Free Energy Landscapes, Diffusion Coefficients, and Kinetic Rates from Transition Paths

Karen Palacio-Rodriguez and Fabio Pietrucci*
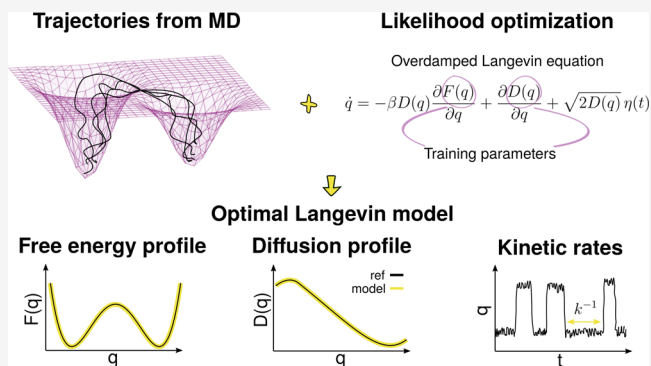
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We address the problem of constructing accurate mathematical models of the dynamics of complex systems projected on a collective variable. To this aim we introduce a conceptually simple yet effective algorithm for estimating the parameters of Langevin and Fokker−Planck equations from a set of short, possibly out-of-equilibrium molecular dynamics trajectories, obtained for instance from transition path sampling or as relaxation from high free-energy configurations. The approach maximizes the model likelihood based on any explicit expression of the short-time propagator, hence it can be applied to different evolution equations. We demonstrate the numerical efficiency and robustness of the algorithm on model systems, and we apply it to reconstruct the projected dynamics of pairs of C_{60} and C_{240} fullerene molecules in explicit water. Our methodology allows reconstructing the accurate thermodynamics and kinetics of activated processes, namely free energy landscapes, diffusion coefficients, and kinetic rates. Compared to existing enhanced sampling methods, we directly exploit short unbiased trajectories at a competitive computational cost.

Trajectories from MD

Likelihood optimization

Overdamped Langevin equation

$$\dot{q} = -\beta D(q)\frac{\partial F(q)}{\partial q} + \frac{\partial D(q)}{\partial q} + \sqrt{2D(q)}\,\eta(t)$$

Training parameters

Optimal Langevin model

Free energy profile | Diffusion profile | Kinetic rates

INTRODUCTION

Atomistic computer simulations of rare events like phase transitions, chemical reactions, or biomolecular interactions and conformational changes have three paramount goals: predicting detailed mechanisms, free energy landscapes, and kinetic rates. All of these tasks are, in many cases, cumbersome and require intensive human and computer effort, the calculation of rates being the most difficult.[1,2]

Usually, practitioners adopt the following strategy: first, a guess of the transformation mechanism is obtained using chemical intuition or acceleration techniques including artificial biasing forces. Next, either the free energy landscape as a function of a few collective variables (CVs) is reconstructed, or the detailed mechanism is investigated with transition path sampling (TPS) techniques.[3,4] Alternative approaches like transition interface sampling[5] or forward flux sampling[6] seek at the same time mechanism, free energy, and kinetic rates; however, their computational cost can be prohibitive and the results could depend on the choice of the order parameter required by these techniques.

Projecting the dynamics of the system on one (or a few) CV q leads to a theoretical framework describing equilibrium thermodynamics through free energy landscapes, and kinetics (including out-of-equilibrium relaxation) through Langevin and Fokker−Planck equations.[7,8] Three main types of Langevin equations that depend on the nature of the physical system and on the observational time resolution τ, appear in the literature. In the non-Markovian case, a generalized

Langevin equation is customarily employed, including a deterministic force (the gradient of the free energy $-dF/dq$) and time-correlated friction and noise, connected by the fluctuation−dissipation theorem.[7,9] Such time correlation, or memory effects, can be significant for the short-time dynamics of small solutes immersed in a liquid bath.[10−12]

However, in many applications in physics, chemistry, or biology an observational time resolution τ coarser than the memory time scale is pertinent, so that a Markovian Langevin equation accurately describes the projected dynamics. Depending on the intensity of the friction, or, in other words, on the extent of its characteristic time compared to τ, which depends on the physical process and on the choice of CVs, a second-order underdamped Langevin equation featuring inertia (and reminiscent of Hamilton's equations) or a first-order overdamped equation is appropriate.[13−15]

A number of algorithms aim at parametrizing Langevin models starting from trajectories of many-particle systems.[11,12,16−43] More in detail, non-Markovian friction (the so-called memory kernel) can be reconstructed in different
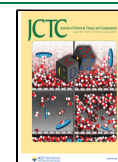
ways, for instance based on a set of time-correlation functions and Volterra integral equations (see, e.g., refs 11, 12, and 17) or via likelihood maximization.[43] In the case of Markovian (overdamped) Langevin equations, the latter approach,[22,27,29] together with direct estimation of Kramers-Moyal coefficients,[29,35] has been employed.

Often, only friction is addressed, assuming that the free-energy profile is known in advance, or that it can be directly estimated from the histogram of the CV based on long equilibrium trajectories reversibly sampling all relevant states (see, e.g., refs 12, 24, and 27). Clearly, this severely limits the scope of the techniques, given that ergodic molecular dynamics (MD) trajectories can be affordably generated only in the case of low kinetic barriers (a few $k_B T$).

In the case of rare events, some approaches are able to exploit enhanced sampling simulations, such as umbrella sampling, to alleviate the time scale problem,[21,29,36] even though the cost of such simulations to estimate free-energy and diffusion profiles remains considerable in the presence of high barriers. Moreover, such acceleration techniques usually rely on an initial guess of CVs to be biased: a bad choice can lead to poor convergence and to explore unfavorable transition mechanisms compared to brute force MD, while a change of CVs requires in general generation of new biased trajectories, multiplying the costs. At variance with previous attempts, in this work we address this issue by basing our Langevin optimization approach on unbiased TPS-like[44] trajectories, spontaneously relaxing from barrier tops, and faithfully reproducing brute force transitions thanks to the reliability of TPS techniques. In our case, the choice of the CV is performed a posteriori, and multiple CVs can be tested at negligible computational cost since no new MD trajectories are required.

Whether the projected dynamics of a given system can be accurately modeled by a specific flavor of Langevin equation depends upon two main observational choices: the CV q and the time resolution τ of the trajectory. In particular, it is expected that memory effects decrease when CVs become similar to the committor function.[15,21,45] For a fixed CV, too small τ can introduce non-Markovian effects. Albeit acknowledged in the literature, these principles have seldom been applied to critically assess how accurately a given Langevin model reproduces the original dynamics for different CVs and different time resolutions.[27,35,40] In many cases a single system is studied, and Langevin predictions at fixed time resolution are compared with independent estimates of free energies and rates. We also remark on a further important point: τ cannot be freely increased to enhance Markovian behavior, since too large values rarefy the sampled trajectory to the point of lacking information on barrier regions.

As a result of all the previous concerns and difficulties, to date CV-based Langevin equations have been used sporadically and most often heuristically, and are not systematically applied to access the thermodynamics and kinetics of a wide range of complex physicochemical processes. It is our aim to improve the state of the art in this respect, with particular care to the assessment of time resolution effects on the accuracy of the reduced-dimensionality model, as detailed below.

In this work we present a conceptually simple and computationally efficient method to construct data-driven Langevin models of rare events in complex systems. Since equilibrium MD data are not necessary for training, the method can be applied irrespective of the height of the barriers separating metastable states. We also provide an explicit

procedure to assess the optimal resolution τ leading to accurate Langevin models, and to predict their reliability. The work is organized as follows: (i) we describe the algorithm used to optimize the models and the computational details of the MD simulations, (ii) we test the algorithm on a double-well potential with known free energy, diffusion profiles, and kinetic rates, (iii) we apply the algorithm on the interaction of C_{60} and C_{240} fullerene dimers in explicit water, a proxy model for protein−protein interactions, allowing also to evaluate the effect of the barrier height. Finally, we discuss the scope of the method, its limitations and future application perspectives.

■ THEORETICAL METHODS

A. Optimization of Langevin Models (Parameter Estimation). Langevin equations can be obtained from Hamilton's equations of motion by projecting the high-dimensional deterministic dynamics on a subset of the phase-space variables.[7,43] In the limit of strong friction, velocity fluctuations away from the equilibrium distribution are damped quickly (compared to the time resolution τ), yielding the widely employed overdamped Langevin equation:

$$\dot{q} = -\beta D(q)\frac{\partial F(q)}{\partial q} + \frac{\partial D(q)}{\partial q} + \sqrt{2D(q)}\,\eta(t) \tag{1}$$

where $F(q) = -k_B T \log \rho_{eq}(q)$ is the free-energy landscape, $D(q)$ is the position-dependent diffusion coefficient, and $\eta(t)$ is a Gaussian white noise. We remark that a realistic description of the dynamics generally requires a nonconstant $D(q)$,[22] and that the use of the overdamped equation can be formally justified (for instance, it provides exact mean first passage times (MFPT)) when q is the optimal CV, that is, any monotonic one-to-one function of the committor.[15]

The corresponding Fokker−Planck equation is the Smoluchowski equation,[46] describing the evolution of the probability density $\rho(q, t)$ from an initial distribution $\rho(q, 0)$ (typically $= \delta(q − q_0)$) under given boundary conditions:

$$\frac{\partial}{\partial t}\rho = \frac{\partial}{\partial q}\left(\beta D\frac{\partial F}{\partial q}\rho - \frac{\partial D}{\partial q}\rho\right) + \frac{\partial^2}{\partial q^2}(D\rho)$$

$$= \frac{\partial}{\partial q}\left[De^{-\beta F}\frac{\partial}{\partial q}(e^{\beta F}\rho)\right] \tag{2}$$

To construct an optimal model of the q-projected dynamics, we adopt an efficient algorithm that maximizes the likelihood $\mathcal{L}(\theta)$ to observe the trajectory $\{q_i \equiv q(t_i)\}_{i=1,...,N}$, $t_{i+1} − t_i = \tau$ given the Langevin model eq 1. The parameters θ encode the shape of $F(q)$ and $D(q)$. To obtain an explicit form for the likelihood, we consider the formal solution of the Fokker−Planck equation $\partial_t\rho = L\rho$ (eq 2), calling L the Fokker−Planck operator (the adjoint of the generator of the stochastic process):[8,47]

$$p(q', t + \tau|q, t) = e^{L\tau}\delta(q' − q)$$

$$= \left[1 + L\tau + \frac{1}{2}L^2\tau^2 + ...\right]\delta(q' − q) \tag{3}$$

For small τ, the short-time transition probability density p, also called propagator, is an ingredient of path-integral formulations of long-time transition probabilities and, neglecting terms of
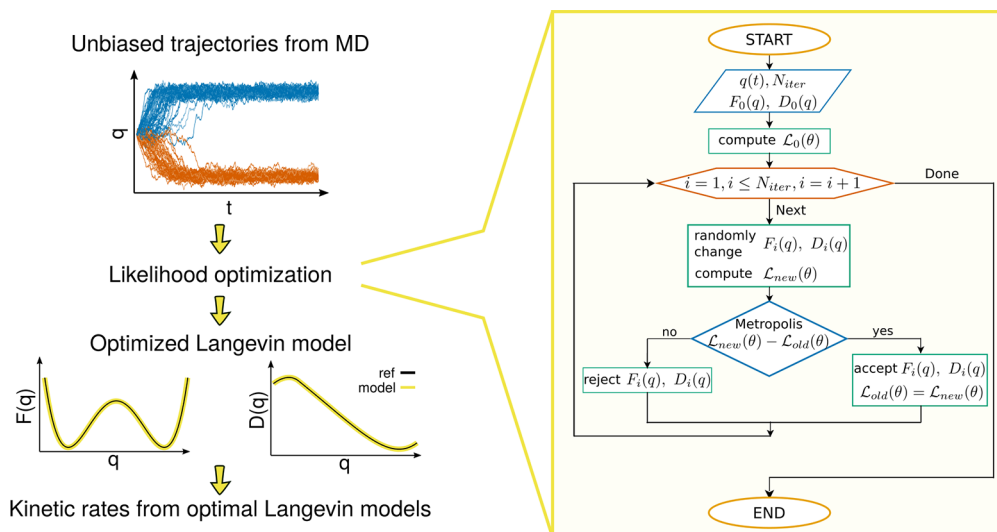
Figure 1. Schematic illustration of the approach proposed to optimize Langevin models of projected dynamics. The input data are formed by unbiased MD trajectories projected on a selected collective variable, $q(t)$. A likelihood function $\mathcal{L}(\theta)$ is maximized (i.e., $-\log \mathcal{L}(\theta)$ is minimized) by randomly varying the parameters encoding the free energy $F(q)$ and diffusion profile $D(q)$ of an overdamped Langevin equation, until optimally modeling the original MD data. At each iteration, the new parameters are accepted or rejected according to a Metropolis criterion. The optimal Langevin model corresponds to the last accepted $F(q)$ and $D(q)$ profiles, and is used for further estimation of kinetic rates (see Methods for details).

the order of τ^2, can be explicitly obtained (using Fourier transforms[8,29]) in Gaussian form:

$$p(q', t + \tau | q, t) \approx \frac{1}{\sqrt{2\pi\mu}} e^{-(q'-q-\phi)^2/2\mu} \tag{4}$$

$$\phi = a\tau = (-\beta D F' + D')\tau, \quad \mu = 2D\tau \tag{5}$$

where the prime indicates d/dq, and a is the drift term in eq 1 (note that this propagator is exact for any $\tau > 0$ if F' and D are constant). A better approximation, still retaining the Gaussian form eq 4, can be obtained by keeping terms up to order τ^2 in eq 3, applying for instance Drozdov's approach based on the cumulant-generating function:[48]

$$\phi = a\tau + \frac{1}{2}(aa' + Da'')\tau^2,$$

$$\mu = 2D\tau + (aD' + 2a'D + DD'')\tau^2 \tag{6}$$

Armed with this explicit expression for the short-time propagator, it is possible to write the likelihood of the observed trajectory $\{q_i\}_{i=1,\dots,N}$:

$$-\log \mathcal{L}(\theta) = \sum_{i=1}^{N-1} \frac{1}{2}\log[2\pi\mu_i(\tau)]$$

$$+ \sum_{i=1}^{N-1} \frac{[q_{i+1} - q_i - \phi_i(\tau)]^2}{2\mu_i(\tau)} \tag{7}$$

where $\phi_i(\tau) \equiv \phi(q_i, \tau)$, $\mu_i(\tau) \equiv \mu(q_i, \tau)$. The optimization of the Langevin model is achieved by minimizing $-\log \mathcal{L}(\theta)$ as a function of the parameters, directly yielding the free energy and diffusion profiles. Equation 7 is minimized using a simple iterative stochastic algorithm: starting from an initial guess for the free energy and diffusion profiles (represented on a discrete grid of 1000 points), there are two types of trial moves (with probabilities 20% and 80%, respectively): (i) adding to the profiles a Gaussian at a random position, with a random width between $(q_{max} - q_{min})/20$ and $(q_{max} - q_{min})$ and height

between $-k_B T/2$ and $k_B T/2$ (with 20% probability to modify either profile, and 80% to modify both, at each optimization step), or (ii) scaling the profiles by a Gaussian-distributed random factor with mean 1 and standard deviation 0.3. The moves are accepted or rejected on the basis of a Metropolis criterion on $(-\log \mathcal{L}_{new} + \log \mathcal{L}_{old})/\tilde{T}$. The effective temperature $\tilde{T}$ is automatically scaled during the iterations to keep the acceptance close to a target value of 5%.

The algorithm we propose for the optimization of the likelihood function (Figure 1) is straightforward:

1. Start with trajectories $q(t)$ from MD and with an initial guess for the free energy and diffusion profiles.

2. Compute $-\log \mathcal{L}(\theta)$ for the initial data.

3. Make a random change of the free energy and diffusion profiles and compute the new $-\log \mathcal{L}(\theta)$.

4. Use a Metropolis criterion to accept or reject the new parameters.

5. Repeat N_{iter} times the procedure.

Note that alternatives to stochastic Monte Carlo optimization could be used; however, we appreciate the simplicity and generality of this approach, that does not require estimation of the derivatives $\partial L/\partial \theta$. In all cases, we set $N_{iter} = 10^6$.

The initial guess of the parameters can facilitate the optimization of the likelihood function: without requiring any additional data besides the projected MD trajectories, an approximate value of D in a free-energy minimum can be estimated from the variance of q divided by its autocorrelation time.[22] As sketched in Figure 1 the input trajectories $q(t)$ typically contain a final part fluctuating in the bottom of a well, that can be employed in the latter formula. The profile $D(q)$ is simply set equal to the constant estimated value. As for the initial $F(q)$, we found that the specific form is largely irrelevant. For the sake of simplicity we start from a flat profile $F(q) = 0$.

The code and input files are freely available on GitHub (https://github.com/physix-repo/optLE).

B. Integration of the Langevin Equation of Motion.
To generate input data for the benchmark systems, as well as to test the quality of the approximate propagator used in the likelihood function (see previous section), we numerically integrate Langevin trajectories using the Milstein algorithm, an order 1 strong Taylor scheme superior to Euler-Maruyama in the case of multiplicative noise (i.e., position-dependent diffusion):[49]

$$\Delta q = \left[-\beta D(q)F'(q) + \frac{1}{2}D'(q) \right]\Delta t + \sqrt{2D(q)\Delta t}\, G$$
$$+ \frac{1}{2}D'(q)\Delta t G^2 \tag{8}$$

where G is a Gaussian-distributed random number with zero mean and unit variance. The integration time step Δt is chosen after testing the observed long-time histogram of q against $e^{-\beta F(q)}$, where $F(q)$ is the exact landscape used in the Langevin equation.

C. Optimal Time Resolution of the Reduced Model.
The accuracy of the optimized Langevin model with respect to the original MD trajectory $q(t)$ crucially depends on the time resolution τ adopted, $q(t) \equiv q(k\tau)$, $k = 1, ..., N$. In fact, it is expected that only for a sufficiently large τ can the overdamped equation model accurately the MD data. To guide the choice of τ we introduce diagnostics able to predict what choices lead to accurate models.

In general, for given profiles $F(q)$ and $D(q)$, a numerical Langevin integrator can be "inverted" to estimate the value of an effective noise $G(t)$ corresponding to each observed displacement $\Delta q = q(t + \tau) - q(t)$ in the input MD trajectory. The simplest formula corresponds to the Euler-Maruyama integrator:

$$G = \frac{\Delta q + (\beta DF' - D')\tau}{\sqrt{2D\tau}} \tag{9}$$

Another option is to invert the Milstein integrator, eq 8:

$$G = \frac{1}{D'\tau}\left(-\sqrt{2D\tau} + \sqrt{(2D\tau) - (2D'\tau)\left(\frac{1}{2}D'\tau - \beta DF'\tau - \Delta q\right)} \right) \tag{10}$$

If the resulting effective noise would have been generated by the model, the $G(t)$ trajectory would be random, with zero mean, unit variance, and no time correlation: $\langle G(t)G(t')\rangle = \delta_{tt'}$ (the latter symbol indicating the Kronecker delta, since time is discretized). Therefore, to estimate the expected accuracy of the optimal Langevin model it is possible to inspect the mean, variance, and correlation time τ_{noise} of the effective noise.[27,40] For the latter quantity, we adopt here a simple operative definition, as the first time where the autocorrelation function $C(t) = \langle G(0)G(t)\rangle/\langle G^2\rangle$ drops below $1/100$. Note that in case of oscillatory behavior of $C(t)$ it is important to inspect the latter to assess if the estimate is acceptable. We find that the expressions for the noise from Euler-Maruyama (eq 9) or Milstein (eq 10) integrators yield similar results (see Supporting Information Figure S1).

As a second metric to diagnose problems in model accuracy we generate (using the integrator eq 8) $M = 100$ Langevin trajectories of duration τ and time step $\Delta t \ll \tau$ starting from each point of the input trajectory $\{q_i\}_{i=1,...,N}$. The resulting final displacements $\tilde{q}_{i+1}^k - q_i$ (with $k = 1, ..., M$) are shifted and scaled according to the theoretical propagator of the model, eq

4, by subtracting ϕ_i and dividing by $\sqrt{\mu_i}$. In this way, their distribution would be Gaussian with zero mean and unit variance if the propagator was exact, and an average likelihood can be evaluated to estimate deviations from such ideal behavior:

$$-\overline{\log \mathcal{L}_{\text{prop}}} = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{1}{M}\sum_{k=1}^{M}\frac{1}{2}\left[\log 2\pi + \left(\frac{\tilde{q}_{i+1}^k - q_i - \phi_i}{\sqrt{\mu_i}}\right)^2\right] \tag{11}$$

which yields $\frac{1}{2}[\log 2\pi + 1] \approx 1.419$ in the ideal case. For increasing τ, the propagator becomes unreliable, reducing the likelihood and hence allowing a prediction, once again, of the quality of the model (see Figures 2 and 4).
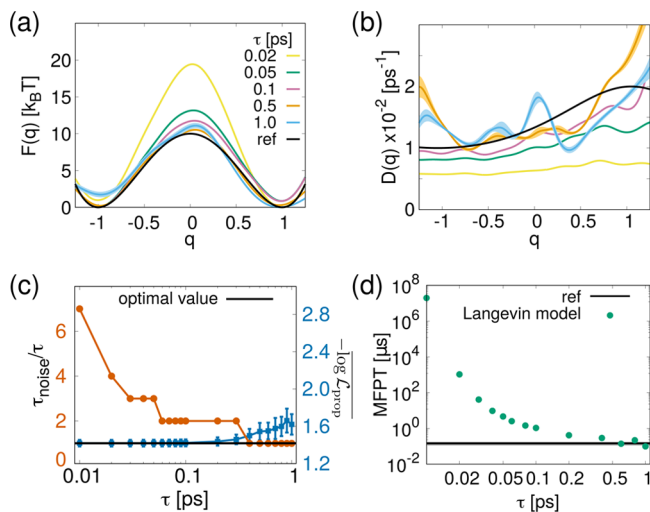
Figure 2. Optimal overdamped Langevin models (eq 1) for different resolutions τ starting from underdamped trajectories on a double-well free-energy landscape. (a) $F(q)$ and (b) $D(q)$ profiles (exact ones are shown with black lines; vertical bars correspond to standard error over 10 independent stochastic optimizations. (c) Autocorrelation time of the effective noise τ_{noise} (eq 10) normalized by τ (i.e., number of correlated steps), and average likelihood quantifying the accuracy of the propagator (eq 11). The black line represents the ideal values. Vertical bars are the standard deviation over all trajectory steps Δq. (d) MFPTs for the optimal overdamped Langevin models shown in panels a and b, computed with eq 13: the reference value in black is computed with reactive flux on the exact profiles in black in panels a and b. Vertical bars are the standard error over 10 independent Langevin models (hardly visible at this scale).

D. Kinetic Rates. In general cases, the kinetic rate for a transition A → B can be computed by means of the reactive flux (Bennett-Chandler) technique, that includes recrossing effects beyond transition state theory[50]

$$k(t) = \frac{\langle \dot{q}(0)\delta[q(0) - q^*]h_B[q(t)]\rangle}{\langle h_A\rangle}$$

where $h_A(q)$ and $h_B(q)$ are the indicator functions of metastable states A and B. Under the hypothesis of rare transition events, there is a plateau in the values of k for a broad interval of t values larger than the typical microscopic time scale (e.g., of molecular vibrations) and smaller than the MFPT. q^* is any point close to the separatrix: only efficiency is affected if they differ significantly.

The rate can be computed numerically by shooting N_s MD simulations from q^* with random atomic velocities drawn for

the Maxwell–Boltzmann distribution: for a given t, $k(t)$ is the product of the sum of initial velocities for trajectories ending up in B, divided by the total number of shots N_s, and of a term containing information about the free energy barrier

$$k(t) = \frac{\sum_{j=1}^{N_s} \dot{q}_j(0) h_B[q_j(t)]}{N_s} \frac{e^{-\beta F(q^*)}}{\int_{\Omega_A} dq\, e^{-\beta F(q)}} \qquad (12)$$

We remark that for rare transitions with overdamped one-dimensional dynamics it is also possible to compute the MFPT (the inverse of the rate) as an integral over the free energy and diffusion landscapes:

$$k^{-1}(q_0) = \int_{q_0}^{b} dy \frac{e^{\beta F(y)}}{D(y)} \int_{a}^{y} dz\, e^{-\beta F(z)} \qquad (13)$$

where, in the typical case of the escape from a well (even though the formula is general), q_0 is the initial position in the well, a is the position of a reflective boundary located on the opposite side as the transition barrier, and b is the position of an absorbing boundary located beyond the barrier.[50] In the following, we adopt as location of the absorbing boundary b the first minimum in $(d/dq)k^{-1}$ beyond the free energy barrier.

E. Molecular Dynamics Simulations. *Fullerene C_{60} Dimer.* We performed MD simulations to study the interaction between two fullerene C_{60} molecules solvated by 2398 water molecules, in a simulation box of $3.607 \times 3.607 \times 3.607$ nm with periodic boundary conditions. The initial fullerene coordinates and topology are taken from ref 51. MD simulations are carried out using GROMACS v2019.4[52,53] patched with PLUMED 2.5.3.[54] We adopted the SPC water model[55] and the OPLS-AA force-field[56] for carbon. Geometry minimization exploited the steepest descent algorithm, stopped when the maximum force was ≤50 kJ/mol·nm. We used the leapfrog algorithm to propagate the equations of motion and the nonbonding interactions were calculated using a PME scheme with a 1.2 nm cutoff for the part in real space. We performed a 100 ps equilibration in an NVT ensemble with a stochastic velocity rescaling scheme[57] followed by a 100 ps equilibration in an NPT ensemble using the Parrinello–Rahman barostat[58] with a time step of 1 fs. We generated MD production trajectories without restraints, with a time step of 1 fs in the NPT ensemble at 298 K and 1 atm. The reference free-energy profile of the association/dissociation of the fullerene C_{60} dimer in water as a function of the distance between the centers of mass (d) was computed from five unbiased simulations of 500 ns each, from the population histogram: $F(d) = -k_B T \log \rho_{eq}(q)$. We estimated the standard deviation of each histogram bin using the five simulations. The resulting profile is consistent with the free-energy profile in ref 51.

The diffusion profile used for comparison was computed from umbrella sampling simulations in which we restrict d with a harmonic potential around a reference value d_i over different windows i,

$$U(d; d_i) = \frac{c}{2}(d - d_i)^2 \qquad (14)$$

Then, the diffusion coefficient in the window i is estimated as the ratio between the variance of the variable $(\sigma_{d_i}^2)$ and the autocorrelation time of the variable itself (τ^{corr}):[22]

$$D(d_i) = \frac{\sigma_i^2}{\tau_i^{corr}} \qquad (15)$$

We setup a window every 0.1 nm in a range of d between 1.0 and 1.8 nm. We performed a 1 ns MD simulation for equilibration and a 10 ns MD simulation production in each window with a spring constant $c = 1000$ kJ/mol. For each window, we split the 10 ns MD simulations in 10 independent blocks of 1 ns each. We estimated the variance of d and the autocorrelation time in each block. Finally, we report the average $D(d_i)$ value and its standard error over the blocks. We note that $D(d)$ does not show large fluctuations with respect to the choice of c (see Figure S2).

To generate the input trajectories for the construction of Langevin models, we employed aimless-shooting (AS)[59,60] simulations. We obtained a first reactive trajectory by shooting from randomly picked configurations of the unbiased trajectories with d between 1.2 and 1.3 nm. We then performed AS using a separation $\delta t = 0.1$ ps between successive shooting points, and a total length of 20 ps for each trajectory. We obtained a total of 1110 accepted trajectories relaxing from the transition state region (see Figure 3b) with an acceptance rate of 14%. The script we developed to perform AS with GROMACS is publicly available at https://github.com/physix-repo/aimless-shooting.
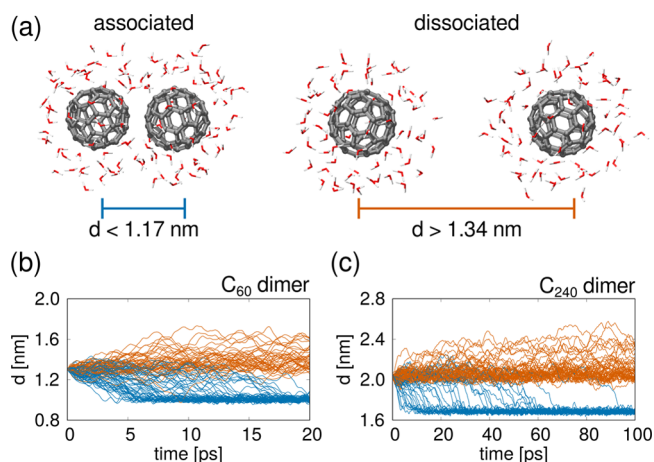
Figure 3. (a) Characteristic atomic configurations for the C_{60} dimer in bulk water solution; only waters distant less than 5 Å from carbon atoms are shown. The associated and dissociated states are defined in terms of the distance d between the centers of mass of the fullerenes. (b) 100 MD trajectories (from aimless shooting) starting from the transition state ensemble and featuring the association (blue) or dissociation (orange) of the C_{60} dimer, and (c) of the C_{240} dimer, respectively.

To estimate the dissociation rate we use the reactive flux formalism over 1000 aimless shooting trajectories. We define the dissociated state as $d \geq 1.34$ nm and the associated state as $d \leq 1.17$ nm (see Figure 3a). The value of the MFPT obtained with reactive flux was validated with the brute force estimate from 100 unbiased MD trajectories launched from the bound state: 6.1 ± 1.2 ns versus 6.5 ± 0.6 ns, respectively. Finally, we estimated the transmission coefficient as[50]

$$\kappa = \frac{k_{A \to B}^{RF}}{k_{A \to B}^{TST}} \qquad (16)$$

where $k_{A \to B}^{RF}$ is the dissociation rate constant from reactive flux (eq 12), and $k_{A \to B}^{TST}$ is the dissociation rate constant from transition state theory. The estimated transmission coefficient for the dissociation of the fullerene C_{60} dimer is 0.37, meaning that the number of recrossings before relaxation is small.

Fullerene C_{240} Dimer. We also simulated the interaction between bigger fullerenes, featuring a more stable complex: two C_{240} molecules in water solution. We used the structure of the fullerene molecule from ref 61. We built the topology automatically using the OBGMX web server,[62] solvating the fullerenes with 5375 water molecules in a simulation box of $5.22 \times 5.22 \times 5.22$ nm^3, conserving the same parameters for the Lennard-Jones potentials used in the simulation of the C_{60} fullerenes. The parameters of the MD simulations also remain the same as for C_{60}.

Given the stability of the bound state, to estimate the reference free energy profile we used well-tempered metadynamics simulations,[63] adding bias over the distance between the center of mass (d) with an initial Gaussian height of 1.0 kJ/mol and width of 0.01 nm. The bias factor was set to 8, and Gaussians were deposited every 1 ps. Five replicas of 500 ns each were simulated. We estimate the reference free energy profile and its statistical uncertainty from the average and standard deviation of the free energy profiles from all simulations. We obtained the diffusion profiles for comparison from umbrella sampling simulations following the same procedure and with the same parameters used for C_{60} fullerene dimer. We placed a window every 0.1 nm in a range of d from 1.6 to 2.6 nm. We performed AS simulations following the same procedure used for C_{60} fullerenes. We obtained 1515 accepted trajectories of 100 ps each, with an acceptance rate of 15%. We calculate the dissociation rate using 1000 trajectories and reactive flux: 9.4 ± 1.1 μs. We define the dissociated state as $d \geq 2.01$ nm and the associated state as $d \leq 1.9$ nm. The estimated transmission coefficient is 0.6.

■ RESULTS AND DISCUSSION

In the new approach, the Langevin model is parametrized from unbiased MD trajectories spanning the relevant CV region. Such trajectories can be easily obtained as spontaneous out-of-equilibrium relaxation from initial high free-energy configurations, e.g., the top of a barrier. Several effective techniques are available to discover transition states and reactive paths at a moderate computational cost:[44,64−68] although not trivial, this task is generally much less involved than reconstructing accurate free energy profiles. In the following we use as input CV-projected MD trajectories $q(t)$ obtained from TPS or by shootings from the barrier top and relaxing into the minima. Note that such trajectories are short, since in rare events the transition path time is typically very fast compared to the waiting time in a minimum, that is, the MFPT.

As an initial benchmark we verify whether our method yields Langevin models accurately reproducing the correct $F(q)$ and $D(q)$ starting from underdamped Langevin trajectories. The input data are composed by 100 short underdamped Langevin trajectories $q(t)$, each one 20 ps long, relaxing from the barrier top of a double-well landscape (see Figure S3). The exact free energy and diffusion profiles are shown in black in Figure 2.

Since the initial trajectories are generated with an underdamped integrator, building the overdamped model entails projecting out the momentum p conjugated to the CV position q. As a result, a q-based Langevin model is expected to display

memory (time correlation) in the friction and noise for τ comparable to the integration time step $\Delta t = 10^{-4}$ ps. Adopting coarse-enough τ resolutions, on the other hand, should lead to memory-less behavior, well reproduced by an overdamped model.

To test this hypothesis, in Figure 2 we generate Langevin models using 0.01 ps $\leq \tau \leq 1$ ps: for the smallest τ value, the model strongly overestimates the barrier and underestimates the diffusion coefficient. As τ increases the reconstructed model becomes more accurate (see Figure 2) until at large values of τ the error increases again. We found that there is a relatively narrow range of τ values for which the Langevin overdamped model is accurate: to identify the optimal range in real-world applications we propose two complementary tests that allow locating the upper and lower boundaries of τ bracketing the high-quality Langevin models.

The lower acceptable τ is the smallest time displaying no memory effects, measured through the autocorrelation of the effective Langevin noise, that is, the noise extracted a posteriori by analyzing the original trajectory with the optimal Langevin model, eq 10 (see Figure S1 for details). The upper limit on τ corresponds to the onset of a significant deviation between the approximate and the exact propagators (the latter estimated from accurate numerical integration, see methods section C and eq 11). Both diagnostics are reported in Figure 2c, and together they predict that the Langevin model should attain the best accuracy for τ values close to 0.5 ps: large enough to avoid memory, but small enough for the short-time propagator to remain reliable.

Overall, when selecting τ based on the proposed diagnostics, models converge smoothly and rapidly ($<10^6$ Monte Carlo steps) to a good approximation of the exact results: as few as 100 reference trajectories are sufficient to reconstruct the free energy profile to within 1 $k_B T$ and $D(q)$ to within 15% error (we remark that fluctuations in $D(q)$ appear converged with respect to the different parameters used in the reconstruction of the models and irrespective of the input trajectories). The same conclusions can be derived regardless of the shape of the diffusion and free energy profiles (see Figure S4). We remark that, when the input trajectories are generated with an overdamped integrator, the reconstructed profiles deviate less than 1 $k_B T$ and 10%, respectively, from the exact $F(q)$ and $D(q)$, regardless of τ, unless when it is too large for our approximate propagator (see Figure S5).

Besides estimating free energy and diffusion profiles, a key objective of the new method is the accurate prediction of kinetic rates at a moderate computational cost. Figure 2d shows the MFPTs obtained with eq 13 using the optimal Langevin models. Too small values of τ lead to strongly overestimated MFPT, as expected from the overestimated barriers, while the rate becomes accurate for τ values within the predicted optimal window (Figure 2c). On the other hand, for τ larger than optimal, MFPTs in the same order of magnitude as the reference value are still obtained.

We note that, alternatively, we can compute the MFPT using the reactive flux technique,[69,70] free from transition state approximations. This traditionally requires two distinct sets of expensive simulations, one to compute the free energy barrier and another to compute a correlation function accounting for recrossings (see the methods section D and Table S1). However, this technique becomes inexpensive if combined with our approach, since the free energy barrier is directly obtained from likelihood optimization on short trajectories,
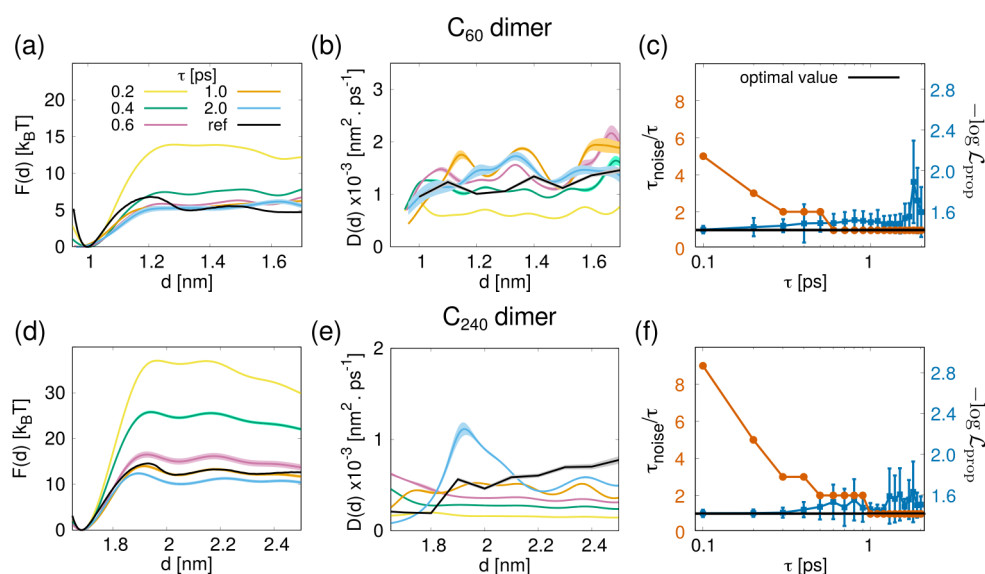
Figure 4. Optimal Langevin models for the interaction of C_{60} and C_{240} fullerene dimers in explicit water, for different values of the time resolution τ. The input data set consists of 100 MD trajectories (aimless shooting) relaxing from the transition state ensemble in Figure 3. (a,d) Free-energy profiles $F(q)$, compared to the reference ones in black: vertical bars correspond to standard deviations over 10 independent stochastic optimizations. (b,e) Diffusion profiles $D(q)$, compared to the umbrella sampling ones in black: vertical bars correspond to standard deviations over 10 independent stochastic optimizations. (c,f) Autocorrelation time of the effective noise (eq 10) normalized by τ (i.e., number of correlated steps), and average likelihood quantifying the accuracy of the propagator (eq 11). The black line represents the ideal value for both quantities; vertical bars refer to the standard deviation over all trajectory steps Δq.

while recrossing statistics can be accurately estimated from inexpensive synthetic Langevin trajectories. We remark that the rates computed from eq 13 are similar to those of reactive flux (see Table S2).

The double-well potentials in the previous section allowed us to benchmark the scope and limitations of the method while having full control over the system. To assess the method in realistic complex systems, we modeled the dynamics of two fullerene molecules in explicit water, considering two different sizes C_{60} and C_{240} (Figure 3). The free-energy landscape as a function of the distance d between fullerenes' centers of mass features a minimum for the bound complex (with different depth for the two sizes) and a relatively flat region for the dissociated state, see black lines in Figure 4. The resulting rare dissociation events and diffusion-controlled association events[51,71−73] are analogous to other processes such as protein−protein interaction. We obtained reference free-energy profiles using 2.5 μs-long brute-force MD trajectories for the C_{60} dimer, while we employed well-tempered metadynamics[63] for the C_{240} dimer due to its very low dissociation rate. The diffusion profiles are obtained from sets of 10 ns umbrella sampling simulations every 0.1 nm in distance between the center of mass (see the methods section E for details).

We maximized the likelihood of Langevin models starting from short aimless shooting[59] TPS trajectories (20/100 ps each, cumulative duration 2/10 ns for C_{60}/C_{240}, see Figure 3). An example of the convergence of $-\log \mathcal{L}(\theta)$ as a function of the number of iterations is shown in Figure S6. We adopted a range of time resolutions τ between 0.1 and 2 ps: for fast time scales, memory is expected to arise from the projection of the solvated fullerenes many-body dynamics on a single CV. Sizable memory effects are evident in the effective noise (eq 10) for $\tau < 0.5$ ps (see also Figure S1), severely affecting the accuracy of $F(d)$. For larger τ values, similarly to the case of the

analytic double-well potential, the Langevin models become increasingly accurate: memory decays after about 0.5 ps for the smaller fullerenes and after about 1.0 ps for the bigger ones. As for the diffusion profile, all models show $D(d)$ in the same order of magnitude, in good agreement with the diffusion coefficients estimated from umbrella sampling[22] (see methods section E). We remark that generating 100 short TPS trajectories for Langevin optimization can be computationally faster than generating a sufficient amount of umbrella sampling trajectories to estimate the diffusion profile.

On the other hand, the quality of the approximate short-time propagator eq 6 decreases as τ increases. Accurate results are clearly identified in the range of τ that minimizes both non-Markovian behavior and error propagation (see Figure 4). Interestingly, optimizing the models based on 1000 reference trajectories instead of 100 yields minimal improvements (see Figure S7).

Finally, we assessed the accuracy of kinetic predictions given by the models. The dissociation MFPT is on the nanosecond and microsecond scales, respectively, for C_{60} and C_{240} dimers; that is, 2 and 5 orders of magnitude slower than the out-of-equilibrium MD trajectories used for training. For each value of τ, we computed MFPTs using eq 13 on the optimal models, yielding Figure 5. We emphasize that the latter approach allows computation of the MFPT without generating any additional expensive MD trajectories.

For too small values of τ, the barrier is overestimated (see Figure 4), leading to overestimation of the MFPTs by several orders of magnitude. For the C_{60} dimer the MFPT becomes accurate (compared to the brute force result) for $\tau > 0.5$ ps, while for the C_{240} dimer this level of accuracy requires $\tau > 1$ ps: once again, these are the optimal time resolutions predicted by the diagnostics in Figure 4. Taken together, all these results demonstrate that the approximations inherent in the Langevin
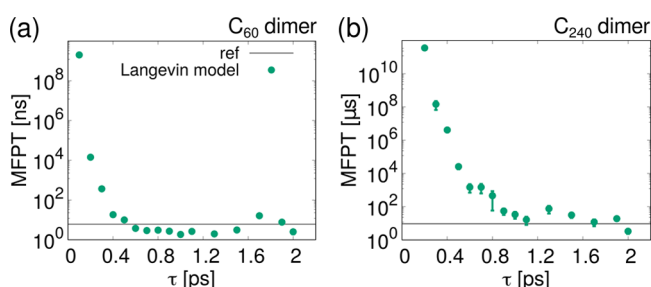
Figure 5. MFPTs computed from the optimal Langevin models generated with different time resolutions τ, for the dissociation of fullerene (a) C_{60} and (b) C_{240} dimer, in water solution. MFPTs are estimated from eq 13 (reference value in black). Vertical bars are the standard error over 10 independent Langevin models.

models are under control, leading to accurate predictions about the thermodynamics and kinetics of complex systems.

CONCLUDING REMARKS

The calculation of free energy landscapes and kinetic rates is a key task of computer simulations of complex systems. Even though these two tasks are usually tackled using different ad hoc techniques,[2] the main contribution of the present work consists in demonstrating that they can be simultaneously achieved in a conceptually simple way, by estimating Langevin models starting from a relatively inexpensive set of about one hundred TPS-like trajectories, regardless of the barrier height. Likelihood maximization—an efficient parameter estimation technique—is combined with a double test bracketing the time resolution τ, granting control over the accuracy of the model. The choice of a suitable τ is especially important for accurately estimating the free energy, because the rate has an exponential dependence on the barrier while only having a linear dependence on D.

The resulting Markovian Langevin equations reproduce well the quantitative thermodynamic and dynamic properties of the original many-body system, including accurate kinetic rates, despite a gap of many orders of magnitude with respect to the short MD trajectories used for training. Moreover, the short time transition probability $p(q', t + \tau|q, t)$ predicted by the optimal Langevin model (eq 4) is in good agreement with a distribution collected from short MD shootings of length τ (see example in Figure S8). Incidentally, we note that our approach is not restricted to transition path sampling: in principle, any set of unbiased trajectories spanning the regions of interest could be used to train the Langevin model (see example in Figure S9).

We also remark that equilibrium properties are systematically recovered from out-of-equilibrium data: standard transition path sampling trajectories are the golden standard for the study of transformation mechanisms; however, since such a data set is a small subset of all possible (reactive and nonreactive) pathways, lacking Boltzmann distribution of the configurations, it cannot be used for the direct estimate of equilibrium histograms (i.e., free energies) and rate matrices (e.g., in Markov state models) by simple averaging.[4,74] Here we show that the contrary is true, provided bare transition paths are employed to train a suitable stochastic model (see also refs 21, 22, 29, and 75−77 for related or alternative ideas). Note that Langevin equations of motion, compared to other machine-learning tools, retain a direct physical interpretation, with the separation between a systematic average force and

friction/noise effects describing the projected-out degrees of freedom (commonly referred to as "the bath").

In future applications of the new approach, two main issues have to be taken into consideration. First, different kinds of Langevin equation (generalized, standard, or overdamped) can be necessary to faithfully reproduce the projected dynamics of a complex system, depending on the physical process, the choice of the CV, and the observational time scale. For chemical reactions in water the memory time scale could be comparable to the transition path time, requiring a non-Markovian equation, whereas for crystal nucleation or protein folding Markovian models are customarily invoked. In the present work, we demonstrated the approach using over-damped Langevin equations: in principle, this can be generalized to other stochastic models by replacing the propagator expression (eq 6) with a different suitable approximation, based on ideas in refs 43 and 48.

A second issue concerns the effect of projecting the dynamics on suboptimal CVs, rather than on the ideal reaction coordinate, commonly identified with the committor function.[15,78,79] In principle, the CV definition can be systematically optimized with an iterative scheme, in which a single initial set of MD trajectories can be employed to build optimal Langevin models for progressively improving CVs. A related idea was recently proposed in the case of discrete Markov state models.[80,81] Therefore, our new approach represents a starting point both for the systematic modeling of MD data through physically motivated stochastic equations as well as for machine-learning approaches to reaction coordinate optimization.

ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c00324.

> Detailed results for different shapes of the double-well benchmark systems, autocorrelation functions of the effective noise, convergence tests, and kinetic rate estimates using reactive flux (PDF)

AUTHOR INFORMATION

Corresponding Author

Fabio Pietrucci − *Muséum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, Sorbonne Université, F-75005 Paris, France;* ⓘ orcid.org/0000-0002-4892-2667; Email: fabio.pietrucci@sorbonne-universite.fr

Author

Karen Palacio-Rodriguez − *Muséum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, Sorbonne Université, F-75005 Paris, France;* ⓘ orcid.org/0000-0003-4327-8478

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jctc.2c00324

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Pietrucci, F. Strategies for the exploration of free energy landscapes: unity in diversity and challenges ahead. *Rev. Phys.* **2017**, *2*, 32−45.

(2) Camilloni, C.; Pietrucci, F. Advanced simulation techniques for the thermodynamic and kinetic characterization of biological systems. *Adv. Phys. X* **2018**, *3*, 1477531.

(3) Dellago, C.; Bolhuis, P. G. *Transition path sampling and other advanced simulation techniques for rare events*; Springer: Berlin, Heidelberg, 2008.

(4) Bolhuis, P. G.; Swenson, D. W. Transition Path Sampling as Markov Chain Monte Carlo of Trajectories: Recent Algorithms, Software, Applications, and Future Outlook. *Advanced Theory and Simulations* **2021**, *4*, 2000237.

(5) Van Erp, T. S.; Bolhuis, P. G. Elaborating transition interface sampling methods. *J. Comput. Phys.* **2005**, *205*, 157−181.

(6) Allen, R. J.; Warren, P. B.; Ten Wolde, P. R. Sampling rare switching events in biochemical networks. *Phys. Rev. Lett.* **2005**, *94*, 018104.

(7) Zwanzig, R. *Nonequilibrium statistical mechanics*; Oxford University Press, 2001.

(8) Risken, H. *The Fokker-Planck Equation*; Springer, 1996.

(9) Łuczka, J. Non-Markovian stochastic processes: Colored noise. *Chaos* **2005**, *15*, 026107.

(10) Grote, R. F.; Hynes, J. T. The stable states picture of chemical reactions. II. Rate constants for condensed and gas phase reaction models. *J. Chem. Phys.* **1980**, *73*, 2715−2732.

(11) Lee, H. S.; Ahn, S.-H.; Darve, E. F. *Building a Coarse-Grained Model Based on the Mori-Zwanzig Formalism*; MRS Online Proceedings Library Archive, 2015; p 1753.

(12) Daldrop, J. O.; Kappler, J.; Brünig, F. N.; Netz, R. R. Butane dihedral angle dynamics in water is dominated by internal friction. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, 5169−5174.

(13) Van Kampen, N. G. *Stochastic processes in physics and chemistry*; Elsevier, 1992; Vol. *1*.

(14) Coffey, W.; Kalmykov, Y. P. *The Langevin equation: with applications to stochastic problems in physics, chemistry and electrical engineering*; World Scientific, 2012.

(15) Lu, J.; Vanden-Eijnden, E. Exact dynamical coarse-graining without time-scale separation. *J. Chem. Phys.* **2014**, *141*, 044109.

(16) Straub, J. E.; Borkovec, M.; Berne, B. J. Calculation of dynamic friction on intramolecular degrees of freedom. *J. Phys. Chem.* **1987**, *91*, 4995−4998.

(17) Tuckerman, M.; Berne, B. Vibrational relaxation in simple fluids: Comparison of theory and simulation. *J. Chem. Phys.* **1993**, *98*, 7301−7318.

(18) Timmer, J. Parameter estimation in nonlinear stochastic differential equations. *Chaos, Solitons & Fractals* **2000**, *11*, 2571−2578.

(19) Gradišek, J.; Siegert, S.; Friedrich, R.; Grabec, I. Analysis of time series from stochastic processes. *Phys. Rev. E* **2000**, *62*, 3146.

(20) Chorin, A. J.; Hald, O. H.; Kupferman, R. Optimal prediction with memory. *Phys. D* **2002**, *166*, 239−257.

(21) Hummer, G.; Kevrekidis, I. G. Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations. *J. Chem. Phys.* **2003**, *118*, 10762−10773.

(22) Hummer, G. Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J. Phys.* **2005**, *7*, 34.

(23) Best, R. B.; Hummer, G. Diffusive model of protein folding dynamics with Kramers turnover in rate. *Phys. Rev. Lett.* **2006**, *96*, 228104.

(24) Lange, O. F.; Grubmüller, H. Collective Langevin dynamics of conformational motions in proteins. *J. Chem. Phys.* **2006**, *124*, 214903.

(25) Horenko, I.; Hartmann, C.; Schütte, C.; Noe, F. Data-based parameter estimation of generalized multidimensional Langevin processes. *Phys. Rev. E* **2007**, *76*, 016706.

(26) Darve, E.; Solomon, J.; Kia, A. Computing generalized Langevin equations and generalized Fokker−Planck equations. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 10884−10889.

(27) Micheletti, C.; Bussi, G.; Laio, A. Optimal Langevin modeling of out-of-equilibrium molecular dynamics simulations. *J. Chem. Phys.* **2008**, *129*, 074105.

(28) Berezhkovskii, A.; Szabo, A. Time scale separation leads to position-dependent diffusion along a slow coordinate. *J. Chem. Phys.* **2011**, *135*, 074108.

(29) Zhang, Q.; Brujić, J.; Vanden-Eijnden, E. Reconstructing free energy profiles from nonequilibrium relaxation trajectories. *J. Stat. Phys.* **2011**, *144*, 344−366.

(30) Crommelin, D.; Vanden-Eijnden, E. Diffusion estimation from multiscale data by operator eigenpairs. *Multiscale Modeling & Simulation* **2011**, *9*, 1588−1623.

(31) Crommelin, D. Estimation of space-dependent diffusions and potential landscapes from non-equilibrium data. *J. Stat. Phys.* **2012**, *149*, 220−233.

(32) Schaudinnus, N.; Bastian, B.; Hegger, R.; Stock, G. Multi-dimensional Langevin modeling of nonoverdamped dynamics. *Phys. Rev. Lett.* **2015**, *115*, 050602.

(33) Schaudinnus, N.; Lickert, B.; Biswas, M.; Stock, G. Global Langevin model of multidimensional biomolecular dynamics. *J. Chem. Phys.* **2016**, *145*, 184114.

(34) Lesnicki, D.; Vuilleumier, R.; Carof, A.; Rotenberg, B. Molecular hydrodynamics from memory kernels. *Phys. Rev. Lett.* **2016**, *116*, 147804.

(35) Meloni, R.; Camilloni, C.; Tiana, G. Properties of low-dimensional collective variables in the molecular dynamics of biopolymers. *Phys. Rev. E* **2016**, *94*, 052406.

(36) Biswas, M.; Lickert, B.; Stock, G. Metadynamics enhanced Markov modeling of protein dynamics. *J. Phys. Chem. B* **2018**, *122*, 5508−5514.

(37) Freitas, F. C.; Lima, A. N.; Contessoto, V. d. G.; Whitford, P. C.; Oliveira, R. J. d. Drift-diffusion (DrDiff) framework determines kinetics and thermodynamics of two-state folding trajectory and tunes diffusion models. *J. Chem. Phys.* **2019**, *151*, 114106.

(38) Nüske, F.; Boninsegna, L.; Clementi, C. Coarse-graining molecular systems by spectral matching. *J. Chem. Phys.* **2019**, *151*, 044116.

(39) Baldovin, M.; Cecconi, F.; Vulpiani, A. Effective equations for reaction coordinates in polymer transport. *Journal of Statistical Mechanics: Theory and Experiment* **2020**, *2020*, 013208.

(40) Lickert, B.; Stock, G. Modeling non-Markovian data using Markov state and Langevin models. *J. Chem. Phys.* **2020**, *153*, 244112.

(41) Wang, S.; Ma, Z.; Pan, W. Data-driven coarse-grained modeling of polymers in solution with structural and dynamic properties conserved. *Soft Matter* **2020**, *16*, 8330−8344.

(42) Ayaz, C.; Tepper, L.; Brünig, F. N.; Kappler, J.; Daldrop, J. O.; Netz, R. R. Non-Markovian modeling of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*.

(43) Vroylandt, H.; Goudenège, L.; Monmarché, P.; Pietrucci, F.; Rotenberg, B. Likelihood-based non-Markovian models from molecular dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119*, No. e2117586119.

(44) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291−318.

(45) Tiwary, P.; Limongelli, V.; Salvalaglio, M.; Parrinello, M. Kinetics of protein−ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, E386−E391.

(46) Carmeli, B.; Nitzan, A. Theory of activated rate processes: Position dependent friction. *Chem. Phys. Lett.* **1983**, *102*, 517−522.

(47) Leimkuhler, B.; Matthews, C. *Molecular Dynamics*; Springer, 2016.

(48) Drozdov, A. N. High-accuracy discrete path integral solutions for stochastic processes with noninvertible diffusion matrices. *Phys. Rev. E* **1997**, *55*, 2496.

(49) Kloeden, P. E.; Platen, E.; Schurz, H. *Numerical solution of SDE through computer experiments*; Springer Science & Business Media, 2012.

(50) Jungblut, S.; Dellago, C. Pathways to self-organization: Crystallization via nucleation and growth. *Eur. Phys. J. E* **2016**, *39*, 1−38.

(51) Monticelli, L. On atomistic and coarse-grained models for C 60 fullerene. *J. Chem. Theory Comput.* **2012**, *8*, 1370−1378.

(52) Berendsen, H. J.; van der Spoel, D.; van Drunen, R. GROMACS: a message-passing parallel molecular dynamics implementation. *Computer physics communications* **1995**, *91*, 43−56.

(53) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to super-computers. *SoftwareX* **2015**, *1*, 19−25.

(54) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **2014**, *185*, 604−613.

(55) Berendsen, H. J.; Postma, J. v.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(56) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(57) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(58) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.

(59) Peters, B.; Trout, B. L. Obtaining reaction coordinates by likelihood maximization. *J. Chem. Phys.* **2006**, *125*, 054108.

(60) Peters, B.; Beckham, G. T.; Trout, B. L. Extensions to the likelihood maximization approach for finding reaction coordinates. *J. Chem. Phys.* **2007**, *127*, 034109.

(61) Noël, Y.; De La Pierre, M.; Zicovich-Wilson, C. M.; Orlando, R.; Dovesi, R. Structural, electronic and energetic properties of giant icosahedral fullerenes up to C6000: insights from an ab initio hybrid DFT study. *Phys. Chem. Chem. Phys.* **2014**, *16*, 13390−13401.

(62) Garberoglio, G. OBGMX: A web-based generator of GROMACS topologies for molecular and periodic systems using the universal force field. *Journal of computational chemistry* **2012**, *33*, 2204−2208.

(63) Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters* **2008**, *100*, 020603.

(64) Izrailev, S.; Stepaniants, S.; Isralewitz, B.; Kosztin, D.; Lu, H.; Molnar, F.; Wriggers, W.; Schulten, K. *Computational Molecular Dynamics*; Springer: Berlin, Heidelberg, 1999; pp 39−65.

(65) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562−12566.

(66) Best, R. B.; Hummer, G. Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6732−6737.

(67) Samanta, A.; Tuckerman, M. E.; Yu, T.-Q.; Weinan, E. Microscopic mechanisms of equilibrium melting of a solid. *Science* **2014**, *346*, 729−732.

(68) Jung, H.; Okazaki, K.-i.; Hummer, G. Transition path sampling of rare events by shooting from the top. *J. Chem. Phys.* **2017**, *147*, 152716.

(69) Chandler, D. Statistical mechanics of isomerization dynamics in liquids and the transition state approximation. *J. Chem. Phys.* **1978**, *68*, 2959−2970.

(70) Hänggi, P.; Talkner, P.; Borkovec, M. Reaction-rate theory: fifty years after Kramers. *Rev. Mod. Phys.* **1990**, *62*, 251−341.

(71) Banerjee, S. Molecular dynamics study of self-agglomeration of charged fullerenes in solvents. *J. Chem. Phys.* **2013**, *138*, 044318.

(72) Li, L.; Bedrov, D.; Smith, G. D. A molecular-dynamics simulation study of solvent-induced repulsion between C 60 fullerenes in water. *J. Chem. Phys.* **2005**, *123*, 1−8.

(73) Chiu, C. C.; DeVane, R.; Klein, M. L.; Shinoda, W.; Moore, P. B.; Nielsen, S. O. Coarse-grained potential models for phenyl-based molecules: II. Application to fullerenes. *J. Phys. Chem. B* **2010**, *114*, 6394−6400.

(74) Bolhuis, P. G.; Dellago, C. Trajectory-Based Rare Event Simulations. *Rev. Comput. Chem.* **2010**, *27*, 111.

(75) Sriraman, S.; Kevrekidis, I. G.; Hummer, G. Coarse master equation from Bayesian analysis of replica molecular dynamics simulations. *J. Phys. Chem. B* **2005**, *109*, 6479−6484.

(76) Innerbichler, M.; Menzl, G.; Dellago, C. State-dependent diffusion coefficients and free energies for nucleation processes from Bayesian trajectory analysis. *Mol. Phys.* **2018**, *116*, 2987−2997.

(77) Brotzakis, Z. F.; Bolhuis, P. G. Approximating free energy and committor landscapes in standard transition path sampling using virtual interface exchange. *J. Chem. Phys.* **2019**, *151*, 174111.

(78) Banushkina, P. V.; Krivov, S. V. Optimal reaction coordinates. *WIREs: Comput. Mol. Sci.* **2016**, *6*, 748−763.

(79) Peters, B. Reaction Coordinates and Mechanistic Hypothesis Tests. *Annu. Rev. Phys. Chem.* **2016**, *67*, 669−690.

(80) Tiwary, P.; Berne, B. J. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 2839−2844.

(81) Chen, W.; Tan, A. R.; Ferguson, A. L. Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design. *J. Chem. Phys.* **2018**, *149*, 072312.