic transition network.

Eigendecomposition methods encounter problems with numerical precision in the rare event regime that is usually of interest, as expected.<sup>38,39</sup> The graph transformation procedure provides access to the MFPT even for strongly metastable systems,<sup>38</sup> but does not conserve higher moments of the first passage time distribution. Kinetic path sampling (kPS), which is based on local graph transformations to compute regional escape times,<sup>43,44</sup> is more numerically robust than methods based on eigendecomposition, and enables higher moments of the FPT distribution to be calculated even for metastable transition networks. However, the sampling effort required to converge higher moments of the FPT distribution, which are strongly influenced by rare extremal values at the tail of the distribution, may be prohibitive for complex systems. Nonetheless, when eigendecomposition fails kPS is the method of choice for obtaining the FPT distribution, and hence insightful mechanistic information, for dynamical systems featuring rare events.<sup>67</sup>

We investigated the extent to which regrouping<sup>48–57</sup> the states in a kinetic transition network is helpful for improving the numerical performance of eigendecomposition. Dimensionality reduction of kinetic transition networks via cluster is attractive because eigendecomposition has cubic scaling with system size,<sup>38,39</sup> and the dynamics of the lumped network may be more readily interpreted in terms of the key features of the underlying energy landscape. Many enhanced sampling algorithms for explicit simulation of the dynamics on transition networks, including kPS, are based on a partitioning of the network. Inappropriate choice of the groups will lead to a lumped transition network that does not properly preserve the kinetics of the original network, and will have a severe adverse effect on the efficiency of dynamical simulations.<sup>67</sup>

We considered the groups determined by a recursive regrouping scheme<sup>60</sup> as a function of the regrouping threshold, and compared the performance of three different formulations for the intergroup rates in reproducing the reference MFPTs and rate constants. The simplest intergroup rate scheme assumes local equilibrium within each group on the timescale of interest,<sup>60</sup> which is expected to work well if the eigenspectrum of the transition network exhibits a spectral gap,<sup>98</sup> provided that the regrouping preserves this structure. We find that the recursive regrouping scheme permits access to lower temperatures than eigendecomposition, but the improvement is relatively small. Recursive regrouping with the local equilibrium approximation also provides a simple and effective means to pre-process a transition network and eliminate the internode transitions that are responsible for the most severe flickering effects, which may otherwise preclude dynamical simulations.<sup>67</sup> Alternative formulations of the intergroup rate constants, based on extended matching of correlation functions for the occupation probabilities of the states,<sup>101,111</sup> reproduce the kinetics of the original transition network more accurately than the local equilibrium approximation. However, evaluation of these expressions requires matrix inversion operations, which result in loss of numerical precision in the metastable regime that is usually of interest.

We have described a new scheme for approximating the dynamics between metastable sets of nodes, based on spectral analysis with a hierarchical truncation of the relaxation modes. The preliminary results from this approach

are promising, and we will provide systematic benchmarking results in a future report. Finally, we note that using graph transformation to eliminate nodes provides yet another approach to the dimensionality reduction of transition networks, which is quite distinct from methods based on the aggregation of sets of nodes and subsequent estimation of intergroup transition rates. An analysis of this approach will also be presented elsewhere.

The FPT distribution contains significantly more information than the MFPT alone. In particular, complex features may manifest in the FPT distributions even for systems where the Arrhenius plots are deceptively simple. Although the MFPT is the usual dynamical observable, detailed comparison of the computed FPT distribution with results for more advanced (for example, single-molecule) experiments is feasible. Eigendecomposition of a kinetic transition network yields detailed mechanistic information by separating the FPT distribution into contributions from orthogonal dynamical eigenmodes, and kPS simulations allow trajectory information to be obtained even for strongly metastable transition networks.<sup>67</sup> Therefore, although computation of the FPT distribution is significantly more challenging than calculation of the MFPT, analysis of the FPT distribution yields rich insight into the dynamics of a transition network, which makes the additional effort very worthwhile.

## VII. ACKNOWLEDGEMENTS

DJW gratefully acknowledges financial support from the Engineering and Physical Sciences Research Council. TDS recognises support from the Agence Nationale de Recherche, via the MEMOPAS project ANR-19-CE46-0006-1. DJS gratefully acknowledges the Cambridge Commonwealth, European and International Trust for a PhD scholarship. DK gratefully recognises support from the Marshall Scholarship.

## VIII. SOFTWARE AND DATA AVAILABILITY

The graph transformation and dimensionality reduction analysis used the freely available PyGT python package<sup>61</sup>, whilst the kinetic path sampling simulations presented were performed with DISCOTRESS simulation Suite<sup>62</sup>. The data that support the findings of this study are available from the corresponding author upon reasonable request.

### Appendix A: Form of reduced rate matrix following spectral reduction

We recall that the full master equation matrix can then be reconstructed, to within the approximation induced by the truncation, via

$$\begin{aligned} \mathbf{M} &\simeq \begin{bmatrix} \mathbb{I}_Q & \\ & \mathbf{X}_1^{\mathcal{P}} \end{bmatrix} \mathbf{M}_1 \begin{bmatrix} \mathbb{I}_Q & \\ & \mathbf{Y}_1^{\mathcal{P}} \end{bmatrix} \equiv \tilde{\mathbf{X}}_1 \mathbf{M}_1 \tilde{\mathbf{Y}}_1, \\ \mathbf{M}_1 &= \begin{bmatrix} \mathbf{M}_{QQ} & \mathbf{M}_{QP} \mathbf{X}_1^{\mathcal{P}} \\ \mathbf{Y}_1^{\mathcal{P}} \mathbf{M}_{PQ} & \mathbf{Y}_1^{\mathcal{P}} \mathbf{M}_{PP} \mathbf{X}_1^{\mathcal{P}} \end{bmatrix}, \end{aligned} \quad (\text{A1})$$

where the omitted off-diagonal elements are rectangular matrices of zeros, with appropriate dimensions, and the first equation defines the rectangular matrices  $\tilde{\mathbf{X}}_1 \in \mathbb{R}^{(N_Q+N_P) \times (N_Q+n_1^{\mathcal{P}})}$  and  $\tilde{\mathbf{Y}}_1 \in \mathbb{R}^{(N_Q+n_1^{\mathcal{P}}) \times (N_Q+N_P)}$ .

When  $\mathbf{Y}_1^{\mathcal{P}}, \mathbf{X}_1^{\mathcal{P}}$  are constructed from eigenvectors of  $\mathbf{M}_{PP}$  we have the additional simplification  $\mathbf{Y}_1^{\mathcal{P}} \mathbf{M}_{PP} \mathbf{X}_1^{\mathcal{P}} = -\boldsymbol{\lambda}_1^{\mathcal{P}}$ , where  $\boldsymbol{\lambda}_1^{\mathcal{P}}$  is the diagonal matrix containing the  $n_1^{\mathcal{P}}$  eigenvalues  $\lambda_l^{\mathcal{P}}, l \in \mathcal{S}_{\mathcal{P}}$ . In full dimensionality  $\mathbf{Y}_1^{\mathcal{P}} \mathbf{X}_1^{\mathcal{P}} = \mathbb{I}_{N_P}$ , and  $\mathbf{Y}_1^{\mathcal{P}} = (\tilde{\mathbf{X}}_1^{\mathcal{P}})^{-1}$ , so that  $\tilde{\mathbf{X}}_1^{\mathcal{P}} \mathbf{Y}_1^{\mathcal{P}} = \mathbb{I}_{N_P}$ . In reduced dimensionality  $n_1^{\mathcal{P}} < N_P$  we have  $\mathbf{Y}_1^{\mathcal{P}} \mathbf{X}_1^{\mathcal{P}} = \mathbb{I}_{n_P}$ , and the accuracy of the reconstruction (A1) depends on the approximation  $\tilde{\mathbf{X}}_1^{\mathcal{P}} \mathbf{Y}_1^{\mathcal{P}} \approx \mathbb{I}_{N_P}$ .

### Appendix B: Spectral reduction for multiple clusters

The treatment of multiple disjoint clusters  $\mathcal{P}, \mathcal{R}, \dots$  in a single iteration is identical to the presentation above with one block matrix per cluster. With multiple clusters the only change is to the projection matrices  $\tilde{\mathbf{X}}_1, \tilde{\mathbf{Y}}_1$ , which become (for two clusters  $\mathcal{P}, \mathcal{R}$ )

$$\tilde{\mathbf{Y}}_1 \rightarrow \begin{bmatrix} \mathbb{I}_Q & & \\ & \mathbf{Y}_1^{\mathcal{P}} & \\ & & \mathbf{Y}_1^{\mathcal{R}} \end{bmatrix}, \quad \tilde{\mathbf{X}}_1 \rightarrow \begin{bmatrix} \mathbb{I}_Q & & \\ & \mathbf{X}_1^{\mathcal{P}} & \\ & & \mathbf{X}_1^{\mathcal{R}} \end{bmatrix}, \quad (\text{B1})$$

with new dimensions  $\tilde{\mathbf{X}}_1 \in \mathbb{R}^{(N_Q+N_P+N_R) \times (N_Q+n_1^P+n_1^R)}$  and  $\tilde{\mathbf{Y}}_1 \in \mathbb{R}^{(N_Q+n_1^P+n_1^R) \times (N_Q+N_P+N_R)}$ . As before, the accuracy of the approximation is controlled by  $\mathbf{X}_1^P \mathbf{Y}_1^P \simeq \mathbb{I}_P$ ,  $\mathbf{X}_1^R \mathbf{Y}_1^R \simeq \mathbb{I}_R$ , etc. The hierarchy of matrices used for (25) is thus unchanged.

In general, the clustering is a disjoint partitioning of the minima into sets  $\mathcal{P}$ ,  $\mathcal{Q}$ ,  $\mathcal{R}$ , etc. A single iteration of the spectral reduction then involves matrices

$$\tilde{\mathbf{Y}}_1 \rightarrow \begin{bmatrix} \mathbf{Y}_1^P & & \\ & \mathbf{Y}_1^Q & \\ & & \mathbf{Y}_1^R \end{bmatrix}, \quad \tilde{\mathbf{X}}_1 \rightarrow \begin{bmatrix} \mathbf{X}_1^P & & \\ & \mathbf{X}_1^Q & \\ & & \mathbf{X}_1^R \end{bmatrix},$$

$$\mathbf{M}_1 \rightarrow \begin{bmatrix} \mathbf{Y}_1^P \mathbf{M}_{PP} \mathbf{X}_1^P & \mathbf{Y}_1^P \mathbf{M}_{PQ} \mathbf{X}_1^Q & \mathbf{Y}_1^P \mathbf{M}_{PR} \mathbf{X}_1^R \\ \mathbf{Y}_1^Q \mathbf{M}_{QP} \mathbf{X}_1^P & \mathbf{Y}_1^Q \mathbf{M}_{QQ} \mathbf{X}_1^Q & \mathbf{Y}_1^Q \mathbf{M}_{QR} \mathbf{X}_1^R \\ \mathbf{Y}_1^R \mathbf{M}_{RP} \mathbf{X}_1^P & \mathbf{Y}_1^R \mathbf{M}_{RQ} \mathbf{X}_1^Q & \mathbf{Y}_1^R \mathbf{M}_{RR} \mathbf{X}_1^R \end{bmatrix}.$$

For any cluster that is not being reduced we simply replace  $\mathbf{Y}_1$  and  $\mathbf{X}_1$  by the identity matrix in this scheme. Eigendecomposition of  $\mathbf{M}_1 = -\tilde{\mathbf{X}}_2 \boldsymbol{\lambda}_2 \tilde{\mathbf{Y}}_2$  now involves a matrix of dimension  $N_m \times N_m$ , where  $N_m = n_1^P + n_1^Q + n_1^R$ , and produces the first passage time distribution

$$p(t) \simeq \mathbf{1}_N \tilde{\mathbf{X}}_1 \tilde{\mathbf{X}}_2 \boldsymbol{\lambda}_2 \exp(-\boldsymbol{\lambda}_2 t) \tilde{\mathbf{Y}}_2 \tilde{\mathbf{Y}}_1 \mathbf{P}(0), \quad (\text{B2})$$

in the same way as for equation (25).

## REFERENCES

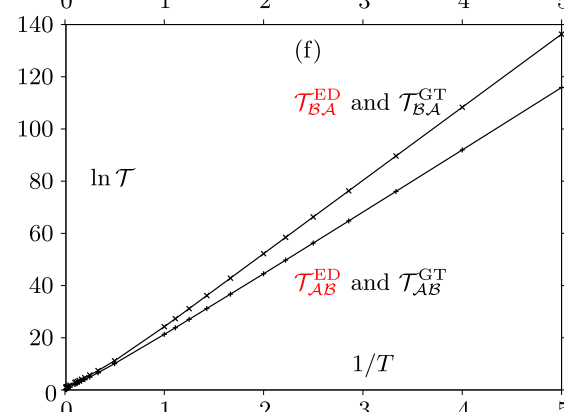
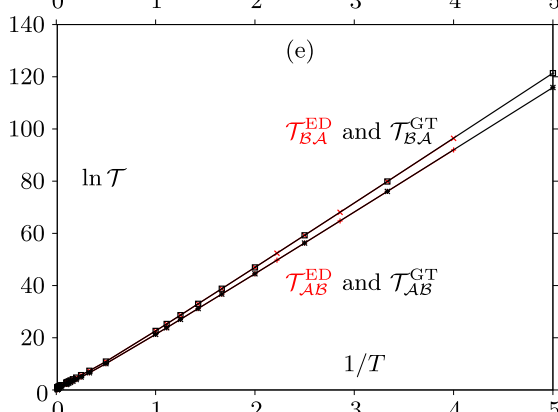
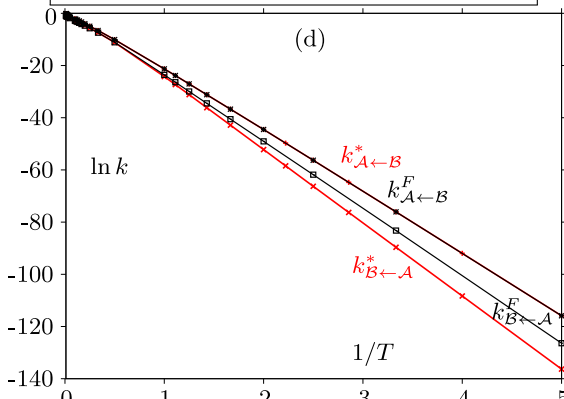
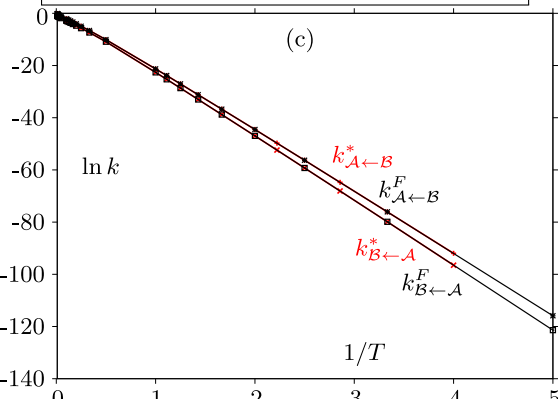
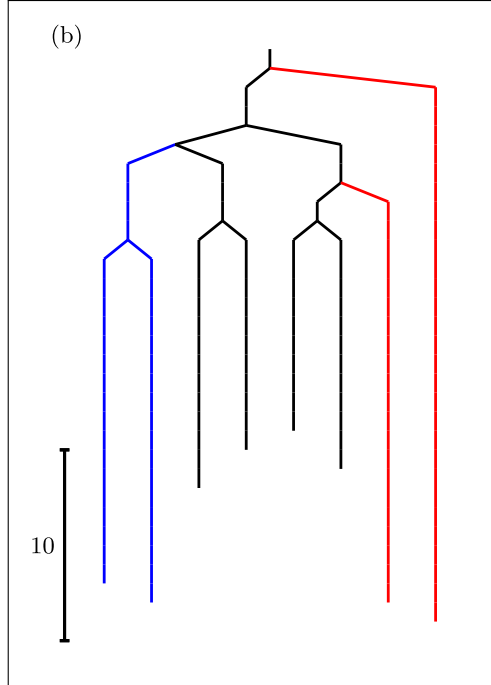
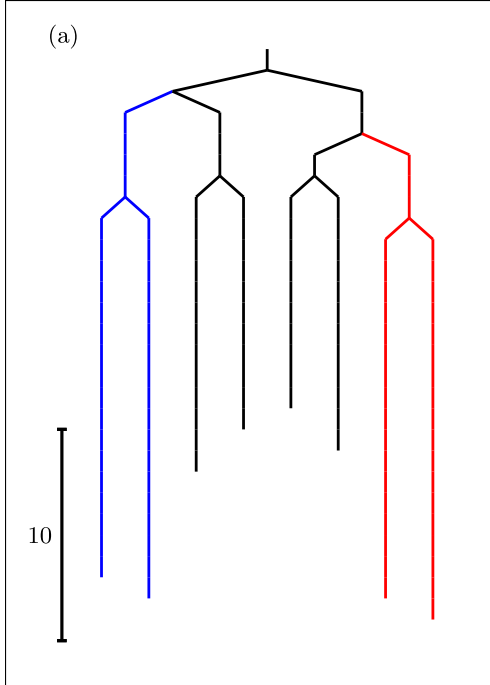
- <sup>1</sup>F. Noé and S. Fischer, *Curr. Opin. Struct. Biol.* **18**, 154 (2008).
- <sup>2</sup>D. Prada-Gracia, J. Gómez-Gardenes, P. Echenique, and F. Fernando, *PLoS Comput. Biol.* **5**, e1000415 (2009).
- <sup>3</sup>D. J. Wales, *Curr. Opin. Struct. Biol.* **20**, 3 (2010).
- <sup>4</sup>D. J. Wales, *Mol. Phys.* **100**, 3285 (2002).
- <sup>5</sup>D. J. Wales, *Mol. Phys.* **102**, 891 (2004).
- <sup>6</sup>D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, 2003).
- <sup>7</sup>W. Forst, *Theory of Unimolecular Reactions* (Academic Press, New York, 1973).
- <sup>8</sup>K. J. Laidler, *Chemical Kinetics* (Harper & Row, New York, 1987).
- <sup>9</sup>N. G. van Kampen, *Stochastic processes in physics and chemistry* (North-Holland, Amsterdam, 1981).
- <sup>10</sup>R. E. Kunz, *Dynamics of First-Order Phase Transitions* (Deutsch, Thun, 1995).
- <sup>11</sup>C. Schütte, A. Fischer, W. Huisinga, and P. Deuffhard, *J. Comput. Phys.* **151**, 146 (1999).
- <sup>12</sup>M. Shirts and V. S. Pande, *Science* **290**, 1903 (2000).
- <sup>13</sup>N. Singhal, C. D. Snow, and V. S. Pande, *J. Chem. Phys.* **121**, 415 (2004).
- <sup>14</sup>W. C. Swope, J. W. Pitera, and F. Suits, *J. Phys. Chem. B* **108**, 6571 (2004).
- <sup>15</sup>J. D. Chodera, K. A. Dill, N. Singhal, V. S. Pande, W. C. Swope, and J. W. Pitera, *J. Chem. Phys.* **126**, 155101 (2007).
- <sup>16</sup>V. S. Pande, K. Beauchamp, and G. R. Bowman, *Methods* **52**, 99 (2010).
- <sup>17</sup>J. H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schutte, and F. Noé, *J. Chem. Phys.* **134**, 174105 (2011).
- <sup>18</sup>B. E. Husic and V. S. Pande, *J. Am. Chem. Soc.* **140**, 2386 (2018).
- <sup>19</sup>T. D. Swinburne and D. Perez, *Phys. Rev. Materials* **2**, 053802 (2018).
- <sup>20</sup>J. A. Joseph, K. Röder, D. Chakraborty, R. G. Mantell, and D. J. Wales, *Chem. Commun.* **53**, 6974 (2017).
- <sup>21</sup>D. J. Wales, *Ann. Rev. Phys. Chem.* **69**, 401 (2018).
- <sup>22</sup>K. Röder, J. A. Joseph, B. E. Husic, and D. J. Wales, *Adv. Theory Simul.* **2**, 1800175 (2019).
- <sup>23</sup>G. A. Huber and S. Kim, *Biophys. J.* **70**, 97 (1996).
- <sup>24</sup>C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, *J. Chem. Phys.* **108**, 1964 (1998).
- <sup>25</sup>P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, *Annu. Rev. Phys. Chem.* **53**, 291 (2002).
- <sup>26</sup>T. S. van Erp, D. Moroni, and P. G. Bolhuis, *J. Chem. Phys.* **118**, 7762 (2003).
- <sup>27</sup>A. K. Faradjian and R. Elber, *J. Chem. Phys.* **120**, 10880 (2004).
- <sup>28</sup>A. M. A. West, R. Elber, and D. Shalloway, *J. Chem. Phys.* **126**, 145104 (2007).
- <sup>29</sup>G. Bussi, F. L. Gervasio, A. Laio, and M. Parrinello, *J. Amer. Chem. Soc.* **128**, 13435 (2006).
- <sup>30</sup>R. J. Allen, D. Frenkel, and P. R. t. Wolde, *J. Chem. Phys.* **124**, 024102 (2006).
- <sup>31</sup>E. Vanden-Eijnden, M. Venturoli, G. Ciccotti, and R. Elber, *J. Chem. Phys.* **129**, 174102 (2008).
- <sup>32</sup>K. Kuczera, G. S. Jas, and R. Elber, *J. Phys. Chem. A* **113**, 7461 (2009).



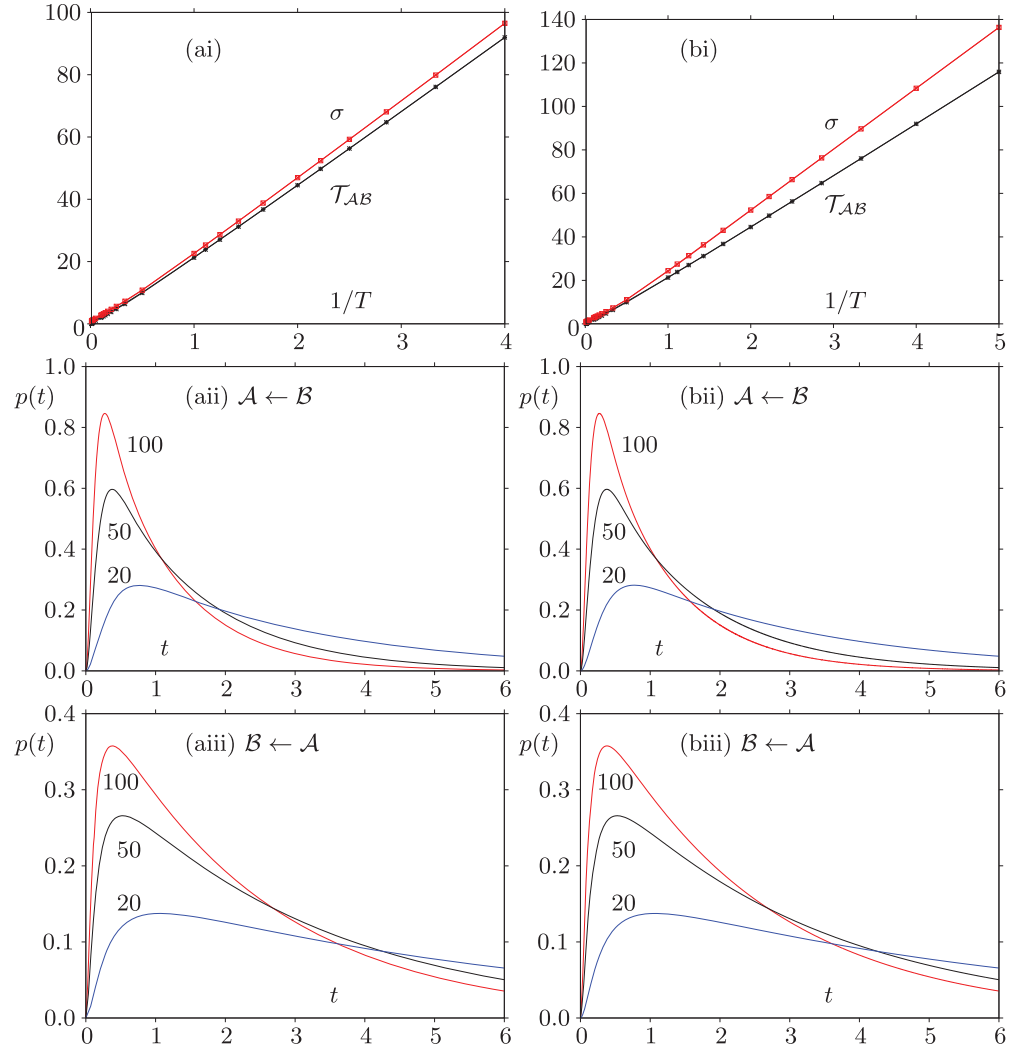
- <sup>33</sup>B. W. Zhang, D. Jasnow, and D. M. Zuckerman, *J. Chem. Phys.* **132**, 054107 (2010).
- <sup>34</sup>E. Vanden-Eijnden and M. Venturoli, *J. Chem. Phys.* **130**, 194101 (2009).
- <sup>35</sup>R. Elber, J. Bello-Rivas, P. Ma, A. Cardenas, and A. Fathizadeh, *Entropy* **19**, 219 (2017).
- <sup>36</sup>D. J. Wales, *Int. Rev. Phys. Chem.* **25**, 237 (2006).
- <sup>37</sup>D. J. Wales, *J. Chem. Phys.* **130**, 204111 (2009).
- <sup>38</sup>J. D. Stevenson and D. J. Wales, *J. Chem. Phys.* **141**, 041104 (2014).
- <sup>39</sup>T. J. Frankcombe and S. C. Smith, *Theor. Chem. Acc.* **124**, 303 (2009).
- <sup>40</sup>S. A. Trygubenko and D. J. Wales, *J. Chem. Phys.* **124**, 234110 (2006).
- <sup>41</sup>R. S. MacKay and J. D. Robinson, *Phil. Trans. R. Soc. A* **376**, 20170232 (2018).
- <sup>42</sup>T. D. Swinburne and D. J. Wales, *Journal of Chemical Theory and Computation* **0**, null (2020).
- <sup>43</sup>M. Athènes and V. V. Bulatov, *Phys. Rev. Lett.* **113**, 230601 (2014).
- <sup>44</sup>M. Athènes, S. Kaur, G. Adjanor, T. Vanacker, and T. Jourdan, *Phys. Rev. Materials* **3**, 103802 (2019).
- <sup>45</sup>H. S. Chung, K. McHale, J. M. Louis, and W. A. Eaton, *Science* **335**, 981 (2012).
- <sup>46</sup>F. H. Stillinger and T. A. Weber, *Science* **225**, 983 (1984).
- <sup>47</sup>J. P. K. Doye and D. J. Wales, *J. Chem. Phys.* **116**, 3777 (2002).
- <sup>48</sup>J. G. Kemeny, *Linear Algebra Appl.* **38**, 193 (1981).
- <sup>49</sup>P. Deuffhard, W. Huisinga, A. Fischer, and C. Schütte, *Linear Algebra Appl.* **315**, 39 (2000).
- <sup>50</sup>P. Deuffhard and M. Weber, *Linear Algebra Appl.* **398**, 161 (2005).
- <sup>51</sup>S. Kube and M. Weber, *J. Chem. Phys.* **126**, 024103 (2007).
- <sup>52</sup>W. E, T. Li, and E. Vanden-Eijnden, *Proc. Natl. Acad. Sci. USA* **105**, 7907 (2008).
- <sup>53</sup>N. Lempesis, D. G. Tsalikis, G. C. Boulougouris, and D. N. Theodorou, *J. Chem. Phys.* **135**, 204507 (2011).
- <sup>54</sup>G. R. Bowman, L. Meng, and X. Huang, *J. Chem. Phys.* **139**, 121905 (2013).
- <sup>55</sup>S. Orioli and P. Faccioli, *J. Chem. Phys.* **145**, 124120 (2016).
- <sup>56</sup>W. Wang, T. Liang, F. K. Sheong, X. Fan, and X. Huang, *J. Chem. Phys.* **149**, 072337 (2018).
- <sup>57</sup>J. A. Ward and M. López-García, *Appl. Netw. Sci.* **4**, 108 (2019).
- <sup>58</sup>D. J. Wales and P. Salamon, *Proc. Natl. Acad. Sci. USA* **111**, 617 (2014).
- <sup>59</sup>D. A. Evans and D. J. Wales, *J. Chem. Phys.* **118**, 3891 (2003).
- <sup>60</sup>J. M. Carr and D. J. Wales, *J. Phys. Chem. B* **112**, 8760 (2008).
- <sup>61</sup>T. D. Swinburne and D. Kannan, "Pygt: Graph transformation and dimensionality reduction in python," <https://github.com/tomswinburne/PyGT> (2020).
- <sup>62</sup>D. J. Sharpe, "Discotress: The discrete state continuous time rare event simulation suite," <https://github.com/danieljsharpe/DISCOTRESS> (2020).
- <sup>63</sup>M. Manhart, W. Kion-Crosby, and A. V. Morozov, *J. Chem. Phys.* **143**, 214106 (2015).
- <sup>64</sup>M. Manhart and A. V. Morozov, *Phys. Rev. Lett.* **111**, 088102 (2013).
- <sup>65</sup>M. Manhart and A. V. Morozov, *Proc. Natl. Acad. Sci. USA* **112**, 1797 (2015).
- <sup>66</sup>D. J. Sharpe and D. J. Wales, *J. Chem. Phys.* **151**, 124101 (2019).
- <sup>67</sup>D. J. Sharpe and D. J. Wales, "Efficient and exact sampling of transition path ensembles on markovian networks," (submitted).
- <sup>68</sup>N.-V. Buchete and G. Hummer, *J. Phys. Chem. B* **112**, 6057 (2008).
- <sup>69</sup>E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed. (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1999).
- <sup>70</sup>R. Lehoucq, K. Maschhoff, D. Sorensen, and C. Yang, "Arpack, a collection of fortran77 subroutines designed to solve large scale eigenvalue problems," <https://www.caam.rice.edu/software/ARPACK/>.
- <sup>71</sup>S. A. Trygubenko and D. J. Wales, *Mol. Phys.* **104**, 1497 (2006).
- <sup>72</sup>P. Metzner, C. Schütte, and E. Vanden-Eijnden, *Multiscale Model. Simul.* **7**, 1192 (2009).
- <sup>73</sup>A. J. Ballard, S. Martiniani, J. D. Stevenson, S. Somani, and D. J. Wales, *WIREs Comput. Mol. Sci.* **5**, 273 (2015).
- <sup>74</sup>M. A. Novotny, *Phys. Rev. Lett.* **74**, 1 (1995).
- <sup>75</sup>M. A. Novotny, in *Annual Reviews of Computational Physics: Vol. 9*, edited by D. Stauffer (World Scientific, Singapore, 2001) pp. 153-210.
- <sup>76</sup>M. A. Novotny, *Comput. Phys. Commun.* **147**, 659 (2002).
- <sup>77</sup>O. M. Becker and M. Karplus, *J. Chem. Phys.* **106**, 1495 (1997).
- <sup>78</sup>D. J. Wales, M. A. Miller, and T. R. Walsh, *Nature* **394**, 758 (1998).
- <sup>79</sup>J. E. Jones and A. E. Ingham, *Proc. R. Soc. A* **107**, 636 (1925).
- <sup>80</sup>M. A. Miller, J. P. K. Doye, and D. J. Wales, *J. Chem. Phys.* **110**, 328 (1999).
- <sup>81</sup>J. P. K. Doye, M. A. Miller, and D. J. Wales, *J. Chem. Phys.* **110**, 6896 (1999).
- <sup>82</sup>D. J. Wales and T. V. Bogdan, *J. Phys. Chem. B* **110**, 20765 (2006).
- <sup>83</sup>D. J. Wales and J. P. K. Doye, *J. Phys. Chem. A* **101**, 5111 (1997).

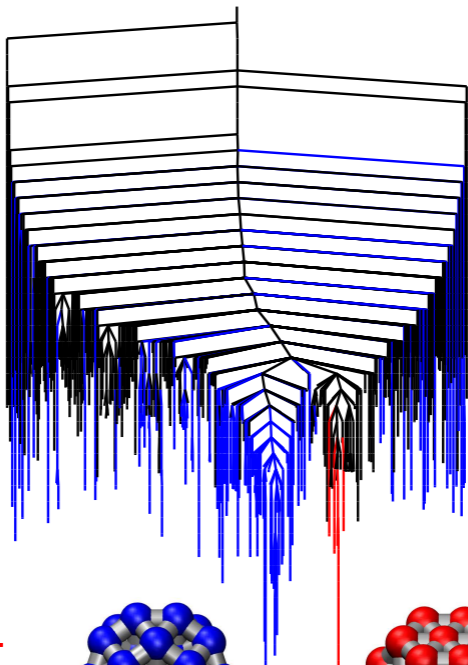
- <sup>84</sup>D. J. Wales and H. A. Scheraga, *Science* **285**, 1368 (1999).
- <sup>85</sup>M. T. Oakley, R. L. Johnston, and D. J. Wales, *Phys. Chem. Chem. Phys.* **15**, 3965 (2013).
- <sup>86</sup>M. Dittner and B. Hartke, *Comput. and Theor. Chem.* **1107**, 7 (2017).
- <sup>87</sup>J. P. K. Doye, D. J. Wales, and M. A. Miller, *J. Chem. Phys.* **109**, 8143 (1998).
- <sup>88</sup>F. Calvo, J. P. Neirotti, D. L. Freeman, and J. D. Doll, *J. Chem. Phys.* **112**, 10350 (2000).
- <sup>89</sup>J. P. Neirotti, F. Calvo, D. L. Freeman, and J. D. Doll, *J. Chem. Phys.* **112**, 10340 (2000).
- <sup>90</sup>P. A. Frantsuzov and V. A. Mandelshtam, *Phys. Rev. E* **72**, 037102 (2005).
- <sup>91</sup>C. Predescu, P. A. Frantsuzov, and V. A. Mandelshtam, *J. Chem. Phys.* **122**, 154305 (2005).
- <sup>92</sup>H. B. Liu and K. D. Jordan, *J. Phys. Chem. B* **109**, 5203 (2005).
- <sup>93</sup>G. Adjanor, M. Athenes, and F. Calvo, *Euro. Phys. J. B* **53**, 47 (2006).
- <sup>94</sup>V. A. Sharapov and V. A. Mandelshtam, *J. Phys. Chem. A* **111**, 10284 (2007).
- <sup>95</sup>V. A. Sharapov, D. Meluzzi, and V. A. Mandelshtam, *Phys. Rev. Lett.* **98**, 105701 (2007).
- <sup>96</sup>F. Calvo, *Phys. Rev. E* **82**, 046703 (2010).
- <sup>97</sup>M. Picciani, M. Athenes, J. Kurchan, and J. Tailleur, *J. Chem. Phys.* **135**, 034108 (2011).
- <sup>98</sup>M. Cameron and E. Vanden-Eijnden, *J. Stat. Phys.* **156**, 427 (2014).
- <sup>99</sup>D. J. Wales, *J. Chem. Phys.* **142**, 130901 (2015).
- <sup>100</sup>D. J. Wales, The Cambridge Landscapes Database, URL <http://www-wales.ch.cam.ac.uk/CCD.html>.
- <sup>101</sup>G. Hummer and A. Szabo, *J. Phys. Chem. B* **119**, 9029 (2015).
- <sup>102</sup>M. E. J. Newman, *Eur. Phys. J. B* **38**, 321 (2004).
- <sup>103</sup>L. Danon, J. Duch, A. Díaz-Guilera, and A. Arenas, *J. Stat. Mech.* **9**, P09008 (2005).
- <sup>104</sup>A. Lancichinetti and S. Fortunato, *Phys. Rev. E* **80**, 056117 (2009).
- <sup>105</sup>S. Fortunato, *Phys. Rep.* **486**, 75 (2010).
- <sup>106</sup>F. D. Malliaros and M. Vazirgiannis, *Phys. Rep.* **533**, 95 (2013).
- <sup>107</sup>S. Fortunato and D. Hric, *Phys. Rep.* **659**, 1 (2016).
- <sup>108</sup>Z. Yang, R. Algesheimer, and C. J. Tessone, *Sci. Rep.* **6**, 30750 (2016).
- <sup>109</sup>S. Emmons, S. Kobourov, M. Gallant, and K. Börner, *PLoS ONE* **11**, e0159161 (2016).
- <sup>110</sup>X. Zhang, Z. Ma, Z. Zhang, Q. Sun, and J. Yan, *J. Phys.: Conf. Ser.* **1069**, 012123 (2018).
- <sup>111</sup>A. Kells, V. Koskin, E. Rosta, and A. Annibale, *J. Chem. Phys.* **152**, 104108 (2020).
- <sup>112</sup>S. Pigolotti and A. Vulpiani, *J. Chem. Phys.* **128**, 154114 (2008).
- <sup>113</sup>E. W. Dijkstra, *Numerische Math.* **1**, 269 (1959).
- <sup>114</sup>J. M. Carr, S. A. Trygubenko, and D. J. Wales, *J. Chem. Phys.* **122**, 234903 (2005).
- <sup>115</sup>A. Kells, Z. E. Mihálka, A. Annibale, and E. Rosta, *J. Chem. Phys.* **150**, 134107 (2019).
- <sup>116</sup>L. Martini, A. Kells, R. Covino, G. Hummer, N.-V. Buchete, and E. Rosta, *Phys. Rev. X* **7**, 031060 (2017).



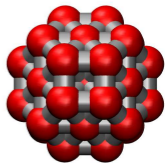
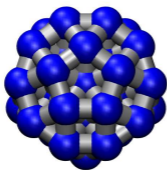


This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.  
PLEASE CITE THIS ARTICLE AS DOI:10.1063/1.50016244

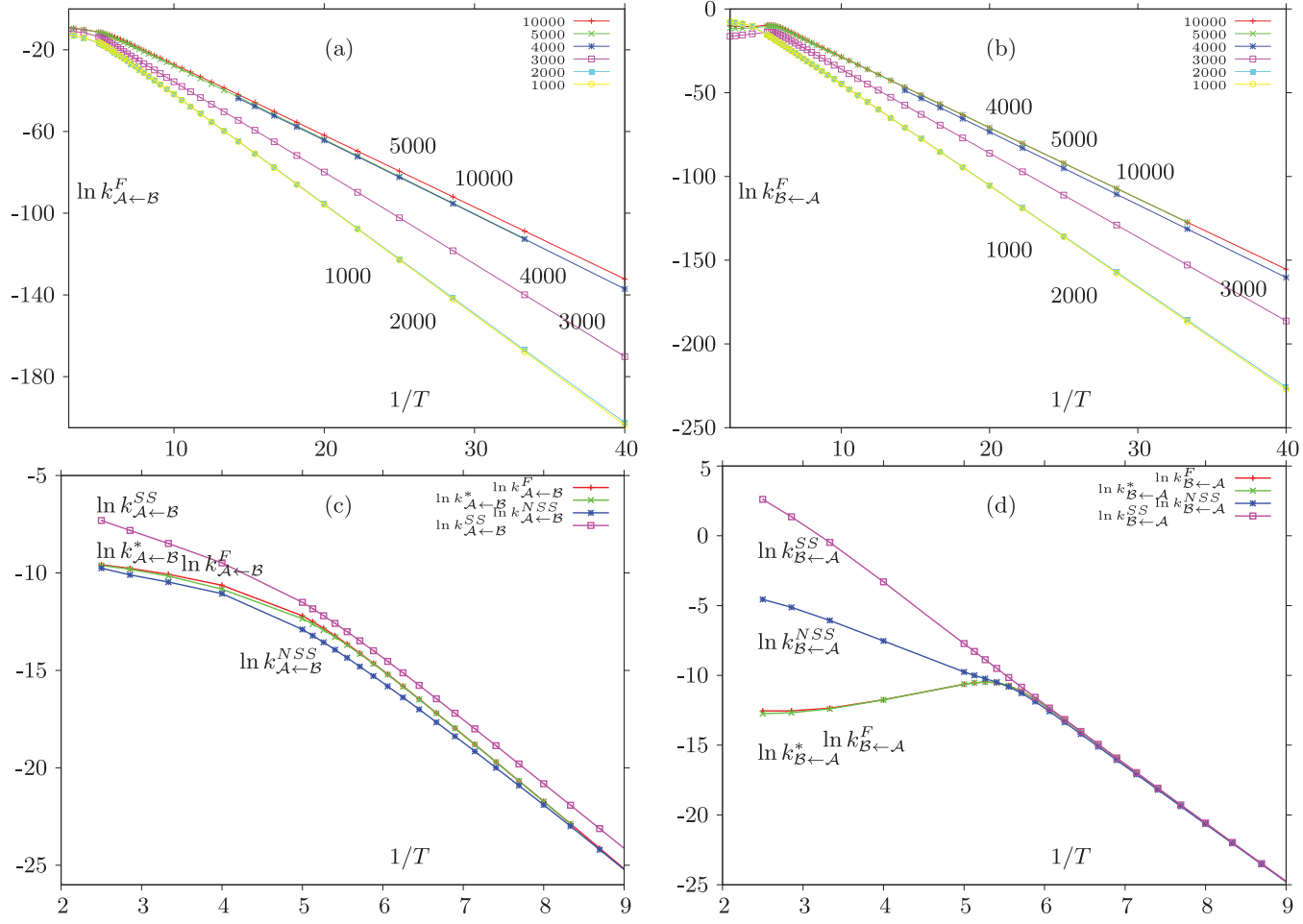


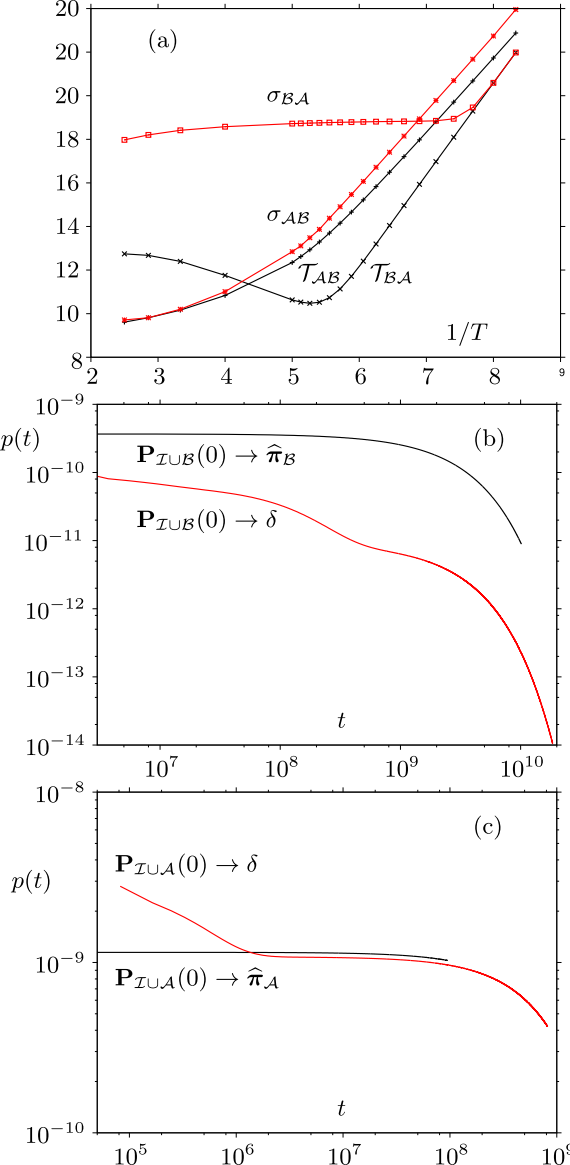


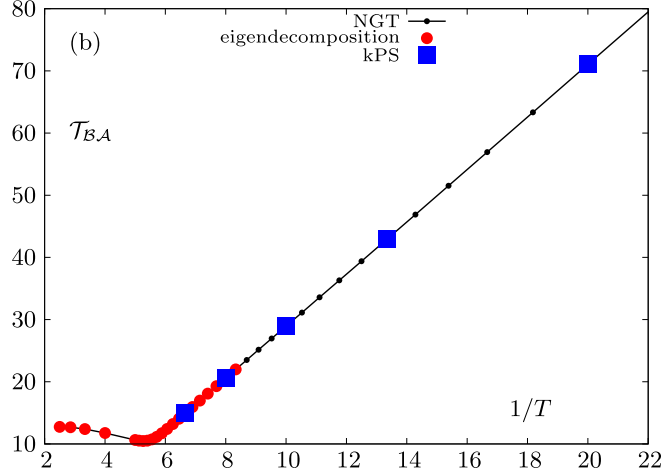
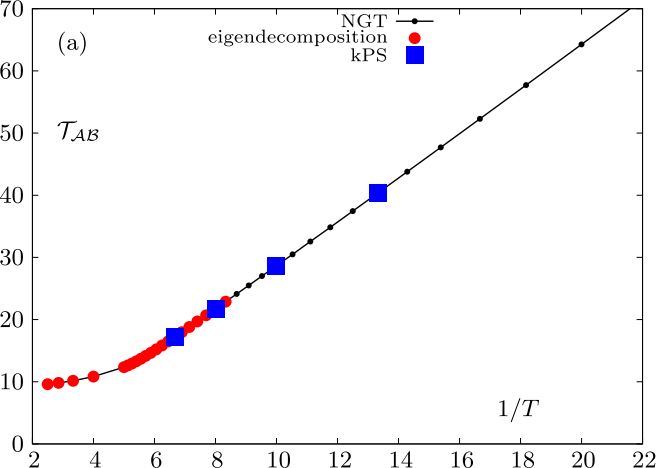
$\epsilon$



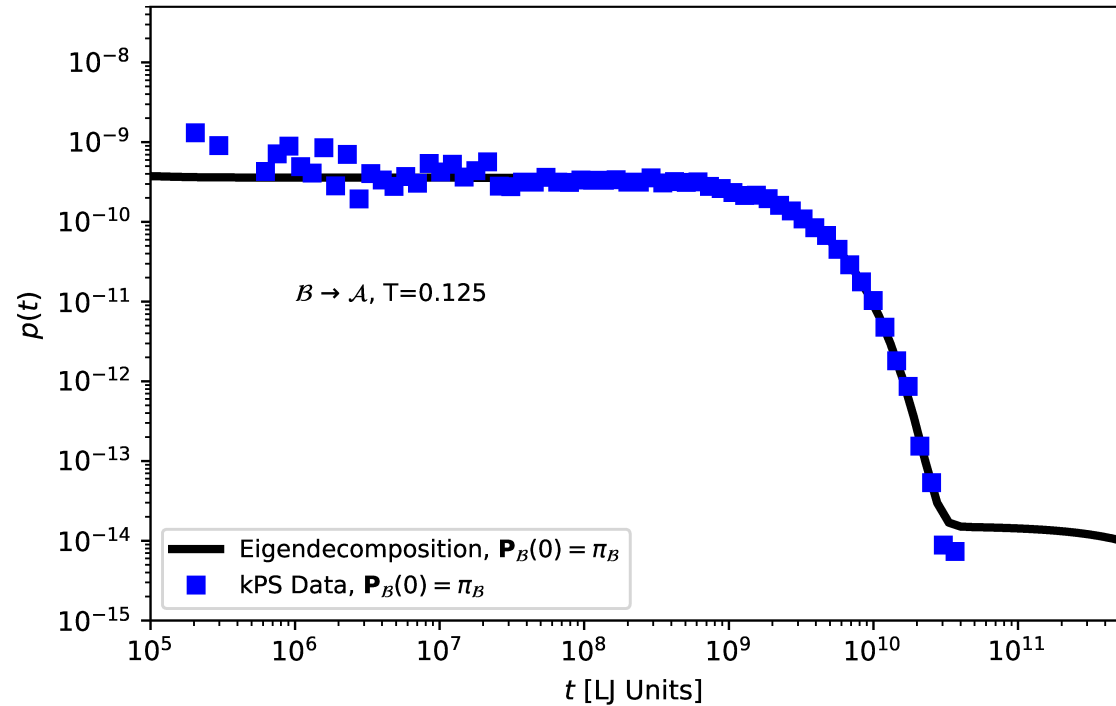
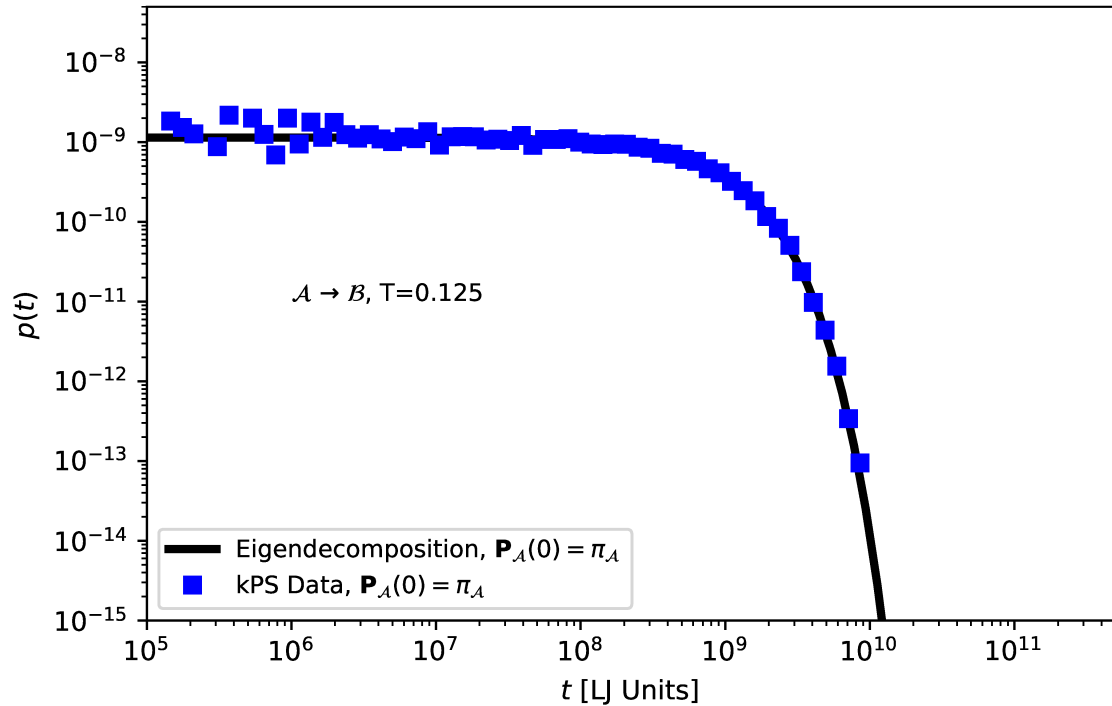
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.  
PLEASE CITE THIS ARTICLE AS DOI:10.1063/1.50016244

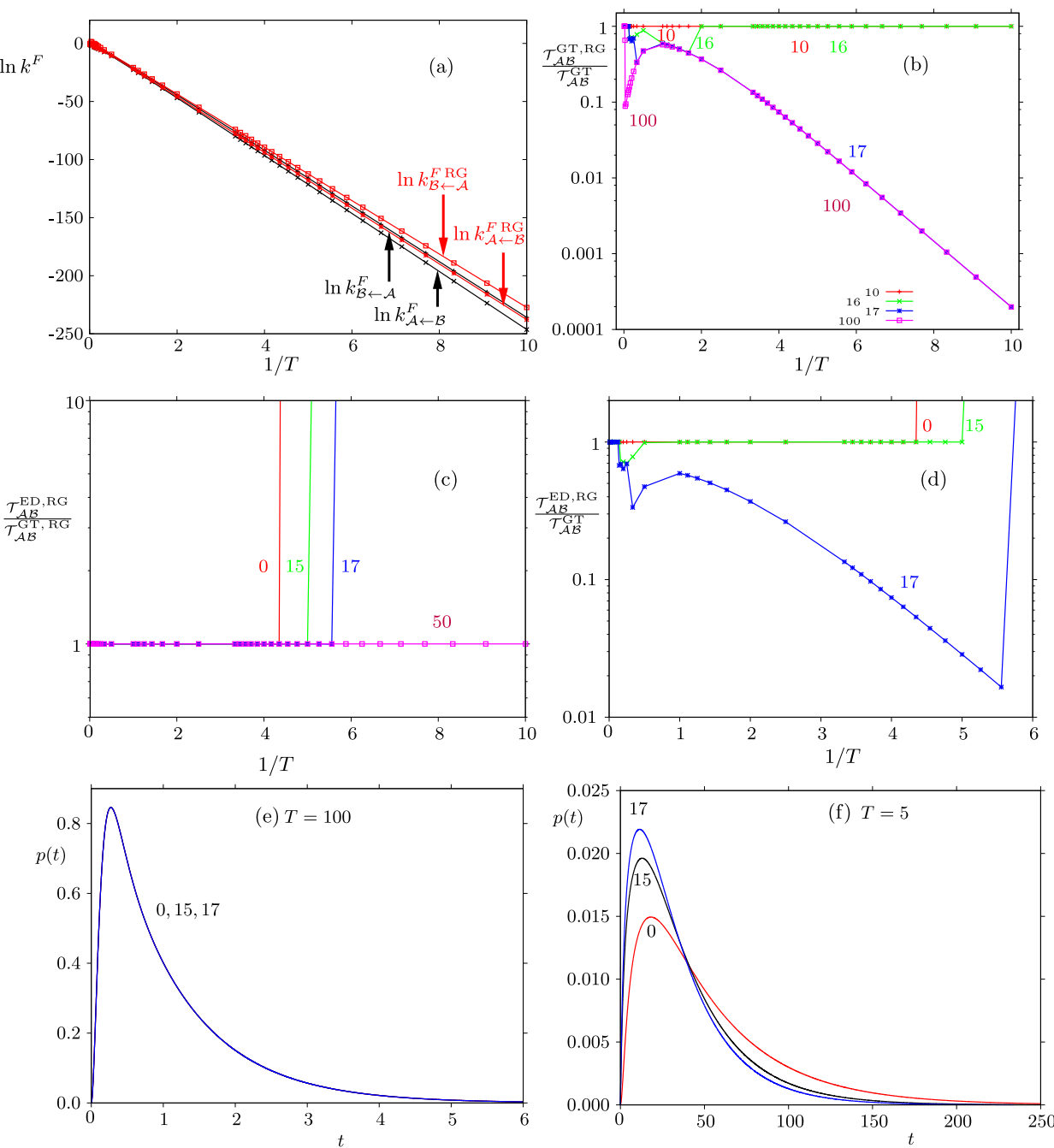












This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.  
PLEASE CITE THIS ARTICLE AS DOI:10.1063/1.50016244

