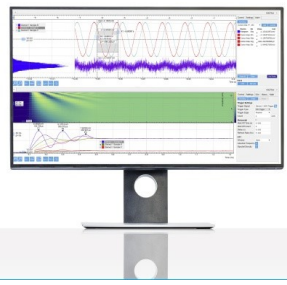
The Journal of Chemical Physics

Path probability ratios for Langevin dynamics—Exact and approximate

Cite as: J. Chem. Phys. **154**, 094102 (2021); https://doi.org/10.1063/5.0038408
Submitted: 22 November 2020 . Accepted: 09 February 2021 . Published Online: 01 March 2021

S. Kieninger, and B. G. Keller

View Online Export Citation CrossMark

Challenge us.
What are your needs for periodic signal detection?
Watch

Zurich Instruments

© 2021 Author(s).

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Path probability ratios for Langevin dynamics—Exact and approximate

Cite as: J. Chem. Phys. 154, 094102 (2021); doi: 10.1063/5.0038408
Submitted: 22 November 2020 • Accepted: 9 February 2021 •
Published Online: 1 March 2021

View Online Export Citation CrossMark

S. Kieninger and B. G. Keller[a)]

AFFILIATIONS

Department of Biology, Chemistry, Pharmacy, Freie Universität Berlin, Arnimallee 22, D-14195 Berlin, Germany

Note: This paper is part of the JCP Special Collection in Honor of Women in Chemical Physics and Physical Chemistry.
[a)]**Author to whom correspondence should be addressed:** bettina.keller@fu-berlin.de

ABSTRACT

Path reweighting is a principally exact method to estimate dynamic properties from biased simulations—provided that the path probability ratio matches the stochastic integrator used in the simulation. Previously reported path probability ratios match the Euler–Maruyama scheme for overdamped Langevin dynamics. Since molecular dynamics simulations use Langevin dynamics rather than overdamped Langevin dynamics, this severely impedes the application of path reweighting methods. Here, we derive the path probability ratio M_L for Langevin dynamics propagated by a variant of the Langevin Leapfrog integrator. This new path probability ratio allows for exact reweighting of Langevin dynamics propagated by this integrator. We also show that a previously derived approximate path probability ratio M_{approx} differs from the exact M_L only by $\mathcal{O}(\xi^4\Delta t^4)$ and thus yields highly accurate dynamic reweighting results. (Δt is the integration time step, and ξ is the collision rate.) The results are tested, and the efficiency of path reweighting is explored using butane as an example.

Published under license by AIP Publishing. https://doi.org/10.1063/5.0038408

I. INTRODUCTION

Molecular dynamics are astonishingly complex and occur in a wide range of length and timescales.[1–3] To elucidate the mechanisms by which different parts of a molecular system interact and how macroscopic properties arise from these interactions, molecular dynamics (MD) simulations have become an indispensable tool.[4–9] Because the timescales covered by MD simulations are often orders of magnitude lower than the slowest timescale of the system, a wide variety of enhanced sampling techniques have been developed, which distort the dynamics of the simulation such that rare molecular transitions occur more frequently. This can be achieved by raising the temperature or by adding a bias to the potential energy function.[10,11] How to extract the correct values of dynamical properties (mean-first passage times, residence times, binding rates, or transition probabilities) from these accelerated dynamics is an open question and a very active field of research.

The goal of dynamical reweighting methods is to estimate dynamical properties of the system at a target state $\widetilde{S}$ from a trajectory generated at simulation state S. S could correspond to a higher temperature or to a biased potential. Starting points for the derivation of dynamical reweighting methods are Kramers rate theory,[12–15] the likelihood function for estimating the transition probabilities from MD trajectories,[16–19] or a discretization of the Fokker–Planck equation.[7,20–22] The methods differ in the ease of use and the severity of the assumptions they make.[23]

A principally exact formalism to reweight dynamic properties is path reweighting methods, which have been reported already early in Refs. 24–28. In path reweighting methods, the trajectory generated at state S is split into short paths ω. Then, the path probability $\widetilde{P}_L(\omega;\Delta t|(x_0,v_0))$ of a given ω at the target state $\widetilde{S}$ is calculated by reweighting the path probability $P_L(\omega;\ \Delta t|(x_0,\ v_0))$ of ω at the simulation state S,

$$\widetilde{P}_L(\omega;\Delta t|(x_0,v_0)) \approx M\cdot P_L(\omega;\Delta t|(x_0,v_0)). \tag{1}$$

$(x_0,\ v_0)$ is the initial state of the path ω, and Δt is the integration time step. $M(\omega)$ is the path probability ratio or reweighting factor. Equation (1) is exact if the path probability ratio $M = \widetilde{P}_L(\omega;\Delta t|(x_0,v_0))/P_L(\omega;\Delta t|(x_0,v_0))$ is derived from the numerical integration scheme used to generate ω. The mathematical basis

Published under license by AIP Publishing

for path reweighting methods is the Girsanov theorem,[29,30] or else, they can be derived from the Onsager–Machlup action.[24–27,31] A prerequisite for path reweighting is that a stochastic integrator is used in the MD simulation, e.g., a Langevin thermostat.

However, it has been challenging to apply path reweighting to simulations of large molecular systems. For example, the variance of the reweighting estimators increases rapidly with increasing path length such that for long paths, reweighting becomes inefficient compared to direct simulation of the target state. Combining path reweighting techniques with Markov state models (MSMs) alleviates this problem.[32–35] In MSMs,[36–42] the dynamics of the system is represented by transitions between discrete states in the conformational space of the molecular system, where the lag time τ of the transition is much shorter than the slow timescales of the system. Thus, only short paths of length τ are needed to estimate and reweight the transition probabilities.

Second, a number of technical difficulties arise. The path probability ratio M decreases exponentially with the path length τ such that the standard numerical accuracy is quickly exceeded. This problem can be solved by using high precision arithmetic libraries.[35] To calculate the path probability ratio M, one needs to know the trajectory and the random numbers of the stochastic integrator at every integration time step. Writing this information to disk at every integration time step is not a workable option. We, therefore, proposed to calculate the path reweighting factor "on-the-fly" during the simulation and to write out intermediate results at regular intervals, e.g., whenever the positions are written to disk. The additional storage requirements and computational costs for the "on-the-fly"-calculations are negligible compared to the overall cost of the simulation.[34,35] Having solved the technical challenges, we tested the path reweighting method on several peptides using path lengths of up to $\tau = 600$ ps.[34,35] Applications to larger systems and longer path lengths are likely within reach.

Yet, the equation for the path probability ratio M poses a barrier to a more widespread use of path reweighting methods. Because M is derived from the stochastic integration scheme used to simulate the system, one cannot readily apply a path probability ratio derived for one integration scheme to a simulation generated by another integration scheme.

In temperature reweighting, i.e., when simulation and target state differ in the temperature, only the random term of the stochastic integrator is affected by the change in temperature. Path probability ratios for temperature reweighting have been constructed by rescaling the normal distributions of the random or noise terms of the stochastic integration scheme.[32,43]

In potential reweighting, i.e., when simulation and target state differ in the potential energy function, one needs to account for changes in the drift terms of the stochastic integration scheme. The path probability ratio M_o for the Euler–Maruyama scheme for overdamped Langevin dynamics has been reported multiple times.[24–26,33] However, the dynamics of large molecular systems is better reproduced by Langevin dynamics, and MD programs implement a wide variety of Langevin integration schemes.[44–53] The time-continuous Onsager–Machlup action for Langevin dynamics has been reported,[27] but to the best of our knowledge, path probability ratios for Langevin integration schemes M_L have not yet been reported. Thus, exact path reweighting for Langevin dynamics has not been possible, so far.

In Refs. 34 and 35, we demonstrated that path reweighting can be applied to biased simulations of large molecular systems, nonetheless. We used an approximate path probability ratio M_{approx} that is based on the path probability ratio for the Euler–Maruyama scheme but uses the random numbers that are generated during the Langevin MD simulation. We tested M_{approx} extensively, and for low-dimensional model systems and for molecular systems, this approximate path probability ratio yielded very accurate results. In these two publications, we used a variant of the Langevin Leapfrog integration scheme developed by Izaguirre, Sweet, and Pande[49] to propagate the system. Both the Langevin Leapfrog integration scheme and its variant are implemented in OpenMM[54] (see Appendix A). We will abbreviate the variant by the "ISP scheme."

In this contribution, we derive the path probability ratio M_L for Langevin dynamics propagated by a variant of the Langevin Leapfrog integrator.[49] M_L allows for exact reweighting of Langevin dynamics (Sec. IV). We analyze why M_{approx} is an excellent approximation to M_L (Sec. VI), and we discuss whether there are scenarios in which M_o is a viable approximation to M_L (Sec. V). The general framework of the path reweighting equations and the corresponding equations for the Euler–Maruyama scheme are summarized in Secs. II and III. Section VIII reports the computational details.

II. PATH REWEIGHTING

The path probability $P(\omega; \Delta t|(x_0, v_0))$ is the probability to generate a time-discretized path $\omega = (x_0, x_1, \ldots, x_n)$ starting in a predefined initial state (x_0, v_0) at the simulation potential $V(x)$. The notation emphasizes that the probability is conditioned on an initial state (x_0, v_0) and that the path has been generated with a fixed time step Δt, whereas ω is the argument of the function. In short, $P(\omega; \Delta t|(x_0, v_0))$ maps a path in position space to a probability. Its functional form depends on the integration scheme used to generate ω and the potential energy function.

The path probability ratio is the ratio between the probability $\widetilde{P}(\omega; \Delta t|(x_0, v_0))$ to generate a path ω at a target potential,

$$\widetilde{V}(x) = V(x) + U(x), \tag{2}$$

and the probability $P(\omega; \Delta t|(x_0, v_0))$ to generate the same path ω at the simulation potential $V(x)$,

$$M(\omega; \Delta t|(x_0, v_0)) = \frac{\widetilde{P}(\omega; \Delta t|(x_0, v_0))}{P(\omega; \Delta t|(x_0, v_0))}. \tag{3}$$

The potential energy function $U(x)$ is usually called perturbation or bias.

In integration schemes for stochastic dynamics, random numbers are used to propagate the system. If a single random number is drawn per integration step, then the probability to generate ω is equal to the probability $P(\eta)$ to generate the corresponding random number sequence $\eta = (\eta_0, \eta_1, \ldots, \eta_{n-1})$,

$$P(\omega; \Delta t|(x_0, v_0)) = P(\eta), \tag{4}$$

where ω and η are linked by the equations for the integration scheme. Since the random numbers η_k are drawn from a Gaussian

Published under license by AIP Publishing

TABLE I. References to the equations for the properties introduced in Sec. II.

		Overdamped Langevin	Langevin
Equation of motion		Eq. (11)	Eq. (19)
Integration scheme		Eq. (12)	Eqs. (20) and (21)
Path probability	$P(\omega; \Delta t \vert (x_0, v_0))$	Eq. (13)	Eq. (22)
Path probability ratio	$M(\omega; \Delta t \vert (x_0, v_0))$	Eq. (14)	Eq. (23)
Random number	η_k	Eq. (15)	Eq. (24)
Random number difference	$\Delta\eta_k$	Eq. (17)	Eq. (26)
Random number probability ratio	$M(\omega, \eta; \Delta t \vert (x_0, v_0))$	Eq. (18)	Eq. (27)

distribution with zero mean and unit variance, the functional form of $P(\eta)$ is

$$P(\eta) = N \exp\left(-\frac{1}{2}\sum_{k=0}^{n-1}\eta_k^2\right), \quad N = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}}. \tag{5}$$

$P(\eta)$ is a function that maps a random number sequence to a probability. One can interpret Eq. (4) as a change in variables from ω to η, where the change is defined by the equations for the integration scheme.

Suppose that η is the random number sequence needed to generate ω at a simulation potential $V(x)$. To generate the same path at a target potential $\widetilde{V}(x)$, one would need a different random number sequence $\widetilde{\eta} = (\widetilde{\eta}_0, \widetilde{\eta}_1, \ldots, \widetilde{\eta}_{n-1})$, with

$$\widetilde{\eta}_k = \eta_k + \Delta\eta_k. \tag{6}$$

$\Delta\eta_k$ is the random number difference, and it depends on the integration scheme and the difference between the two potentials. The random number probability ratio is the ratio between the probability of drawing η and the probability of drawing $\widetilde{\eta}_k$,

$$\begin{aligned}\frac{P(\widetilde{\eta})}{P(\eta)} &= \frac{N \exp\left(-\frac{1}{2}\sum_{k=0}^{n-1}(\eta_k + \Delta\eta_k)^2\right)}{N \exp\left(-\frac{1}{2}\sum_{k=0}^{n-1}\eta_k^2\right)} \\ &= \exp\left(-\sum_{k=0}^{n-1}\eta_k \cdot \Delta\eta_k\right) \cdot \exp\left(-\frac{1}{2}\sum_{k=0}^{n-1}(\Delta\eta_k)^2\right).\end{aligned} \tag{7}$$

Mathematically, the following has happened in the previous paragraph. The path ω remained unchanged. The functional form of the path probability has changed as $\widetilde{P}(\omega; \Delta t \vert (x_0, v_0))$ because the potential energy enters the equations for the integration scheme. Likewise, the change in variables from ω to $\widetilde{\eta}$ has changed. The functional form of the random number probability remains the same [Eq. (5)]. The analogon to Eq. (4) for the target potential is

$$\widetilde{P}(\omega; \Delta t \vert (x_0, v_0)) = P(\widetilde{\eta}), \tag{8}$$

where ω and $\widetilde{\eta}$ are linked by the equations for the integration scheme using $\widetilde{V}(x)$. Given the two changes in variables for the simulation and the target potential, the path probability ratio [Eq. (3)] and the random number probability ratio [Eq. (7)] are equal. Note that Eq. (3) is a ratio of two different functions that have the same argument ω, whereas Eq. (7) is the ratio of the same function with different arguments η and $\widetilde{\eta}$.

Equation (7) is of little practical use because $\widetilde{\eta}$ is not available from a simulation at the simulation state. However, the random number difference $\Delta\eta_k$ can be expressed as a function of ω, and the random number probability ratio can thus be expressed as a function of ω and η,

$$M(\omega, \eta; \Delta t \vert (x_0, v_0)) = \frac{P(\widetilde{\eta})}{P(\eta)}. \tag{9}$$

For a path ω and the corresponding random number sequence η that was used to generate this path, we will use the following equality:

$$M(\omega, \eta; \Delta t \vert (x_0, v_0)) = M(\omega; \Delta t \vert (x_0, v_0)). \tag{10}$$

The functional form and the value of the properties introduced in this section depend strongly on the integration scheme. In Sec. III, we summarize the equations for the Euler–Maruyama scheme for overdamped Langevin dynamics. In Sec. IV, we derive the corresponding equations for the ISP integration scheme for Langevin dynamics (see Table I). Throughout this manuscript, properties associated with Langevin dynamics are subscripted with L, and properties associated with overdamped Langevin dynamics are subscripted with o.

III. OVERDAMPED LANGEVIN DYNAMICS

A. Equation of motion and integration scheme

Consider a one particle system that moves in a one-dimensional position space with temperature T and potential energy function V. The overdamped Langevin equation of motion is

$$\dot{x}(t) = -\frac{\nabla V(x(t))}{\xi m} + \sqrt{\frac{2k_BT}{\xi m}}\,\eta(t), \tag{11}$$

with particle mass m, position x, velocity $v = \dot{x}$, and Boltzmann constant k_B. $x(t) \in \Omega_o$ is the state of the system at time t, where $\Omega_o \subset \mathbb{R}$ is the state space of the system. The collision rate ξ (in units of s^{-1}) models the interaction with the thermal bath. $\eta(t) \in \mathbb{R}$ describes an uncorrelated Gaussian white noise with unit variance centered at zero, which is scaled by the volatility $\sqrt{\frac{2k_BT}{\xi m}}$.

A numerical algorithm to calculate an approximate solution to Eq. (11) is the Euler–Maruyama integration scheme,[30,55]

Published under license by AIP Publishing

$$x_{k+1} = x_k - \frac{\nabla V(x_k)}{\xi m}\Delta t + \sqrt{\frac{2k_BT}{\xi m}}\sqrt{\Delta t}\,\eta_{o,k}, \tag{12}$$

where Δt is the time step, x_k is the position, and $\eta_{o,k}$ is the random number at iteration k. The random numbers are drawn from a Gaussian distribution with zero mean and unit variance. For $k = 0, \ldots, n-1$, Eq. (12) yields a time-discretized overdamped Langevin path $\omega_o = (x_0, x_1, \ldots, x_n)$, which starts at the pre-defined initial position x_0. Note that while the state of the system at iteration k is defined by the position x_k, the progress to x_{k+1} depends on x_k and on the value of the random number $\eta_{o,k}$. The random number sequence that was used to generate a specific ω_o is denoted by $\eta_o = (\eta_{o,0}, \ldots, \eta_{o,n-1})$.

B. Path probability and path probability ratio

The probability to observe a path ω_o generated by the Euler–Maruyama scheme [Eq. (12)] is[28,34,56,57]

$$P_o(\omega_o;\Delta t|x_0) = \left[\sqrt{\frac{\xi m}{4\pi k_BT\Delta t}}\right]^n \cdot \exp\left(-\frac{\xi m}{4k_BT\Delta t}\sum_{k=0}^{n-1}\left(x_{k+1} - x_k + \frac{\Delta t}{\xi m}\nabla V(x_k)\right)^2\right). \tag{13}$$

For the Euler–Maruyama scheme, the path probability $P_o(\omega_o;\Delta t|x_0)$ does not depend on the initial velocity; hence, we dropped v_0 in the notation. However, it does depend on the potential energy function $V(x)$ that has been used in Eq. (12) to generate the path ω_o.

The path probability that the same path ω_o has been generated at a target potential $\widetilde{V}(x)$ [Eq. (2)] is $\widetilde{P}_o(\omega_o;\Delta t|x_0)$, which is obtained by replacing the potential $V(x)$ with $\widetilde{V}(x)$ in Eq. (13). The ratio between the two path probabilities is

$$M_o(\omega_o;\Delta t|x_0) = \frac{\widetilde{P}_o(\omega_o;\Delta t|x_0)}{P_o(\omega_o;\Delta t|x_0)} = \exp\left(-\frac{\sum_{k=0}^{n-1}(x_{k+1}-x_k)\left(\nabla\widetilde{V}(x_k) - \nabla V(x_k)\right)}{2k_BT}\right) \times \exp\left(-\frac{\sum_{k=0}^{n-1}\left(\nabla\widetilde{V}^2(x_k) - \nabla V^2(x_k)\right)\Delta t}{4k_BT\xi m}\right). \tag{14}$$

Equation (14) is a function of the path ω_o and does not depend on the random number sequence η_o explicitly. It is equivalent to Eq. (B4) in Ref. 34.

C. Random numbers and random number probability ratio

Given ω_o, the sequence of random numbers η_o that was used to generate ω_o at the simulation potential $V(x)$ can be back-calculated by rearranging Eq. (12) for $\eta_{o,k}$,

$$\eta_{o,k} = \sqrt{\frac{\xi m}{2k_BT\Delta t}}\left(x_{k+1} - x_k + \frac{\nabla V(x_k)}{\xi m}\Delta t\right). \tag{15}$$

We remark that the path probability [Eq. (13)] can formally be derived by inserting Eq. (15) into Eq. (5). Since Eq. (15) defines a coordinate transformation from x_k to $\eta_{o,k}$, one needs to normalize with respect to the new coordinates in order to obtain the correct normalization constant. The random number sequence $\widetilde{\eta}_o$ needed to generate ω_o at a target potential $\widetilde{V}(x)$ is calculated by inserting Eq. (2) into Eq. (15),

$$\widetilde{\eta}_{o,k} = \sqrt{\frac{\xi m}{2k_BT\Delta t}}\left(x_{k+1} - x_k + \frac{\nabla V(x_k)}{\xi m}\Delta t\right) + \sqrt{\frac{\Delta t}{2k_BT\xi m}}\nabla U(x_k) = \eta_{o,k} + \Delta\eta_{o,k}. \tag{16}$$

Equation (15) defines the change in variables from ω to η_o for the Euler–Maruyama scheme at the simulation potential. Likewise, Eq. (16) defines the change in variables from ω to $\widetilde{\eta}_o$ at the target potential. The random number difference is

$$\Delta\eta_{o,k} = \sqrt{\frac{\Delta t}{2k_BT\xi m}}\nabla U(x_k). \tag{17}$$

It depends on the perturbation $U(x)$, but not on the simulation potential $V(x)$. Inserting $\Delta\eta_{o,k}$ [Eq. (17)] into Eq. (7) yields the random number probability ratio,

$$M_o(\omega_o,\eta_o;\Delta t|x_0) = \exp\left(-\sum_{k=0}^{n-1}\sqrt{\frac{\Delta t}{2k_BT\xi m}}\nabla U(x_k)\cdot\eta_{o,k}\right) \cdot \exp\left(-\frac{1}{2}\sum_{k=0}^{n-1}\frac{\Delta t}{2k_BT\xi m}\left(\nabla U(x_k)\right)^2\right). \tag{18}$$

Because of Eq. (10), Eqs. (14) and (18) are equal. However, the two probability ratios use different time-series and different information on the system to evaluate the path probability ratio. To evaluate Eq. (14), one needs the path ω_o, the simulation potential $V(x)$, and the target potential $\widetilde{V}(x)$. To evaluate Eq. (18), one needs the path ω_o, the random number sequence for the simulation potential η_o, and the perturbation $U(x)$. Because $U(x)$ often only affects a few coordinates of the systems, i.e., it is low-dimensional, Eq. (18) is computationally more efficient. Besides the force calculation $-\nabla V(x)$ needed to generate the path ω_o, it requires an additional force calculation $-\nabla U(x)$ only along the coordinates that are affected by the perturbation. By contrast, Eq. (14) requires an additional force calculation on the entire system $-\nabla\widetilde{V}(x)$.

IV. LANGEVIN DYNAMICS

A. Equation of motion and integration scheme

Consider a one particle system that moves in a one-dimensional position space with temperature T and potential energy function V. The Langevin equation of motion is

$$m\ddot{x}(t) = -\nabla V(x(t)) - \xi m\dot{x}(t) + \sqrt{2k_BT\xi m}\,\eta(t), \tag{19}$$

with particle mass m, position x, velocity $v = \dot{x}$, acceleration $a = \ddot{x}$, and Boltzmann constant k_B. The state of the system at time t is deter-

Published under license by AIP Publishing

mined by the position and the velocity $(x(t), \dot{x}(t)) \in \Omega_L$, where $\Omega_L \subset \mathbb{R}^2$ is the state space of the system. The collision rate ξ (in units of s^{-1}) models the interaction with the thermal bath. $\eta \in \mathbb{R}$ describes an uncorrelated Gaussian white noise with unit variance centered at 0, which is scaled by the volatility $\sqrt{2k_BT\xi m}$.

A numerical algorithm to calculate an approximate solution to Eq. (19) is the ISP scheme,[49]

$$x_{k+1} = x_k + \exp(-\xi\,\Delta t)\, v_k \Delta t - [1 - \exp(-\xi\,\Delta t)]\,\frac{\nabla V(x_k)}{\xi m}\Delta t + \sqrt{\frac{k_BT}{m}\,[1 - \exp(-2\xi\,\Delta t)]}\,\eta_{L,k}\,\Delta t, \tag{20}$$

$$v_{k+1} = \frac{x_{k+1} - x_k}{\Delta t}, \tag{21}$$

where Δt is the time step, x_k is the position, v_k is the velocity, and $\eta_{L,k}$ is the random number at iteration k (see Appendix A). The random numbers are drawn from a Gaussian distribution with zero mean and unit variance. For $k = 0, \ldots, n-1$, Eqs. (20) and (21) yield a time-discretized Langevin path $\omega_L = ((x_0, v_0), (x_1, v_1), \ldots, (x_n, v_n))$, which starts at the pre-defined initial state (x_0, v_0). Note that while the state of the system at iteration k is defined by the tuple $(x_k, v_k) \in \Omega_L$, the progress to (x_{k+1}, v_{k+1}) depends on (x_k, v_k) and on the value of the random number $\eta_{L,k}$. The random number sequence that was used to generate a specific ω_L is denoted by $\eta_L = (\eta_{L,0}, \ldots, \eta_{L,n-1})$.

The position x_{k+1} is treated as a random variable because it directly depends on a random number [Eq. (20)], while the velocity v_{k+1} is calculated from the new position x_{k+1} and the preceding position x_k. Because the velocity v_k in Eq. (20) is determined by the positions x_k and x_{k-1} [Eq. (21)], it carries a small memory effect into the time-evolution of x.

B. Path probability and path probability ratio

The probability to generate a path ω_L by the ISP scheme [Eqs. (20) and (21)] at the simulation potential $V(x)$ is

$$P_L(\omega_L; \Delta t|(x_0, v_0)) = \left[\prod_{k=0}^{n-1} \delta\left(v_{k+1} - \frac{x_{k+1} - x_k}{\Delta t}\right)\right] \cdot \left[\sqrt{\frac{m}{2\pi k_BT\Delta t^2(1 - \exp(-2\xi\Delta t))}}\right]^n \times \exp\left(-\sum_{k=0}^{n-1} \frac{m\left(x_{k+1} - x_k - \exp(-\xi\Delta t)v_k\Delta t + (1 - \exp(-\xi\Delta t))\frac{\nabla V(x_k)}{\xi m}\Delta t\right)^2}{2k_BT(1 - \exp(-2\xi\Delta t))\Delta t^2}\right). \tag{22}$$

The derivation of Eq. (22) is shown in Appendixes B and C. Appendix B explains the strategy for the derivation, and Appendix C shows how to solve the integrals that appear in the derivation.

The path probability $\widetilde{P}_L(\omega_L; \Delta t|(x_0, v_0))$ to generate a path ω_L by the ISP scheme at the target potential is obtained by inserting $\widetilde{V}(x)$ [Eq. (2)] into Eq. (22). The path probability ratio for overdamped Langevin dynamics is

$$\begin{aligned} M_L(\omega_L; \Delta t|(x_0, v_0)) &= \frac{\widetilde{P}_L(\omega_L; \Delta t|(x_0, v_0))}{P_L(\omega_L; \Delta t|(x_0, v_0))} \\ &= \exp\left(-\frac{\sum_{k=0}^{n-1}(x_{k+1} - x_k)\left(\nabla\widetilde{V}(x_k) - \nabla V(x_k)\right)}{k_BT\xi(1 + \exp(-\xi\Delta t))\Delta t}\right) \\ &\quad \cdot \exp\left(\frac{\sum_{k=0}^{n-1} v_k\left(\nabla\widetilde{V}(x_k) - \nabla V(x_k)\right)}{k_BT\xi(1 + \exp(\xi\Delta t))}\right) \\ &\quad \cdot \exp\left(-\frac{\exp(\xi\Delta t) - 1}{\exp(\xi\Delta t) + 1} \cdot \frac{\sum_{k=0}^{n-1}\left(\nabla\widetilde{V}^2(x_k) - \nabla V^2(x_k)\right)}{2k_BT\xi^2 m}\right). \end{aligned} \tag{23}$$

Analogous to Eq. (14), Eq. (23) is a function of the path ω_L and does not depend on the random number sequence η_L.

C. Random numbers and random number probability ratio

Given ω_L, the sequence of random numbers η_L, which was used to generate ω_L at the simulation potential $V(x)$, can be back-calculated by rearranging Eq. (20) for $\eta_{L,k}$,

$$\eta_{L,k} = \sqrt{\frac{m}{k_BT(1 - \exp(-2\xi\Delta t))\Delta t^2}} \times \left(x_{k+1} - x_k - \exp(-\xi\Delta t)v_k\Delta t + (1 - \exp(-\xi\Delta t))\frac{\nabla V(x_k)}{\xi m}\Delta t\right). \tag{24}$$

The random number sequence $\widetilde{\eta}_L$ needed to generate ω_L at a target potential $\widetilde{V}(x)$ is calculated by inserting Eq. (2) into Eq. (24),

$$\begin{aligned} \widetilde{\eta}_{L,k} &= \sqrt{\frac{m}{k_BT(1 - \exp(-2\xi\Delta t))\Delta t^2}} \times \left(x_{k+1} - x_k - \exp(-\xi\Delta t)(x_k - x_{k-1}) + (1 - \exp(-\xi\Delta t))\frac{\nabla V(x_k)}{\xi m}\Delta t\right) \\ &\quad + \sqrt{\frac{1}{k_BT\xi^2 m}} \cdot \frac{1 - \exp(-\xi\Delta t)}{\sqrt{1 - \exp(-2\xi\Delta t)}}\nabla U(x_k) \\ &= \eta_{L,k} + \Delta\eta_{L,k}. \end{aligned} \tag{25}$$

Published under license by AIP Publishing

Equation (24) defines the change in variables from ω to η_L for the ISP scheme at the simulation potential. Likewise, Eq. (25) defines the change in variables from ω to $\widetilde{\eta}_L$ at the target potential. The random number difference is

$$\Delta\eta_{L,k} = \sqrt{\frac{1}{k_B T \xi^2 m}} \cdot \frac{1-\exp(-\xi\Delta t)}{\sqrt{1-\exp(-2\xi\Delta t)}} \nabla U(x_k). \tag{26}$$

Again, the random number difference depends on the perturbation potential $U(x)$, but not on the simulation potential $V(x)$. Inserting $\Delta\eta_{L,k}$ [Eq. (26)] into Eq. (7) yields the random number probability ratio,

$$\begin{aligned} &M_L(\omega_L, \eta_L; \Delta t|(x_0, v_0)) \\ &= \exp\left(-\frac{1-\exp(-\xi\Delta t)}{\sqrt{1-\exp(-2\xi\Delta t)}} \cdot \frac{\sum_{k=0}^{n-1} \nabla U(x_k)\, \eta_{L,k}}{\sqrt{k_B T \xi^2 m}} \right) \\ &\times \exp\left(-\frac{(1-\exp(-\xi\Delta t))^2}{1-\exp(-2\xi\Delta t)} \cdot \frac{\sum_{k=0}^{n-1} \nabla U^2(x_k)}{2 k_B T \xi^2 m} \right). \end{aligned} \tag{27}$$

Analogous to the path probability ratio for overdamped Langevin dynamics, $M_L(\omega_L; \Delta t|(x_0, v_o))$ [Eq. (23)] and $M_L(\omega_L, \eta_L; \Delta t|(x_0, v_0))$ [Eq. (27)] yield the same path probability ratio for a given path ω_L that has been generated using the random number sequence η_L, but they use different arguments. Again, the path probability from random numbers $M_L(\omega_L, \eta_L; \Delta t|(x_0, v_0))$ requires an additional force calculation $-\nabla U(x)$ only along the coordinates that are affected by the perturbation, making it computationally more efficient than $M_L(\omega_L; \Delta t|(x_0, v_0))$ in most cases.

V. COMPARING LANGEVIN AND OVERDAMPED LANGEVIN DYNAMICS

A. Test system

Our test system is a one-dimensional one particle system at the simulation potential $V(x)$ (Fig. 1, orange line) and at the target potential $\widetilde{V}(x)$ (Fig. 1, black line). The trajectories generated at $V(x)$ will be reweighted to the target potential $\widetilde{V}(x)$. The black lines in Fig. 4(b) represent the first three dominant MSM eigenfunctions[40] associated with the target potential. The implied timescales[37] are $t_0 = \infty$, $t_1 = 20.5$ s, and $t_2 = 6.0$ s, which are shown as black lines in Fig. 4(c). Computational details are reported in Sec. VIII.

B. From random numbers η to paths ω_o and ω_L

Given a random number sequence $\eta = (\eta_0, \ldots, \eta_{n-1})$ and a starting state (x_0, v_0), one can use the Euler–Maruyama scheme to generate an overdamped Langevin path ω_o, or else, one can use the ISP scheme to generate a Langevin path ω_L. We discuss briefly how the difference between ω_o and ω_L depends on the combined parameter $\xi\Delta t$, which can be interpreted as the number of collisions per time step.

In the limit of high friction $\xi m\dot{x} \gg m\ddot{x}$, the Langevin dynamics [Eq. (19)] approaches the overdamped Langevin dynamics

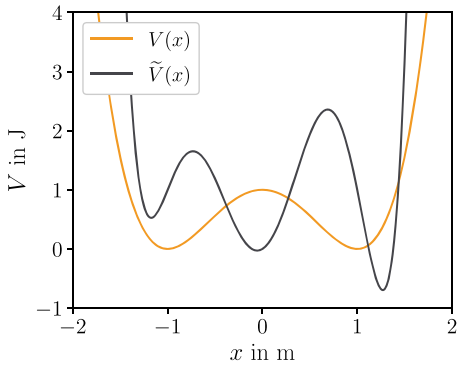


FIG. 1. Simulation potential $V(x)$ (orange) and target potential $\widetilde{V}(x)$ (black).

[Eq. (11)]. More specifically, in Eq. (19), set $m\ddot{x} = 0$, and rearranging yields Eq. (11). However, even though the equation of motion for Langevin dynamics converges to the equation of motion for overdamped Langevin dynamics, the ISP scheme [Eqs. (20) and (21)] does not converge to the Euler–Maruyama scheme [Eq. (12)] in the limit of high friction. By "high friction," we denote the range of collision rates ξ for which $e^{-\xi\Delta t} \approx 0$ in Eq. (20), but $\left|\frac{\nabla V}{\xi m}\right| > 0$. (As reference, $e^{-0.1} = 0.904$, $e^{-1} = 0.368$, and $e^{-5} = 0.007$.) If $e^{-\xi\Delta t} \approx 0$, then also $e^{-2\xi\Delta t} \approx 0$, and Eq. (20) becomes

$$x_{k+1} \approx x_k - \frac{\nabla V(x_k)}{\xi m}\Delta t + \sqrt{\frac{k_B T}{m}}\, \eta_{L,k}\, \Delta t. \tag{28}$$

The first two terms on the right-hand side are identical to the Euler–Maruyama scheme [Eq. (12)], but the random number term differs from the Euler–Maruyama scheme. Thus, even in the limit of high friction, the two algorithms yield different paths for a given random number sequence η. The difference between a Langevin path ω_L and an overdamped Langevin path ω_o can be scaled by the combined parameter $\xi\Delta t$. For some value $\xi\Delta t > 1$, the difference between the two paths becomes minimal before increasing again, but for no value of $\xi\Delta t$, the two paths fully coincide.

When Langevin integration schemes are used as a thermostat in MD simulations, the optimal friction coefficient should reproduce the expected temperature fluctuations and therefore depends on the system and the simulation box.[58] Reported collision rates[49,50,59] (while keeping the time step at $\Delta t = 0.002$ ps) range from 0.1 ps^{-1} to ~100 ps^{-1}, corresponding to $\xi\Delta t = 0.0002$ to $\xi\Delta t = 0.2$. However, even for a large collision rate of 100 ps^{-1}, $e^{-\xi\Delta t} = e^{-0.2} = 0.819 \not\approx 0$. For these two reasons—MD simulations are not conducted in the high-friction regime, and even in the high-friction regime, ω_o differs from ω_L—a simulation with the ISP scheme yields a materially different path ensemble than a simulation with the Euler–Maruyama scheme.

C. From a path ω to random numbers η_o and η_L

In Sec. V B, we showed that given a random number sequence η, the path generated by the Euler–Maruyama integration scheme

Published under license by AIP Publishing

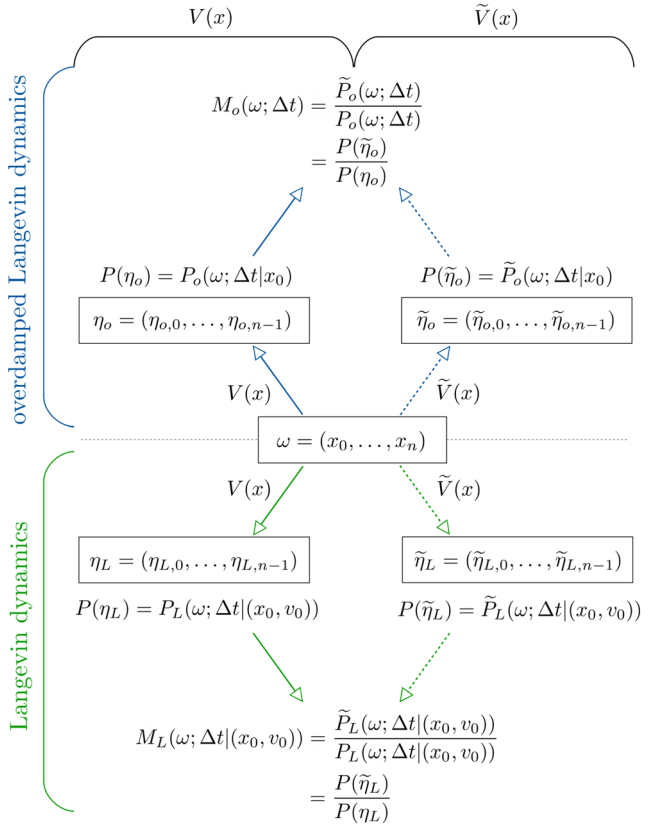


FIG. 2. Overview of path probabilities and path probability ratios for a sample path $\omega = (x_0, \ldots x_n)$.

for overdamped Langevin dynamics differs from the path generated by the ISP integration scheme for Langevin dynamics. More relevant for path reweighting is the reverse situation: Given a sample path $\omega = (x_0, \ldots, x_n)$ in position space and the parameters of the dynamics (m, V, T, ξ, k_B, and Δt), how does the random number sequence η_o needed to generate ω with the Euler–Maruyama scheme [Eq. (12)] differ from the random number sequence η_L needed to generate the same ω with the ISP scheme [Eqs. (20) and (21)]? An equivalent question is as follows: How does the path probability that ω has been generated by the Euler–Maruyama scheme differ from the path probability that ω has been generated by the ISP scheme? How does this difference affect the path probability ratios between the simulation and a target potential? Figure 2 gives an overview of the quantities we will compare. Note that we dropped the index o or L from the path ω because ω is a given dataset, which will be analyzed using various approaches to calculate the path probabilities.

First, we need to discuss whether such a comparison between the ISP scheme and the Euler–Maruyama scheme is even possible. From an algorithmic view point, this is clearly possible because both integrators [Eqs. (12) and (20)] use a single random number per integration time step. The path probabilities are then equal to the probability of the different random number sequences η_L and η_o needed to generate ω. From a physical view point, the answer is not as clear because overdamped Langevin dynamics evolves in position space (x_k), whereas Langevin dynamics evolves in phase space (x_k, v_k). The velocity v_k enters the integration scheme [Eq. (20)] and the path probability [Eq. (22)]. However, v_k is fully determined by the current position x_k and the previous position x_{k-1} [Eq. (21)]. Thus, if the initial velocity v_0 is known, the position trajectory is enough to evaluate the path probability [Eq. (22)], and the comparison to overdamped Langevin dynamics is possible.

We consider the test system described in Sec. V A at the simulation potential $V(x)$ (double-well potential) simulated by the ISP scheme for Langevin dynamics. With $\xi = 50\ \mathrm{s}^{-1}$ and $\Delta t = 0.01$ s, we have $e^{-\xi\Delta t} = e^{-0.5} = 0.607 \not\approx 0$, meaning that the system is not in the high-friction limit. Figure 3(a) additionally shows that with these parameters $\mathcal{O}(\xi m\dot{x}) \approx \mathcal{O}(m\ddot{x})$ and also according to the criterion for the stochastic differential equation, the system is not in the high-friction limit.

Figure 3(b) shows a sample path $\omega = (x_0, x_1, \ldots, x_{10})$. Figure 3(c) shows the random numbers η_o needed to generate ω with the Euler–Maruyama scheme [blue solid line, calculated using Eq. (15)] and the random numbers η_L needed to generate ω with the ISP scheme [green solid line, calculated using Eq. (24)]. As expected for the low-friction regime, these two random number sequences differ markedly.

Consequently, the path probabilities differ. Figure 3(d) shows the unnormalized path probability for generating ω with the Euler–Maruyama scheme (blue solid line),

$$P_o(\omega;\Delta t|x_0) \sim \exp\left(-\frac{\xi m}{4k_BT\Delta t}\sum_{k=0}^{n-1}\left(x_{k+1} - x_k + \frac{\Delta t}{\xi m}\nabla V(x_k)\right)^2\right), \tag{29}$$

and for generating ω with the ISP scheme (green solid line),

$$P_L(\omega;\Delta t|(x_0,v_0)) \sim \exp\left(-\sum_{k=0}^{n-1}\frac{m\left(x_{k+1} - x_k - \exp(-\xi\Delta t)v_k\Delta t + (1-\exp(-\xi\Delta t))\frac{\nabla V(x_k)}{\xi m}\Delta t\right)^2}{2k_BT(1-\exp(-2\xi\Delta t))\Delta t^2}\right), \tag{30}$$

where we omitted those factors from Eqs. (13) and (22) that cancel in the path probability ratio. We checked that the path probabilities are consistent with $P(\eta_o)$ and $P(\eta_L)$. The two path probabilities diverge from the first simulation step on. After ten integration time steps, they differ by two orders of magnitude. Clearly, $P_L(\omega;\Delta t|(x_0, v_0))$ cannot be used as an approximation for $P_o(\omega;\Delta t|x_0)$.

However, an interesting observation arises when we consider reweighting ω to the target potential $\widetilde{V}(x)$ (triple-well potential).

Published under license by AIP Publishing

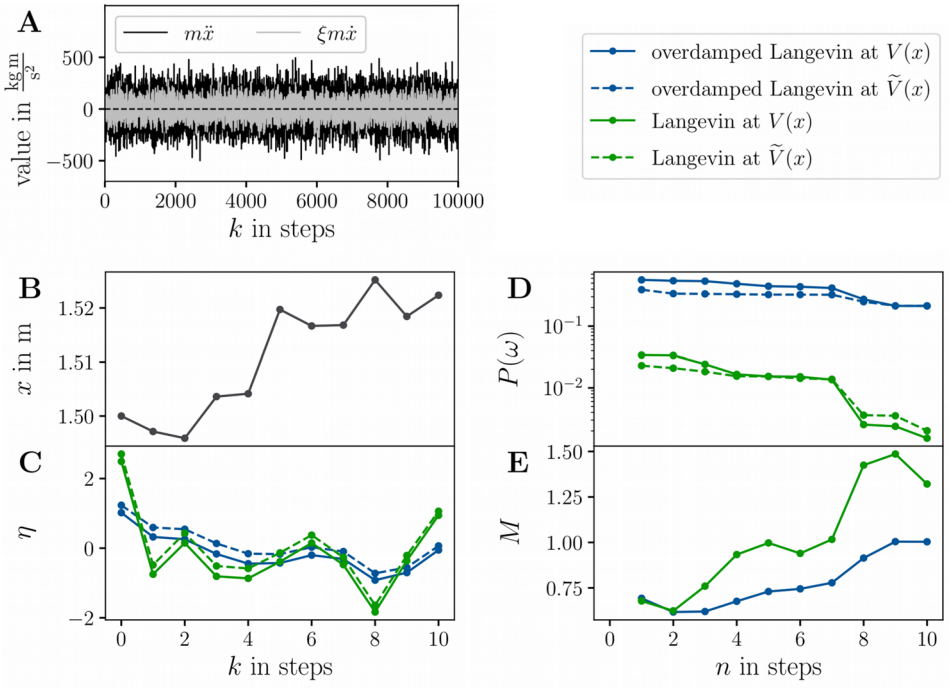


FIG. 3. (a) The acceleration term $m\ddot{x}$ and the friction $\xi m\dot{x}$ for the test system at $V(x)$. (b) Example path ω of length $n = 10$. (c) Random number sequences η_L (green solid), η_o (blue solid), $\widetilde{\eta}_L$ (green dashed), and $\widetilde{\eta}_o$ (blue dashed) that correspond to ω. (d) Path probabilities $P_L(\omega; \Delta t|(x_0, v_0))$ (green solid), $P(\omega; \Delta t|x_0)$ (blue solid), $\widetilde{P}_L(\omega; \Delta t|(x_0, v_0))$ (green dashed), and $\widetilde{P}_o(\omega; \Delta t|x_0)$ (blue dashed). (e) Path probability ratios: $M_L(\omega, \Delta t|(x_0, v_0))$ (green) and $M_o(\omega; \Delta t|x_0)$ (blue).

Figure 3(c) shows the random numbers $\widetilde{\eta}_o$ needed to generate ω with the Euler–Maruyama scheme at $\widetilde{V}(x)$ [blue dashed line, calculated using Eq. (16)] and the random numbers $\widetilde{\eta}_L$ needed to generate ω with the ISP scheme at $\widetilde{V}(x)$ [green dashed line, calculated using Eq. (25)]. The corresponding unnormalized path probabilities $\sim \widetilde{P}_o(\omega; \Delta t|x_0)$ and $\sim \widetilde{P}_L(\omega; \Delta t|(x_0, v_0))$ are shown as dashed lines in Fig. 3(d). Strikingly, a change in the integration scheme from the Euler–Maruyama scheme to ISP has a much stronger influence on the random numbers and the path probability than the modification of the potential energy function. Figure 3(e) shows the path probability ratios, i.e., the ratio between the dashed and the solid lines in Fig. 3(d), for the Euler–Maruyama scheme $M_o = M_o(\omega; \Delta t|x_0) = M_o(\omega, \eta_o; \Delta t|x_0)$ (blue line) and the ISP scheme $M_L = M_L(\omega; \Delta t|(x_0, v_0)) = M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$ (green line). Because, within an integration scheme, the path probability does not change drastically when going from the simulation potential $V(x)$ to the target potential $\widetilde{V}(x)$, both path probability ratios remain at ≈1 throughout the path and follow similar curves, that is, the path probability ratios for Langevin and overdamped Langevin dynamics are much more similar than the underlying path probabilities.

D. Path reweighting

We return to the scenario described in the Introduction and ask the following: are the two path probability ratios similar enough that we can use M_o as an approximation to M_L in Eq. (1)? Figure 4(a) compares different ways to calculate the path probability $\widetilde{P}_L(\omega; \Delta t|(x_0, v_0))$, i.e., the probability with which an example path ω would have been generated at the target potential $\widetilde{V}(x)$. The black line is the reference solution calculated by inserting $\widetilde{V}(x)$ into Eq. (22). It is identical to the green dashed line in Fig. 3(d). The green line in Fig. 4(a) shows the reweighted path probability, where we used the exact path probability ratio for the ISP scheme, $M_L(\omega; \Delta t|(x_0, v_0))$ [Eq. (23)], in Eq. (1). As expected, this reweighted path probability coincides with the directly calculated path probability. The blue line shows the reweighted path probability, where we used the path probability ratio for the Euler–Maruyama scheme, $M_o(\omega; \Delta t|x_0)$ [Eq. (14)], as an approximation to M_L in Eq. (1). The path probability deviates from the reference solution, but overall follows a similar curve.

Figure 4(a) merely serves to illustrate the concepts. With only ten steps, the example path ω is far too short to judge the accuracy of the two path probability ratios for reweighting dynamic properties. We, therefore, constructed MSMs for the target potential $\widetilde{V}(x)$. The reference solution has been generated from simulations at the target potential $\widetilde{V}(x)$ using the ISP scheme. The dominant MSM eigenfunctions and associated implied timescales are shown as black lines in Figs. 4(b) and 4(c). Next, we ran simulations at the simulation potential $V(x)$ using the ISP scheme and constructed a reweighted MSM using the exact reweighting factor $M_L(\omega; \Delta t|(x_0, v_0))$ [Eq. (23)]. The dominant MSM eigenfunctions are shown as green lines in Fig. 4(b). They exactly match the reference solution. The reweighted implied timescales are shown as green lines in Fig. 4(c) and are in good agreement with the reference solution. Finally, we used the simulation at $V(x)$ to construct a reweighted MSM using the reweighting factor for the Euler–Maruyama scheme $M_o(\omega; \Delta t|x_0)$ [Eq. (14)]. The dominant MSM eigenfunctions are shown as blue lines in Fig. 4(b). The eigenfunctions differ considerably from the reference solution. Most notably,

Published under license by AIP Publishing

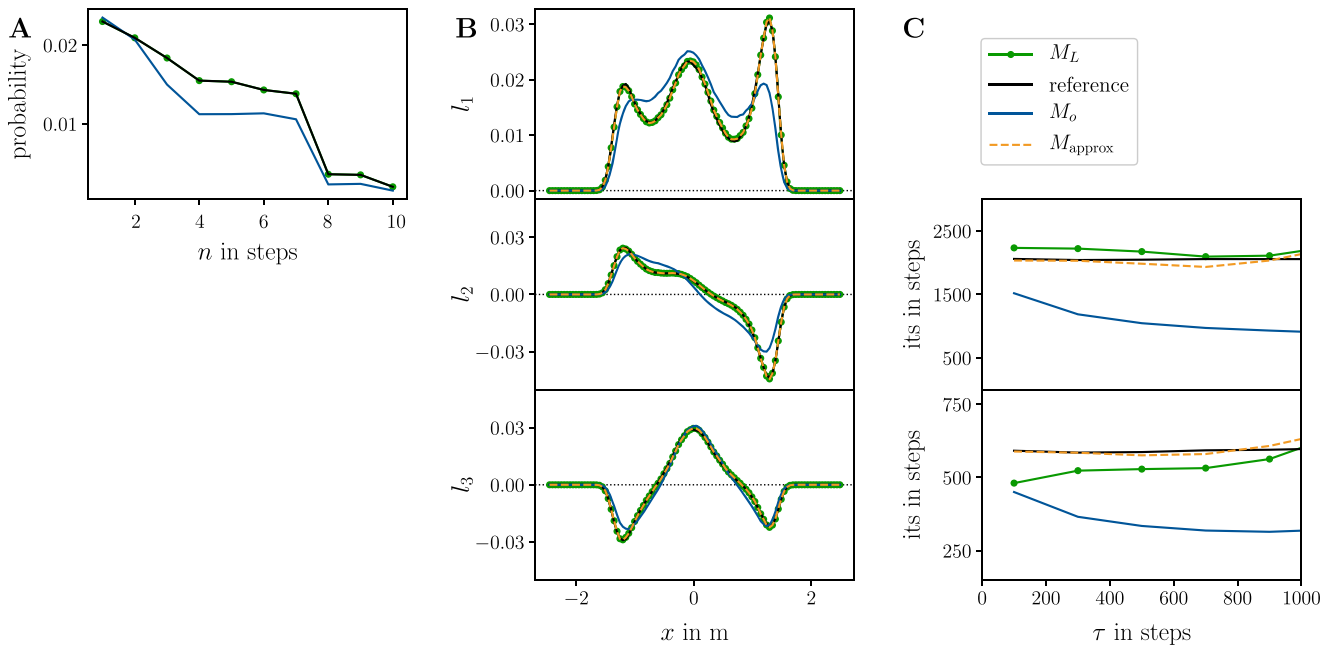


FIG. 4. (a) Reference and reweighted path probabilities for ω for Langevin dynamics. (b) Reference and reweighted first three dominant MSM left eigenfunctions l_1, l_2, and l_3 associated with $\widetilde{V}(x)$ for Langevin dynamics. (c) Reference and reweighted implied timescales corresponding to l_2 and l_3.

the stationary distribution is not reproduced correctly [blue line in the upper panel of Fig. 4(b)]. The left peak is reduced to a shoulder of the central peak, and the relative heights of central peak and the right peak do not match those of the reference solution. Likewise, the implied timescales [blue line in Fig. 4(c)] are severely underestimated. This indicates that using the path probability ratio for overdamped Langevin dynamics, $M_o(\omega;\, \Delta t|x_0)$, to reweight Langevin trajectories does not yield acceptable results.

VI. APPROXIMATE PATH PROBABILITY RATIO

A. Derivation and numerical results

With the results from Sec. IV, the exact random number probability ratio $M_L(\omega,\, \eta_L;\, \Delta t|(x_0,\, v_0))$ [Eq. (7)] for the ISP scheme is straightforward to evaluate from a simulation at $V(x)$: the random number sequence $\eta = \eta_L$ can be recorded during the simulation, and the random number difference $\Delta\eta = \Delta\eta_L$ is given by Eq. (26). Inserting η_L and $\Delta\eta_L$ into Eq. (7) yields $M_L(\omega,\, \eta_L;\, \Delta t|(x_0,\, v_0))$. However, $\Delta\eta_{L,k}$ in Eq. (26) is specific to the ISP scheme. If one uses a different Langevin integration scheme to simulate the dynamics at $V(x)$, one needs to adapt Eq. (26) via the strategy outlined in Sec. IV.

Fortunately, the random number difference for overdamped Langevin dynamics $\Delta\eta_{o,k}$ [Eq. (17)] is approximately equal to $\Delta\eta_{L,k}$ for any given perturbation $U(x)$. Figure 3(c) already suggests that. In Appendix D, we show that the difference between $\Delta\eta^2_{L,k}$ and $\Delta\eta^2_{o,k}$ is, in fact, only of $\mathcal{O}(\xi^4\Delta t^4)$ so that for $\xi\Delta t < 1$, we can assume with high accuracy that

$$\Delta\eta_{L,k} \approx \Delta\eta_{o,k},$$
$$\sqrt{\frac{1}{k_BT\xi^2 m}}\,\frac{1-\exp(-\xi\Delta t)}{\sqrt{1-\exp(-2\xi\Delta t)}}\cdot\nabla U(x_k) \approx \sqrt{\frac{\Delta t}{2k_BT\xi m}}\cdot\nabla U(x_k). \tag{31}$$

The difference between $\Delta\eta_{L,k}$ and $\Delta\eta_{o,k}$ is determined by the prefactors in front of $\nabla U(x_k)$ in Eq. (31), which are shown as a function of $\xi\Delta t$ in Fig. 5(b). For $\xi\Delta t < 1$, the two curves are virtually identical.

With the approximation in Eq. (31), we can derive an approximate random number probability ratio, by using the recorded η_L, but substituting $\Delta\eta_{L,k}$ [Eq. (26)] by $\Delta\eta_{o,k}$ [Eq. (17)] in Eq. (7), we obtain

$$\begin{aligned} M_L(\omega,\eta_L;\Delta t|(x_0,v_0)) &\approx M_{\text{approx}}(\omega,\eta_L;\Delta t|x_0) \\ &= \exp\left(-\sum_{k=0}^{n-1}\sqrt{\frac{\Delta t}{2k_BT\xi m}}\nabla U(x_k)\cdot\eta_{L,k}\right) \\ &\quad\cdot\exp\left(-\frac{1}{2}\sum_{k=0}^{n-1}\frac{\Delta t}{2k_BT\xi m}(\nabla U(x_k))^2\right). \end{aligned} \tag{32}$$

Equation (32) has the same functional form as the random number probability ratio for the Euler–Maruyama scheme $M_o(\omega,\, \eta_o;\, \Delta t|x_0)$ [Eq. (18)], but it uses η_L, the random numbers generated during the ISP simulation, instead of η_o. Equation (32) is the approximation

Published under license by AIP Publishing

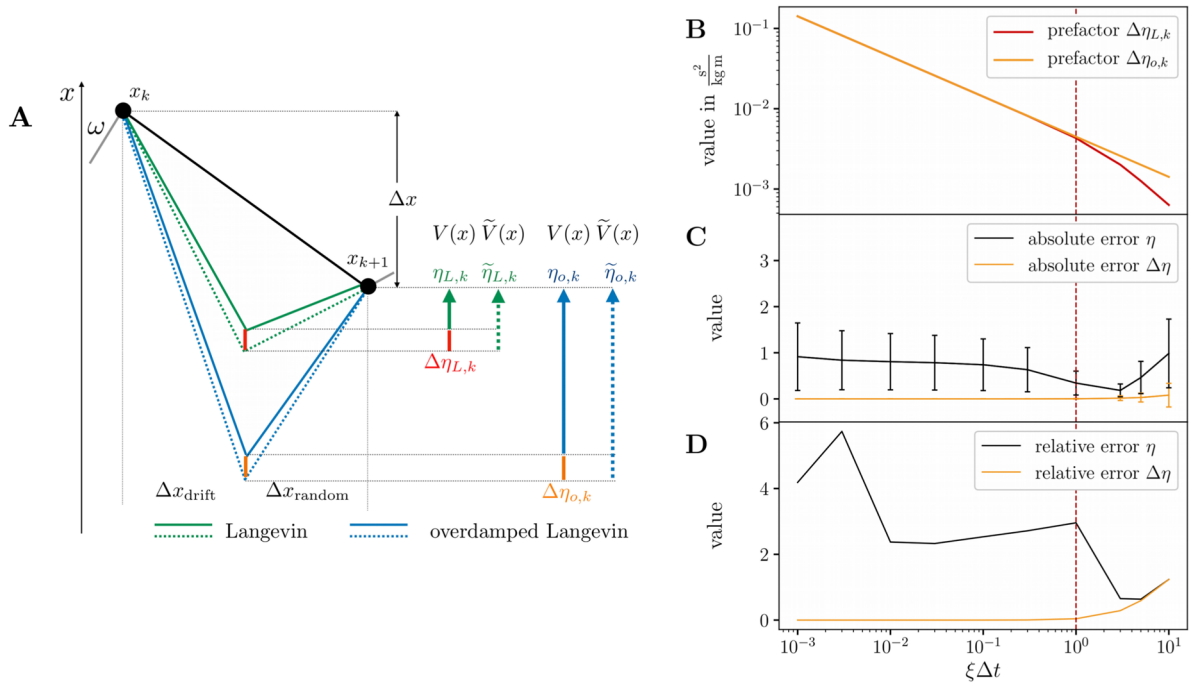


FIG. 5. (a) Sketch of a step $x_k \to x_{k+1}$ and the quantities of influence for Langevin and overdamped Langevin dynamics. (b) Prefactors of $\Delta\eta_{L,k}$ and $\Delta\eta_{o,k}$ as a function of $\xi\Delta t$. (c) Absolute difference (absolute error) between the random numbers $\langle|\eta_{o,k} - \eta_{L,k}|\rangle$ and the random number differences $\langle|\Delta\eta_{o,k} - \Delta\eta_{L,k}|\rangle$ as a function of $\xi\Delta t$. (d) Relative difference (relative error) between the random numbers $\langle|(\eta_{o,k} - \eta_{L,k})/\eta_{L,k}|\rangle$ and the random number differences $\langle|(\Delta\eta_{o,k} - \Delta\eta_{L,k})/\Delta\eta_{L,k}|\rangle$ as a function of $\xi\Delta t$.

that we used in Refs. 34 and 35 because we had not yet derived $M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$ [Eqs. (23) and (27)].

Figure 4 demonstrates the accuracy of the approximate random number probability ratio $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ [Eq. (32)] for our test system. The orange dashed line in Fig. 4(a) shows the reweighted path probability for the short example path, where we used $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ [Eq. (32)], in Eq. (1). It exactly matches the reference solution (black line).

Next, we constructed a reweighted MSM for the target potential $\widetilde{V}(x)$ based on our simulations at the simulation potential $V(x)$ using $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ [Eq. (32)] to reweight the transition counts. The dominant MSM eigenfunctions of the reweighted MSM are shown as orange dashed lines in Fig. 4(b). They exactly match the reference solution. The reweighted implied timescales are shown as orange dashed lines in Fig. 4(c) and seem to match the reference solution even better than the ones calculated using the exact path probability ratio [green line in Fig. 4(c)]. However, the difference between the orange dashed line and the green line is likely within statistical uncertainty. In summary, $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ is a highly accurate approximation to $M_L(\omega, \eta_L; \Delta t|x_0)$ for $\xi\Delta t < 1$. Using $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ instead of $M_L(\omega, \eta_L; \Delta t|x_0)$ could even have the following advantages: (*i*) the implementation is less error-prone because the functional form of M_{approx} is simpler than that of M_L and (*ii*) M_{approx} might be numerically more stable because the calculation of exponential function on the left-hand side of Eq. (31) is avoided.

B. Intuition

We discuss why $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ is a better approximation to $M_L(\omega, \eta_L; \Delta t|x_0)$ than $M_o(\omega; \Delta t|x_0) = M_o(\omega, \eta_o; \Delta t|x_0)$. Figure 5(a) shows one integration time step of a stochastic integration scheme from x_k to x_{k+1} (black line). From k to $k+1$, the system has progressed by $\Delta x = x_{k+1} - x_k$. In the ISP scheme, this progress is composed of a progress

$$\Delta x_{\mathrm{drift},L} = \exp(-\xi\,\Delta t)\, v_k \Delta t - [1 - \exp(-\xi\,\Delta t)]\,\frac{\nabla V(x_k)}{\xi m}\Delta t \qquad (33)$$

due to the drift force and the velocity of the system [second and third terms on the right-hand side of Eq. (20)] and a progress

$$\Delta x_{\mathrm{random},L} = \sqrt{\frac{k_B T}{m}\,[1 - \exp(-2\xi\,\Delta t)]}\,\eta_{L,k}\,\Delta t \qquad (34)$$

due to the random force [fourth term on the right-hand side of Eq. (20)] such that $\Delta x = \Delta x_{\mathrm{drift},L} + \Delta x_{\mathrm{random},L}$. $\Delta x_{\mathrm{drift},L}$ and $\Delta x_{\mathrm{random},L}$ are illustrated as green solid lines in Fig. 5(a). The probability of generating the step $x_k \to x_{k+1}$ is determined by $\Delta x_{\mathrm{random},L}$, which is proportional to the random number $\eta_{L,k}$ (green solid arrow).

With a different potential energy function $\widetilde{V}(x)$ at x_k, the displacement due to the drift force differs from the original $\Delta x_{\mathrm{drift},L}$. To achieve the same overall displacement Δx, $\Delta x_{\mathrm{random},L}$ needs to

Published under license by AIP Publishing

be adjusted (green dotted line). The corresponding random number $\widetilde{\eta}_{L,k}$ is shown as a green dotted arrow, and the difference between the two random numbers $\Delta\eta_{L,k}$ is shown as a red line. In path reweighting, one constructs $\widetilde{\eta}_{L,k}$ by adding $\Delta\eta_{L,k}$ to $\eta_{L,k}$,

$$\widetilde{\eta}_{L,k} = \eta_{L,k} + \Delta\eta_{L,k} \tag{35}$$

[analogous to Eq. (6)], which then yields the general form of the random number probability ratio in Eq. (7).

An analogous analysis applies to the Euler-Maruyama scheme, where the progress due to the drift force is

$$\Delta x_{\mathrm{drift},o} = -\frac{\nabla V(x_k)}{\xi m}\,\Delta t \tag{36}$$

[second term on the right-hand side of Eq. (12)], and the progress due to the random force is

$$\Delta x_{\mathrm{random},o} = \sqrt{\frac{2k_BT}{\xi m}}\,\sqrt{\Delta t}\,\eta_{o,k} \tag{37}$$

[third term on the right-hand side of Eq. (12)]. In Fig. 5(a), $\Delta x_{\mathrm{drift},o}$ and $\Delta x_{\mathrm{random},o}$ are illustrated as blue solid lines, and the random number is represented as a blue solid arrow. With a different potential energy function $\widetilde{V}(x)$ at x_k, the progress due to the drift force differs from the original $\Delta x_{\mathrm{drift},o}$. To achieve the same overall progress Δx, $\Delta x_{\mathrm{random},o}$ needs to be adjusted (blue dotted line). The corresponding random number $\widetilde{\eta}_{o,k}$ is shown as a blue dotted arrow, and the difference between the two random numbers $\Delta\eta_{o,k}$ is shown as an orange line.

In Sec. VI A, we have shown that $\Delta\eta_{L,k} \approx \Delta\eta_{o,k}$ (for $\xi\Delta t < 1$). Thus, approximating $\Delta\eta_{L,k}$ by $\Delta\eta_{o,k}$ in Eq. (35), or, visually, approximating the red line by the orange line in Fig. 5(a), is valid. However, the displacement due to the drift $\Delta x_{\mathrm{drift},o}$ in the Euler–Maruyama scheme can differ strongly from $\Delta x_{\mathrm{drift},L}$ in the ISP scheme, and consequently, the random numbers needed to generate the same overall progress Δx differ

$$\eta_{L,k} \not\approx \eta_{o,k} \tag{38}$$

[blue solid and green solid arrow in Fig. 5(a)]. Consequently, approximating $\eta_{L,k}$ by $\eta_{o,k}$ in Eq. (35), or visually approximating the green solid arrow by the blue solid arrow in Fig. 5(a), is not valid.

The exact random number probability ratio $M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$ [Eq. (27)] uses the exact η_L recorded during the simulation and the exact $\Delta\eta_L$ [Eq. (26)]. It therefore yields results that exactly match the reference solutions (green lines in Fig. 4). $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ uses the exact η_L recorded during the simulation but approximates $\Delta\eta_{L,k}$ by $\Delta\eta_{o,k}$. This introduces only a small error but still yields excellent reweighting results in our test system (orange dashed lines in Fig. 4). However, in $M_o(\omega; \Delta t|x_0) = M_o(\omega, \eta_o; \Delta t|x_0)$, one additionally approximates η_L by η_o. The difference between η_L and η_o is much larger than the difference between $\Delta\eta_L$ and $\Delta\eta_o$, and this additional approximation leads to the distorted reweighting results we observed as the blue lines in Fig. 4.

The proportions in Fig. 5(a) are not exaggerated. The black line in Fig. 5(c) shows the average absolute difference between the random numbers $\langle|\eta_{o,k} - \eta_{L,k}|\rangle$ as a function of $\xi\Delta t$. Visually, this is the difference between the green solid arrow and the blue solid arrow in Fig. 5(a). The orange line in Fig. 5(c) shows the average absolute difference between the random number differences $\langle|\Delta\eta_{o,k} - \Delta\eta_{L,k}|\rangle$, i.e., the difference between the orange and the red line in Fig. 5(a). The graph has been calculated by averaging over a path with 10^6 time steps. The standard deviations are shown as vertical bars. $\langle|\Delta\eta_{o,k} - \Delta\eta_{L,k}|\rangle$ is close to 0 for all values of $\xi\Delta t$, whereas there is a substantial difference between η_L and η_o. $\langle|\eta_{o,k} - \eta_{L,k}|\rangle$ has a minimum at $\xi\Delta t \approx 2$ because the difference between the Euler–Maruyama scheme and the ISP scheme is minimal for $\xi\Delta t \approx 2$ (see Sec. V B). Figure 5(d) shows the corresponding average relative errors. For $\xi\Delta t > 1$, $\langle|(\eta_o - \eta_L)/\eta_L|\rangle$ (black line) decreases in accordance with the decrease in the absolute difference $\langle|(\eta_o - \eta_L)|\rangle$ and $\langle|(\Delta\eta_o - \Delta\eta_L)/\Delta\eta_L|\rangle$ (orange line) increases, reflecting the fact that the approximation [Eq. (31)] does not hold for $\xi\Delta t > 1$. However, for $\xi\Delta t < 1$, the region in which MD simulations are conducted, the relative error for the random numbers is much larger than the relative error for the random number difference. This reinforces that the random numbers $\eta_{L,k}$ should not be approximated in the path probability ratio but, instead, should be recorded from the simulation at $V(x)$. By contrast, the random number difference $\Delta\eta_{L,k}$ can reliably be approximated by Eq. (31).

VII. MOLECULAR EXAMPLE: BUTANE

The slowest degree of freedom in butane is the torsion around the C_2–C_3 bond, which exhibits three metastable states: the trans-conformation at $\phi = \pi$ and the two gauche-conformations at $\phi = \pm\frac{1}{3}\pi$. Consequently, butane has three dominant MSM eigenvectors, where l_1 corresponds to the stationary density and l_2 and l_3 represent slow transitions along ϕ [Fig. 6(a)]. Because the two gauche-conformations are equally populated, l_2 and l_3 are degenerate [Fig. 6(b)]. We simulated butane in implicit water at three different temperatures, $T = 300$ K, $T = 200$ K, and $T = 150$ K, using direct and biased simulations. As we lower the temperature, we expect that the relative population of the trans-conformation increases, but that otherwise, the overall shape and sign-structure of the eigenvectors remain unchanged.

At $T = 300$ K and $T = 200$ K, the reweighting results using $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ [Eq. (32), orange dashed line] or $M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$ [Eq. (23), green solid line] match the MSM obtained by direct simulation. In particular, the eigenvectors are reproduced with very high precision. By contrast, the reweighted results using $M_o(\omega; \Delta t|x_0)$ [Eq. (14), blue line] deviate considerably from the reference MSMs obtained by direct simulations. The stationary distribution l_1 is not reproduced correctly, which then leads to further errors in the dominant eigenvectors l_2 and l_3. The associated implied timescales are underestimated. Moreover, for $T = 200$ K and $T = 150$ K, the use of $M_o(\omega; \Delta t|x_0)$ yielded numerically instable transition matrices for lag times of $\tau > 100$ ps. This demonstrates that path reweighting with an appropriate path probability ratio, such as $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ or $M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$, yields accurate results. However, $M_o(\omega; \Delta t|x_0)$ should not be used as an approximation for the exact path probability ratio $M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$.

Note that reweighting results using the approximate probability ratio $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ are virtually indistinguishable from the results using the exact probability ratio $M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$ for all three temperatures. This confirms our analysis that $M_{\mathrm{approx}}(\omega, \eta_L; \Delta t|x_0)$ can be used as highly accurate approximation to $M_L(\omega, \eta_L; \Delta t|(x_0, v_0))$.

Published under license by AIP Publishing

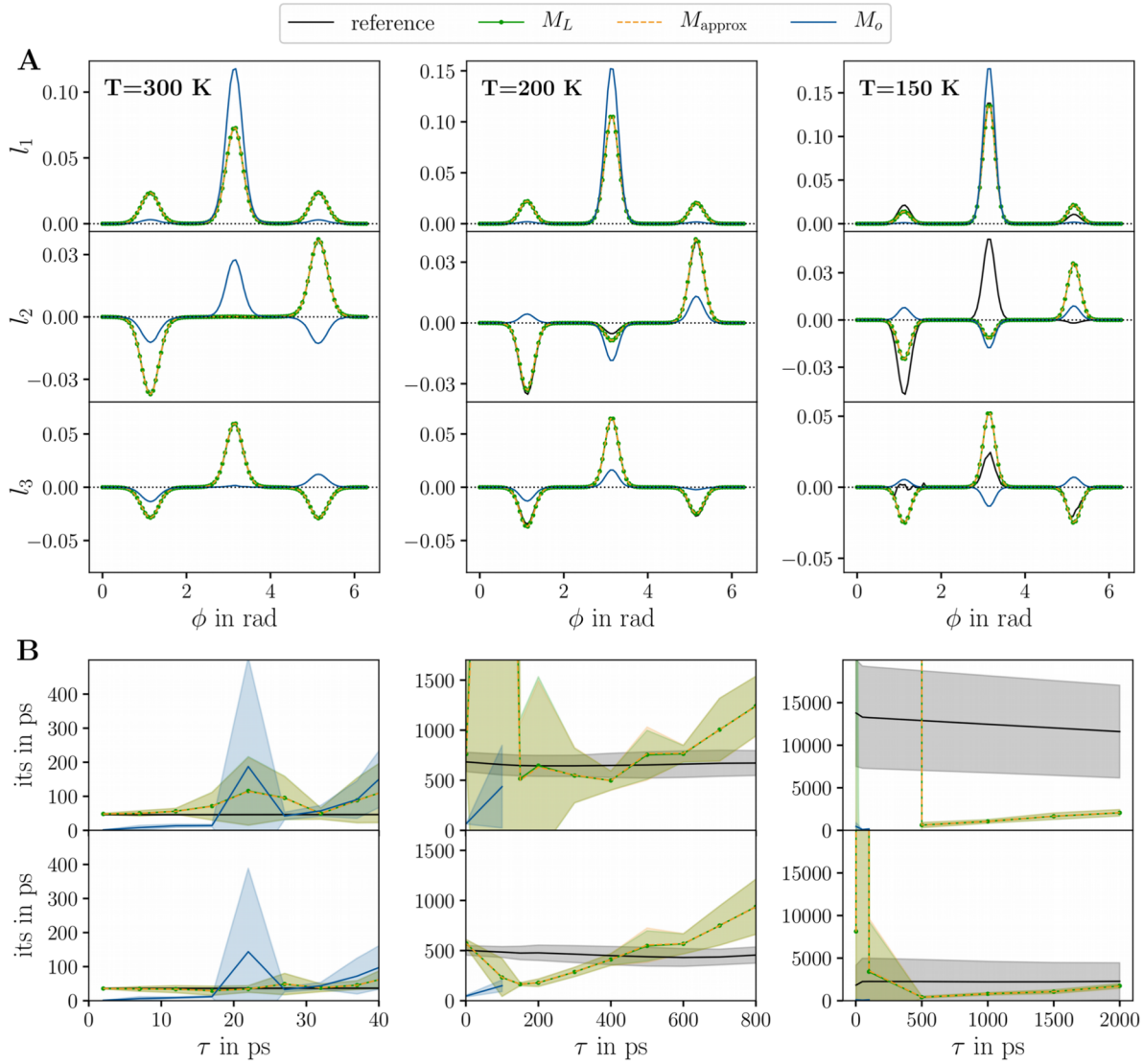


FIG. 6. Dynamics of the torsion angle in butane at T = 300 K, 200 K, and 150 K. (a) Dominant left eigenfunctions l_1, l_2, and l_3 of the MSM along the torsion angle ϕ, obtained by evaluating direct simulations at the target potential, as well as by reweighting biased simulations. (b) Implied timescales corresponding to l_2 and l_3 in (a). Solid lines: mean and shaded area: standard deviation. Standard deviations for the eigenvectors are too small to be shown.

The variation of the temperature from 300 K to 200 K and 150 K illustrates under which circumstances path reweighting is an efficient method. At T = 300 K, many transitions across the torsion angle barriers are observed in the direct simulation. Path reweighting and direct simulation yield identical results. However, path reweighting has a larger statistical uncertainty. At T = 200 K, fewer transitions are observed in the direct simulations, which results in an increased statistical uncertainty in the direct MSM. Finally, at T = 150 K, the transitions in the direct simulation are insufficient to correctly sample the stationary density. The MSM of the direct simulation predicts a higher population for the gauche-conformation at $\phi = +\frac{1}{3}\pi$ than for the gauche-conformation at $\phi = -\frac{1}{3}\pi$, which is clearly a sampling error. This error in the stationary density then leads to vastly incorrect estimates for l_2 and l_3. Additionally, the direct MSM predicts that the degeneracy is lifted. By contrast, the results of the reweighted MSM are in line with what we expect: the gauche-conformations are equally populated, the overall shapes of the dominant eigenvectors correspond to those of the eigenvectors at higher temperatures, and l_2 and l_3 are degenerate. In conclusion, path reweighting in combination with enhanced sampling techniques is a promising tool in situations, where the stationary density cannot be sampled accurately by direct simulation.

Published under license by AIP Publishing

VIII. METHODS

A. Simulations of the test system

The test system is a one-dimensional one particle system with mass $m = 1$ kg and $k_BT = 2.494$ J (corresponding to $k_B = 0.008\,314$ J/K and $T = 300$ K). The simulation potential (orange line in Fig. 1) is

$$V(x) = (x^2 - 1)^2, \tag{39}$$

and the target potential (black line in Fig. 1) is

$$\widetilde{V}(x) = 4\left(x^3 - \frac{3}{2}x\right)^2 - x^3 + x. \tag{40}$$

For the results in Figs. 3–5, we simulated the system using the ISP scheme [Eqs. (20) and (21)] with a time step of $\Delta t = 0.01$ s. The initial conditions were $x_0 = 1.50$ m, $v_0 = 0$ m/s. The number of time steps N_t, the collision rate ξ, and the potential energy function used are summarized in Table II.

In Fig. 3(a), we computed the acceleration $\ddot{x} = a$ as $a_{k+1} = \frac{v_{k+1} - v_k}{\Delta t}$. Figure 3(b) displays the first ten steps of the simulation as an example path ω, and all quantities displayed in Figs. 3(c)–3(e) are calculated from this short path.

The absolute and relative differences of the random numbers in Fig. 5 were calculated as

$$\langle|\eta_{o,k} - \eta_{L,k}|\rangle = \frac{1}{N_t - 1}\sum_{k=0}^{N_t-1}|\eta_{o,k} - \eta_{L,k}| \tag{41}$$

and

$$\left\langle\left|\frac{\eta_{o,k} - \eta_{L,k}}{\eta_{L,k}}\right|\right\rangle = \frac{1}{N_t - 1}\sum_{k=0}^{N_t-1}\left|\frac{\eta_{o,k} - \eta_{L,k}}{\eta_{L,k}}\right|. \tag{42}$$

Analogous equations were used for $\langle|\Delta\eta_{o,k} - \Delta\eta_{L,k}|\rangle$ and $\langle|(\Delta\eta_{o,k} - \Delta\eta_{L,k})/\Delta\eta_{L,k}|\rangle$. $\eta_{L,k}$ was recorded during the simulation. We used Eq. (26) to calculate $\Delta\eta_{L,k}$, Eq. (15) to calculate $\eta_{o,k}$, and Eq. (17) to calculate $\Delta\eta_{o,k}$.

The reference MSM in Figs. 4(b) and 4(c) has been constructed from the simulation at the target potential $\widetilde{V}(x)$. The state space has been discretized using a regular grid of 100 microstates $(S_1, \ldots, S_{100})$ in the range $-1.7 \le x \le 1.6$. Transition counts between microstates were calculated as

$$c_{ij}(\tau) = \frac{1}{N_t - \tau}\sum_{k=0}^{N_t-\tau}\chi_i(x_k)\chi_j(x_{k+\tau}), \tag{43}$$

TABLE II. Simulation parameters.

Figures	N_t	ξ	Potential
3(a)	10^5	50 s^{-1}	$V(x)$
4(b) and 4(c)	10^7	50 s^{-1}	$V(x)$
4(b) and 4(c)	10^7	50 s^{-1}	$\widetilde{V}(x)$
5(c) and 5(d)	10^7	0.1 s^{-1}–1000 s^{-1}	$V(x)$

with

$$\chi_i(x) = \begin{cases} 1 & \text{if } x \in S_i \\ 0 & \text{else,} \end{cases} \tag{44}$$

where x_k is the trajectory and lag time $\tau = 200$ steps. The resulting count matrix $\mathbf{C}(\tau)$ was symmetrized as $\mathbf{C}(\tau) + \mathbf{C}^\top(\tau)$ to enforce detailed balance and row-normalized to obtain the MSM transition matrix $\mathbf{T}(\tau)$. The dominant MSM eigenvectors l_i and associated eigenvalues $\lambda_i(\tau)$ were calculated from $\mathbf{T}(\tau)$ using a standard eigenvalue solver, and the implied timescales were calculated as $t_i = -\tau/\ln(\lambda_i(\tau))$.

The reweighted MSMs in Figs. 4(b) and 4(c) have been constructed from the simulation at the simulation potential $V(x)$ using the same grid and lag time as for the reference MSM. Transition counts between microstates were counted and reweighted as[34,35]

$$\widetilde{c}_{ij}(\tau) = \frac{1}{N_t - \tau}\sum_{k=0}^{N_t-\tau} W((x_k, x_{k+1}, \ldots, x_{k+\tau}); \Delta t|(x_k, v_k))\,\chi_i(x_k)\chi_j(x_{k+\tau}). \tag{45}$$

The weight W is defined as

$$W((x_k, x_{k+1}, \ldots, x_{k+\tau}); \Delta t|(x_k, v_k)) = g(x_k)\cdot M((x_k, x_{k+1}, \ldots, x_{k+\tau}); \Delta t|(x_k, v_k)), \tag{46}$$

with M being the path probability ratio [Eq. (3)] and g being

$$g(x_k) = \exp\left(-\frac{U(x_k)}{k_BT}\right), \tag{47}$$

where the perturbation U is defined in Eq. (2). The remaining procedure was analogous to the reference MSM.

B. Butane—Direct simulations

We performed all-atom MD simulations of *n*-butane in implicit water using the OpenMM 7.4.1[54] simulation package. The GAFF (Generalized Amber Force Field)[60] was used to model butane, and the GBSA (Generalized Born Surface Area) model[61] was used to model implicit water. Interactions beyond 1 nm were truncated. The trajectory was propagated according to the ISP integration scheme for a $3N$-dimensional system,

$$x^i_{k+1} = x^i_k + \exp(-\xi\,\Delta t)\,v^i_k\Delta t - [1 - \exp(-\xi\,\Delta t)]\,\frac{\nabla_i V(\mathbf{x}_k)}{\xi m_i}\Delta t + \sqrt{\frac{k_BT}{m_i}[1 - \exp(-2\xi\,\Delta t)]}\,\eta^i_{L,k}\,\Delta t, \tag{48}$$

$$v^i_{k+1} = \frac{x^i_{k+1} - x^i_k}{\Delta t}, \tag{49}$$

with $i = 1, 2, \ldots, 3N$ and N being the number of atoms. x^i_k, v^i_k, and η^i_k are the position, velocity, and random number along dimension i at iteration step k, m_i is the mass of dimension i, and $\nabla_i V(\mathbf{x}_k)$ denotes the gradient of $V(\mathbf{x}_k)$ along dimension i measured at the position $\mathbf{x}_k$, with $\mathbf{x} \in \mathbb{R}^{3N}$. We implemented the ISP integration scheme using

Published under license by AIP Publishing

the simtk.openmm.openmm.CustomIntegrator[62] class of OpenMM. The collision rate was $\xi = 10\ \mathrm{ps}^{-1}$. The simulation time step was $\Delta t = 0.002$ ps. Positions were written to disk every `txout` = 50 steps = 0.1 ps. We generated three trajectories with 500 ns each at $T = 300$ K, $T = 200$ K, and $T = 150$ K. These direct simulations correspond to simulations at the target potential $\widetilde{V}(\mathbf{x})$.

For the analysis, we cut each trajectory into five pieces of length 100 ns. For each 100-ns-trajectory, we constructed a MSM following the procedure outlined in Sec. VIII A. As state space we chose the C_2–C_3 dihedral angle ϕ, which we discretized using a regular grid of 100 microstates in the range $0 \le \phi \le 2\pi$. This resulted in five MSMs for each temperature. Figure 6 shows the mean and the standard deviation of the first three left MSM eigenvectors (evaluated at lag time $\tau = 1$ ps) and the mean and the standard deviations of the associated implied timescale.

C. Butane—Path reweighting

We biased the simulations along the C_2–C_3 dihedral angle ϕ. To generate the bias potential $U(\phi)$, we constructed a histogram of the free-energy function $\widetilde{F}(\phi)$,

$$\widetilde{F}(\phi) = -k_B T \ln(\widetilde{p}(\phi)), \tag{50}$$

where $\widetilde{p}(\phi)$ is the stationary density along ϕ as measured from the 500 ns direct simulations at $T = 300$ K. Fitting the histogram with a third order Fourier series yielded

$$\begin{aligned}\widetilde{F}_{300\,\mathrm{K}}(\phi) = {} & 8.985 + 3.122\cos(\omega\phi) + 0.959\cos(2\omega\phi) \\ & + 7.742\cos(3\omega\phi) + 0.095\sin(\omega\phi) \\ & + 0.047\sin(2\omega\phi) + 0.002\sin(3\omega\phi),\end{aligned} \tag{51}$$

with $\omega = 0.989$. The same procedure for the simulation at $T = 200$ K yielded

$$\begin{aligned}\widetilde{F}_{200\,\mathrm{K}}(\phi) = {} & 8.311 + 2.847\cos(\omega\phi) + 0.841\cos(2\omega\phi) \\ & + 7.697\cos(3\omega\phi) + 0.046\sin(\omega\phi) \\ & + 0.026\sin(2\omega\phi) + 0.004\sin(3\omega\phi),\end{aligned} \tag{52}$$

with $\omega = 0.989$. $\widetilde{F}_{300\,\mathrm{K}}(\phi)$ and $\widetilde{F}_{200\,\mathrm{K}}(\phi)$ are almost identical. The simulation at $T = 150$ K did not yield a converged stationary density, and thus, no free-energy function was constructed for this temperature, and instead, $\widetilde{F}_{300\,\mathrm{K}}(\phi)$ was used.

The biased simulations were carried out with the potential

$$V_\alpha(\mathbf{x}) = \widetilde{V}(\mathbf{x}) - \alpha \cdot \widetilde{F}(\phi(\mathbf{x})), \tag{53}$$

where $\widetilde{V}(\mathbf{x})$ is the target potential and $\alpha \in [0, 1]$ specifies the bias strength. $V_\alpha(\mathbf{x})$ corresponds to the "simulation potential" within the terminology of this paper; thus,

$$U(\phi(\mathbf{x})) = \alpha \cdot \widetilde{F}(\phi(\mathbf{x})). \tag{54}$$

α was set to 0.1 in all biased simulations, corresponding to "10% of the full metadynamics potential." We carried out biased simulations at three temperatures $T = 300$ K, $T = 200$ K, and $T = 150$ K, with bias potentials $U_{300\,\mathrm{K}}(\phi) = 0.1 \cdot \widetilde{F}_{300\,\mathrm{K}}(\phi)$, $U_{200\,\mathrm{K}}(\phi) = 0.1 \cdot \widetilde{F}_{200\,\mathrm{K}}(\phi)$, and $U_{150\,\mathrm{K}}(\phi) = 0.1 \cdot \widetilde{F}_{300\,\mathrm{K}}(\phi)$. All other simulation parameters were as described in Sec. VIII B.

The path probability ratios for the biased simulations were calculated on-the-fly[34,35] and were written to disk at the same frequency `txout` as the positions. For the approximate path probability ratio M_{approx}, we calculated

$$\begin{aligned}\mathbb{M}_{\mathrm{approx}}(b) = \sum_{i=1}^{3N} \sum_{k=(b-1)\cdot\mathrm{txout}}^{b\cdot\mathrm{txout}-1} \Bigg(& -\sqrt{\frac{\Delta t}{2k_B T \xi m_i}} \nabla_i U(\mathbf{x}_k)\, \eta_{L,k}^i \\ & - \frac{\Delta t}{4k_B T \xi m_i} (\nabla_i U(\mathbf{x}_k))^2 \Bigg)\end{aligned} \tag{55}$$

and constructed the complete path probability ratio as

$$M_{\mathrm{approx}}(\boldsymbol{\omega}, \boldsymbol{\eta}_L; \Delta t | \mathbf{x}_0) = \exp\left(\sum_{b=1}^{A} \mathbb{M}_{\mathrm{approx}}(b) \right) \tag{56}$$

during the construction of the MSM, where $A \in \mathbb{N}$ such that $\tau = A \cdot \mathtt{txout} \cdot \Delta t$.

For the Langevin path probability ratio M_L, we calculated the terms

$$\mathbb{M}_{L,1}(b) = \sum_{i=1}^{3N} \sum_{k=(b-1)\cdot\mathrm{txout}}^{b\cdot\mathrm{txout}-1} (x_{k+1}^i - x_k^i)\left(\nabla_i \widetilde{V}(\mathbf{x}_k) - \nabla_i V(\mathbf{x}_k)\right), \tag{57}$$

$$\mathbb{M}_{L,2}(b) = \sum_{i=1}^{3N} \sum_{k=(b-1)\cdot\mathrm{txout}}^{b\cdot\mathrm{txout}-1} v_k^i \left(\nabla_i \widetilde{V}(\mathbf{x}_k) - \nabla_i V(\mathbf{x}_k)\right), \tag{58}$$

$$\mathbb{M}_{L,3}(b) = \sum_{i=1}^{3N} \sum_{k=(b-1)\cdot\mathrm{txout}}^{b\cdot\mathrm{txout}-1} \frac{\left(\left(\nabla_i \widetilde{V}(\mathbf{x}_k)\right)^2 - \left(\nabla_i V(\mathbf{x}_k)\right)^2\right)}{m_i} \tag{59}$$

and constructed the complete path probability ratio as

$$\begin{aligned}M_L(\boldsymbol{\omega}, \Delta t | \mathbf{x}_0) & \\ = \exp\Bigg[\sum_{b=1}^{A} \Bigg(& -\frac{\mathbb{M}_{L,1}(b)}{k_B T \xi (1 + \exp(-\xi \Delta t)) \Delta t} \\ & + \frac{\mathbb{M}_{L,2}(b)}{k_B T \xi (1 + \exp(\xi \Delta t))} - \frac{\exp(\xi \Delta t) - 1}{\exp(\xi \Delta t) + 1} \frac{\mathbb{M}_{L,3}(b)}{2k_B T \xi^2} \Bigg) \Bigg]\end{aligned} \tag{60}$$

during the construction of the MSM, where $A \in \mathbb{N}$ such that $\tau = A \cdot \mathtt{txout} \cdot \Delta t$.

For the overdamped Langevin path probability ratio M_o, we calculated the terms

$$\mathbb{M}_{o,1}(b) = \sum_{i=1}^{3N} \sum_{k=(b-1)\cdot\mathrm{txout}}^{b\cdot\mathrm{txout}-1} (x_{k+1}^i - x_k^i)\left(\nabla_i \widetilde{V}(\mathbf{x}_k) - \nabla_i V(\mathbf{x}_k)\right), \tag{61}$$

$$\mathbb{M}_{o,2}(b) = \sum_{i=1}^{3N} \sum_{k=(b-1)\cdot\mathrm{txout}}^{b\cdot\mathrm{txout}-1} \frac{\left(\left(\nabla_i \widetilde{V}(\mathbf{x}_k)\right)^2 - \left(\nabla_i V(\mathbf{x}_k)\right)^2\right)}{m_i} \tag{62}$$

and constructed the complete path probability ratio as

$$M_o(\boldsymbol{\omega}, \Delta t | \mathbf{x}_0) = \exp\left[\sum_{b=1}^{A} \left(-\frac{\mathbb{M}_{o,1}(b)}{2k_B T} - \frac{\mathbb{M}_{o,2}(b)\, \Delta t}{4k_B T \xi} \right) \right] \tag{63}$$

Published under license by AIP Publishing

during the construction of the MSM, where $A \in \mathbb{N}$ such that $\tau = A \cdot \texttt{txout} \cdot \Delta t$.

For the analysis, we cut each trajectory into five pieces of length 100 ns. For each 100-ns-trajectory, we constructed a MSM following the procedure outlined in Sec. VIII A. As state space we chose the C_2–C_3 dihedral angle ϕ, which we discretized using a regular grid of 100 microstates in the range $0 \le \phi \le 2\pi$. Transition counts between microstates were counted and reweighted as described in Eq. (45) with $x_k = \phi_k$ and

$$g(\phi_k) = \exp\left(-\frac{U(\phi_k)}{k_B T}\right) = \exp\left(-\frac{0.1 \cdot \widetilde{F}(\phi_k)}{k_B T}\right), \tag{64}$$

where ϕ_k is the first entry in the path of length τ. This resulted in five reweighted MSMs for each temperature. Figure 6 shows the mean and the standard deviation of the first three left MSM eigenvectors (evaluated at lag time $\tau = 1$ ps) and the mean and the standard deviations of the associated implied timescale.

Example scripts for simulation and analysis are included as the supplementary material.

IX. CONCLUSION AND OUTLOOK

We have presented two strategies to derive the path probability ratio M_L for the ISP scheme. In the first strategy, the correctly normalized path probability is derived by integrating out the random number η_k from the one-step transition probability. In the second strategy, the equations for the ISP scheme are solved for η_k, and the resulting transformation is used as a change in variables on the Gaussian probability density of the random numbers. This yields an unnormalized path probability. The path probability ratio M_L is then calculated as the ratio between the path probability at the target potential $\widetilde{P}_L(\omega_L; \Delta t|(x_0, v_0))$ and the path probability at the simulation potential $P_L(\omega_L; \Delta t|(x_0, v_0))$.

With M_L, we are now able to perform exact path reweighting for trajectories generated by the ISP integration scheme. Moreover, the two strategies serve as a blueprint for deriving path probability ratios for other Langevin integration schemes, which use Gaussian white noise.[44–47,49–53] Thus, path reweighting can now readily be applied to MD simulation conducted at the NVT ensemble thermostatted with a stochastic thermostat.

We compared the approximate path probability ratio M_{approx} that we used in earlier publications[34,35] to the exact path probability ratio M_L, both analytically and numerically. We showed that the two expressions only differ by $\mathcal{O}(\xi^4 \Delta t^4)$. Thus, M_{approx} is an excellent approximation to M_L for Langevin MD simulations. To understand why the approximation is so good, we showed that the random number η_k needed to generate a given step $x_k \to x_{k+1}$ is highly dependent on the integration scheme. However, $\Delta\eta_k$, the difference between the random number $\widetilde{\eta}_k$ at $\widetilde{V}(x)$ and the random number η_k at $V(x)$, has about the same value in the ISP scheme and in the Euler–Maruyama scheme.

In M_{approx}, one uses the random numbers directly recorded during the simulation at $V(x)$, which does not introduce any error and approximates $\Delta\eta_k$ by the expression from the Euler–Maruyama scheme $\Delta\eta_{o,k}$ to construct $\widetilde{\eta}_k$.

We have chosen the ISP algorithm for the present analysis in order to be consistent with our previous work.[34,35] However, the same strategy can be used to derive the path probability ratio for other Langevin integrators.[44–47,49–53] Specifically, solve the integrator equations for the random number η_k; from there, derive an expression for $\Delta\eta_k$, record η_k during the simulation at $V(x)$ and calculate $\Delta\eta_k$ on the fly, and insert η_k and $\Delta\eta_k$ into Eq. (7). For a large application of path reweighting, using a modern Langevin integrator is likely worthwhile, such as the the BAOAB method[51] (or alternatively the VRORV method[53]). This method is exceptionally efficient at sampling the configurational stationary distribution, which allows for increasing the time step.[51,53]

It is tempting to speculate that $\Delta\eta_k$ for other Langevin integration schemes could also have about the same value as $\Delta\eta_{o,k}$ for the Euler–Maruyama scheme. This would open up a route to a general approximate path probability ratio $M_{\approx}$ and would eliminate the problem that the path probability needs to be adapted for each integration scheme. On the other hand, the structure of the ISP scheme is closer to that of the Euler–Maruyama scheme than most other Langevin integrators. Whether the approximate path probability can indeed be generalized to these integrators is, therefore, not yet obvious and needs to be checked carefully.

Our one-dimensional test system and our molecular system showed that the accuracy of the reweighting sensitively depends on an accurate representation of η_k in the path probability ratio. For example, reweighting a Langevin path by the path probability ratio for the Euler–Maruyama scheme yielded very distorted results. Neither the MSM eigenvectors nor the implied timescales were reproduced correctly. It is, however, possible that the distortion is less severe in the limit of infinite sampling of the combined space of molecular states and random numbers (probably less relevant to actual applications) or if the dynamics is projected onto a reaction coordinate before the reweighted dynamical properties are evaluated (probably very relevant to actual applications).

We used path reweighting to reweight MSMs. The dynamical property that is reweighted to estimate a transition probability is a correlation function. It is important to point out that correlation functions are a combination of path ensemble averages, where the path is conditioned on a particular initial state (x_0, v_0) and a phase-space ensemble average for the initial states. Thus, the total reweighting factor for MSMs is combined of the path probability ratio M for the path ensemble average and the Boltzmann probability ratio for the phase-space ensemble average $g(x)$ [Eq. (47)].[27,32–34] Even though the reweighting of the path ensemble average can be made exact, by averaging over the initial states within a microstate, one assumes local equilibrium within this microstate.[23] Beyond local equilibrium, the formalism has been extended to reweighting transition probabilities from non-equilibrium steady-state simulations.[63]

When is the combination of enhanced sampling and path reweighting more efficient than a direct simulation? This depends on the uncertainty of the transition counts estimated from a direct simulation [Eq. (43)] compared to the uncertainty of the reweighted transition counts [Eq. (45)]. The molecular example demonstrated that path reweighting is particularly useful if the stationary density cannot be sampled accurately by direct simulation with the available computer resources. Furthermore, the efficiency of path reweighting increases if the number of transitions at the enhanced sampling simulation is large compared to the direct simulation and if the weights $W = g \cdot M$ are not too small. The path probability ratio M decreases

Published under license by AIP Publishing

with the path length τ and with the dimensionality of the bias potential U. The path length is kept short by combining path reweighting with MSMs and can be further limited by using advanced MSM discretization techniques.[64–66] The bottleneck for the dimensionality U already occurs at the stage of sampling because most enhanced sampling techniques[10] are limited to very low-dimensional biases in practice. Note that increasing the dimensionality of the overall system does not lower the efficiency of the path reweighting. The question of how strong the bias should be is more difficult to answer. Strong biases increase the transitions in the biased simulation but reduce both g and M. In Ref. 35, we empirically found that a bias of ca. 10% of the full metadynamics biasing potential yielded optimal results, but this will likely depend on the system. Here, we have restricted ourselves to systems with low barriers in the order of k_BT so that we could generate reference solutions by direct simulation. However, we believe that path reweighting is most useful for systems with large barriers that cannot be sampled by direct simulation. An example is the β-hairpin folding equilibrium in Ref. 35.

Path reweighting is closely related to path sampling techniques, in particular path sampling techniques that aim at optimizing the path action.[67–70] The combination of enhanced sampling, path sampling, and path reweighting might change the way we explore the molecular state space and investigate rare events.

SUPPLEMENTARY MATERIAL

See the supplementary material for an example OpenMM script and the corresponding Python3 scripts to construct a reweighted MSM.

DEDICATION

This paper is dedicated to Dr. Irina V. Gopich, a master of stochastic processes. Her work has influenced the way scientists in the field think about the dynamics of molecules—in simulation and in experiment.

ACKNOWLEDGMENTS

The authors would like to thank Luca Donati and Marcus Weber for helpful comments on this manuscript. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2008—390540038—UniSysCat and through grant CRC 1114 "Scaling Cascades in Complex Systems," Project Number 235221301, Project B05" Origin of scaling cascades in protein dynamics."

APPENDIX A: LANGEVIN LEAPFROG AND THE ISP SCHEME

Izaguirre, Sweet, and Pande developed the following Langevin Leapfrog algorithm:

$$v_{k+\frac{1}{2}} = \exp\left(-\xi\frac{\Delta t}{2}\right)v_k - \left[1-\exp\left(-\xi\frac{\Delta t}{2}\right)\right]\frac{\nabla V(x_k)}{\xi m} + \sqrt{\frac{k_BT}{m}\left[1-\exp(-\xi\Delta t)\right]}\,\eta_k, \tag{A1}$$

$$x_{k+1} = x_k + v_{k+\frac{1}{2}}\Delta t, \tag{A2}$$

$$v_{k+1} = \exp\left(-\xi\frac{\Delta t}{2}\right)v_{k+\frac{1}{2}} - \left[1-\exp\left(-\xi\frac{\Delta t}{2}\right)\right]\frac{\nabla V(x_{k+1})}{\xi m} + \sqrt{\frac{k_BT}{m}\left[1-\exp(-\xi\Delta t)\right]}\,\eta_{k+1} \tag{A3}$$

[Eqs. (14)–(16) in Ref. 49]. First, the velocity $v_{k+\frac{1}{2}}$ is updated by a half step using v_k, x_k, and a random number η_k [Eq. (A1)]. Then, the position update to x_{k+1} is computed from x_k, assuming a constant velocity $v_{k+\frac{1}{2}}$ in the interval $[k, k+1]$ [Eq. (A2)]. Finally, the remaining half step of the velocities to v_{k+1} is computed using x_{k+1}, $v_{k+\frac{1}{2}}$, and a new random number η_{k+1} [Eq. (A3)].

This Langevin Leapfrog algorithm has been converted to the following full-step scheme in the C^{++} CpuLangevinDynamics class of OpenMM:[71]

$$v_{k+1} = \exp(-\xi\Delta t)v_k - \left[1-\exp(-\xi\Delta t)\right]\frac{\nabla V(x_k)}{\xi m} + \sqrt{\frac{k_BT}{m}\left[1-\exp(-2\xi\Delta t)\right]}\,\eta_k, \tag{A4}$$

$$x_{k+1} = x_k + v_{k+1}\Delta t, \tag{A5}$$

where the velocities are propagated by a full step [i.e., $\Delta t/2$ in Eq. (A1) is replaced by Δt, and Δt in Eq. (A1) is replaced by $2\Delta t$] and the position update is based on v_k rather than on $v_{k+\frac{1}{2}}$. The second half step for the velocities [Eq. (A3)] is omitted. This integration scheme only uses a single random number per iteration. Equations (A4) and (A5) are the integration scheme we used in Refs. 34 and 35. To distinguish it from the original Langevin Leapfrog scheme [Eqs. (A1)–(A3)], we will refer to Eqs. (A4) and (A5) as the "ISP scheme."

To be able to analyze the path probability as a function of the positions, we rearrange Eqs. (A4) and (A5) such that we first update the positions using a stochastic step [replace v_{k+1} in Eq. (A5) by Eq. (A4)] and then update the velocity as finite difference [rearrange Eq. (A5) with respect to v_{k+1}]. This yields Eqs. (20) and (21).

APPENDIX B: PATH PROBABILITY FOR LANGEVIN DYNAMICS

We derive the closed-form expression for $P_L(\omega_L; \Delta t|(x_0, v_0))$ in Eq. (22) from the integration scheme [Eqs. (20) and (21)] by following the approach in Ref. 57. As a first step, we derive a closed-form expression for the one-step probability $P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k))$ of observing a step $(x_k, v_k) \to (x_{k+1}, v_{k+1})$. According to Eqs. (20) and (21), the tuple (x_{k+1}, v_{k+1}) at iteration step $k+1$ is entirely determined by the tuple (x_k, v_k) at iteration step k if additionally the random number η_k is known. Thus, $P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k, \eta_k))$, i.e., the one-step probability with fixed random number η_k, is a Dirac delta function centered at (x_{k+1}, v_{k+1}). Our strategy is to derive a closed-form expression for this Dirac delta function

Published under license by AIP Publishing

using Eqs. (20) and (21) and to integrate out the dependency on η_k. In this Appendix, we omit the index L in $\eta_{L,k}$ to simplify the notation.

We reformulate the two-dimensional probability $P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k, \eta_k))$ as a product of two one-dimensional probabilities,

$$P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k, \eta_k)) = P_L(v_{k+1}; \Delta t|(x_{k+1}, x_k, v_k, \eta_k)) \cdot P_L(x_{k+1}; \Delta t|(x_k, v_k, \eta_k)) \quad \text{(B1)}$$

using the rule $P(A, B|C) = P(A|B, C) \cdot P(B|C)$, with $A = v_{k+1}$, $B = x_{k+1}$, and $C = (x_k, v_k, \eta_k)$. This rule is the extension of the conditional probability $P(A, B) = P(A|B) \cdot P(B)$ to an additional condition C. The first factor is a Dirac delta function constrained to Eq. (21),

$$P_L(v_{k+1}; \Delta t|(x_{k+1}, x_k, v_k, \eta_k)) = P_L(v_{k+1}; \Delta t|(x_{k+1}, x_k)) = \delta\left(v_{k+1} - \frac{x_{k+1} - x_k}{\Delta t}\right), \quad \text{(B2)}$$

where the first equality emphasizes that v_{k+1} does not depend on η_k or v_k in Eq. (21). Note that the probability of the velocity v_{k+1} [Eq. (B2)] does not depend on a random number, which mirrors our previous observation that v_{k+1} is not treated as a random variable in Eq. (21). The second factor in Eq. (B1) is a Dirac delta function constrained to Eq. (20),

$$P_L(x_{k+1}; \Delta t|(x_k, v_k, \eta_k)) = \delta\Big(x_{k+1} - x_k - \exp(-\xi\,\Delta t)\, v_k \Delta t + [1 - \exp(-\xi\,\Delta t)]\, \frac{\nabla V(x_k)}{\xi m} \Delta t - \sqrt{\frac{k_B T}{m}\,[1 - \exp(-2\xi\,\Delta t)]}\, \eta_k \Delta t\Big). \quad \text{(B3)}$$

Reinserting the two factors into Eq. (B1) yields the desired closed-form expression for $P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k, \eta_k))$. Since we know that the random numbers η_k are drawn from a Gaussian distribution $P(\eta_k)$ with zero mean and unit variance

$$P(\eta_k) = N^{-1} \exp\left(-\frac{\eta_k^2}{2}\right), \qquad N = \sqrt{2\pi}, \quad \text{(B4)}$$

we can average out the random number dependency in Eq. (B1) to obtain the one-step probability,

$$\begin{aligned} P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k)) &= \int_{-\infty}^{\infty} \mathrm{d}\eta_k P(\eta_k)\, P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k, \eta_k)) \\ &= \delta\left(v_{k+1} - \frac{x_{k+1} - x_k}{\Delta t}\right) \cdot \int_{-\infty}^{\infty} \mathrm{d}\eta_k P_\eta(\eta_k)\, P_L(x_{k+1}; \Delta t|(x_k, v_k, \eta_k)). \end{aligned} \quad \text{(B5)}$$

The challenge lies in solving the integral in this equation. The solution, which is detailed in Appendix C, yields the closed-form expression for the one-step probability,

$$P_L(x_{k+1}, v_{k+1}; \Delta t|(x_k, v_k)) = \delta\left(v_{k+1} - \frac{x_{k+1} - x_k}{\Delta t}\right) \cdot \sqrt{\frac{m}{2\pi k_B T \Delta t^2 (1 - \exp(-2\xi\Delta t))}} \times \exp\left(-\frac{m\left(x_{k+1} - x_k - \exp(-\xi\Delta t) v_k \Delta t + (1 - \exp(-\xi\Delta t))\frac{\nabla V(x_k)}{\xi m}\Delta t\right)^2}{2k_B T(1 - \exp(-2\xi\Delta t))\Delta t^2}\right). \quad \text{(B6)}$$

Applying the Chapman–Kolmogorov equation[72] recursively to the one-step probability yields the closed-form expression for the path probability $P_L(\omega_L; \Delta t|(x_0, v_0))$, shown in Eq. (22).

APPENDIX C: SOLVING THE DOUBLE INTEGRAL

We compute the integral

$$P_L(x_{k+1}; \Delta t|(x_k, v_k)) = \int_{-\infty}^{\infty} \mathrm{d}\eta_k\, P(\eta_k)\, P_L(x_{k+1}; \Delta t|(x_k, v_k, \eta_k)) \quad \text{(C1)}$$

from Eq. (B5). First, we replace $P(\eta_k)$ according to Eq. (B4). Second, we substitute $P_L(x_{k+1}; \Delta t|(x_k, v_k, \eta_k))$, which is a δ-function [Eq. (B3)], with its Fourier transform

$$\delta(z - z') = \int_{-\infty}^{+\infty} \frac{\mathrm{d}w}{2\pi} \exp\big(iw(z - z')\big), \quad \text{(C2)}$$

where $z = x_{k+1}$ and z' is equal to the right-hand side of Eq. (20). This yields a double integral whose outer integral is with respect to w,

Published under license by AIP Publishing

while the inner integral is with respect to η_k,

$$\begin{aligned} &P_L(x_{k+1};\Delta t|(x_k,v_k)) \\ &= \int_{-\infty}^{+\infty}\frac{\mathrm{d}w}{2\pi}\int_{-\infty}^{+\infty}\frac{\mathrm{d}\eta_k}{N}\exp\left(-\frac{\eta_k^2}{2}\right) \\ &\quad\times\exp\Bigg(iw\Bigg[x_{k+1}-x_k-\exp(-\xi\Delta t)\,v_k\Delta t \\ &\quad+[1-\exp(-\xi\Delta t)]\,\frac{\nabla V(x_k)}{\xi m}\Delta t \\ &\quad-\sqrt{\frac{k_BT}{m}\,[1-\exp(-2\xi\Delta t)]}\,\eta_k\Delta t\Bigg]\Bigg) \\ &= \int_{-\infty}^{+\infty}\frac{\mathrm{d}w}{2\pi}\exp(iwB)\int_{-\infty}^{+\infty}\frac{\mathrm{d}\eta_k}{N}\exp\left(-\frac{\eta_k^2}{2}-iwR\,\eta_k\right), \end{aligned} \tag{C3}$$

where we moved all terms that do not depend on η_k out of the inner integral and defined the abbreviations,

$$\begin{aligned} B &= \Bigg[x_{k+1}-x_k-\exp(-\xi\Delta t)\,v_k\Delta t \\ &\quad+[1-\exp(-\xi\Delta t)]\,\frac{\nabla V(x_k)}{\xi m}\Delta t\Bigg], \\ R &= \Delta t\sqrt{\frac{k_BT}{m}\,[1-\exp(-2\xi\Delta t)]}. \end{aligned} \tag{C4}$$

Both integrals in Eq. (C3) can be solved with the completing-the-square technique for Gaussian integrals. The goal of this technique is to expand and rearrange the inner integral such that we can use the analytic solution

$$\int_{-\infty}^{\infty}\mathrm{d}x\,\exp\left(-a(x\pm b)^2\right)=\sqrt{\frac{\pi}{a}}\qquad\text{for } a,b\in\mathbb{R}. \tag{C5}$$

This can be achieved by a systematic step-to-step procedure that can be applied to all Gaussian integrals of this type,

$$\begin{aligned} &\int_{-\infty}^{+\infty}\frac{\mathrm{d}\eta_k}{N}\exp\left(-\frac{\eta_k^2}{2}-iwR\,\eta_k\right) \\ &= \int_{-\infty}^{+\infty}\frac{\mathrm{d}\eta_k}{N}\exp\left(-\frac{1}{2}\left[\eta_k^2+2iwR\,\eta_k\overbrace{+i^2w^2R^2-i^2w^2R^2}^{=0}\right]\right) \\ &= \exp\left(-\frac{w^2R^2}{2}\right)\int_{-\infty}^{+\infty}\frac{\mathrm{d}\eta_k}{N}\exp\left(-\frac{1}{2}(\eta_k+iwR)^2\right) \\ &= \exp\left(-\frac{w^2R^2}{2}\right)\frac{1}{N}\sqrt{2\pi} \\ &= \exp\left(-\frac{w^2R^2}{2}\right). \end{aligned} \tag{C6}$$

In the first line, we isolate η_k^2 by factoring out $-\frac{1}{2}$ and complete the first binomial formula by adding a zero. Then, we separate the exponent into the binomial formula and the term $\exp\left(-\frac{w^2R^2}{2}\right)$, which can be moved in front of the integral because it does not depend on η_k. In the third line, we solve the remaining integral using Eq. (C5), which can be further simplified by inserting the normalization constant of the Gaussian distribution: $N=\sqrt{2\pi}$.

Inserting Eq. (C6) into Eq. (C3) yields the outer integral

$$\int_{-\infty}^{+\infty}\frac{\mathrm{d}w}{2\pi}\exp(iwB)\,\exp\left(-\frac{w^2R^2}{2}\right)=\int_{-\infty}^{+\infty}\frac{\mathrm{d}w}{2\pi}\exp\left(-\frac{w^2R^2}{2}+iwB\right),$$

which is solved using the same procedure,

$$\begin{aligned} &\int_{-\infty}^{+\infty}\frac{\mathrm{d}w}{2\pi}\exp\left(-\frac{w^2R^2}{2}+iwB\right) \\ &= \int_{-\infty}^{+\infty}\frac{\mathrm{d}w}{2\pi}\exp\left(-\frac{R^2}{2}\left[w^2+\frac{2iwB}{R^2}+\overbrace{\frac{i^2B^2}{R^4}-\frac{i^2B^2}{R^4}}^{=0}\right]\right) \\ &= \exp\left(-\frac{B^2}{2R^2}\right)\int_{-\infty}^{\infty}\frac{\mathrm{d}w}{2\pi}\exp\left(-\frac{R^2}{2}\left(w+\frac{iB}{R^2}\right)^2\right) \\ &= \exp\left(-\frac{B^2}{2R^2}\right)\frac{1}{2\pi}\sqrt{\frac{2\pi}{R^2}} \\ &= \sqrt{\frac{1}{2\pi R^2}}\,\exp\left(-\frac{B^2}{2R^2}\right). \end{aligned} \tag{C7}$$

Inserting the expressions for the constants R and B [Eq. (C4)] yields

$$\begin{aligned} &P_L(x_{k+1};\Delta t|(x_k,v_k)) \\ &= \sqrt{\frac{m}{2\pi k_BT\Delta t^2(1-\exp(-2\xi\Delta t))}} \\ &\quad\times\exp\left(-\frac{m\left(x_{k+1}-x_k-\exp(-\xi\Delta t)v_k\Delta t+(1-\exp(-\xi\Delta t))\frac{\nabla V(x_k)}{\xi m}\Delta t\right)^2}{2k_BT(1-\exp(-2\xi\Delta t))\Delta t^2}\right). \end{aligned} \tag{C8}$$

This is inserted into Eq. (B5) to yield Eq. (B6).

APPENDIX D: PROOF OF EQ. (31)

$$\begin{aligned} \frac{\left(1-e^{-x}\right)^2}{x\cdot\left(1-e^{-2x}\right)} &= \frac{1}{2}-\frac{x^2}{24}+\frac{x^4}{240}\pm\mathcal{O}(x^5), \\ \left(1-e^{-x}\right)^2 &= \frac{x}{2}\cdot\left(1-e^{-2x}\right)-\frac{x^2}{24}\cdot x\cdot\left(1-e^{-2x}\right) \\ &\quad\pm\mathcal{O}(x^4)\cdot x\cdot\left(1-e^{-2x}\right), \\ \left(1-e^{-x}\right)^2 &= \frac{x}{2}\cdot\left(1-e^{-2x}\right)-\mathcal{O}(x^4). \end{aligned} \tag{D1}$$

The first line shows the Taylor expansion of the expression on the right-hand side. To obtain the second line, we multiplied by $x\cdot\left(1-e^{-2x}\right)$. In the third line, we used the fact that the leading term of the Taylor expansion of $x\cdot\left(1-e^{-2x}\right)$ is $2x^2$, thus yielding an error of $\mathcal{O}(x^4)$. Substituting $x=\xi\Delta t$ yields

Published under license by AIP Publishing

$$\left(1-e^{-\xi\Delta t}\right)^2 = \frac{\xi\Delta t}{2}\cdot\left(1-e^{-2\xi\Delta t}\right) - \mathcal{O}(\xi^4\Delta t^4),$$
$$\left(1-e^{-\xi\Delta t}\right)^2 \approx \frac{\xi\Delta t}{2}\cdot\left(1-e^{-2\xi\Delta t}\right), \tag{D2}$$

and multiplying by $\frac{1}{k_BT\xi^2 m\left(1-e^{-2\xi\Delta t}\right)}(\nabla U(x_k))^2$ yields

$$\frac{1}{k_BT\xi^2 m}\frac{\left(1-e^{-\xi\Delta t}\right)^2}{1-e^{-2\xi\Delta t}}(\nabla U(x_k))^2 \approx \frac{\Delta t}{2k_BT\xi m}(\nabla U(x_k))^2, \tag{D3}$$
$$\Delta\eta_{L,k}^2 \approx \Delta\eta_{o,k}^2.$$

Thus, the difference between $\Delta\eta_{L,k}^2$ [Eq. (26)] and $\Delta\eta_{o,k}^2$ [Eq. (17)] is of order $\mathcal{O}(\xi^4\Delta t^4)$. Equation (D3) is Eq. (31) squared. □

DATA AVAILABILITY

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

REFERENCES

[1] I. V. Gopich, "Multisite reversible association in membranes and solutions: From non-Markovian to Markovian kinetics," J. Chem. Phys. **152**, 104101 (2020).
[2] T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande, "To milliseconds and beyond: Challenges in the simulation of protein folding," Curr. Opin. Struct. Biol. **23**, 58 (2013).
[3] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw, "Biomolecular simulation: A computational microscope for molecular biology," Annu. Rev. Biophys. **41**, 429 (2012).
[4] E. P. Barros, L. Casalino, Z. Gaieb, A. C. Dommer, Y. Wang, L. Fallon, L. Raguette, K. Belfon, C. Simmerling, and R. E. Amaro, "The flexibility of ACE2 in the context of SARS-CoV-2 infection," Biophys. J. **120**, 1 (2020).
[5] T. J. Harpole and L. Delemotte, "Conformational landscapes of membrane proteins delineated by enhanced sampling molecular dynamics simulations," Biochim. Biophys. Acta **1860**, 909 (2018).
[6] Z. Cournia, B. Allen, and W. Sherman, "Relative binding free energy calculations in drug discovery: Recent advances and practical considerations," J. Chem. Inf. Model. **57**, 2911 (2017).
[7] M. Badaoui, A. Kells, C. Molteni, C. J. Dickson, V. Hornak, and E. Rosta, "Calculating kinetic rates and membrane permeability from biased simulations," J. Phys. Chem. B **122**, 11571 (2018).
[8] J.-O. Joswig, J. Anders, H. Zhang, C. Rademacher, and B. G. Keller, "Molecular mechanism of the pH-dependent calcium affinity in Langerin," bioRxiv:986851 (2020).
[9] A. S. J. S. Mey, B. Allen, H. E. B. Macdonald, J. D. Chodera, M. Kuhn, J. Michel, D. L. Mobley, L. N. Naden, S. Prasad, A. Rizzi, J. Scheen, M. R. Shirts, G. Tresadern, and H. Xu, "Best practices for alchemical free energy calculations," Living J. Comput. Mol. Sci. Living J. Comp. Mol. Sci. ASAP Version, pages **2**, 1, available at https://www.livecomsjournal.org/article/18378-best-practices-for-alchemical-free-energy-calculations-article-v1-0.
[10] M. Tuckerman, *Monte Carlo Statistical Mechanics: Theory and Molecular Simulation* (Oxford University Press, Inc., New York, 2010), pp. 300–304.
[11] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 1st ed. (Academic Press, San Diego, San Francisco, NY, Boston, London, Sydney, Tokyo, 2002).
[12] C. A. F. de Oliveira, D. Hamelberg, and J. A. McCammon, "Estimating kinetic rates from accelerated molecular dynamics simulations: Alanine dipeptide in explicit solvent as a case study," J. Chem. Phys. **127**, 175105 (2007).
[13] P. Tiwary and M. Parrinello, "From metadynamics to dynamics," Phys. Rev. Lett. **111**, 230602 (2013).
[14] O. Valsson, P. Tiwary, and M. Parrinello, "Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint," Annu. Rev. Phys. Chem. **67**, 159 (2016).
[15] R. Casasnovas, V. Limongelli, P. Tiwary, P. Carloni, and M. Parrinello, "Unbinding kinetics of a p38 MAP kinase type II inhibitor from metadynamics simulations," J. Am. Chem. Soc. **139**, 4780 (2017).
[16] H. Wu, A. S. J. S. Mey, E. Rosta, and F. Noé, "Statistically optimal analysis of state-discretized trajectory data from multiple thermodynamic states," J. Chem. Phys. **141**, 214106 (2014).
[17] A. S. J. S. Mey, H. Wu, and F. Noé, "xTRAM: Estimating equilibrium expectations from time-correlated simulation data at multiple thermodynamic states," Phys. Rev. X **4**, 041018 (2014).
[18] H. Wu, F. Paul, C. Wehmeyer, and F. Noé, "Multiensemble Markov models of molecular thermodynamics and kinetics," Proc. Natl. Acad. Sci. U. S. A. **113**, E3221 (2016).
[19] L. S. Stelzl, A. Kells, E. Rosta, and G. Hummer, "Dynamic histogram analysis to determine free energies and rates from biased simulations," J. Chem. Theory Comput. **13**, 6328 (2017).
[20] D. J. Bicout and A. Szabo, "Electron transfer reaction dynamics in non-Debye solvents," J. Chem. Phys. **109**, 2325 (1998).
[21] E. Rosta and G. Hummer, "Free energies from dynamic weighted histogram analysis using unbiased markov state model," J. Chem. Theory Comput. **11**, 276 (2014).
[22] L. Donati, M. Heida, B. G. Keller, and M. Weber, "Estimation of the infinitesimal generator by square-root approximation," J. Phys.: Condens. Matter **30**, 425201 (2018).
[23] S. Kieninger, L. Donati, and B. G. Keller, "Dynamical reweighting methods for Markov models," Curr. Opin. Struct. Biol. **61**, 124 (2020).
[24] D. M. Zuckerman and T. B. Woolf, "Dynamic reaction paths and rates through importance-sampled stochastic dynamics," J. Chem. Phys. **111**, 9475 (1999).
[25] T. B. Woolf, "Path corrected functionals of stochastic trajectories: Towards relative free energy and reaction coordinate calculations," Chem. Phys. Lett. **289**, 433 (1998).
[26] D. M. Zuckerman and T. B. Woolf, "Efficient dynamic importance sampling of rare events in one dimension," Phys. Rev. E **63**, 016702 (2000).
[27] C. Xing and I. Andricioaei, "On the calculation of time correlation functions by potential scaling," J. Chem. Phys. **124**, 034110 (2006).
[28] A. B. Adib, "Stochastic actions for diffusive dynamics: Reweighting, sampling, and minimization," J. Phys. Chem. B **112**, 5910 (2008).
[29] I. V. Girsanov, "On transforming a certain class of stochastic processes by absolutely continuous substitution of measures," Theory Probab. Appl. **5**, 285 (1960).
[30] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, 6th ed. (Springer Verlag, Berlin, 2003).
[31] L. Onsager and S. Machlup, "Fluctuations and irreversible processes," Phys. Rev. **91**, 1505 (1953).
[32] J.-H. Prinz, J. D. Chodera, V. S. Pande, W. C. Swope, J. C. Smith, and F. Noé, "Optimal use of data in parallel tempering simulations for the construction of discrete-state Markov models of biomolecular dynamics," J. Chem. Phys. **134**, 244108 (2011).
[33] C. Schütte, A. Nielsen, and M. Weber, "Markov state models and molecular alchemy," Mol. Phys. **113**, 69 (2015).
[34] L. Donati, C. Hartmann, and B. G. Keller, "Girsanov reweighting for path ensembles and Markov state models," J. Chem. Phys. **146**, 244112 (2017).
[35] L. Donati and B. G. Keller, "Girsanov reweighting for metadynamics simulations," J. Chem. Phys. **149**, 072335 (2018).
[36] W. Huisinga, C. Schütte, and A. M. Stuart, "Extracting macroscopic stochastic dynamics: Model problems," Commun. Pure Appl. Math. **56**, 234 (2003).
[37] W. C. Swope, J. W. Pitera, and F. Suits, "Describing protein folding kinetics by molecular dynamics simulations. 1. Theory," J. Phys. Chem. B **108**, 6571 (2004).
[38] N.-V. Buchete and G. Hummer, "Coarse master equations for peptide folding dynamics," J. Phys. Chem. B **112**, 6057 (2008).
[39] B. Keller, X. Daura, and W. F. van Gunsteren, "Comparing geometric and kinetic cluster algorithms for molecular simulation data," J. Chem. Phys. **132**, 074110 (2010).

Published under license by AIP Publishing

[40]J.-H. Prinz, H. Wu, M. Sarich, B. G. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics: Generation and validation," J. Chem. Phys. **134**, 174105 (2011).
[41]J.-H. Prinz, B. Keller, and F. Noé, "Probing molecular kinetics with Markov models: Metastable states, transition pathways and spectroscopic observables," Phys. Chem. Chem. Phys. **13**, 16912 (2011).
[42]B. E. Husic and V. S. Pande, "Markov state models: From an art to a science," J. Am. Chem. Soc. **140**, 2386 (2018).
[43]J. D. Chodera, W. C. Swope, F. Noé, J.-H. Prinz, M. R. Shirts, and V. S. Pande, "Dynamical reweighting: Improved estimates of dynamical properties from simulations at multiple temperatures," J. Chem. Phys. **134**, 244107 (2011).
[44]W. F. van Gunsteren and H. J. C. Berendsen, "Algorithms for Brownian dynamics," Mol. Phys. **45**, 637 (1981).
[45]A. Brünger, C. L. Brooks, and M. Karplus, "Stochastic boundary conditions for molecular dynamics simulations of ST2 water," Chem. Phys. Lett. **105**, 495 (1984).
[46]G. Stoltz, "Path sampling with stochastic dynamics: Some new algorithms," J. Comput. Phys. **225**, 491 (2007).
[47]G. Bussi and M. Parrinello, "Accurate sampling using Langevin dynamics," Phys. Rev. E **75**, 056707 (2007).
[48]M. Ceriotti, G. Bussi, and M. Parrinello, "Langevin equation with colored noise for constant-temperature molecular dynamics simulations," Phys. Rev. Lett. **102**, 020601 (2009).
[49]J. A. Izaguirre, C. R. Sweet, and V. Pande, "Multiscale dynamics of macromolecules using normal mode Langevin," Pac. Symp. Biocomput. **15**, 240 (2010).
[50]N. Goga, A. J. Rzepiela, A. H. de Vries, S. J. Marrink, and H. J. C. Berendsen, "Efficient algorithms for Langevin and DPD dynamics." J. Chem. Theory Comput. **8**, 3637 (2012).
[51]B. Leimkuhler and C. Matthews, "Robust and efficient configurational molecular sampling via Langevin dynamics," J. Chem. Phys. **138**, 174102 (2013).
[52]D. A. Sivak, J. D. Chodera, and G. E. Crooks, "Time step rescaling recovers continuous-time dynamical properties for discrete-time Langevin integration of nonequilibrium systems," J. Phys. Chem. B **118**, 6466 (2014).
[53]J. Fass, D. Sivak, G. Crooks, K. Beauchamp, B. Leimkuhler, and J. Chodera, "Quantifying configuration-sampling error in Langevin simulations of complex molecular systems," Entropy **20**, 318 (2018).
[54]P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande, "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics," PLoS Comput. Biol. **13**, 1 (2017).
[55]N. Bou-Rabee, "Time integrators for molecular dynamics," Entropy **16**, 138 (2014).
[56]C. C. Chow and M. A. Buice, "Path integral methods for stochastic differential equations," J. Math. Neurosci. **5**, 1 (2015).
[57]P. C. Bressloff, *Stochastic Processes in Cell Biology*, 1st ed. (Springer, New York, 2014).
[58]P. H. Hünenberger, "Thermostat algorithms for molecular dynamics simulations," in *Advanced Computer Simulation* (Springer, Berlin, Heidelberg, 2005).
[59]J. E. Basconi and M. R. Shirts, "Effects of temperature control algorithms on transport properties and kinetics in molecular dynamics simulations," J. Chem. Theory Comput. **9**, 2887 (2013).
[60]J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field," J. Comput. Chem. **25**, 1157 (2004).
[61]A. Onufriev, D. Bashford, and D. A. Case, "Exploring protein native states and large-scale conformational changes with a modified generalized born model," J. Comput. Chem. **55**, 383 (2004).
[62]See http://docs.openmm.org/latest/api-python/generated/simtk.openmm.openmm.CustomIntegrator.html for information about the CustomIntegrator Class of the simulation package OpenMM; accessed 25 January 2021.
[63]M. Bause, T. Wittenstein, K. Kremer, and T. Bereau, "Microscopic reweighting for nonequilibrium steady-state dynamics," Phys. Rev. E **100**, 060103 (2019).
[64]G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for Markov model construction," J. Chem. Phys. **139**, 015102 (2013).
[65]F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé, "Variational approach to molecular kinetics," J. Chem. Theory Comput. **10**, 1739–1752 (2014).
[66]O. Lemke and B. G. Keller, "Density-based cluster algorithms for the identification of core sets," J. Chem. Phys. **145**, 164104 (2016).
[67]L. T. Chong, A. S. Saglam, and D. M. Zuckerman, "Path-sampling strategies for simulating rare events in biomolecular systems," Curr. Opin. Struct. Biol. **43**, 88–94 (2017).
[68]G. Grazioli and I. Andricioaei, "Advances in milestoning. I. Enhanced sampling via wind-assisted reweighted milestoning (WARM)," J. Chem. Phys. **149**, 084103 (2018).
[69]P. D. Dixit, J. Wagoner, C. Weistuch, S. Pressé, K. Ghosh, and K. A. Dill, "Perspective: Maximum caliber is a general variational principle for dynamical systems," J. Chem. Phys. **148**, 010901 (2018).
[70]E. K. Peter, J.-E. Shea, and A. Schug, "CORE-MD, a path correlated molecular dynamics simulation method," J. Chem. Phys. **153**, 084114 (2020).
[71]See https://github.com/openmm/openmm/blob/master/platforms/cpu/src/CpuLangevinDynamics.cpp for information about the CpuLangevinDynamics Class of the simulation package OpenMM accessed 15 November 2020.
[72]C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, 2nd ed. (Springer Verlag, Berlin, Heidelberg, 1983).

Published under license by AIP Publishing