A variational method for efficient estimation of diffusion and free-energy profiles along collective variables

Anže Hubman[1, 2] and Franci Merzel[1]
[1)]*Laboratory for molecular structural dynamics, Theory department, National Institute of Chemistry, Hajdrihova 19, 1001 Ljubljana, Slovenia*
[2)]*Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia.*

(*Electronic mail: franci.merzel@ki.si)

(Dated: 24 March 2026)

An efficient variational method is presented for estimating the diffusion coefficients and free-energy profiles along selected collective variables from projected molecular dynamics trajectories under both equilibrium and non-equilibrium conditions. The method is based on the assumption that the short-time transition probability density of the coordinate moves can be approximated by a Gaussian form. Defining a loss function as the sum of Kullback-Leibler divergences between the analytical short-time propagators of an overdamped Langevin model and those estimated directly from the projected trajectories maximizes the agreement between the two and allows for its analytic evaluation. For cases where Gaussian approximation is insufficient we present a robust alternative. To efficiently minimize this loss function by varying diffusion and free-energy profiles along collective variables, we use an adaptive Monte Carlo scheme. The method is applied to two model systems exhibiting diffusive dynamics, as well as to water diffusion across the interface of a biomolecular condensate, demonstrating its robustness and accuracy.

I. INTRODUCTION

Molecular dynamics (MD) simulations have advanced to the point where they can be used to quantitatively model complex phenomena, including protein conformational dynamics[1,2], self-assembly of biomolecules in solution[3], permeation of water and ions through biological membranes[4,5], chemical reactions on catalytic surfaces[6], and phase transitions in crystalline materials[7,8]. In MD simulations, the equations of motion are integrated numerically, which produces trajectories in the $6N$-dimensional phase space (N being the number of particles) that describe the time evolution of particle positions and momenta. The structural, dynamical, and kinetic properties of the system can be obtained from an a posteriori analysis of the simulated trajectories[9].

The high dimensionality of MD trajectories typically requires dimensionality reduction for their analysis[10]. A low-dimensional effective description of the system can be constructed by projecting the trajectories onto a reduced set of collective variables (CVs), which are functions of the generalized coordinates that characterize key states of the system, such as reactants, products, and transition states[11]. The construction of optimal CVs is problem-specific and is guided by physical intuition[12] or data-driven methods[13,14]. Although several collective variables can be used simultaneously[15], we shall focus here on the simplest case of a one-dimensional CV.

The chosen collective variable q uniquely defines the associated free-energy profile $F(q)$, which can be determined accurately and efficiently using enhanced sampling techniques[15–17]. Formally, the projected dynamics follows the non-Markovian generalized Langevin equation (GLE), which is valid over arbitrary timescales[18–20]. In addition to the free-energy profile, the GLE depends on the memory kernel, which encodes memory effects responsible for non-Markovian dynamics. Despite significant progress, accurate parameterization of the memory kernel remains numerically challenging and requires long trajectories sampled at high temporal resolution[21–30].

However, on timescales longer than the characteristic memory time, non-Markovian effects become negligible. In this limit, the GLE reduces to the Markovian underdamped Langevin equation, in which the memory kernel is replaced by a friction term that can depend on both the instantaneous value of q and its corresponding velocity $\dot{q}$[31]. If, in addition, the timescales of interest exceed the autocorrelation time of $\dot{q}$, the temporal evolution of q at inverse temperature β is well approximated by the Markovian overdamped Langevin equation[32]:

$$\dot{q} = v(q) + \sqrt{2D(q)}\,\eta(t), \tag{1}$$

in which the generalized drift $v(q)$ is expressed as the sum of the thermodynamic force due to the free-energy profile $F(q)$ and the drift originating from the position-dependent diffusion coefficient $D(q)$:

$$v(q) = -D(q)\,\beta\,\frac{\partial F(q)}{\partial q} + \frac{\partial D(q)}{\partial q}. \tag{2}$$

Here, $\eta(t)$ denotes Gaussian white noise, which satisfies $\langle \eta(t) \rangle = 0$ and $\langle \eta(t)\eta(t') \rangle = \delta(t-t')$.

Overdamped Langevin models have been successfully used to describe the kinetics of protein folding[33,34], conformational transitions of proteins[35], looping of a polymer chain in solution[36], nucleation of vapor bubbles[37], and the dynamics of self-association between pairs of fullerenes in water[38]. To obtain an accurate description of the kinetics, it is often necessary for the diffusion coefficient to depend on the collective variable. However, extracting the diffusion profile $D(q)$ from the time series of q (trajectories) is challenging.

The primary goal of several existing methods for inferring $D(q)$ is to construct a likelihood function that quantifies the

probability of observing a specific trajectory in the collective-variable space, given the model parameters $F(q)$ and $D(q)$. To construct the likelihood, an explicit expression for the propagator $p(q',\tau|q,0)$ is required. The propagator defines the conditional probability of observing a system in a state q' at time τ, given that it was in a state q at time $t=0$.

Building on the work of Bicout and Szabo[39], Hummer[40] discretized the collective-variable space and expressed the local propagators of an overdamped Langevin model in terms of the lag-time τ and a rate matrix $\mathbf{R}$ that describes transitions between adjacent cells and encodes information on $F(q)$ and $D(q)$. In this framework, the likelihood depends on the number of transitions that occur between neighboring cells within the time interval τ and on the rate matrix $\mathbf{R}$, whose elements are treated as parameters. The elements of $\mathbf{R}$ that best reproduce the observed trajectory along q are obtained by Bayesian inference or by maximizing the likelihood, yielding self-consistent estimates of $F(q)$ and $D(q)$. A key requirement of this approach is that the lag-time τ be chosen such that a sufficient number of nearest-neighbor transitions between cells are observed.

Alternatively, when the chosen τ is sufficiently short, the analytical approximations for the propagators can be used[38,41]. In this case, $F(q)$ and $D(q)$ are directly varied to maximize the likelihood of the observed trajectory. The choice of τ is crucial in this method, as it must be short enough for the analytical approximation to hold, yet long enough for the condition of Markovianity to be satisfied. Although this approach has strong statistical foundations, evaluating the likelihood function at every optimization step becomes computationally expensive for long trajectories. A related approach is to perform the likelihood maximization analytically with respect to the drift $v(q)$ and the diffusion profile $D(q)$, which gives the first and second Kramers-Moyal coefficients that link the conditional averages and variances of the displacements $\Delta q = q(t+\tau)-q(t)$ to the drift and diffusion terms in Eq. (1)[36,42–44]. While this avoids variational optimization, it requires verification that the reconstructed $F(q)$ from $v(q)$ and $D(q)$ is consistent with the true free-energy profile.

Other methods avoid explicit likelihood construction and are generally less sensitive to the choice of τ. However, they are typically more computationally demanding. Liu *et al.*[45] introduced a dual-simulation approach, in which Langevin dynamics simulations are performed with a known free-energy profile, and the diffusion coefficient is adjusted until the simulated survival probability matches that obtained directly from MD. Woolf and Roux[46] estimated $F(q)$ and $D(q)$ from a series of locally restrained simulations. Pérez-Villa and Pietrucci[31] optimized the free-energy, friction, and mass profiles by matching the time-dependent probability density $P(q,\dot{q},t)$ obtained from a few hundred short MD trajectories, initiated at the transition state, to that generated by Langevin dynamics. This approach is particularly appealing because it applies to both underdamped and overdamped regimes. For a comprehensive review of methods for determining $D(q)$, see the work of Domingues et al.[47] and the references therein.

In this work, we present a simple and computationally efficient variational approach for estimating the diffusion and, optionally, free-energy profiles along selected collective variables from projected MD trajectories. The method is based on a loss function designed to maximize the agreement between the analytical approximation for the short-τ propagators of an overdamped Langevin model, defined by a given $F(q)$ and $D(q)$, and the corresponding propagators directly estimated from the projected trajectories. For cases where analytical approximations for the propagator are insufficient, we present a robust, though slightly more computationally expensive alternative. Furthermore, we show that an adaptive Monte Carlo minimization scheme emerges naturally from the construction of the loss function. The robustness of the approach is demonstrated on two model systems exhibiting diffusive dynamics under both equilibrium and non-equilibrium conditions, and by analyzing the diffusion of water between the dense and dilute phases of a biomolecular condensate.

II. METHODS

A. Variational optimization

We outline a variational procedure for determining the position-dependent diffusion coefficient $D(q)$ along a single collective variable q. The method can be extended to multiple dimensions if substantially more data is available. Unless stated otherwise, the free-energy profile is assumed to have been estimated independently.

Consider a discrete time series $q(t_i)=q_i$, sampled at a temporal resolution $\Delta t = t_{i+1}-t_i$. We define the lag-time τ as $\tau = k\Delta t$, where k is a positive integer. If τ is chosen such that the dynamics of q is both overdamped and Markovian, the temporal evolution of q can be described by the overdamped Langevin equation, parameterized in terms of the free-energy profile $F(q)$ and the diffusion profile $D(q)$. Under these conditions, the conditional probability $p(q',\tau|q,0)$ of observing a transition from q to q' in a time τ is given by the propagator of the overdamped Langevin model, which depends only on the free energy and diffusion profile.

Given a recorded time series of q, the propagators can be estimated directly. We refer to these as "empirical" propagators and denote them by $\hat{p}$. To perform the estimation, we assume that $0 \le q(t) < L$ and discretize the collective variable space into N bins of equal width $\delta = L/N$. Along the observed trajectory, we collect the displacements $\Delta q = q(t+\tau)-q(t)$ and assign each displacement to the j-th bin if $j\delta \le q(t) < (j+1)\delta$. The probability density of the collected displacements then corresponds to the propagator for bin j, which becomes exact in the limit $N\to\infty$.

To obtain a self-consistent estimate of $D(q)$ (and optionally $F(q)$), we require that, for a given choice of free-energy and diffusion profiles, the propagators p_j predicted by the Langevin model match the empirical propagators $\hat{p}_j$ as closely as possible. More specifically, we introduce a loss function $\mathscr{L}$ defined as the sum of Kullback-Leibler (KL) divergences $\mathscr{D}_{\mathrm{KL}}$

between p_j and $\hat{p}_j$:

$$\mathscr{L} = \mathscr{L}_{KL} + \mathscr{L}_{\lambda} = \frac{1}{N}\sum_{j=0}^{N-1} \mathscr{D}_{\mathrm{KL}}(p_j \| \hat{p}_j) + \lambda \sum_{j=0}^{N-2} [D(q_{j+1}) - D(q_j)]^2, \tag{3}$$

where the second term imposes smoothness on $D(q)$ and is equivalent to the smoothening prior in the method of Hummer[40]. A similar term can be added to regularize $F(q)$, which becomes important when $F(q)$ and $D(q)$ are estimated simultaneously. A practical procedure of choosing the optimal λ will be given in the next Section. The KL divergence for bin j is defined as:

$$\mathscr{D}_{\mathrm{KL},j} = \int_{-\infty}^{\infty} p_j(q) \log \frac{p_j(q)}{\hat{p}_j(q)}\, dq, \tag{4}$$

and vanishes if $p_j(q) = \hat{p}_j(q)$, meaning that the loss function in Eq. (3) is bounded from below by $\mathscr{L} = 0$ in the case of $\lambda = 0$ or constant $D(q)$. To establish the connection between $F(q)$ and $D(q)$ and the model propagator, we present two approaches.

The first approach is suitable when a sufficiently short lagtime τ is used. In this case, the propagator can be written explicitly and takes a Gaussian form. Adopting the notation of Palacio-Rodriguez and Pietrucci[38], we write:

$$p(q', \tau | q, 0) = \frac{1}{\sqrt{2\pi\mu}} \exp\left[-\frac{(q' - q - \phi)^2}{2\mu}\right]. \tag{5}$$

In the simplest (first-order) approximation, we have:

$$\phi = v(q)\tau, \tag{6a}$$

$$\mu = 2D(q)\tau, \tag{6b}$$

while in the second-order approximation[38,48] one finds:

$$\phi = v(q)\tau + \frac{1}{2}\left[v(q)v'(q) + D(q)v''(q)\right]\tau^2, \tag{7a}$$

$$\mu = 2D(q)\tau + \left[v(q)D'(q) + 2v'(q)D(q) + D(q)D''(q)\right]\tau^2. \tag{7b}$$

As will be shown in the next Section, the second-order approximation permits the use of larger τ than the first-order approximation, as it includes higher-order corrections that account for the curvature of $F(q)$ and $D(q)$. The Gaussian (normal) form of the propagator is particularly convenient, as it is fully determined by the first two moments of the distribution. In practice, this means that one needs only to compute the empirical mean displacement $\hat{\phi}$ and variance $\hat{\mu}$ when estimating the propagators. Furthermore, for two normal distributions $p_j = \mathscr{N}(\phi, \mu)$ and $\hat{p}_j = \mathscr{N}(\hat{\phi}, \hat{\mu})$, the integral in Eq. (4) can be evaluated analytically, yielding

$$\mathscr{D}_{KL,j} = \frac{1}{2}\log\frac{\hat{\mu}_j}{\mu_j} + \frac{\mu_j + (\phi_j - \hat{\phi}_j)^2}{2\hat{\mu}_j} - \frac{1}{2}. \tag{8}$$

The second approach addresses cases where large τ leads to a non-Gaussian propagator. This occurs when the time series q is sampled at low temporal resolution or when a large τ is required to satisfy the Markovian condition. To proceed, we recall that the overdamped Langevin equation corresponds to the Fokker–Planck (Smoluchowski) equation for the probability density $P(q,t)$:

$$\frac{\partial P(q,t)}{\partial t} = \frac{\partial}{\partial q}\left\{ D(q) e^{-\beta F(q)} \frac{\partial}{\partial q}\left[e^{\beta F(q)} P(q,t)\right]\right\}. \tag{9}$$

The Green's function of Eq. (9) is the propagator. In principle, numerically integrating Eq. (9) from a delta-function initial condition yields an exact propagator. In practice, however, a delta function cannot be realized numerically.

To construct a practical numerical scheme, we again estimate N propagators from the observed time series of q. While estimating the first two moments was sufficient in the first approach, we now represent each propagator as a histogram on a grid of M points. The value of M is chosen to provide sufficient spatial resolution for the discretized probability density. Similarly, we approximate each of the N initial conditions from the time series and represent them on the same grid. For a given choice of $F(q)$ and $D(q)$, the Fokker–Planck equation is numerically propagated from each initial condition for a duration τ using the scheme of Bicout and Szabo[39] (see SI for details). This allows us to evaluate the propagator and estimate the KL-divergence by performing numerical integration in Eq. (4).

In both approaches, the loss $\mathscr{L}$ is a functional of $F(q)$ and $D(q)$. We minimize $\mathscr{L}$ using a Monte Carlo (MC) scheme, which is preferred over gradient-based algorithms as it avoids the need to compute the gradient of the loss function. Furthermore, the MC approach easily enforces the strict positivity of $D(q)$. By representing $F(q)$ and $D(q)$ on a grid of N points, the functional minimization is reduced to a multivariate optimization over the grid values. Each MC step consists of proposing a small perturbation to a selected grid value, which is then accepted or rejected based on the Metropolis criterion. In this context, the loss serves as an effective energy function (see SI for details). Importantly, the computational cost of evaluating the loss scales with the number of grid points N when Gaussian propagators are used, whereas maximum-likelihood-based methods scale with the trajectory length[38].

A significant advantage of our loss function is that minimization can be performed very efficiently. We achieve this by converting the set of KL-divergences $\mathscr{D}_{\mathrm{KL},j}$ from a previous optimization step into a discrete probability distribution $\mathscr{P}$:

$$\mathscr{P}_j = \frac{\mathscr{D}_{\mathrm{KL},j}}{\sum_{j=0}^{N-1} \mathscr{D}_{\mathrm{KL},j}}. \tag{10}$$

Rather than choosing move locations randomly, new MC perturbations are sampled from $\mathscr{P}$. This strategy allows the algorithm to adaptively focus on regions of $D(q)$ that contribute most strongly to the loss. The adaptive sampling is particularly effective in the first approach, where a high local KL-divergence typically indicates a locally incorrect $D(q)$.

Another aspect of our optimization procedure is that we typically keep $F(q)$ fixed at its initial estimate, although we

will demonstrate that this is not strictly necessary. We observed that fixing $F(q)$ reduces the number of MC steps required and improves robustness, particularly in the presence of small free energy barriers. A limitation of keeping the free energy profile fixed is that any error in $F(q)$ propagates to the estimate of $D(q)$. This issue can be mitigated by performing additional optimization steps in which both $F(q)$ and $D(q)$ are allowed to vary.

B. Example 1: Diffusion on a periodic domain

We tested the method on a one-dimensional model system undergoing diffusive dynamics on a periodic domain[40]. The free-energy and diffusion profiles were defined as:

$$\beta F(q) = -\cos(2q) + \text{const.}, \tag{11a}$$

$$D(q) = D_0[2 + \sin(q)], \tag{11b}$$

where $0 \leq q < 2\pi$ and $D_0 = 0.1$ rad^2 ps^{-1}. The trajectory used for the analysis was generated by integrating the overdamped Langevin equation for 100 ns with a timestep of 10^{-2} ps using the Euler–Maruyama integrator.

The diffusion profile was reconstructed using $N = 24$ local propagators across a range of lag times τ, employing both first- and second-order Gaussian propagator approximations, as well as an approach based on the Fokker–Planck equation (see Results and Discussion). The minimizations were initialized from a uniform diffusion profile, $D(q) = 0.2$ rad^2 ps^{-1}. The analytical expression for the free-energy was used to isolate the estimation of $D(q)$ from inaccuracies in $F(q)$. The Fokker-Planck equation was solved numerically on a grid of $M = 200$ points with an integration timestep of $\Delta t = 10^{-3}$ ps. Statistical uncertainties were quantified by dividing the trajectory into four equal-sized blocks and estimating $D(q)$ for each block independently.

C. Example 2: Diffusion across the interface of a biomolecular condensate

To demonstrate the applicability of our method to a realistic system, we simulated phase coexistence between the dilute and dense phases of the low-complexity domain of the Fused in Sarcoma protein (FUS-LCD) using the coarse-grained Martini3-IDP force field[49]. An equilibrated configuration consisting of 36 protein chains was taken from the work of Wang *et al.*[49] and placed at the center of an elongated simulation box with dimensions 12 nm × 12 nm × 50 nm. The system was solvated with 44,504 water molecules (see Fig. 1). A salt concentration of 0.1 M was achieved by adding 488 Na^+ and 416 Cl^- ions. After a short equilibration period in the NPT ensemble, the system was simulated for 150 ns in the NVT ensemble at $T = 300$ K using GROMACS 2025.2[50]. Trajectories were recorded every 4 ps. The default parameters recommended for the Martini3-IDP force field and periodic boundary conditions were used.

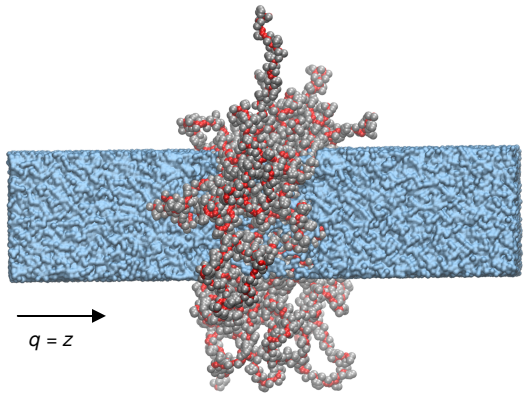

FIG. 1. Coexistence between the dilute and dense phases of the FUS-LCD condensate simulated using the slab method. Because of the relatively short simulation time, no exchange of proteins between the protein-rich (dense) phase (shown in red) and the water-rich (dilute) phase (shown in blue) was observed. Backbone and side-chain beads are shown in red and gray, respectively, and the solvent is depicted in blue. Sodium and chloride ions are omitted.

We focused on the position-dependent diffusion coefficient of water molecules along the z direction, i.e., perpendicular to the interface. The collective variable q was defined as the z coordinate of water molecules. The free-energy profile was obtained from:

$$F(q) = -\beta^{-1} \ln \rho(q), \tag{12}$$

where $\rho(q)$ is the normalized histogram of q obtained from the MD trajectory. The position-dependent diffusion coefficient, $D(q)$, was estimated using $N = 30$ local propagators and a lag-time of $\tau = 200$ ps. Minimizations were initiated from a uniform diffusion profile $D(q) = 0.2$ Å^{2}ps^{-1}. Statistical uncertainties were quantified by dividing the trajectory into three 50-ns blocks, with the free-energy and diffusion profiles estimated independently for each block.

To assess the accuracy of the estimated $F(q)$ and $D(q)$, the distributions of first-passage times obtained from molecular dynamics trajectories were compared with those predicted by a Langevin model. Assuming that the dense phase is centered at $q = 0$, a first-passage event was defined as the time required for a molecule initially located within $q < \pm L/10$ to exit the dense phase, whose boundaries were set at $q = \pm L/4$, where L is the length of the simulation box along the z direction. First-passage times were extracted from the MD trajectories by analyzing the motion of water molecules along the z coordinate. Langevin dynamics simulations, with initial conditions randomly sampled near $q = 0$, were performed using the estimated $F(q)$ and $D(q)$, and the resulting trajectories were analyzed identically to the MD data.

D. Example 3: Relaxation from the free-energy barrier

In the third example, we simulated diffusive dynamics in a double-well potential. The free-energy and diffusion profiles were given by:

$$\beta F(q) = 8(q^2-1)^2, \tag{13a}$$

$$D(q) = D_0\left[2+\exp\left(-\frac{q^2}{2\sigma^2}\right)\right], \tag{13b}$$

where $D_0 = 0.01$ ps^{-1} and $\sigma = 0.3$. In contrast to the first model system, where the collective-variable space was easily explored, this potential contains a much higher free-energy barrier of $8k_BT$, making ergodic sampling difficult. To explore the region between the two minima, we followed the approach of Ref.[38] and relaxed short trajectories from the top of the barrier. The trajectories were generated using the Euler–Maruyama integrator with a timestep of 10^{-3} ps and propagated for 2×10^4 steps, sufficient to observe relaxation into one of the minima.

This example was designed to test whether a relatively small set of non-equilibrium trajectories is sufficient to recover $F(q)$ and $D(q)$ with our method. Since the free-energy profile cannot be estimated by histogramming, the minimization was initialized with $F(q) = 0$ and $D(q) = 0.02$ ps^{-1}. In each Monte Carlo step, the decision to perturb either $F(q)$ or $D(q)$ was made randomly. We used $N = 36$ local propagators and a lag-time of $\tau = 10^{-2}$ ps. Statistical uncertainties were quantified by performing ten independent sets of short simulations and estimating the diffusion profile (and the free-energy) for each set. Each set contained an ensemble of 50 trajectories.

III. RESULTS AND DISCUSSION

We assess the performance of our method using three representative examples introduced in the previous Section. The diffusion profiles were estimated independently for each trajectory block. The free-energy profiles were estimated per block as well, but only in the second and third examples. All reported results correspond to averages over the blocks, with error bars representing $\pm\sigma$, where σ is the standard deviation across blocks. Unless stated otherwise, it is assumed that $\lambda = 0$.

A. Diffusion on a periodic domain

First, we examine diffusion on a periodic domain for three different values of the lag-time τ. As shown in Fig. 2, the simplest (first-order) Gaussian approximation of the propagator accurately reproduces the true diffusion profile only at very short lag-times, $\tau = 0.05$ ps, while for larger τ, its accuracy deteriorates significantly (Figs. 2B and 2C). Because of the same level of approximation used in the Kramers–Moyal (KM) analysis, and no regularization ($\lambda = 0$) applied during variational fitting, both approaches exhibit similar performance.

Using a second-order Gaussian approximation extends the validity of the method to substantially larger τ (Fig. 2B), because it accounts for the curvature of the free-energy and diffusion profile by including up to third-order derivatives of $F(q)$ and $D(q)$. In our tests (not shown), discrepancies between the true and recovered $D(q)$ begin to appear for $\tau > 0.2$ ps near $q = \pi/2$. In contrast, the approximation-free approach based on the Fokker–Planck (FPE) equation is insensitive to the choice of τ and correctly recovers the true diffusion profile even when the Gaussian approximations fail (Fig. 2C).

As stated in the Methods section, the number of Monte Carlo (MC) steps required to reach convergence can be reduced by using an adaptive (biased) scheme for proposing new MC moves. Importantly, we confirmed that the choice of scheme does not affect the inferred $D(q)$. In Fig. 3A, we show that biased moves reduce the number of optimization steps by approximately a factor of 1.6. Furthermore, Fig. 3B demonstrates that the algorithm correctly identifies regions where deviations between the initial guess and the true diffusion profile are the largest. In addition, the local KL divergences used to construct weights for proposing new MC move locations become progressively lower and more uniform as the optimization proceeds.

Next, we illustrate how regularization can be systematically applied during variational fitting, which is not possible in the Kramers-Moyal approach. For demonstration, we generated a relatively short trajectory of 0.5 ns. As shown in Fig. 4B, using the first-order Gaussian propagator with non-regularized fitting ($\lambda = 0$) on such a short trajectory produces a noisy estimate of $D(q)$. A similar result would be obtained using the standard Kramers–Moyal analysis. Applying the regularization requires selecting an optimal regularization strength λ in order to obtain solutions that are as smooth as possible while remaining consistent with the data. We determined λ using L-curve analysis[51], where, for several trial values, the loss function is plotted on the x-axis and the regularization term on the y-axis (Fig. 4A). The optimal λ is chosen from the region where the curve exhibits a characteristic kink (red dot in Fig. 4A). In our case, visual inspection indicates $\lambda \approx 0.2$. To quantify the effect of regularization, we computed the root-mean-squared deviation (RMSD) between the estimated and true diffusion profiles:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(D_i^{\mathrm{est.}} - D_i^{\mathrm{true}})^2}, \tag{14}$$

where i runs over the bins used to discretize $D(q)$. As shown in Fig. 4C, applying optimal regularization reduces the RMSD by a factor of 1.6 and yields significantly smoother $D(q)$ (Fig. 4B).

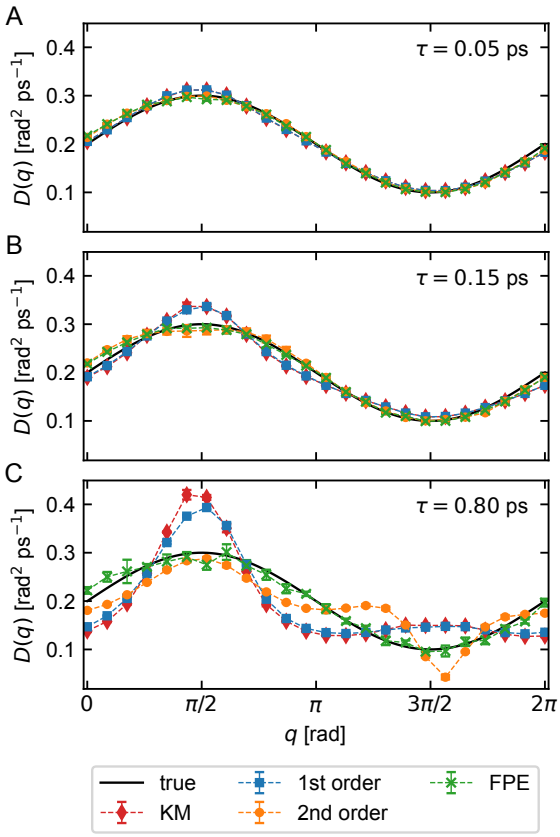

FIG. 2. Diffusion profiles estimated using the Kramers–Moyal expansion (red), the variational method with first- (blue) and second-order (orange) Gaussian approximations of the propagator, and the variational method based on the Fokker–Planck equation (green), for (a) $\tau = 0.05$ ps, (b) $\tau = 0.15$ ps, and (c) $\tau = 0.80$ ps. Dotted lines are shown as guides to the eye.

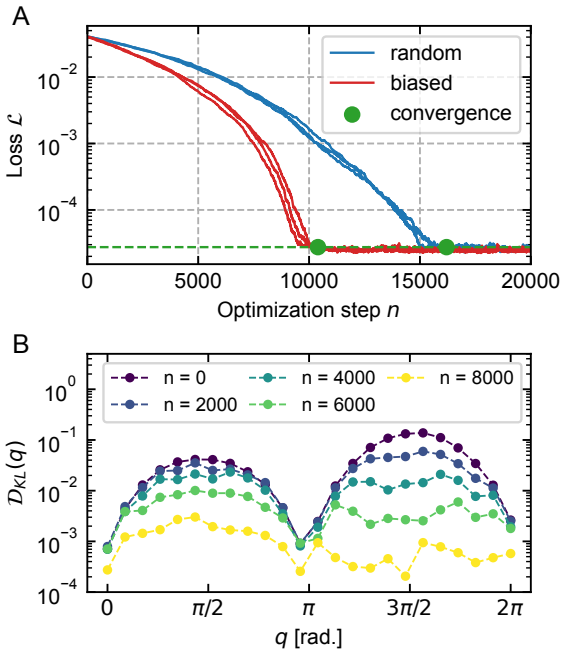

FIG. 3. (a) Comparison of the convergence rates for random (blue) and biased (red) Monte Carlo move locations. Each line represents a separate realization of the minimization procedure. (b) An example evolution of local KL-divergences during Monte Carlo minimization of the loss.

B. Diffusion across the interface of a biomolecular condensate

In the second example, we examine the diffusion of water across the interface of a biomolecular condensate. Figure 5A shows the free-energy profile along the z-direction, which exhibits a small barrier of approximately $0.5k_BT$. This is consistent with the rapid exchange of water molecules between the coexisting phases of biomolecular condensates[52].

To select an appropriate lag-time τ that satisfies the condition that the dynamics of q is both overdamped and Markovian, we first computed the velocity autocorrelation function (VACF) for water beads in the dense and dilute phases separately (see SI for details). Both VACFs decay to zero for $\tau > 2$ ps, indicating that the dynamics is overdamped beyond this threshold. Next, we computed $D(q)$ for several values of τ ranging between 8.0 ps and 400 ps (Fig. S3), and observed that the diffusion profile fully converges for $\tau > 80$ ps. Importantly, convergence occurs already at lower τ in the region of $D(q)$ corresponding to the dilute phase. This is expected because the dense phase creates a crowded environment where non-Markovian effects are typically stronger. Larger τ values are therefore required to satisfy Markovian condition.

Figure 5B presents the estimated diffusion profile using $\tau = 200$ ps. The final estimates of $D(q)$ are independent of the specific Monte Carlo move proposal scheme. We obtain $D_{\text{wat}} = 0.196 \pm 0.003$ Å^2/ps in the dilute phase and $D_{\text{wat}} \approx 0.037 \pm 0.001$ Å^2/ps in the dense phase. These values are in excellent agreement with those of Wang *et al.*[49], who reported $D_{\text{wat}} = 0.199 \pm 0.007$ Å^2/ps and 0.032 ± 0.012 Å^2/ps for the dilute and dense phases, respectively. Their analysis was based on the local mean-squared displacements, which cannot be applied in regions where the curvature of the free-energy is nonzero, whereas our approach remains valid throughout the entire interfacial region.

Fig. 6 compares the efficiency of optimization schemes differing in how new MC moves are proposed. Although both approaches converge to effectively the same minimum loss, biased proposals require approximately 2.7 times fewer optimization steps. The exact performance gain depends on the number of bins used to discretize $D(q)$ and on the fraction of bins where $D(q)$ can be given a good initial guess. This makes our method particularly suitable for interfacial systems, where the diffusion coefficient can be accurately determined far from the interface using the Einstein relation, allowing the algorithm to focus optimization on the interfacial region.

Finally, Fig. 7 compares first-passage-time (FPT) distributions obtained directly from MD simulations with those predicted by the Langevin model. This serves as a self-consistency test, because both $F(q)$ and $D(q)$ influence FPT

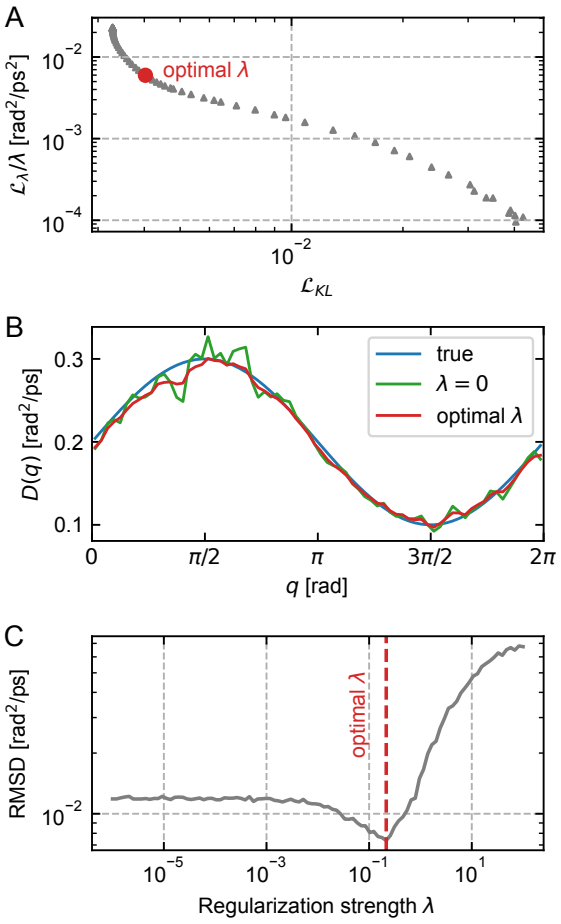

FIG. 4. (a) Non-regularized loss versus regularization term for different values of the regularization strength λ. (b) Comparison of the inferred $D(q)$ from a short trajectory using non-regularized (green) and optimally regularized (red) variational fitting. (c) Root-mean-squared deviation (RMSD) between the inferred and true diffusion profiles for a range of regularization strengths.

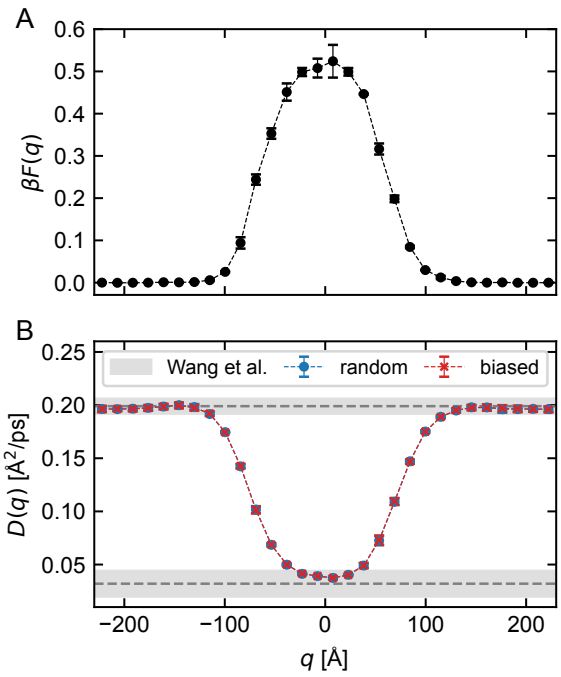

FIG. 5. (a) Free-energy profile along the z-coordinate of water molecules estimated using the standard histogramming technique. (b) The corresponding position-dependent diffusion profiles computed using random (blue symbols) and biased (red symbols) Monte Carlo move proposals. Reference values of diffusion coefficient, including error bars, for Martini3 water beads (denoted by the grey interval) refer to the work of Wang et al.[49]. Dotted lines are shown as guides to the eye.

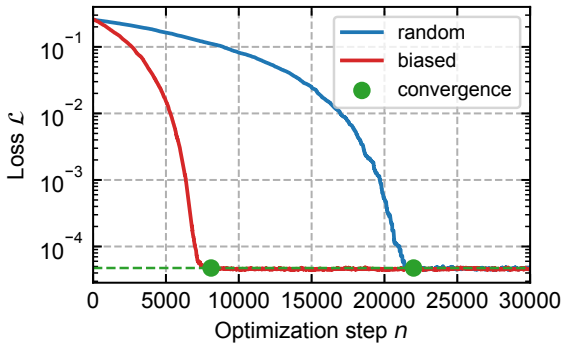

FIG. 6. Comparison of convergence rates for random (blue) and biased (red) Monte Carlo move proposal schemes applied to water diffusion across the interface of a FUS-LCD condensate.

distributions, especially when the free-energy barrier is small. The Langevin model accurately reproduces the MD results as evidenced by a close overlap of both FPT distributions. Furthermore, the mean first-passage time from MD is $\langle\tau_{\mathrm{FP}}\rangle = 41 \pm 3$ ns, while the Langevin model yields $\langle\tau_{\mathrm{FP}}\rangle = 37.7 \pm 0.7$ ns.

C. Relaxation from the free-energy barrier

The third example differs fundamentally from the preceding two because both $F(q)$ and $D(q)$ are inferred from non-equilibrium trajectories. Two cases are considered. In the first case (setup 1, blue symbols in Fig. 8B), the analytical form of the free energy is fixed while only the diffusion profile is optimized. This setup represents a situation in which the free-energy profile has been determined independently using one of the enhanced sampling techniques. In the second case (setup 2, red symbols in Figs. 8A and 8B), both $F(q)$ and $D(q)$ are optimized simultaneously, corresponding to a situation where only the transition-state location is known in advance, but not the detailed shape of $F(q)$.

In both cases, the true diffusion profile is accurately recovered. The free-energy profile is also well reproduced when both $F(q)$ and $D(q)$ are optimized. As expected, the second case requires substantially more optimization steps, with the exact number depending on the maximum allowed Monte

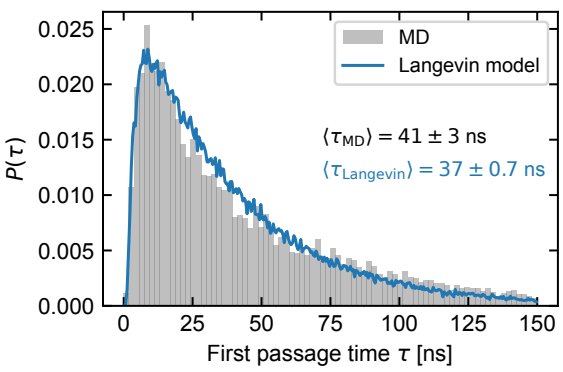

FIG. 7. Comparison between the first-passage-time distributions of water molecules leaving the dense phase computed from MD simulations (gray) and those predicted by a Langevin model (blue).

Carlo step size. Overall, this example demonstrates that even a limited set of non-equilibrium trajectories contains sufficient information to reliably extract both $F(q)$ and $D(q)$ using our approach, in agreement with Refs.[31,38].

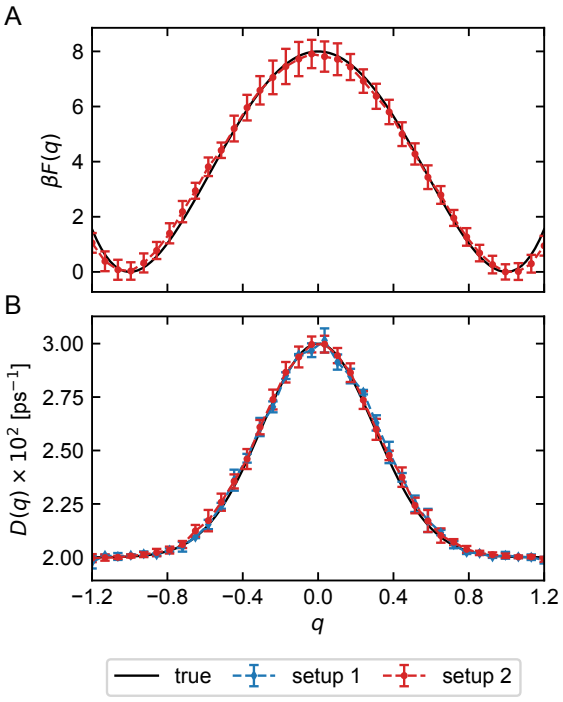

FIG. 8. (a) Comparison between the true (black line) and reconstructed free energy profile (blue symbols). (b) Comparison between the true diffusion profile (black line) and the reconstructed profiles obtained by optimizing either only $D(q)$ (red symbols) or both $F(q)$ and $D(q)$ simultaneously. Dotted lines are shown as guides to the eye.

IV. CONCLUSIONS

In summary, we have presented a variational method for an efficient parameterization of overdamped Langevin models from projected molecular dynamics trajectories in terms of free-energy and diffusion profiles. When a Markovian time series with high temporal resolution is available, the method is particularly efficient because analytical approximations for the propagators can be employed. For more coarsely sampled trajectories, we proposed a robust but computationally more demanding alternative. A further advantage of our approach is that evaluating the loss function is computationally inexpensive. This efficiency gain is particularly relevant for recently developed methods for the automatic learning of optimal reaction coordinates, which require re-parameterization of the Langevin model at every iteration of the learning procedure[53].

We also emphasized efficient minimization of the loss function and demonstrated how effective Monte Carlo moves can be generated within our framework. Similar speed-ups could be achieved in likelihood-based methods by retaining successful Monte Carlo proposals to guide future updates, although this would require additional tuning of parameters that control how past proposals are replaced by new ones.

Applying the method on three diverse representative systems encompassing both equilibrium and non-equilibrium conditions demonstrates its robustness, efficiency, and accuracy. Our aim for future work is to extend the present framework to multidimensional collective variable spaces.

SUPPLEMENTARY MATERIAL

Supplementary material includes additional detailes about the Monte Carlo optimization, water diffusion dependence on the lag-time in the biomolecular condensate and discretization of the Fokker-Planck equation.

ACKNOWLEDGMENTS

The authors acknowledge the financial support under Grant Nos. P1-0010 and J1-50033 from the Slovenian Research and Innovation Agency. The use of ChatGPT-5 and Gemini 3.0 for assistance in improving the clarity and readability of the manuscript text is also acknowledged.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Anže Hubman: Conceptualization (equal); Formal analysis (lead); Methodology (equal); Validation (lead); Visualization (lead); Writing - original draft (lead). **Franci Merzel**: Conceptualization (lead); Formal analysis (equal); Funding acquisition (lead); Methodology (equal); Supervision (lead); Validation (equal); Writing - original draft (supporting); Writing - review and editing (equal).

DATA AVAILABILITY STATEMENT

The computer codes enabling full reproduction of Example 1 in the manuscript are available at `https://github.com/AnzeHubman/Position-Diffusion`. All other data are available from the authors upon request.

[1] K. Arora and C. L. Brooks, PNAS **104**, 18496 (2007).
[2] N. Plattner and F. Noé, Nat. Commun. **6**, 7653 (2015).
[3] Z. Benayad, S. von Bülow, L. S. Stelzl, and G. Hummer, J. Chem. Theory Comput. **17**, 525 (2021).
[4] M. Sever and F. Merzel, Int. J. Mol. Sci. **24**, 10528 (2023).
[5] Y. Zhuang, C. M. Noviello, R. E. Hibbs, R. J. Howard, and E. Lindahl, PNAS **119**, e2208081119 (2022).
[6] S. Perego, L. Bonati, S. Tripathi, and M. Parrinello, ACS Catal. **14**, 14652 (2024).
[7] J. G. Lee, C. J. Pickard, and B. Cheng, J. Chem. Phys. **156**, 074106 (2022).
[8] C. Verdi, F. Karsai, P. Liu, R. Jinnouchi, and G. Kresse, Npj Comput. Mater. **7**, 1 (2021).
[9] D. Frenkel and B. Smit, *Understanding molecular simulation* (Academic Press, 2023).
[10] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, and A. Laio, Chem. Rev. **121**, 9722 (2021).
[11] C. Abrams and G. Bussi, Entropy **16**, 163 (2013).
[12] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, Comp. Phys. Comm. **185**, 604 (2014).
[13] A. France-Lanord, H. Vroylandt, M. Salanne, B. Rotenberg, A. M. Saitta, and F. Pietrucci, J. Chem. Theory Comput. **20**, 3069 (2024).
[14] W. Chen, A. R. Tan, and A. L. Ferguson, J. Chem. Phys. **149**, 072312 (2018).
[15] A. Laio and M. Parrinello, PNAS **99**, 12562 (2002).
[16] G. Torrie and J. Valleau, J. Comp. Phys. **23**, 187 (1977).
[17] E. Darve and A. Pohorille, J. Chem. Phys. **115**, 9169 (2001).
[18] H. Mori, Prog. Theor. Phys. **33**, 423 (1965).
[19] R. Zwanzig, Phys. Rev. **124**, 983 (1961).
[20] D. D. Girardier, H. Vroylandt, S. Bonella, and F. Pietrucci, J. Chem. Phys. **159**, 164111 (2023).
[21] E. Darve, J. Solomon, and A. Kia, PNAS **106**, 10884 (2009).
[22] G. Jung, M. Hanke, and F. Schmid, J. Chem. Theory Comput. **13**, 2481 (2017).
[23] O. F. Lange and H. Grubmüller, J. Chem. Phys. **124**, 214903 (2006).
[24] C. Hijón, P. Español, E. Vanden-Eijnden, and R. Delgado-Buscalioni, Faraday Discuss. **144**, 301 (2010).
[25] H. Vroylandt, L. Goudenège, P. Monmarché, F. Pietrucci, and B. Rotenberg, PNAS **119**, e2117586119 (2022).
[26] V. Klippenstein, M. Tripathy, G. Jung, F. Schmid, and N. F. A. van der Vegt, J. Phys. Chem. B **125**, 4931 (2021).
[27] P. Xie and W. E, J. Chem. Theory Comput. **20**, 7708 (2024).
[28] P. Xie, R. Car, and W. E, PNAS **121**, e2308668121 (2024).
[29] B. A. Dalton, A. Klimek, H. Kiefer, F. N. Brünig, H. Colinet, L. Tepper, A. Abbasi, and R. R. Netz, Annu. Rev. Phys. Chem. **76**, 431 (2025).
[30] L. Tepper, B. Dalton, and R. R. Netz, J. Chem. Theory Comput. **20**, 3061 (2024).
[31] A. Pérez-Villa and F. Pietrucci, (2018), arXiv:1810.00713 [cond-mat.stat-mech].
[32] H. Risken, *The Fokker-Planck Equation* (Springer, 1996).
[33] R. B. Best and G. Hummer, Phys. Rev. Lett. **96**, 228104 (2006).
[34] R. B. Best and G. Hummer, PNAS **107**, 1088 (2010).
[35] R. Hegger and G. Stock, J. Chem. Phys. **130**, 034106 (2009).
[36] C. Micheletti, G. Bussi, and A. Laio, J. Chem. Phys. **129**, 074105 (2008).
[37] M. Innerbichler, G. Menzl, and C. Dellago, Mol. Phys. **116**, 2987 (2018).
[38] K. Palacio-Rodriguez and F. Pietrucci, J. Chem. Theory Comput. **18**, 4639 (2022).
[39] D. J. Bicout and A. Szabo, J. Chem. Phys. **109**, 2325 (1998).
[40] G. Hummer, New J. Phys. **7**, 34 (2005).
[41] J. Comer, C. Chipot, and F. D. González-Nilo, J. Chem. Theory Comput. **9**, 876 (2013).
[42] F. Sicard, V. Koskin, A. Annibale, and E. Rosta, J. Chem. Theory Comput. **17**, 2022 (2021).
[43] M. Baldovin, F. Cecconi, and A. Vulpiani, "Effective equations for reaction coordinates in polymer transport," Journal of Statistical Mechanics: Theory and Experiment **2020**, 013208 (2020).
[44] A. Vulpiani and M. Baldovin, "Effective equations in complex systems: from langevin to machine learning," Journal of Statistical Mechanics: Theory and Experiment **2020**, 014003 (2020).
[45] P. Liu, E. Harder, and B. J. Berne, J. Phys. Chem. B **108**, 6595 (2004).
[46] T. B. Woolf and B. Roux, J. Am. Chem. Soc. **116**, 5916 (1994).
[47] T. S. Domingues, R. Coifman, and A. Haji-Akbari, "Estimating position-dependent and anisotropic diffusivity tensors from molecular dynamics trajectories: Existing methods and future outlook," Journal of Chemical Theory and Computation **20**, 4427–4455 (2024).
[48] A. N. Drozdov, "High-accuracy discrete path integral solutions for stochastic processes with noninvertible diffusion matrices," Physical Review E **55**, 2496–2508 (1997).
[49] L. Wang, C. Brasnett, L. Borges-Araújo, P. C. T. Souza, and S. J. Marrink, Nat. Commun. **16**, 2874 (2025).
[50] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," SoftwareX **1–2**, 19–25 (2015).
[51] P. C. Hansen, "Analysis of discrete ill-posed problems by means of the l-curve," SIAM Review **34**, 561–580 (1992).
[52] W. Zheng, G. L. Dignon, N. Jovic, X. Xu, R. M. Regy, N. L. Fawzi, Y. C. Kim, R. B. Best, and J. Mittal, "Molecular details of protein condensates probed by microsecond long atomistic simulations," The Journal of Physical Chemistry B **124**, 11671–11679 (2020).
[53] L. Mouaffac, K. Palacio-Rodriguez, and F. Pietrucci, "Optimal reaction coordinates and kinetic rates from the projected dynamics of transition paths," Journal of Chemical Theory and Computation **19**, 5701–5711 (2023).