

PAPER • OPEN ACCESS

Markov models from the square root approximation of the Fokker–Planck equation: calculating the grid-dependent flux

To cite this article: Luca Donati *et al* 2021 *J. Phys.: Condens. Matter* **33** 115902

View the [article online](#) for updates and enhancements.



IOP | ebooks™

Bringing together innovative digital publishing with leading authors from the global scientific community.

Start exploring the collection—download the first chapter of every title for free.

Markov models from the square root approximation of the Fokker–Planck equation: calculating the grid-dependent flux

Luca Donati¹ , Marcus Weber²  and Bettina G Keller^{1,*} 

¹ Department of Biology, Chemistry, Pharmacy, Freie Universität Berlin, Takustraße 3, D-14195 Berlin, Germany

² Zuse Institute Berlin, Takustr. 7, 14195 Berlin, Germany

E-mail: bettina.keller@fu-berlin.de

Received 8 October 2020, revised 26 November 2020

Accepted for publication 22 December 2020

Published 7 January 2021



Abstract

Molecular dynamics (MD) are extremely complex, yet understanding the slow components of their dynamics is essential to understanding their macroscopic properties. To achieve this, one models the MD as a stochastic process and analyses the dominant eigenfunctions of the associated Fokker–Planck operator, or of closely related transfer operators. So far, the calculation of the discretized operators requires extensive MD simulations. The square-root approximation of the Fokker–Planck equation is a method to calculate transition rates as a ratio of the Boltzmann densities of neighboring grid cells times a flux, and can in principle be calculated without a simulation. In a previous work we still used MD simulations to determine the flux. Here, we propose several methods to calculate the exact or approximate flux for various grid types, and thus estimate the rate matrix without a simulation. Using model potentials we test computational efficiency of the methods, and the accuracy with which they reproduce the dominant eigenfunctions and eigenvalues. For these model potentials, rate matrices with up to $\mathcal{O}(10^6)$ states can be obtained within seconds on a single high-performance compute server if regular grids are used.

Keywords: molecular dynamics, Markov state model, Smoluchowski equation, square-root approximation, molecular dynamics simulations, Fokker–Planck equation

 Supplementary material for this article is available [online](#)


(Some figures may appear in colour only in the online journal)

1. Introduction

The dynamics of molecular systems is astonishingly complex. Only a small fraction of their high-dimensional state

space is actually accessible at room temperature. Yet finding out which regions of the state space are accessible, requires sophisticated computer simulations, i.e. molecular dynamics (MD) simulations. MD can be very sensitive to small changes in some variables of the system or the environment, but can also be remarkably robust with respect to changes in other variables. Humanly understandable models of the MD are therefore essential for the elucidation of complex molecular systems.

* Author to whom any correspondence should be addressed

 Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Markov state models (MSMs) represent the conformational dynamics of molecular system as transition probabilities between states in the conformational space [1–6]. From the dominant eigenvectors and eigenvalues of the transition matrix $\mathbf{T}(\tau)$, one can deduce a wealth of useful information on the molecular system, such as the long-lived conformations, the dynamic processes that govern the dynamic equilibrium between them, transition networks and pathways in these networks, and one can quantify the sensitivity of experimental observables with respect to the dynamic processes [7, 8]. MSMs are now a well-established and valuable tool for the elucidation of large molecular systems, and in particular biomolecular systems [9–15].

In the construction of MSMs, one assumes that the MD is a stochastic process. The time-evolution of the probability density is governed by the associated Fokker–Planck equation, or equivalently: the infinitesimal generator of the stochastic process \mathcal{Q} . By formally integrating the Fokker–Planck equation one obtains a transfer operator, whose discretized version is the MSM transition matrix $\mathbf{T}(\tau)$. The matrix elements of $\mathbf{T}(\tau)$ can conveniently be estimated from MD simulations as correlation functions. On the other hand, this means that the accuracy of the MSM stands and falls with the quality of this simulation.

Because MD simulations are costly and slow to converge, enhanced sampling techniques have been developed to speed up the exploration of state space and the convergence of ensemble averages [16–19]. With recently developed dynamic reweighting methods one can additionally recover the correlation functions and thus the MSM of the unbiased system from these biased simulations [20–23]. But despite enhanced sampling techniques, there is usually no way to be certain whether an MD simulation has explored all of the accessible state space, and even assessing whether the sampling within the explored state space has converged can be difficult [24, 25]. Thus, there is ample motivation to investigate avenues to obtain an MSM of a molecular system without generating an MD simulation.

Square root approximation (SqRA) is a technique that approximates the Fokker–Planck equation by a rate matrix [26, 27]. Given a discretization of the state space, the rate from cell Ω_i to cell Ω_j is

$$Q_{ij,\text{adjacent}} = \Phi \frac{S_{ij}}{\mathcal{V}_i} \sqrt{\frac{\pi(x_j)}{\pi(x_i)}},$$

where $\Phi \frac{S_{ij}}{\mathcal{V}_i}$ is the flux of the probability density through the intersecting surface S_{ij} in the absence of any potential energy function, \mathcal{V}_i is the volume of cell Ω_i , and $\pi(x_i)$ and $\pi(x_j)$ are the Boltzmann densities at the centers of cell Ω_i and Ω_j , respectively. We recently derived the SqRA for N_D -dimensional systems by exploiting Gauss’s flux theorem, and showed that for infinitely small grid cells the geometric average of the Boltzmann weights converges to the Smoluchowski diffusion equation, i.e. the Fokker–Planck equation associated to overdamped Langevin dynamics [27, 28]. Previously an analogous formula for one-dimensional systems has been derived from the one-dimensional Smoluchowski equation [29] and

using the maximum caliber (maximum path entropy) approach [30–32]. In addition the geometric average of the Boltzmann weights has been used as reweighting factor in the dynamic histogram analysis method (DHAM) to reweight transition probabilities [21].

The SqRA opens up a way to calculate the transition rates without having to resort to rare-event simulations, at least for systems with not too many degrees of freedom. The ratio of the Boltzmann densities $\pi(x_i)/\pi(x_j)$ can be readily calculated from the potential energy function. The grid volume \mathcal{V}_i and the intersecting surface S_{ij} can be calculated from the discretization of the state space. However, how to best calculate $\Phi \frac{S_{ij}}{\mathcal{V}_i}$ is an open question. In our previous work [27], we assumed that the factor is $\frac{S_{ij}}{\mathcal{V}_i}$ is constant for all grid cells. This is true for hyper-cubic grids and approximately true for Voronoi grids with very small grid cells. We then estimated the factor $\Phi \frac{S_{ij}}{\mathcal{V}_i}$ by comparing the rate matrix to an MSM transition matrix, the construction of which required an MD simulation.

In this contribution, we derive the exact expression for Φ from the equation of the overdamped Langevin dynamics with constant potential, and show that it depends on the diffusion constant and on the discrete Laplace operator. We then compare several methods to calculate $\frac{S_{ij}}{\mathcal{V}_i}$ for different types of discretizations. For regular grids, this ratio can be calculated analytically. For Voronoi grids, we use the quickhull algorithm [33] to calculate $\frac{S_{ij}}{\mathcal{V}_i}$ numerically, and we approximate the ratio by interpolating between all neighbors of the cell Ω_i [34]. We additionally propose a method to calculate $\Phi \frac{S_{ij}}{\mathcal{V}_i}$ by comparing to the analytically known transition probability of a Wiener process (i.e. diffusion at a constant potential energy function). With these methods, we can construct the rate matrix without any MD simulation. We test the methods on model potentials with respect to computational efficiency, the dimensionality of the systems, and accuracy of the resulting rate matrix.

2. Theory

We consider a system of n_p particles that move in the three-dimensional Cartesian space, i.e. in a state space with $N_D = 3n_p$ dimensions: $\Omega \subset \mathbb{R}^{N_D}$. Its dynamics is described by the overdamped Langevin dynamics:

$$dx(t) = -\xi^{-1} \mathbf{M}^{-1} \nabla V(x(t)) dt + \sigma dB(t), \quad (1)$$

where $x(t) \in \Omega$ is the state vector at time t , ξ is a friction parameter with units of 1/s, \mathbf{M} is a diagonal $3n_p \times 3n_p$ -mass matrix, \mathbf{M}^{-1} is its inverse, $V(x)$ is the potential energy function, and $B(t)$ is an N_D -dimensional Wiener process scaled by the diagonal matrix $\sigma = \sqrt{2k_B T \xi^{-1} \mathbf{M}^{-1}}$, where T is the temperature, and k_B is the Boltzmann constant. Equation (1) generates a Markovian, ergodic and reversible process [1, 35].

The time-evolution of the associated probability density $\rho(x, t)$ is given by the following Fokker–Planck equation

$$\begin{aligned} \partial_t \rho(x, t) &= \frac{\sigma^2}{2} \Delta \rho(x, t) + \nabla (\rho(x, t) \cdot \xi^{-1} \mathbf{M}^{-1} \nabla V(x)) \\ &= \mathcal{Q} \rho(x, t), \end{aligned} \quad (2)$$

which is also known as the Smoluchowski diffusion equation. The symbol ∇ denotes the gradient of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, and $\Delta = \nabla \times \nabla$ is the corresponding Laplacian. The factor in front of the Laplacian can be interpreted as the matrix of the diffusion coefficients $\mathbf{D} = \frac{1}{2} \sigma^2$, which are assumed to be independent of the particle positions [35]. Equation (2) introduces the Fokker–Planck operator \mathcal{Q} . \mathcal{Q} can also be interpreted as the infinitesimal generator of a transfer operator (or propagator) with lag time τ : $\mathcal{T}(\tau) = \exp(\mathcal{Q}\tau)$. The operator $\mathcal{T}(\tau)$ propagates $\rho(x, t)$ forward in time by a time interval τ : $\mathcal{T}(\tau)\rho(x, t) = \rho(x, t + \tau)$. The stationary solution of equation (2) is the Boltzmann density

$$\pi(x) = \frac{\exp\left(-\frac{1}{k_B T} V(x)\right)}{Z}, \quad (3)$$

where $Z = \int_{\Omega} \exp\left(-\frac{1}{k_B T} V(x)\right) dx$ is the classical partition function, i.e. $\partial_t \pi(x) = \mathcal{Q} \pi(x) = 0$.

2.1. Square root approximation

The SqRA of the infinitesimal generator is a method to discretize \mathcal{Q} , and to calculate the corresponding matrix elements [26, 27]. We will briefly review its derivation in the following section.

Consider a disjoint decomposition of the state space Ω into N Voronoi cells Ω_i , such that $\Omega = \cup_{i=1}^N \Omega_i$. The characteristic function associated to each Voronoi cell Ω_i is

$$\chi_i(x) = \begin{cases} 1 & \text{if } x \in \Omega_i \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

We introduce the following scalar product $\langle u|v \rangle_{\pi} = \int_{\Omega} u(x)v(x)\pi(x)dx = \int_{\Omega} u(x)v(x)\pi(x)dx$. For disjoint sets, the Galerkin discretization of \mathcal{Q} is computed via $\mathcal{Q}_{ij} = (\langle \chi_j | \chi_i \rangle_{\pi})^{-1} \langle \chi_j | \mathcal{Q} \chi_i \rangle_{\pi}$ which reduces to

$$\mathcal{Q}_{ij} = \frac{1}{\pi_i} \langle \chi_j | \mathcal{Q} \chi_i \rangle_{\pi}, \quad (5)$$

if we use equation (4) as ansatz functions. We remark that $\mathcal{Q} \chi_i(x)$ is formally not defined, because \mathcal{Q} is a differential operator, but $\chi_i(x)$ is not differentiable. In the mathematical treatment of Markov processes the possibility to replace the non-differentiable functions χ_i with differentiable approximations has been discussed [36]. Different possibilities for this replacement are imaginable, for example using the Fourier transform. The term $\pi_i = \langle \chi_i | \chi_i \rangle_{\pi} = \int_{\Omega_i} \pi(x) dx$ denotes the stationary probability of cell Ω_i .

Equation (5) defines an $N \times N$ transition rate matrix \mathbf{Q} with elements Q_{ij} , where Q_{ij} , for $i \neq j$, denotes the rate from cell Ω_i to cell Ω_j . The discretization is analogous to the discretization of the transfer operator in the derivation of MSMs [1, 5], which yields a transition matrix $\mathbf{T}(\tau)$ with matrix elements

$T_{ij}(\tau) = (\langle \chi_j | \chi_i \rangle_{\pi})^{-1} \langle \chi_j | \mathcal{T}(\tau) \chi_i \rangle_{\pi}$. Just as in MSMs, the state space Ω is usually so high-dimensional that solving the integral in equation (5) is not a viable option. However, in contrast to MSMs, the numerator in equation (5) cannot be estimated from correlation functions obtained by simulating the stochastic process in equation (1) [5, 37]. The square root approximation provides a solution to this impasse, which neither requires solving the high-dimensional integral nor sampling the stochastic process.

The derivation starts by noting that for time-homogeneous processes the rate matrix and the transition matrix are related by $\mathbf{Q} := \left. \frac{\partial \mathbf{T}(\tau)}{\partial \tau} \right|_{\tau=0}$. For a detailed discussion on the relation between \mathcal{Q} , \mathbf{Q} , $\mathcal{T}(\tau)$, and $\mathbf{T}(\tau)$, see references [26, 27, 28]. For infinitesimally small lag times τ , the transition rates between cells which do not share a common boundary is certainly zero. Thus, we can set the rate matrix elements for non-adjacent cells to

$$Q_{ij} = 0 \quad \text{if } i \neq j, \quad \text{and } \Omega_i \text{ is not adjacent to } \Omega_j. \quad (6)$$

Because the matrix elements $T_{ij}(\tau)$ represent transition probabilities, we can use the Gauss theorem to show that the rate matrix elements for adjacent cells satisfy [26, 27]

$$Q_{ij} = \frac{1}{\pi_i} \oint_{\partial \Omega_i \partial \Omega_j} \Phi(z) \pi(z) dS(z), \quad (7)$$

where \oint denotes a surface integral. Furthermore, $\partial \Omega_i \partial \Omega_j$ is the common surface between the cell Ω_i and Ω_j . $\Phi(z) = \delta_{\Omega_i = \Omega_j} \mathbf{v} \cdot \mathbf{n}$ denotes the flux of the configurations z through the surface $\partial \Omega_i \partial \Omega_j$. The vector \mathbf{v} is the velocity field associated to the time-dependent probability density. This is analogous to the fluid velocity in fluid dynamics, which describes the velocity of a small element of fluid such that the mass is conserved.

To approximate the surface integral in equation (7), we introduce the first of two assumptions of SqRA:

- (a) The flux does not depend on the position in state space: $\Phi(x) = \Phi$. Then

$$Q_{ij} = \frac{1}{\pi_i} \Phi \int_{\partial \Omega_i \partial \Omega_j} \pi(z) dS(z). \quad (8)$$

The remaining surface integral in equation (8) represents the stationary density on the intersecting surface $\partial \Omega_i \partial \Omega_j$.

To approximate it, we formulate our second assumption:

- (b) Each cell is small such that the potential energy $V(x)$ is almost constant within the cell: $V(x)|_{\Omega_i} \approx V_i$.

It follows that the stationary density $\pi(x)$, and, by extension, also the time-dependent density $\rho(x, t)$, is constant within a given cell Ω_i . The continuous and the discretized probabilities are related by

$$\begin{aligned} \pi_i &= \int_{\Omega_i} \pi(x) dx \approx \pi(x_i) \mathcal{V}_i \\ \rho_i(t) &= \int_{\Omega_i} \rho(x, t) dx \approx \rho(x_i, t) \mathcal{V}_i \end{aligned} \quad (9)$$

where $\int_{\Omega_i} 1 dx = \mathcal{V}_i$ is the volume of the cell Ω_i , and in particular we have

$$\pi(x_i) = \frac{1}{Z} \exp\left(-\frac{1}{k_B T} V_i\right), \quad (10)$$

where x_i is the center of Ω_i . Likewise, we can assume that the potential energy function on $\partial\Omega_i\partial\Omega_j$ is essentially constant, and that it can be approximated by some average of V_i and V_j . We choose the arithmetic mean $V(x)|_{\partial\Omega_i\partial\Omega_j} \approx \frac{V_i+V_j}{2}$, because for this type of mean-value calculation one can show that the resulting discretized operator \mathbf{Q} converges to the Fokker–Planck-operator \mathcal{Q} in the limit of infinitesimally small cells [27, 28]. The surface integral in equation (8) then becomes

$$\begin{aligned} \oint_{\partial\Omega_i\partial\Omega_j} \pi(z) dS(z) &= \oint_{\partial\Omega_i\partial\Omega_j} \frac{1}{Z} \exp\left(-\frac{1}{k_B T} \frac{V_i+V_j}{2}\right) dS(z) \\ &= \mathcal{S}_{ij} \sqrt{\pi(x_i)\pi(x_j)}, \end{aligned} \quad (11)$$

where $\oint_{\partial\Omega_i\partial\Omega_j} 1 dS(z) = \mathcal{S}_{ij}$ is the area of the intersecting surface. Note that an arithmetic mean of the potential energy function results in a geometric mean of the stationary densities: $\sqrt{\pi(x_i)\pi(x_j)}$.

With this approximation of the surface integral and with equation (9), we obtain the following expression for rates between adjacent cells [equation (7)]

$$Q_{ij,\text{adjacent}} = \frac{1}{\pi_i} \Phi \mathcal{S}_{ij} \sqrt{\pi(x_i)\pi(x_j)} = \Phi \frac{\mathcal{S}_{ij}}{\mathcal{V}_i} \sqrt{\frac{\pi(x_j)}{\pi(x_i)}}, \quad (12)$$

and the following rate matrix

$$Q_{ij} = \begin{cases} \Phi \frac{\mathcal{S}_{ij}}{\mathcal{V}_i} \sqrt{\frac{\pi(x_j)}{\pi(x_i)}} & \text{if } i \neq j, \text{ and } \Omega_i \text{ is adjacent to } \Omega_j \\ 0 & \text{if } i \neq j, \text{ and } \Omega_i \text{ is not adjacent to } \Omega_j \\ -\sum_{j=1, j \neq i}^n Q_{ij} & \text{if } i = j. \end{cases} \quad (13)$$

This is the SqRA of the Fokker–Planck operator \mathcal{Q} . Note that in our previous publication [27], we did not write the factor $\mathcal{S}_{ij}/\mathcal{V}_i$ explicitly, because we assumed that it is approximately the same for all pairs of adjacent cells and can be incorporated into $\hat{\Phi} = \Phi \frac{\mathcal{S}_{ij}}{\mathcal{V}_i}$.

The discretization of the Fokker–Planck equation [equation (2)] then is

$$\partial_t \boldsymbol{\rho}^\top(t) = \boldsymbol{\rho}^\top(t) \mathbf{Q}, \quad (14)$$

where $\boldsymbol{\rho}(t)$ is the vector-representation of the continuous probability density $\rho(x, t)$ with elements $\rho_i(t) = \int_{\Omega_i} \rho(x, t) dx = \int_{\Omega} \rho(x, t) \chi_i(x) dx$, and $\boldsymbol{\rho}^\top(t)$ denotes the transpose of $\boldsymbol{\rho}(t)$. Equation (14) can be rewritten as an evolution equation for the individual vector elements

$$\begin{aligned} \partial_t \rho_j(t) &= \sum_{i=1}^N \rho_i(t) Q_{ij} = \left[\sum_{i=1, i \neq j}^N \rho_i(t) Q_{ij} \right] \\ &+ \rho_j(t) \left[-\sum_{k=1, k \neq j}^n Q_{jk} \right], \end{aligned} \quad (15)$$

which is often written more concisely as a master equation

$$\partial_t \rho_j(t) = \sum_{i \sim j} [\rho_i(t) Q_{ij} - \rho_j(t) Q_{ji}], \quad (16)$$

where $\sum_{i \sim j}$ denotes the sum over all adjacent cells Ω_i of cell Ω_j .

The great appeal of the SqRA of the Fokker–Planck operator is that, apart from the grid-dependent flux $\Phi \frac{\mathcal{S}_{ij}}{\mathcal{V}_i}$, it only requires the Boltzmann-density $\pi(x_i)$ at the cell centers [equation (13)], which are readily available from the potential energy surface of the system. In principle, no time-series are required to calculate the rate matrix. The challenge lies in estimating $\Phi \frac{\mathcal{S}_{ij}}{\mathcal{V}_i}$. In the following, we introduce two different approaches to calculate $\Phi \frac{\mathcal{S}_{ij}}{\mathcal{V}_i}$ that do not rely on a realization of equation (1).

2.2. $\Phi \frac{\mathcal{S}_{ij}}{\mathcal{V}_i}$ by discretizing the Laplacian

If the flux Φ does not depend on the potential energy function (assumption a), one should be able to determine Φ by analyzing the overdamped Langevin dynamics on a constant potential $V(x) = \text{const.}$,

$$dx_t = \sigma dB_t, \quad (17)$$

and the associated Fokker–Planck equation

$$\partial_t \rho(x, t) = \frac{\sigma^2}{2} \Delta \rho(x, t) = \mathcal{Q} \rho(x, t). \quad (18)$$

This has two advantages. First, the differential operator in equation (18) essentially consists of the Laplacian, whose discretization is known. Second, the stationary density [equation (3)] of this process is constant, which simplifies the expression for the rates [equation (12)].

Applying the Gauss theorem, the Laplacian of the probability density $\rho(x, t)$ over a small region with volume \mathcal{V} and surface \mathcal{S} , is written as [38]

$$\Delta \rho(x, t) = \lim_{\mathcal{V} \rightarrow 0} \frac{1}{\mathcal{V}} \oint_{\mathcal{S}} \nabla \rho(z, t) \cdot \mathbf{n} dS(z), \quad (19)$$

where \mathbf{n} is the unit vector orthogonal to the surface \mathcal{S} . It follows, that on a Voronoi tessellation of the space, the discrete Laplacian on a small Voronoi cell Ω_i is [39, 40]

$$\Delta \rho(x, t)|_{x=x_j} = \frac{1}{\mathcal{V}_j} \sum_{i \sim j} \nabla \rho(x, t)|_{x=x_j} \cdot \mathbf{n}_{ji} \mathcal{S}_{ji}. \quad (20)$$

The term $\nabla \rho(x, t)|_{x=x_j} \cdot \mathbf{n}_{ji}$ is the gradient in the direction $j \rightarrow i$ (directional derivative), which can be approximated by the finite difference

$$\nabla \rho(x, t)|_{x=x_j} \cdot \mathbf{n}_{ji} \approx \frac{\rho(x_i, t) - \rho(x_j, t)}{h_{ji}}, \quad (21)$$

where $h_{ji} = x_j - x_i$ is the distance between the centers of the cells Ω_j and Ω_i . Inserting this finite difference into equation (20) yields

$$\Delta\rho(x, t)|_{x=x_j} = \frac{1}{\mathcal{V}_j} \sum_{i \sim j} \frac{\rho(x_i, t) - \rho(x_j, t)}{h_{ji}} S_{ji}. \quad (22)$$

Assuming that the density $\rho(x, t)$ is approximately constant within cell Ω_i (assumption b), we have $\rho_i(t) \approx \int_{\Omega_i} \rho(x_i, t) dx = \rho(x_i, t) \mathcal{V}_i$. Substituting $\rho(x_i, t) = \frac{\rho_i(t)}{\mathcal{V}_i}$ in equation (22) and inserting into equation (18) yields

$$\partial_t \frac{\rho_j(t)}{\mathcal{V}_j} = \frac{\sigma^2}{2} \frac{1}{\mathcal{V}_j} \sum_{i \sim j} \frac{\frac{\rho_i(t)}{\mathcal{V}_i} - \frac{\rho_j(t)}{\mathcal{V}_j}}{h_{ij}} S_{ij}, \quad (23)$$

and we obtain the discrete Fokker–Planck equation [equation (18)] at constant potential

$$\begin{aligned} \partial_t \rho_j(t) &= \frac{\sigma^2}{2} \sum_{i \sim j} \frac{\frac{\rho_i(t)}{\mathcal{V}_i} - \frac{\rho_j(t)}{\mathcal{V}_j}}{h_{ij}} S_{ij} \\ &= \sum_{i \sim j} \frac{\sigma^2}{2} \frac{1}{h_{ij}} \frac{S_{ij}}{\mathcal{V}_i} \rho_i(t) - \frac{\sigma^2}{2} \frac{1}{h_{ij}} \frac{S_{ij}}{\mathcal{V}_j} \rho_j(t). \end{aligned} \quad (24)$$

Comparing equation (24) to the master equation [equation (16)] and to the definition of rates between adjacent cell within the SqRA [equation (12)] we obtain the following equality

$$Q_{i,j,\text{adjacent}} = \Phi \frac{S_{ij}}{\mathcal{V}_i} = \frac{\sigma^2}{2} \frac{1}{h_{ij}} \frac{S_{ij}}{\mathcal{V}_i}, \quad (25)$$

where we used that $\sqrt{\frac{\pi(x_j)}{\pi(x_i)}} = 1$ at constant potential. Thus,

$$\Phi = \frac{\sigma^2}{2h_{ij}}. \quad (26)$$

We have obtained an analytical expression for Φ between adjacent cells that only depends on the distance h_{ij} between the cell centers. Appendix B contains an alternative derivation of equation (26) using Fick’s first law of diffusion.

The rate matrix [equation (13)] can now be written more concretely as

$$Q_{ij} = \begin{cases} \frac{\sigma^2}{2} \frac{1}{h_{ij}} \frac{S_{ij}}{\mathcal{V}_i} \sqrt{\frac{\pi(x_j)}{\pi(x_i)}} & \text{if } i \neq j, \text{ and } \Omega_i \text{ is adjacent to } \Omega_j \\ 0 & \text{if } i \neq j, \text{ and } \Omega_i \text{ is not adjacent to } \Omega_j \\ -\sum_{j=1, j \neq i}^n Q_{ij} & \text{if } i = j. \end{cases} \quad (27)$$

Section 3 introduces formulas to evaluate $\frac{1}{h_{ij}} \frac{S_{ij}}{\mathcal{V}_i}$ for various grid types.

2.3. $\Phi \frac{S_{ij}}{\mathcal{V}_i}$ by analyzing the transition probability density

Our starting point is again equation (13), and we introduce a third assumption:

- (c) The volumes of all cells are approximately equal ($\mathcal{V}_i \approx \mathcal{V}, \forall \Omega_i$), and the intersecting surfaces areas are approximately equal for all adjacent cells ($S_{ij} \approx S, \forall \Omega_i$ adjacent to Ω_j).

In this case, the factor $\Phi \frac{S_{ij}}{\mathcal{V}_i} \approx \Phi \frac{S}{\mathcal{V}} = \Phi_{\text{grid}}$ has approximately the same value for each pair of adjacent cells. Φ_{grid} is thus a flux value which is characteristic for a given grid rather than for a specific pairs of cells. This is the assumption we used in reference [27].

Every grid can be represented as an unweighted graph, in which nodes correspond to the grid cells Ω_i , and two nodes are connected by an edge if the corresponding grid cells are adjacent. At constant potential, the rate matrix [equation (13)] can then be written as the Laplacian matrix of the graph \mathbf{L} multiplied by the grid flux

$$\mathbf{Q} = -\Phi \frac{S}{\mathcal{V}} \mathbf{L} = -\Phi_{\text{grid}} \mathbf{L}, \quad (28)$$

where the Laplacian matrix of the graph is defined as

$$L_{ij} = \begin{cases} -1 & \text{if } i \neq j, \text{ and } \Omega_i \text{ is adjacent to } \Omega_j \\ 0 & \text{if } i \neq j, \text{ and } \Omega_i \text{ is not adjacent to } \Omega_j \\ -\sum_{j=1, j \neq i}^n L_{ij} & \text{if } i = j. \end{cases} \quad (29)$$

Note that $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where \mathbf{A} is the adjacency matrix of the graph, with elements $A_{ij} = 1$ if Ω_i and Ω_j are neighbors, and $A_{ij} = 0$ otherwise. \mathbf{D} is the degree matrix of the graph, a diagonal matrix whose diagonal entries contain the number of neighbors for each cell, i.e. $D_{ii} = \sum_j A_{ij}$.

The transition matrix $\mathbf{T}(\tau)$ and the rate matrix \mathbf{Q} are related by

$$\mathbf{T}(\tau) = \exp(\tau \mathbf{Q}) = \exp(-\tau \Phi_{\text{grid}} \mathbf{L}). \quad (30)$$

If one knows the transition probability of a $T_{ij}(\tau)$ of single pair of adjacent cells at constant potential, one can calculate Φ_{grid} by comparing $T_{ij}(\tau)$ to the matrix element $[\exp(-\tau \Phi_{\text{grid}} \mathbf{L})]_{ij}$. The transition probability is defined as the integral transition probability density $p(x, y, \tau)$ over the initial and final cell

$$T_{ij}(\tau) = \frac{1}{\pi_i} \int_{\Omega_i} \int_{\Omega_j} p(x, y, \tau) \pi(x) dx dy, \quad (31)$$

where $\pi(x)$ is the unconditional probability density of finding the system at point x at time t , and $p(x, y, \tau)$ is the conditional probability density of finding the system in dy at time $t + \tau$ given that it started in point x at time t . In reference [27] we obtained the transition probability by constructing an MSM based on a simulation of the dynamic process.

Here, we propose a different approach. We again use the idea that the flux, and by extension Φ_{grid} , does not depend on the potential energy function. Therefore, it can be determined from an overdamped Langevin dynamics at constant potential energy [equation (17)], for which the transition probability

density is:

$$p(x, y, \tau) = \left(\sqrt{\frac{1}{2\pi\sigma^2\tau}} \right)^{N_D} \exp\left(-\frac{(y-x)^2}{2\sigma^2\tau}\right), \quad (32)$$

where N_D is the dimension of the state space. If the cells are small (assumption b), the distance from any point in cell Ω_i to any other point in Ω_j is approximately equal to the distance of the centers of the two cells, i.e. $y-x \approx x_j-x_i$ for all $x \in \Omega_i, y \in \Omega_j$. With this assumption equation (31) becomes

$$\begin{aligned} T_{ij}(\tau) &= \frac{1}{\pi_i} \left(\sqrt{\frac{1}{2\pi\sigma^2\tau}} \right)^{N_D} \\ &\times \exp\left(-\frac{h_{ij}^2}{2\sigma^2\tau}\right) \int_{\Omega_i} \pi(x) dx \int_{\Omega_j} 1 dy \\ &= \left(\sqrt{\frac{1}{2\pi\sigma^2\tau}} \right)^{N_D} \exp\left(-\frac{h_{ij}^2}{2\sigma^2\tau}\right) \mathcal{V}, \end{aligned} \quad (33)$$

where $h_{ij} = x_j - x_i$ is the distance between the centers of Ω_i and Ω_j . Besides N_D, σ , and h_{ij} , one only needs the average cell volume \mathcal{V} to calculate $T_{ij}(\tau)$. The lag time τ can in principle be chosen freely.

Now that we have a closed-form approximation for $T_{ij}(\tau)$ at constant potential, we can use equation (30) to calculate Φ_{grid} . Because the Laplacian matrix \mathbf{L} is not invertible, we cannot determine Φ_{grid} by rearranging equation (30). Instead we use Φ as a parameter that minimizes the difference between $T(\tau)$ and $\exp(-\tau\Phi\mathbf{L})$, i.e. we minimize the function:

$$f(\Phi) = \left(T_{ij}(\tau) - [\exp(-\tau\Phi\mathbf{L})]_{ij} \right)^2. \quad (34)$$

and $\Phi_{\text{grid}} = \arg \max_{\Phi} f(\Phi)$. This approach requires calculating the matrix exponential of the potentially large but sparse matrix \mathbf{L} and is tested in the result section.

It is tempting to avoid the minimization of $f(\Phi)$ [equation (34)] by approximating the matrix exponential as a truncated Taylor series, and solving for Φ . Mathematically this is possible. But the resulting equation for Φ_{grid} is a poor approximation to the true value of Φ_{grid} , and we do not recommend using this approach. Appendix C discusses the details.

3. Methods

In section 2.2 we showed that the grid-dependent flux factor can be expressed in terms of the known parameters σ and h_{ij} : $\Phi \frac{S_{ij}}{\mathcal{V}_i} = \frac{\sigma^2}{2h_{ij}} \frac{S_{ij}}{\mathcal{V}_i}$. In this section, we summarize methods to evaluate S_{ij}/\mathcal{V}_i for different grid types.

3.1. Arbitrary grid/exact method

‘The quickhull algorithm’ [33], implemented in the MATLAB function ‘convhulln()’, computes the convex hull of a set of multidimensional points and can be used to numerically calculate the surface S_{ij} and the volume \mathcal{V}_i of an arbitrarily shaped

cell. We will call this the ‘exact method’, because it directly calculates $\frac{S_{ij}}{\mathcal{V}_i}$ without any assumptions on the grid geometry. However, the algorithm requires not only the centers of the cells, but also the vertices of each cell, which makes it computationally expensive for high-dimensional spaces.

3.2. Hyper-rectangular grid

On a (hyper-)rectangular grid, the ratio between interface surface and cell volume is simply given by the cell length in direction $i \rightarrow j$, i.e. $\frac{S_{ij}}{\mathcal{V}_i} = \frac{1}{h_{ij}}$, and

$$Q_{ij,\text{adjacent,rectangular}} = \frac{\sigma^2}{2h_{ij}^2} \sqrt{\frac{\pi(x_j)}{\pi(x_i)}}. \quad (35)$$

Note that on a (hyper-)cubic grid $h_{ij} = h$ is the same in all grid dimensions, while for a one-dimensional grid one obtains the equation derived in reference [29] from the one-dimensional reaction–diffusion equation. Appendix B shows that equation (35) converges to the Fokker–Planck equation in the limit of infinitesimally small cells [27, 28].

3.3. Hexagonal grid

The apothem a of a cell is the distance from the cell center to one of the midpoint of its sides. On a two-dimensional hexagonal grid, $a = h/2$, where $h = h_{ij}$ is the distance between cell centers. Using the apothem we can calculate the intersecting surface, which is equal to the length of each side of the hexagon $S_{ij} = \frac{2}{\sqrt{3}}a = \frac{h}{\sqrt{3}}$. Thus, $\frac{S_{ij}}{\mathcal{V}_i} = \frac{2}{3h_{ij}}$, and the rate between adjacent hexagonal cells is

$$Q_{ij,\text{adjacent,hexagonal}} = \frac{\sigma^2}{3h_{ij}^2} \sqrt{\frac{\pi(x_j)}{\pi(x_i)}}. \quad (36)$$

3.4. Voronoi grid via the neighbors-method

On arbitrary Voronoi grids, several methods [41] have been proposed to approximate $\frac{S_{ij}}{\mathcal{V}_i}$. For example, from the Taylor expansion of a function on an irregular mesh, the rate between adjacent cells can be expressed as [34]

$$Q_{ij,\text{adjacent,Voronoi}} = \frac{\sigma^2}{2} \frac{4}{n_i \bar{h}_i h_{ij}} \sqrt{\frac{\pi(x_j)}{\pi(x_i)}}, \quad (37)$$

where n_i is the number of neighbors of the cell Ω_i , and \bar{h}_i is the average distance between the cell Ω_i and all the neighbors.

3.5. Method overview

Table 1 summarizes the methods that are now at our disposal to evaluate $\Phi \frac{S_{ij}}{\mathcal{V}_i}$. In the following analysis, we will compare the eigenvalues κ_i and left eigenvectors \mathbf{l}_i

$$\mathbf{l}_i^\top \mathbf{Q} = \kappa_i \mathbf{l}_i^\top \quad (38)$$

of rate matrices \mathbf{Q} constructed using equation (28) in combination with the methods in Table 1. Note that equation (28) implies that the row-sums of each of these matrices \mathbf{Q} are zero, consistent with equation (27).

Table 1. Methods to calculate $\Phi \frac{S_{ij}}{V_i}$ in the SqRA.

Method	Grid	$\Phi \frac{S_{ij}}{V_i}$	Equation
Exact ratio of intersecting surface area and cell volume			
Exact	Arbitrary	$\frac{\sigma^2}{2} \frac{1}{h_{ij}} \frac{S_{ij}}{V_i}$	Equation (27)
Rectangular	Hyper-cube	$\frac{\sigma^2}{2h_{ij}^2}$	Equation (35)
Hexagonal	2D-hexagonal	$\frac{\sigma^2}{3h_{ij}^2}$	Equation (36)
Approximate ratio of intersecting surface area and cell volume			
Neighbors	Voronoi	$\frac{\sigma^2}{2} \frac{4}{n_i h_i h_{ij}}$	Equation (37)
Comparison to transition probability			
Minimization	Arbitrary	$\operatorname{argmin}_{\Phi} \left(T_{ij}(\tau) - [\exp(-\tau \Phi \mathbf{L})]_{ij} \right)^2$	Equation (34)

4. Results and discussion

4.1. Computational efficiency

The usefulness of the SqRA critically depends on how many dimensions N_D the dynamical system in equation (1) may have, before the calculation of \mathbf{Q} via equation (13) becomes computationally intractable. \mathbf{Q} is an $N \times N$ square matrix, where N is the number of cells Ω_i . If each dimension of the dynamical system is discretized into N_{bins} , the number of cells is given as $N = N_{\text{bins}}^{N_D}$, i.e. N grows exponentially with the number of dimensions N_D . Thus, even for low-dimensional systems we have to construct a sparse but extremely high-dimensional rate matrix \mathbf{Q} .

To compare the computational efficiency of the methods to estimate $\Phi \frac{S_{ij}}{V_i}$, we devised a model system, which consists of particles of mass $m = 1$ kg moving in a N_D -dimensional Cartesian space according to equation (1) with $\xi = 1$ s⁻¹ and $\sigma = 2.2$ J^{1/2} kg^{-1/2} s^{-1/2}. The potential energy function consists of uncoupled terms for each Cartesian coordinate x_i

$$V(x_1, \dots, x_{N_D}) = \sum_{i=1}^{N_D} V_i(x_i), \quad (39)$$

defined on the domain $\Omega = \{(x_1, \dots, x_{N_D}) : -\pi < x_i \leq \pi \text{ for } i = 1, \dots, N_D\}$. We applied periodic boundary conditions in each direction, and the one dimensional potential energy term

$$V_i(x_i) = \frac{1}{2} k_i (1 + \cos(m_i \cdot x_i - x_{0i})), \quad (40)$$

is 2π -periodic in direction i . The parameter k_i is the force constant, m_i is the multiplicity and describes the number of barriers and wells of the function, and x_{0i} is the phase. For each direction i , we used the same triplet of parameters $k_i = 2$ kg s⁻¹, $m_i = 2$ and $x_{0i} = 0$ rad. Figure 1(A) shows the potential for the one-dimensional system, which is a periodic double well potential. For an N_D -dimensional system, the potential has 2^{N_D} wells in the N_D -dimensional space. Equation (40) mimics the function that governs torsion angles

in MD force fields. By choosing a Cartesian space with periodic boundary conditions rather than an actual torsion angle, we avoid any complications that arise from the coordinate transformation to the torsion angle space, and a volume element in Ω is simply given as $dV = dx_1 dx_2 \dots dx_{N_D}$.

For the system with $N_D = 1$, we constructed a reference solution with $N_{\text{bins}} = 60$ bins using the method ‘rectangular’. The leading eigenvalues of \mathbf{Q} are $\kappa_0 = 0$, $\kappa_1 = -1.56$; $\kappa_2 = -2.87$; $\kappa_3 = -6.51$. Figure 1(B) shows the eigenvector l_0 , which corresponds to the stationary distribution. Figure 1(C) shows the eigenvector l_1 , which represents a transition between the regions $[0, \pi)$ and $[\pi, 0)$.

We next scanned the number bins between 2 and 60 to find the coarsest possible discretization that still yields accurate results for the dominant processes of the one-dimensional system. For $N_{\text{bins}} = 5$, $\kappa_1 = -1.58$, which is 1.2% lower than the reference value; while $\kappa_2 = -2.69$ and $\kappa_3 = -5.68$ are respectively 6.2% and 11.4% higher than the reference values. A smaller number of bins yield considerable deviations from the reference value. Figures 1(B) and (C) show the approximation of the two leading eigenfunctions with $N_{\text{bins}} = 5$. In spite of the very low resolution of the eigenvector, we can identify the two peaks corresponding to the two wells of the potential.

We constructed grids for up to $N_D = 9$ dimensions. For the hypercubic grids, we discretized each dimension into $N_{\text{bins}} = 5$ equally-sized bins, where the distance between two adjacent states is $h = 2\pi/5 \approx 1.26$. For the Voronoi grids, we discretized each dimension into $N_{\text{bins}} = 5$ bins of random size. The number of states N are: $N_D = 1:5$, $N_D = 2:25$, $N_D = 3:125$, $N_D = 4:625$, $N_D = 5:3125$, $N_D = 6:15625$, $N_D = 7:78125$, $N_D = 8:390625$, and $N_D = 9:1935125$ states. Likewise the memory size of the corresponding rate matrices grows exponentially with N_D . For the case with $N_D = 9$, the full matrix \mathbf{Q} occupies more than 28 TB of memory, but its corresponding sparse matrix is just 578 MB, which is manageable by modern computers.

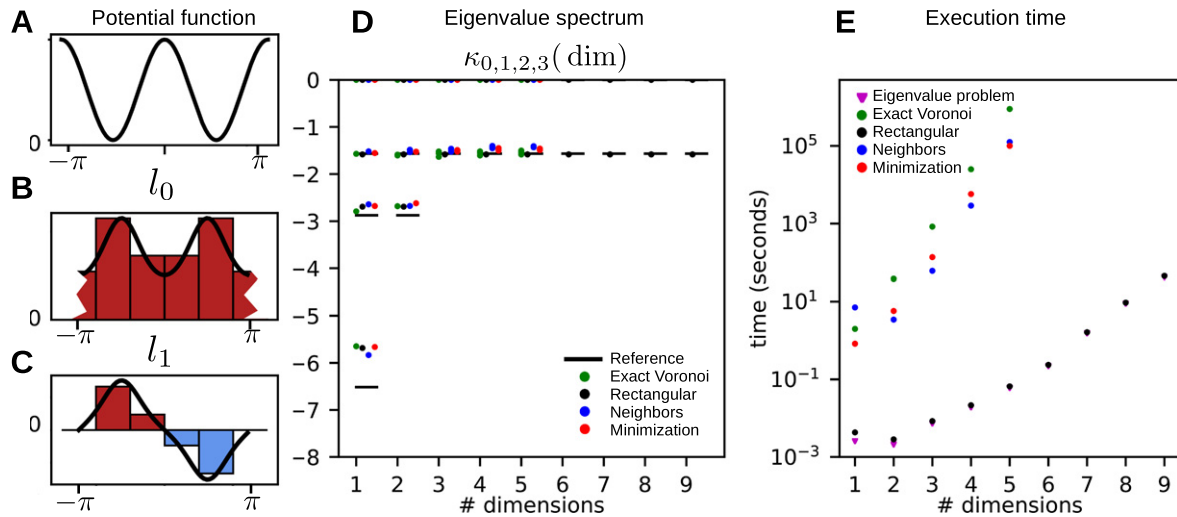


Figure 1. Computational efficiency. (A) One-dimensional periodic potential function, and corresponding first (B) and second (C) eigenfunction of the Fokker–Planck operator \mathcal{Q} (black lines) compared to the eigenvectors of the corresponding rate matrix \mathbf{Q} (histograms). The fringes at $-\pi$ and π indicate that this crosses the periodic boundary. (D) First four eigenvalues for each N_D -dimensional system; (E) execution time for each N_D -dimensional system.

We included the following methods in this scan: ‘rectangular’ on a hypercubic grid, and ‘exact’, ‘neighbors’, and ‘minimization’ on a Voronoi grid. The method ‘hexagonal’ is excluded, because a hexagonal grid can only be constructed on a two-dimensional Cartesian space. The method ‘exact’ for the hypercubic grid is not shown explicitly, because it is identical to the method ‘rectangular’. For each rate matrix, we calculated the four leading eigenvalues. For this system the second eigenvalue has a degeneracy equal to the number of dimensions. Because they perfectly overlap, there appear to be less eigenvalues for the higher-dimensional systems in figure 1(D). All four methods yield eigenvalues that are in excellent agreement with the reference solution. Thus, at least at this level of discretization, approximating the ratio $\frac{S_{ij}}{V_i}$ by these methods does not introduce an error of relevant magnitude.

However, the computational cost varies drastically between the methods [figure 1(E)]. Three separate tasks go into calculating the eigenvectors and eigenvalues of \mathbf{Q} : (i) constructing the adjacency matrix of the grid from which the Laplacian matrix of the grid \mathbf{L} can then be calculated, (ii) calculating $\Phi \frac{S_{ij}}{V_i}$ using one of the four methods, and (iii) calculating the dominant eigenvalue–eigenvector pairs for the resulting matrix \mathbf{Q} . The most efficient method is ‘rectangular’ on a hypercubic grid, for which we could calculate rate matrices for up to nine dimensions on a server with an Intel Xeon CPU (E5-2690 v3 @ 2.60 GHz) and 160 GB of RAM. Using MATLAB, the execution time was 45 s. We provide an example script on hypercubic grids, the distance h_{ij} between neighboring cells is a constant, and the factor $\Phi \frac{S_{ij}}{V_i} = \frac{1}{2} \frac{\sigma^2}{h^2}$ can be calculated at negligible cost. Moreover, one can build adjacency matrices and construct the matrix \mathbf{Q} very efficiently using sparse matrices and the Kronecker product [see supplementary material (<https://stacks.iop.org/JPCM/33/115902/mmedia>)]. Consequently, approximately the 90% of the computational time is used up by the third task: solving the eigenvalue problem

[figure 1(E), magenta triangles]. The time to solve the eigenvalue problem primarily depends on the dimension of the matrix \mathbf{Q} , i.e. the number of cells N . It does not depend on the method of computing the flux, and it only weakly depends on the type of grid. Thus, the computational cost, that is displayed as magenta triangles in figure 1(E) is part of every calculation in figure 1(E). Note that the execution time depends on the algorithm used to solve the eigenvalue problem. The MATLAB function ‘eigs()’ permits to provide the number of eigenvalues–eigenvectors to be calculated. This is particularly useful when one is interested only in the slowest dominant processes, which are associated to the largest eigenvalues.

All three methods to construct the rate matrix on a Voronoi grid are orders of magnitude slower than the ‘rectangular’ method for hypercubic grids, because building adjacency matrices for a Voronoi discretization is computationally difficult and costly. We constructed the adjacency matrices \mathbf{A} using an algorithm based on linear programming as suggested in reference [26]. Note that for Voronoi grids, the computational cost of diagonalizing the rate matrix (magenta triangles) makes up only a small fraction of the total calculation.

Among the methods for a Voronoi grid, the ‘exact’ method [green dots in figure 1(D)] is about an order of magnitude more expensive than the methods ‘neighbors’ or ‘minimization’ [blue and red dots in figure 1(D)], because it requires the exact calculation of cell volumes. Using the same computer as for the ‘rectangular’ method, we were able to build the rate matrix of the five dimensional system. The calculation took 8.9×10^5 s, corresponding to more than ten days of calculations. However, the ‘exact’ method is slightly more accurate than the other three methods for Voronoi grids. The methods ‘neighbors’ and the ‘minimization’ slightly overestimate the eigenvalues, but the execution time reduced to 1.2×10^5 s (33.3 h) and 1.1×10^5 s (30.5 h), respectively.

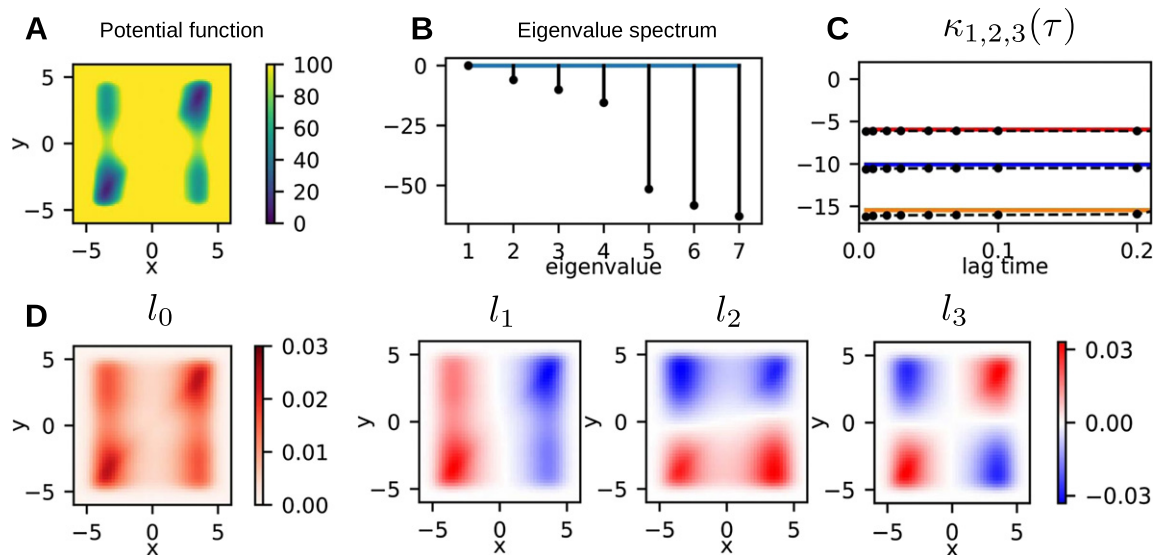


Figure 2. Two dimensional diffusion process in a four well potential. (A) Potential energy function. (B) Eigenvalue spectrum of the corresponding Fokker–Planck operator \mathcal{Q} . (C) First three eigenvalues of the infinitesimal generator as function of the lag time τ : SqRA (solid line), MSM (black dots). (D) First four eigenfunctions of \mathcal{Q} .

4.2. Accuracy

Next, we test whether the methods in table 1 differ in their accuracy. We consider a particle of mass 1 kg which moves on a two-dimensional Cartesian space according to equation (1) with $\xi = 1 \text{ s}^{-1}$ and $\sigma = 15 \text{ J}^{\frac{1}{2}} \text{ kg}^{-\frac{1}{2}} \text{ s}^{-\frac{1}{2}}$. The potential energy function is

$$V(x, y) = k_1(y^2 - a_1^2)^4 + k_2(x^2 - a_2^2)^2 + \frac{k_{12}}{\sqrt{(x - y)^2 + c^2}}, \quad (41)$$

with the parameters $k_1 = 0.003$, $a_1 = 3.3$, $k_2 = 1$, $a_2 = 3.5$, $k_{12} = -50$ and $c = 1$. This potential is composed of a one-dimensional term which describes a slow double-well dynamics along the x axis, a one-dimensional term which describes a fast double-well dynamics along the y axis, and a coupling term [figure 2(A)]. The eigenvalue spectrum of the Fokker–Planck operator for this system [equation (2)] exhibits four dominant eigenvalues at $\kappa_0^{2D} = 0$, $\kappa_1^{2D} = -5.98$, $\kappa_2^{2D} = -10.2$, and $\kappa_3^{2D} = -15.06$ [figure 2(B)]. The corresponding eigenfunctions $l_0(x, y)$ to $l_3(x, y)$ are shown in [figure 2(D)]. The eigenvector $l_0(x, y)$ is equal to the stationary density. Eigenvectors $l_1(x, y)$ and $l_2(x, y)$ describe slow transitions along the x and y axis, respectively. Eigenvector $l_4(x, y)$ represents a dynamic process which mixes x and y , and is due to the coupling term in $V(x, y)$. We constructed this reference solution by evaluating the SqRA [equation (13)] on a quadratic grid [equation (35)] with $N = 50 \times 50 = 2500$ cells on the space $[-6, 6] \times [-6, 6]$.

Additionally, we constructed an MSM on the same grid. We generated a time-discretized trajectory of 1×10^8 time-steps, with a time-step $\Delta t = 0.001$, integrating equation (1) according to the Euler–Maruyama scheme [42]. The MSM has been constructed by counting transitions $C_{ij}(\tau)$ from cell Ω_i to cell Ω_j within a lag time τ varied in a range of [5:500] time-steps [5]. Detailed balance has been enforced by symmetrizing

the resulting 50×50 -count matrix: $\mathbf{C}_{\text{sym}}(\tau) = \mathbf{C}(\tau) + \mathbf{C}^\top(\tau)$, where $\mathbf{C}^\top(\tau)$ denotes the transpose of $\mathbf{C}(\tau)$. The MSM transition matrix $\mathbf{T}(\tau)$ was obtained by row-normalizing $\mathbf{C}_{\text{sym}}(\tau)$.

The eigenvectors of the MSM transition matrix are defined as $\mathbf{l}_i^\top \mathbf{T}(\tau) = \lambda_i(\tau) \mathbf{l}_i^\top$. The MSM yielded the same dominant eigenvectors as the SqRA of the rate matrix. The MSM eigenvalues $\lambda_i(\tau)$ and the eigenvalues of the rate matrix can be interconverted by

$$\lambda_i(\tau) = \exp(\kappa_i \tau) \Leftrightarrow \kappa_i = \frac{\ln(\lambda_i(\tau))}{\tau}, \quad (42)$$

and are in excellent agreement [figure 2(C)]. The fact that the ratio $\frac{\ln(\lambda_i(\tau))}{\tau}$ does not vary with τ indicates that the MSM on this grid has a negligible projection error. Since the SqRA-model and the MSM do not deviate from each other, we can assume that also the SqRA-model has a negligible projection error. We will therefore use the SqRA-model on a regular grid with $N = 2500$ cells as a reference solution for further tests.

To assess whether the method to estimate the flux has influence in the accuracy of the SqRA of the rate matrix, we varied the number of grid cells from $N = 4$ to $N = 225$. We constructed quadratic grids, hexagonal grids and arbitrary Voronoi grids. For the arbitrary Voronoi grids, we randomly placed grid centers in the two-dimensional state space. To account for the variance in these randomly constructed grids, we constructed fifty different grids for each value of N and constructed the corresponding rate matrix. In figure 3, we report the mean and the variance of the dominant eigenvalues for Voronoi grids, that were calculated using the methods ‘exact’ (green), ‘neighbors’ (blue), and ‘minimization’ (red). Figure 3 also shows the dominant eigenvalues for the quadratic grid calculated using the method ‘rectangular’ (black) and the hexagonal grid calculated using the method ‘hexagonal’ (orange), as well as the reference value for the eigenvalues (dashed).

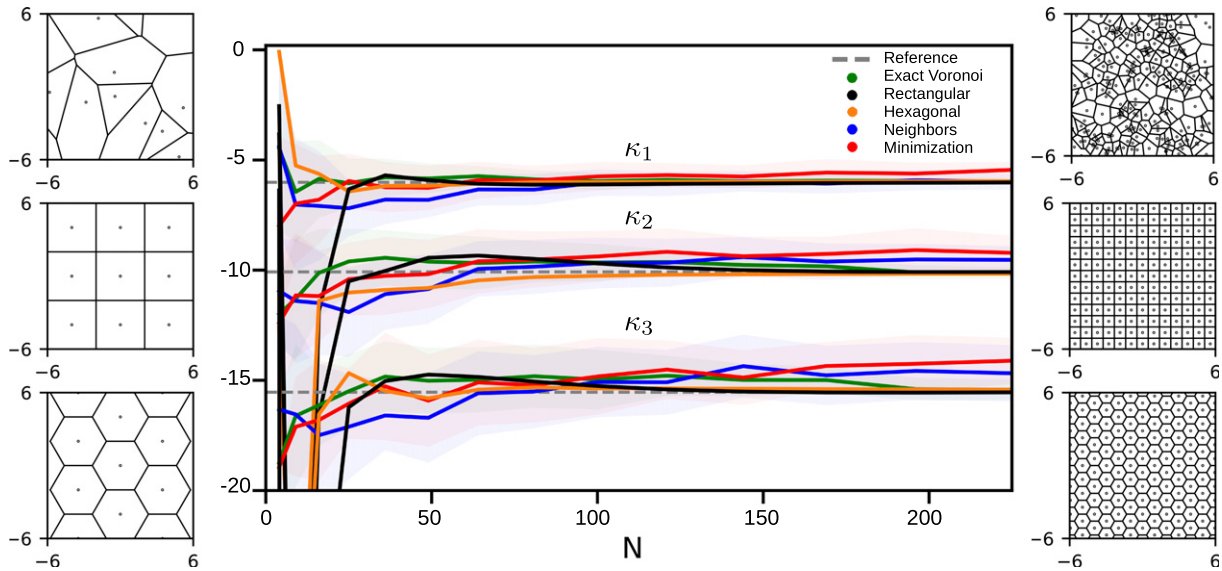


Figure 3. Accuracy of the flux estimated by discretizing the Laplacian: reference (gray); exact Voronoi (green); rectangular grid (black); hexagonal grid (orange); neighbors (blue); minimization (red).

The results for the Voronoi grids seem to converge faster than the results for the regular grid. The mean of the eigenvalues for the Voronoi grids is already reasonably accurate for $N = 9$ or $N = 16$ grid cells. Note however that the variance is sizable at these low numbers of grid cells and that, depending on the exact location of the grid cells, the Voronoi results can also deviate considerably from the reference value. For $N = 25$ all five methods yield results that are close to the reference value, and the accuracy of all five methods increases only slowly with increasing N . In fact, between $N = 100$ and $N = 225$ we do not find a significant improvement for any of the five methods. This means that, at least for this potential energy function, 25 to 100 grid cells are sufficient to discretize the two-dimensional state space. This is an order of magnitude lower than previously reported discretizations of two-dimensional molecular states spaces that relied on the SqRA [21, 43].

For $N > 100$ the eigenvalues obtained from regular grids are almost exactly equal to the reference values, whereas the results from Voronoi grids tend to overestimate the eigenvalues. This indicates that, if one is interested in a highly accurate estimate of the dominant eigenvalues, one should opt for a regular grid.

5. Conclusion

We have derived, from the equation of the overdamped Langevin dynamics with constant potential, the expression of the flux Φ that appears in the SqRA formula [26, 27]. An analogous formula, was previously derived for the one-dimensional Smoluchowski equation discretized on a regular grid [29] and later used to estimate the diffusion coefficients of molecular systems projected on one-dimensional relevant coordinates [44, 45]. Our result generalizes to the case

of N_D -dimensional diffusive systems discretized on multidimensional arbitrary grids. Moreover, we proposed and tested several methods which can be used to calculate the exact or the approximate value of the multiplicative factor $\Phi \frac{S_{ij}}{V_i}$ for different grid types. We now have an approach in place that in principle allows us to calculate MSMs of molecular systems without running MD simulations.

The accuracy with which the dominant eigenvalues of the rate matrix can be estimated is similar for all methods. But our analysis has shown that, depending on the grid and the method to estimate the flux, the relative and absolute computational costs of these three methods vary drastically. The entire computation of the SqRA of the Fokker–Planck equation, from discretization of the state space to the analysis of the dominant eigenvectors, consists of three steps: (i) generate the adjacency matrix, (ii) calculate the rates Q_{ij} , (iii) calculate the eigenvalue and eigenvectors of the rate matrix. On a regular grids, the generation of the adjacency matrix using the algorithm in the supplementary material is computationally cheap. The factor $\Phi \frac{S_{ij}}{V_i}$ is essentially a constant, and the computational cost is dominated by the calculation of the eigenvectors. Thus, calculating the SqRA on a regular grid is by far the most efficient approach, if one aims at discretizing the entire state space.

However, molecules at room temperature only access a small fraction of their state space, and the experience with MSMs has shown that Voronoi grids are useful for discretizing the accessible state space [5]. We therefore do not yet want to rule out Voronoi grids. We have compared three methods to calculate $\Phi \frac{S_{ij}}{V_i}$: an ‘exact’ method that aims at calculating $\frac{S_{ij}}{V_i}$ numerically, and two approximate methods. The ‘neighbor’ method is based on an already known interpolation scheme between all neighbors of a given cell Ω_i . The ‘minimization’ method is an approach that we proposed in this contribution,

and is based on a comparison to the analytically known transition probability at constant potential. On Voronoi grids, the construction of the adjacency matrix is computationally much more demanding than on regular grids, and for the approximate methods, the computational cost is dominated by the construction of the adjacency matrix. However, with the ‘exact’ method the calculation of $\frac{S_{ij}}{V_i}$ is the most costly step, increasing the computational cost of the entire calculation by an order of magnitude. Taking into account that this method only slightly improves the accuracy of the MSM, the ‘exact’ method is not suited for an actual application.

With a few seconds computing time on a single compute server, we could reach $\mathcal{O}(10^6)$ grid cells for regular grids, while we need a few days to reach $\mathcal{O}(10^3)$ for Voronoi grids. Using the ‘neighbors’ or the ‘minimization’ method, $\mathcal{O}(10^4)$ are within reach for Voronoi grids, and moving the calculation to high-performance compute clusters or GPUs will likely push the limit to $\mathcal{O}(10^5)$ states. With grids of this size, the SqRA becomes useful for small molecules. However, a brute-force discretization of the entire state space of larger molecules would require even larger grids. There are two possible remedies: (i) one discretizes only the accessible state space, or (ii) one projects the dynamics on a lower-dimensional space and discretizes this space. In the first approach, the accessible state space will best be formulated in terms of internal coordinates. Then the distances between grid cells and potentially also the cell volumes have to be transformed accordingly. In the second approach, one needs to calculate the free-energy surface and the position-dependent flux on the low-dimensional space, for which MD simulations are needed [29, 35, 44, 46]. Additionally, one needs to adjust the SqRA to account for the position-dependent flux. For one-dimensional regular grids, the adjusted SqRA rates are reported in references [29, 44]. Note that the second approach is currently not limited by the computational cost for the SqRA, but by the computational cost for the MD simulations. In future work we compare these two approaches, and apply the SqRA to molecular systems.

Acknowledgments

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany Excellence Strategy—EXC 2008/1 (UniSysCat)—390540038, and through grant CRC 1114 ‘Scaling Cascades in Complex Systems’, project B05 ‘Origin of the scaling cascades in protein dynamics’. BGK is grateful for a writing retreat funded by ‘Die Junge Akademie’.

Appendix A. Estimating the flux from Fick’s laws

In the following, we provide an alternative derivation of the quantity Φ that appears in equation (13). The Fokker–Planck equation [equation (2)] can be written in the form of a continuity equation [35], which at constant potential energy function reduces to Fick’s second law of diffusion [47]

$$\partial_t \rho(x, t) = \frac{\sigma^2}{2} \Delta \rho(x, t) = -\nabla \mathbf{J}, \quad (\text{A1})$$

where the flux is given by

$$\mathbf{J}(x) = -\frac{\sigma^2}{2} \nabla \rho(x, t). \quad (\text{A2})$$

To discretize the $\nabla \mathbf{J}$ on a Voronoi grid, we apply the Gauss theorem, and discretize the surface integral along the sides of the Voronoi cell Ω_j

$$\nabla \mathbf{J}|_{x=x_j} = \lim_{V_j \rightarrow 0} \frac{1}{V_j} \oint_{\partial \Omega_j} \mathbf{J}(z) d\mathcal{S}(z) = \lim_{V_j \rightarrow 0} \frac{1}{V_j} \sum_{i \sim j} \mathbf{J}_{ji} S_{ji}, \quad (\text{A3})$$

where

$$\mathbf{J}_{ji} = -\frac{\sigma^2}{2} \nabla \rho(x, t)|_{x=x_j} \cdot \mathbf{n}_{ji} \quad (\text{A4})$$

is the flux in direction $\mathbf{n}_{ji} = \frac{x_i - x_j}{h_{ji}}$. Using the same finite difference as in equation (21), we obtain

$$\partial_t \rho(x_j, t) = -\nabla \mathbf{J} = \lim_{V_j \rightarrow 0} \frac{1}{V_j} \sum_{i \sim j} \frac{\sigma^2}{2} \frac{\rho(x_i, t) - \rho(x_j, t)}{h_{ji}} \cdot S_{ji}. \quad (\text{A5})$$

To convert the continuous probability density evaluated at the cell centers $\rho(x_j, t)$ to discrete probability defined on finite cell volumes we use the relation $\rho_j(t) = \int_{\Omega_j} \rho(x, t) dx \approx \rho(x_j, t) V_j$. We obtain

$$\partial_t \rho_j(t) = \frac{\sigma^2}{2} \sum_{i \sim j} \left(\frac{\rho_i(t)}{V_i} - \frac{\rho_j(t)}{V_j} \right) \frac{1}{h_{ji}} \cdot S_{ji} \quad (\text{A6})$$

which is identical to equation (24). And thus also from the view-point of the continuity equation and Fick’s laws of diffusion the flux is given as

$$\Phi = \frac{\sigma^2}{2h_{ij}}. \quad (\text{A7})$$

Appendix B. The limit of infinitesimally small cells

We show that the SqRA on a (hyper-)cubic grid converges to the Fokker–Planck equation in the limit of infinitesimally small cells. On these grids the cell length h is the same in each grid dimension, but the extension to rectangular grids is straight forward.

The rates between adjacent cells are given by equation (35)

$$Q_{ij} = \frac{\sigma^2}{2} \frac{1}{h^2} \sqrt{\frac{\pi(x_j)}{\pi(x_i)}} = \frac{\sigma^2}{2h^2} \exp\left(-\beta \frac{V_j - V_i}{2}\right), \quad (\text{B1})$$

which yields the following the master equation [equation (16)]

$$\begin{aligned} \partial_t \rho_j(t) = & \frac{\sigma^2}{2h^2} \sum_{i \sim j} \left[\rho_i(t) \exp\left(-\beta \frac{V_j - V_i}{2}\right) \right. \\ & \left. - \rho_j(t) \exp\left(-\beta \frac{V_i - V_j}{2}\right) \right]. \end{aligned} \quad (\text{B2})$$

We replace the exponential function in equation (B1) by its first-order Taylor expansion, $\exp\left(-\beta \frac{V_j - V_i}{2}\right) = 1 - \beta \frac{V_j - V_i}{2} + \mathcal{O}((V_j - V_i)^2)$, and obtain

$$\begin{aligned}\partial_t \rho_j &\approx \frac{\sigma^2}{2h^2} \sum_{i \sim j} \left[\rho_i \left(1 - \beta \frac{V_j - V_i}{2} \right) - \rho_j \left(1 - \beta \frac{V_i - V_j}{2} \right) \right] \\ &\approx \frac{\sigma^2}{2h^2} \sum_{i \sim j} \left[(\rho_i - \rho_j) - \rho_i \beta \frac{V_j - V_i}{2} + \rho_j \beta \frac{V_i - V_j}{2} \right],\end{aligned}\quad (\text{B3})$$

where we omitted the t -dependence of $\rho_i(t)$ and $\rho_j(t)$ for the sake of brevity. Next we write $V_j - V_i = -(V_i - V_j)$ and substitute $\rho_j \beta \frac{V_i - V_j}{2} = \rho_j \beta V_i - V_j - \rho_j \beta \frac{V_i - V_j}{2}$

$$\begin{aligned}\partial_t \rho_j &\approx \frac{\sigma^2}{2h^2} \sum_{i \sim j} \left[(\rho_i - \rho_j) + \rho_i \beta \frac{V_i - V_j}{2} \right. \\ &\quad \left. + \rho_j \beta (V_i - V_j) - \rho_j \beta \frac{V_i - V_j}{2} \right] \\ &\approx \frac{\sigma^2}{2h^2} \sum_{i \sim j} \left[(\rho_i - \rho_j) + \rho_j \beta (V_i - V_j) \right. \\ &\quad \left. + (\rho_i - \rho_j) \beta \frac{V_i - V_j}{2} \right].\end{aligned}\quad (\text{B4})$$

We recover the continuous probability density $\rho(x, t)$ from the discrete probabilities using the relation $\rho_i(t) = \rho(x_i, t) \mathcal{V}_i = \rho(x_i, t) h^n$, and the relation for the potential energy function $V_i = V(x_i)$. The cell volume h^n appears linearly on both sides of the equation, and cancels:

$$\begin{aligned}\partial_t \rho(x_j) &\approx \frac{\sigma^2}{2h^2} \sum_{i \sim j} \left[(\rho(x_i) - \rho(x_j)) + \rho(x_j) \beta (V(x_i) - V(x_j)) \right. \\ &\quad \left. + (\rho(x_i) - \rho(x_j)) \beta \frac{V(x_i) - V(x_j)}{2} \right],\end{aligned}\quad (\text{B5})$$

where x_i and x_j are the (still discrete) cell centers, and we omit the t -dependence of $\rho(x, t)$ for the sake of brevity.

We remind the reader that $\sum_{i \sim j}$ denotes a sum over all cells Ω_i which are adjacent to cell Ω_j . On a regular grid, every cell Ω_j has two neighbors in each grid dimension, which are centered at $x_j + h \times n_k$ and $x_j - h \times n_k$, where n_k is the unit vector pointing in direction k , and $h \times n_k$ is the lattice vector along the k th dimension. We will now sort the sum over adjacent cells according to grid dimension k and will take the limit $h \rightarrow 0$ to recover the differential equation. With this approach, the first term in equation (B5) becomes

$$\begin{aligned}\lim_{h \rightarrow 0} \frac{1}{h^2} \sum_{i \sim j} (\rho_i - \rho_j) &= \lim_{h \rightarrow 0} \frac{1}{h^2} \sum_{k=1}^n \left[\rho(x_j - h_k) - \rho(x_j) \right. \\ &\quad \left. + \rho(x_j + h_k) - \rho(x_j) \right] \\ &= \lim_{h \rightarrow 0} \sum_{k=1}^n \frac{\rho(x_i - h_k) - 2\rho(x_i) + \rho(x_i + h_k)}{h^2} \\ &= \sum_{k=1}^n \partial_k^2 \rho(x_j) \\ &= \Delta \rho(x_j, t),\end{aligned}\quad (\text{B6})$$

where ∂_k denotes the derivative with respect to the k th dimension. Similarly,

$$\lim_{h \rightarrow 0} \frac{1}{h^2} \sum_{i \sim j} \rho(x_j) \beta (V(x_i) - V(x_j)) = \beta \rho(x_j, t) \Delta V(x_j). \quad (\text{B7})$$

The third term in equation (B5) has the following limit

$$\begin{aligned}\lim_{h \rightarrow 0} \frac{1}{h^2} \sum_{k=1}^n (\rho(x_i) - \rho(x_j)) \beta \frac{V(x_i) - V(x_j)}{2} \\ = \lim_{h \rightarrow 0} \frac{1}{h^2} \sum_{k=1}^n \left[\frac{\beta \rho(x_j + h_k) - \rho(x_j)}{2} \frac{V(x_j + h_k) - V(x_j)}{h_k} \right. \\ \left. + \frac{\beta \rho(x_j - h_k) - \rho(x_j)}{2} \frac{V(x_j - h_k) - V(x_j)}{h_k} \right] \\ = \beta \sum_{k=1}^n \partial_k \rho(x_j) \partial_k V(x_j) \\ = \beta \nabla \rho(x_j, t) \nabla V(x_j).\end{aligned}\quad (\text{B8})$$

In the limit $h \rightarrow 0$ equation (B4) becomes

$$\partial_t \rho(x, t) = \frac{\sigma^2}{2} [\Delta \rho(x, t) + \beta \rho(x, t) \Delta V(x) + \beta \nabla \rho(x, t) \nabla V(x)]. \quad (\text{B9})$$

Applying the product rule and using $\beta = \frac{2}{\sigma^2}$, we obtain the Fokker–Planck equation as stated in equation (2):

$$\partial_t \rho(x, t) = \frac{\sigma^2}{2} \Delta \rho(x, t) + \nabla (\rho(x, t) \cdot \nabla V(x)). \quad (\text{B10})$$

Appendix C. Estimating Φ_{grid} via Taylor expansion of the propagator

The starting point for the derivation is equation (30). We then express the exponential function in terms of its Taylor series and truncate the series after the linear term

$$\begin{aligned}\mathbf{T}(\tau) &= \exp(\tau \mathbf{Q}) = \exp(-\tau \Phi_{\text{grid}} \mathbf{L}) \\ &= \sum \frac{(-\tau \Phi_{\text{grid}} \mathbf{L})^n}{n!} \approx \mathbf{1} - \tau \Phi_{\text{grid}} \mathbf{L}.\end{aligned}\quad (\text{C1})$$

This approximation is valid at small values of τ , and yields the following approximate expression for the rate matrix

$$\mathbf{Q} = -\Phi_{\text{grid}} \mathbf{L} = \frac{1}{\tau} (\mathbf{T}(\tau) - \mathbf{1}), \quad (\text{C2})$$

where $\mathbf{1}$ is the identity matrix. For adjacent cells $L_{ij} = -1$, and $[\mathbf{1}]_{ij} = 0$, and we obtain the following equation for Φ_{grid}

$$\begin{aligned}\Phi_{\text{grid}} &\approx \frac{1}{\tau} T_{ij}(\tau) = \frac{1}{\tau} \left(\sqrt{\frac{1}{2\pi\sigma^2\tau}} \right)^n \exp\left(-\frac{h_{ij}^2}{2\sigma^2\tau}\right) \mathcal{V} \\ &= \Phi_{\text{grid, approx}}(\tau),\end{aligned}\quad (\text{C3})$$

where we used equation (33) to express $T_{ij}(\tau)$. Given the parameters n , σ , h_{ij} , \mathcal{V} , and setting τ to some fixed value, one can in principle calculate an approximation of Φ_{grid} .

However, $\Phi_{\text{grid,approx}}(\tau)$ is very sensitive to τ . It is positive everywhere. For $\tau \rightarrow 0$ and for $\tau \rightarrow \infty$, it approaches zero, and in between it has a maximum which, depending on the other parameters, can be very steep. Thus, choosing τ arbitrarily leads to very inconsistent results. Let us instead choose the value of τ at which $\Phi_{\text{grid,approx}}(\tau)$ reaches its maximum as the optimal value for τ . The derivative of $\Phi_{\text{grid,approx}}(\tau)$ with respect to τ is

$$\frac{d}{d\tau}\Phi_{\text{grid,approx}}(\tau) = -\frac{1}{\tau} \left[1 + \frac{n}{2} - \left(\frac{h_{ij}^2}{2\sigma^2\tau} \right) \right] \cdot \Phi_{\text{grid,approx}}(\tau). \quad (\text{C4})$$

Setting $\frac{d}{d\tau}\Phi_{\text{grid,approx}}(\tau) = 0$ and solving for τ yields

$$\tau_{\text{opt}} = \left(1 + \frac{n}{2} \right)^{-1} \frac{h_{ij}^2}{2\sigma^2}. \quad (\text{C5})$$

Inserting equation (C5) into equation (C3) yields

$$\begin{aligned} \Phi_{\text{grid,approx}}(\tau_{\text{opt}}) &= \frac{(n+2)\sigma^2}{h_{ij}^2} \cdot \frac{1}{h_{ij}^n} \cdot \left(\frac{(1+\frac{n}{2})}{\pi} \right)^{n/2} \\ &\quad \times \exp\left(-\left(1+\frac{n}{2}\right)\right) \cdot \mathcal{V}. \end{aligned} \quad (\text{C6})$$

To test whether equation (C6) is a useful approximation, we compare $\Phi_{\text{grid,approx}}(\tau_{\text{opt}})$ to the grid flux on a hyper-cubic grid which is given by equation (35) as $\Phi_{\text{grid,hyper-cube}} = \frac{1}{2} \frac{\sigma^2}{h_{ij}^2}$. On a hyper-cubic grid $\mathcal{V} = h_{ij}^n$, and equation (C6) simplifies to

$$\begin{aligned} \Phi_{\text{grid,approx}}(\tau_{\text{opt}}) &= (n+2) \cdot \left(\frac{(1+\frac{n}{2})}{\pi} \right)^{n/2} \\ &\quad \times \exp\left(-\left(1+\frac{n}{2}\right)\right) \cdot \frac{\sigma^2}{h_{ij}^2}. \end{aligned} \quad (\text{C7})$$

Note that equation (C7) scales correctly with σ and h_{ij} . For $n = 1, 2, 3$ and 4 , the ratio $\Phi_{\text{grid,approx}}(\tau_{\text{opt}})/\Phi_{\text{grid,hyper-cube}}$ is respectively equal to $0.92, 0.68, 0.58$ and 0.54 ; for $n > 4$ the ratio grows exponentially as the term $\left(\frac{(1+\frac{n}{2})}{\pi}\right)^{n/2}$ in equation (C7) dominates all other terms. Thus, $\Phi_{\text{grid,approx}}(\tau_{\text{opt}})$ cannot be used as a valid approximation of the characteristic flux of the grid. Since equation (C6) likely shows a similar behavior for arbitrary Voronoi grids, we do not recommend using it, and have not included it in our analysis in the main part of the publication.

ORCID iDs

Luca Donati  <https://orcid.org/0000-0002-6014-6875>
 Marcus Weber  <https://orcid.org/0000-0003-3939-410X>
 Bettina G Keller  <https://orcid.org/0000-0002-7051-0888>

References

- [1] Schütte C, Huisinga W and Deuffhard P 2001 Transfer operator approach to conformational dynamics in biomolecular systems *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems* ed B Fiedler (Berlin: Springer) pp 191–223
- [2] Swope W C *et al* 2004 Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and a β -hairpin peptide† *J. Phys. Chem. B* **108** 6582–94
- [3] Buchete N-V and Hummer G 2008 Coarse master equations for peptide folding dynamics† *J. Phys. Chem. B* **112** 6057–69
- [4] Keller B, Daura X and van Gunsteren W F 2010 Comparing geometric and kinetic cluster algorithms for molecular simulation data *J. Chem. Phys.* **132** 074110
- [5] Prinz J-H, Wu H, Sarich M, Keller B, Senne M, Held M, Chodera J D, Schütte C and Noé F 2011 Markov models of molecular kinetics: generation and validation *J. Chem. Phys.* **134** 174105
- [6] Wang W, Cao S, Zhu L and Huang X 2018 Constructing markov state models to elucidate the functional conformational changes of complex biomolecules *WIREs Comput. Mol. Sci.* **8** e1343
- [7] Prinz J-H, Keller B and Noé F 2011 Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables *Phys. Chem. Chem. Phys.* **13** 16912–27
- [8] Husic B E and Pande V S 2018 Markov state models: from an art to a science *J. Am. Chem. Soc.* **140** 2386–96
- [9] Voelz V A, Bowman G R, Beauchamp K and Pande V S 2010 Molecular simulation of *ab initio* protein folding for a millisecond folder NTL9(1–39) *J. Am. Chem. Soc.* **132** 1526–8
- [10] De Fabritiis G, Stanley N and Esteban-martó S 2014 Kinetic modulation of a disordered protein domain by phosphorylation *Nat. Commun.* **5** 5272
- [11] Bowman G R, Bolin E R, Hart K M, Maguire B C and Marqusee S 2015 Discovery of multiple hidden allosteric sites by combining Markov state models and experiments *Proc. Natl Acad. Sci. USA* **112** 2734–9
- [12] Plattner N and Noé F 2015 Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models *Nat. Commun.* **6** 7653
- [13] Zhang L, Unarta I C, Cheung P P-H, Wang G, Wang D and Huang X 2016 Elucidation of the dynamics of transcription elongation by RNA polymerase II using kinetic network models *Acc. Chem. Res.* **49** 698–4
- [14] Witek J, Keller B G, Blatter M, Meissner A, Wagner T and Riniker S 2016 Kinetic models of cyclosporin A in polar and apolar environments reveal multiple congruent conformational states *J. Chem. Inf. Model.* **56** 1547–62
- [15] Keller B G, Aleksic S and Donati L 2018 Markov state models in drug design *Biomolecular Simulations in Structure-Based Drug Discovery* ed F L Gervasio (New York: Wiley-Interscience) p 67
- [16] Pietrucci F 2017 Strategies for the exploration of free energy landscapes: unity in diversity and challenges ahead *Rev. Phys.* **2** 32–45
- [17] Zuckerman D M and Chong L T 2017 Weighted ensemble simulation: review of methodology, applications, and software *Annu. Rev. Biophys.* **46** 43–57
- [18] Valsson O, Tiwary P and Parrinello M 2016 Enhancing important fluctuations: rare events and metadynamics from a conceptual viewpoint *Annu. Rev. Phys. Chem.* **67** 159–84
- [19] Abrams C and Bussi G 2014 Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration *Entropy* **16** 163–99
- [20] Chodera J D, Swope W C, Noé F, Prinz J-H, Shirts M R and Pande V S 2011 Dynamical reweighting: improved estimates of dynamical properties from simulations at multiple temperatures *J. Chem. Phys.* **134** 244107
- [21] Rosta E and Hummer G 2014 Free energies from dynamic weighted histogram analysis using unbiased Markov state model *J. Chem. Theory Comput.* **11** 276–85

- [22] Donati L and Keller B G 2018 Girsanov reweighting for meta-dynamics simulations *J. Chem. Phys.* **149** 072335
- [23] Kieninger S, Donati L and Keller B G 2020 Dynamical reweighting methods for Markov models *Curr. Opin. Struct. Biol.* **61** 124–31
- [24] Grossfield A, Patrone P N, Roe D R, Schultz A J, Siderius D W and Zuckerman D M 2018 Best practices for quantification of uncertainty and sampling quality in molecular simulations [article v1.0] *Living J. Comput. Mol. Sci.* **1** 5067
- [25] Zuckerman D M 2011 Equilibrium sampling in biomolecular simulations *Annu. Rev. Biophys.* **40** 41–62
- [26] Lie H C, Fackeldey K and Weber M 2013 A square root approximation of transition rates for a Markov state model *SIAM J. Matrix Anal. Appl.* **34** 738–56
- [27] Donati L, Heida M, Keller B G and Weber M 2018 Estimation of the infinitesimal generator by square-root approximation *J. Phys.: Condens. Matter* **30** 425201
- [28] Heida M 2018 Convergences of the squareroot approximation scheme to the Fokker–Planck operator *Math. Models Methods Appl. Sci.* **28** 2599–635
- [29] Bicout D J and Szabo A 1998 Electron transfer reaction dynamics in non-Debye solvents *J. Chem. Phys.* **109** 2325
- [30] Dixit P D, Jain A, Stock G and Dill K A 2015 Inferring transition rates of networks from populations in continuous-time Markov processes *J. Chem. Theory Comput.* **11** 5464–72
- [31] Stock G, Ghosh K and Dill K A 2008 Maximum caliber: a variational approach applied to two-state dynamics *J. Chem. Phys.* **128** 194102
- [32] Otten M and Stock G 2010 Maximum caliber inference of nonequilibrium processes *J. Chem. Phys.* **133** 034119
- [33] Barber C B, Dobkin D P and Huhdanpaa H 1996 The quick-hull algorithm for convex hulls *ACM Trans. Math. Softw.* **22** 469–83
- [34] Oostendorp T F, van Oosterom A and Huiskamp G 1989 Interpolation on a triangulated 3D surface *J. Comput. Phys.* **80** 331–43
- [35] Risken H 1989 The Fokker–Planck equation *Methods of Solution and Applications* 2nd edn (Berlin: Springer)
- [36] Weber M 2006 *Meshless Methods in Conformation Dynamics* (München: Verlag Dr. Hut)
- [37] Nüske F, Keller B G, Pérez-Hernández G, Mey A S J S and Noé F 2014 Variational approach to molecular kinetics *J. Chem. Theory Comput.* **10** 1739–52
- [38] Arfken G B, Weber H J and Harris F E 2001 *Mathematical Methods for Physicists* 5th edn (New York: Academic)
- [39] Sukumar N 2003 Voronoi cell finite difference method for the diffusion operator on arbitrary unstructured grids *Int. J. Numer. Meth. Eng.* **57** 1–34
- [40] Son S-K 2011 Voronoi-cell finite difference method for accurate electronic structure calculation of polyatomic molecules on unstructured grids *J. Comput. Phys.* **230** 160–2173
- [41] Reuter M, Biasotti S, Giorgi D, Patanè G and Spagnuolo M 2009 Discrete Laplace–Beltrami operators for shape analysis and segmentation *Comput. Graph.* **33** 381–90
- [42] Leimkuhler B and Matthews C 2015 *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods (Interdisciplinary Applied Mathematics vol 39)* (Berlin: Springer)
- [43] Wan H, Zhou G and Voelz V A 2016 A maximum-caliber approach to predicting perturbed folding kinetics due to mutations *J. Chem. Theory Comput.* **12** 5768–76
- [44] Hummer G 2005 Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations *New J. Phys.* **7** 34
- [45] Schulz R et al 2017 Data-based modeling of drug penetration relates human skin barrier function to the interplay of diffusivity and free-energy profiles *Proc. Natl Acad. Sci. USA* **114** 3631–6
- [46] Peter H, Talkner P and Borkovec M 1990 Reaction-rate theory: fifty years after Kramers *Rev. Mod. Phys.* **62** 251–341
- [47] Engel T and Reid P 2014 *Physical Chemistry* 3rd edn (Harlow, Essex: Pearson Education)