Non-Markovian modeling of protein folding

Cihan Ayaz[a], Lucas Tepper[a], Florian N. Brünig[a], Julian Kappler[b], Jan O. Daldrop[a], and Roland R. Netz[a,1]

[a]Fachbereich Physik, Freie Universität Berlin, 14195 Berlin, Germany; and [b]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, United Kingdom

Edited by Ken A. Dill, Stony Brook University, Stony Brook, NY, and approved June 12, 2021 (received for review November 17, 2020)

We extract the folding free energy landscape and the time-dependent friction function, the two ingredients of the generalized Langevin equation (GLE), from explicit-water molecular dynamics (MD) simulations of the α-helix forming polypeptide alanine$_9$ for a one-dimensional reaction coordinate based on the sum of the native H-bond distances. Folding and unfolding times from numerical integration of the GLE agree accurately with MD results, which demonstrate the robustness of our GLE-based non-Markovian model. In contrast, Markovian models do not accurately describe the peptide kinetics and in particular, cannot reproduce the folding and unfolding kinetics simultaneously, even if a spatially dependent friction profile is used. Analysis of the GLE demonstrates that memory effects in the friction significantly speed up peptide folding and unfolding kinetics, as predicted by the Grote–Hynes theory, and are the cause of anomalous diffusion in configuration space. Our methods are applicable to any reaction coordinate and in principle, also to experimental trajectories from single-molecule experiments. Our results demonstrate that a consistent description of protein-folding dynamics must account for memory friction effects.

protein folding | non-Markovian processes | mean first-passage times | generalized Langevin equation | memory effects

Biological macromolecular function relies on coupled processes that take place on widely different timescales; this makes the theoretical description of such systems challenging. For proteins, the topic of this paper, folding occurs in the range of microseconds to many minutes or even hours and involves bond vibrations and hydration water motion on subpicosecond times (1–3). In order to enable large-scale simulations as well as meaningful theories, which should concentrate on the essential features of such processes, several methods for the elimination of irrelevant degrees of freedom have been introduced. For the classical dynamics of an interacting many-body system, the rigorous treatment is based on the Liouville equation and employs the projection operator formalism to integrate out all degrees of freedom except one or a few reaction coordinates (4, 5). Instead of $6N$ equations of motion for all positions and momenta of an N-particle system, the dynamics is described by few equations for the observables of interest. This coarse-graining procedure leads from a deterministic Hamiltonian to a stochastic description by the generalized Langevin equation (GLE), which for the case of a one-dimensional coordinate $q(t)$, reads (4–7)

$$m\ddot{q}(t) = -\nabla U[q(t)] - \int_0^t \mathrm{d}s\, \Gamma(t-s)\dot{q}(s) + F_R(t), \qquad [1]$$

where m is the effective mass of the coordinate q. The potential of mean force $U(q)$, which for proteins, corresponds to the folding free energy landscape, is obtained from the equilibrium probability distribution $\rho(q)$ via $U(q) = -k_B\,T \ln \rho(q)$, where $k_B\,T$ is the thermal energy with k_B the Boltzmann constant and T the absolute temperature. The elimination of degrees of freedom introduces non-Markovian effects in terms of the memory function $\Gamma(t)$, which describes time-dependent friction and thereby, couples the present dynamics to the past states, and stochastic effects in terms of the random force $F_R(t)$. In equilibrium, the random force $F_R(t)$ is related to $\Gamma(t)$ via the fluctuation–dissipation theorem $\langle F_R(t)F_R(t')\rangle = k_B\,T\Gamma(|t-t'|)$ (7). The derivation of the GLE in Eq. **1** relies on several approximations (8–11). Thus, for a given reaction coordinate that is a nonlinear function of the microscopic coordinates, the validity of Eq. **1** is not guaranteed and needs to be explicitly checked.

The folding free energy $U(q)$ can be straightforwardly obtained from simulations; it can also be obtained from single-molecule experiments (12–15). Clearly, there is no guarantee that a given reaction coordinate, which could be an experimental observable such as the distance between two attached fluorophores, is a good reaction coordinate, meaning that it leads to a Markovian description of the folding process. Different reaction coordinates have been proposed for the efficient description of protein-folding simulations (16); schemes to construct reaction coordinates that optimally yield the transition state, which separates unfolded and folded basins of attraction from each other, have been developed (17). As an alternative to continuous reaction coordinates, Markov models describe protein dynamics in terms of a set of metastable states (18, 19), for which full access to the underlying microscopic coordinates is typically needed. These works have in common that descriptions are sought that minimize memory effects, so that stochastic Markovian theory applies. In the opposite direction, various methods were developed to extract the memory function $\Gamma(t)$ from time series data for a given reaction coordinate (9, 20–25), but the complexities of the GLE, in particular for a nonlinear protein-folding free energy in combination with a numerically determined memory function, prevented predictions of protein-folding times from the GLE, with the notable exception of dialanine (26). This is why in protein-folding theory, the Markovian Langevin equation (LE), where the memory integral is replaced by an instantaneous

Significance

Protein-folding kinetics is often described as Markovian (i.e., memoryless) diffusion in a one-dimensional free energy landscape, governed by an instantaneous friction coefficient that is fitted to reproduce experimental or simulated folding times. For the α-helix forming polypeptide alanine$_9$ and a specific reaction coordinate that consists of the summed native hydrogen-bond lengths, we demonstrate that the friction extracted from molecular dynamics simulations exhibits significant memory with a decay time that is in the nanosecond range and thus, of the same order as the folding and unfolding times. Our non-Markovian modeling not only reproduces the molecular dynamics simulations accurately but also demonstrates that memory friction effects lead to anomalous and drastically accelerated protein kinetics.

Author contributions: C.A., L.T., F.N.B., J.K., J.O.D., and R.R.N. designed research; C.A. and L.T. performed research; and C.A. and R.R.N. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: rnetz@physik.fu-berlin.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2023856118/-/DCSupplemental.

Published July 29, 2021.

friction term, is predominantly used. Such a Markovian theory yields many useful insights into protein-folding dynamics and culminated in the comparison of transition-path times and mean folding times (27, 28). However, the success of free energy folding theory on the Markovian level relies partly on the fact that the friction, which determines the prefactor of the Kramers folding time, is normally used as a fitting parameter. Even when the friction is allowed to vary with the reaction coordinate and is extracted from simulations, it is typically computed from folding or reconfiguration times, which by construction, leads to self-consistent predictions of the kinetics (29, 30). In fact, recent experiments revealed significant inconsistencies when comparing directly measured free energy barrier heights with those inferred from transition path and folding times (15), which were suggested to be due to memory effects (31, 32). The same inconsistencies are obtained when the friction of a reaction coordinate is not fitted to folding times but rather, extracted directly from simulation trajectories and used in the framework of Markovian theory, as we demonstrate here.

In our approach, instead of searching for a good reaction coordinate, we employ a standard one-dimensional coordinate that consists of the sum of the separations between native contacts. We use accurate tools for extracting all parameters of the GLE from molecular dynamics (MD) simulations for the helix-forming polypeptide Ala_9 in water. The free energy $U(q)$ shows multiple minima separated by low barriers, indicative of the sequential formation of the helix, while the longest decay time of the multiexponential memory function $\Gamma(t)$ is of the order of the unfolding time. These properties render Ala_9 as a very sensitive test of kinetic theory. We simulate the resulting GLE by Markovian embedding techniques. By comparison of the MD and GLE results for the mean folding and unfolding times, we demonstrate that the one-dimensional GLE is an accurate and practical tool for the description of protein-folding dynamics. On the other hand, the Markovian version of the overdamped GLE cannot describe the folding and unfolding kinetics of the peptide as long as the friction is not a fitting parameter but rather, taken as extracted from the MD simulations. This stays true even when the friction coefficient is allowed to depend on the reaction coordinate. As predicted by the Grote–Hynes theory, memory typically accelerates barrier crossing, where the acceleration magnitude depends primarily on the ratio of the memory time and the distance between the minimum and the barrier in reaction coordinate space (33–38). This memory-induced speedup of folding and unfolding is found to be accompanied by pronounced anomalous diffusion in reaction coordinate space. Our results are corroborated by a systematic Kramers–Moyal coefficient (KMC) analysis, which shows that higher-order quartic KMCs are nonnegligible and that the linear and quadratic KMCs vanish in the short time limit, as expected in the presence of non-Markovian effects. This implies that the description of protein folding in terms of the Fokker–Planck equation is only valid above a certain timescale that needs to be suitably chosen. We also find that a spurious reaction coordinate–dependent friction profile arises when non-Markovian protein dynamics is described using a Markovian model.

Results and Discussion

MD Simulations and GLE Parameter Extraction. The effective GLE is constructed from a 10-μs-long MD trajectory for Ala_9 in water, which is the simplest polypeptide that forms an α-helix (39) (*Methods* and *SI Appendix*, section 1 have details). As a reaction coordinate, we use the summed separations between the H-bond donor nitrogen of residue n and the acceptor oxygen of residue $n+4$,

$$q(t)=\frac{1}{3}\sum_{i=2}^{4}\|\mathbf{r}_i^N(t)-\mathbf{r}_{i+4}^O(t)\|, \qquad [2]$$

which characterize the left-handed α-helical conformation. In the α-helical state, q has a value around 0.3 nm, the mean H-bond length between nitrogen and oxygen. We will further also consider the end-to-end distance as an alternative reaction coordinate. The free energies $U(q)$ in Fig. 1*A* for different simulation lengths demonstrate that the simulation is fully converged after about $6\,\mu$s. The free energy displays several metastable states, which are also discernible in the trajectory in Fig. 1*B* and make this simple polypeptide challenging for theoretical description.

Using a generalization of earlier methods (40), we extract the running integral $G(t)=\int_0^t ds\Gamma(s)$ (*SI Appendix*, section 2 has details), from which the memory function $\Gamma(t)$ is obtained via a numerical derivative and fitted using least-square methods to a multiexponential of the form

$$\Gamma(t)=\sum_{n=1}^{5}\frac{\gamma_n}{\tau_n}e^{-t/\tau_n}. \qquad [3]$$

The extracted $G(t)$ (gray line) is compared with the corresponding fit (red line) in Fig. 2*A*; no significant deviations can be discerned. The comparison of the extracted and fitted memory function $\Gamma(t)$ in Fig. 2*B* reveals oscillations below a picosecond, which are not reproduced by the exponential fit function but also do not play a role for the kinetics, as will be shown below. The fitted memory times τ_n and friction coefficients γ_n are presented in Table 1; the typical reconfiguration time, which can be qualitatively inferred from the trajectory in Fig. 1*B*, is of the order of the longest decay time $\tau_5\approx 5$ ns. This means that the reaction coordinate is not particularly good since it exhibits pronounced non-Markovian effects and thus, constitutes a suitable test of our methods.

The effective mass follows from the equipartition theorem according to $m=k_BT/\langle\dot{q}^2\rangle$ and turns out to be independent of q and given by $m=31.3$ u (*SI Appendix*, section 3). The motion described by the GLE is expected to become diffusive after the inertial time $\tau_m=m/\bar{\gamma}$, where the total friction coefficient is given by $\bar{\gamma}=\sum_n\gamma_n=3.5\cdot10^5$ u/ps (Table 1). It follows that $\tau_m=0.1$ fs, even shorter than the MD integration time step; thus, inertial effects are completely negligible. Nevertheless, the

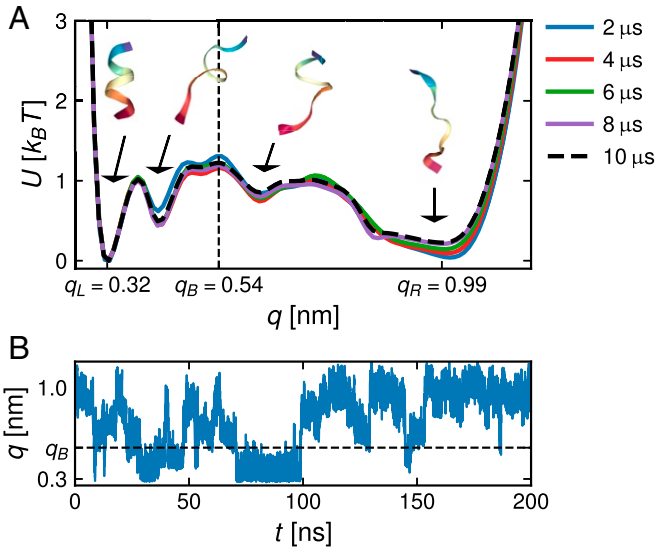

Fig. 1. (*A*) The free energy $U(q)$ for the mean hydrogen-bond distance reaction coordinate of Ala_9 for different simulation lengths; representative snapshots of the polypeptide backbone in all local minima are shown. The barrier used for the calculation of unfolding and folding times is positioned at $q_B=0.54$ nm. (*B*) A 200-ns-long segment of the trajectory is shown.

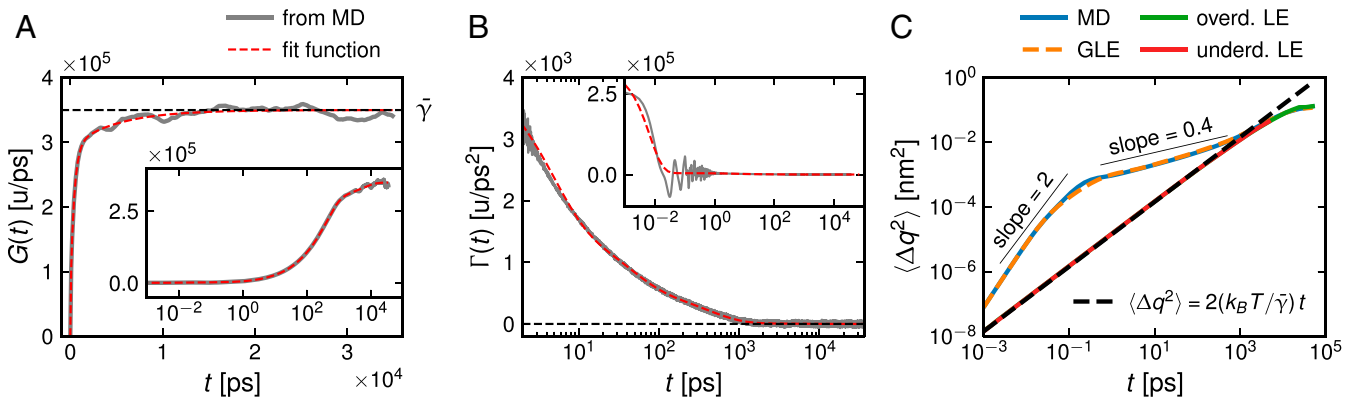

Fig. 2. (*A*) Running integral $G(t)$ over the memory function; *Inset* shows a lin-log plot. The horizontal dashed line denotes the total friction coefficient $\bar{\gamma}$. (*B*) Memory function $\Gamma(t)$; *Inset* includes short times. Gray lines correspond to the numerical data; red lines correspond to the multiexponential fit according to Eq. **3**. (*C*) Mean-square displacement of the reaction coordinate; MD (blue line) and GLE (orange broken line) simulation results agree perfectly and exhibit superdiffusion for times up to 0.1 ps and subdiffusion up to 1 ns. Underdamped (underd.; red line) and overdamped (overd.; green line, underneath the red line) Markovian Langevin simulations agree perfectly with each other but miss the anomalous diffusion.

acceleration term in Eq. **1** is kept in the GLE simulations, as it stabilizes the numerical integration. In order to estimate the importance of memory effects, the memory times τ_n are compared with the diffusion timescale $\tau_D = \beta\bar{\gamma}L^2/2$ (36), which is the time it takes a free Brownian particle to diffuse over a length L in reaction coordinate space where $\beta = 1/k_B T$ is the inverse thermal energy. For $L = 0.22$ nm, the distance between the folded minimum at $q = 0.32$ nm and the barrier at $q = 0.54$ nm in Fig. 1, one obtains $\tau_D = 6.8$ ns, which is of the order of the longest memory time τ_5. This places the system in the so-called memory-acceleration regime, where memory effects are relevant and significantly accelerate barrier crossing (36–38).

Comparison of MD and GLE Simulations. Numerical integration of the GLE is straightforwardly achieved by Markovian embedding (i.e., by transforming the GLE into a system of linearly coupled LEs) (22) (*SI Appendix*, section 4).

In Fig. 3*A*, we show profiles of the mean first-passage time (MFPT) $\tau_{\mathrm{MFPT}}(q_S, q_F)$ for unfolding (start position $q_S = q_L = 0.32$ nm; solid lines) and folding kinetics (start position $q_S = q_R = 0.99$ nm; broken lines) as a function of the final position q_F. Statistical errors are determined accounting for data correlations (41) (*SI Appendix*, section 5) and are smaller than the line thickness. MD and GLE simulation results (blue and orange lines, respectively) agree nicely; this demonstrates that GLE-based non-Markovian modeling of protein folding is feasible and accurate. Even first-passage time distributions from GLE and MD simulations agree satisfactorily with each other, as shown in *SI Appendix*, section 6.

Beyond reproducing MD results, the GLE is a diagnostic tool that allows us to quantify the importance of memory effects. In order to modulate memory effects in the GLE, we rescale the memory times according to $\tau_n \to \alpha\tau_n$ for $n = 2, 3, 4, 5$ while keeping the memory time τ_1 of the fastest exponential contribution fixed. Since $\tau_1 = 7$ fs is above the simulation time step of 1 fs, this ensures that in the limit $\alpha \to 0$, we obtain a regularized model that, as we will show below, corresponds to the Markovian limit. In Fig. 3*B*, we show MFPTs between the three positions $q_L = 0.32$ nm, $q_B = 0.54$ nm, and $q_R = 0.99$ nm as a function of the rescaling factor α from GLE simulations. The six different MFPTs are illustrated in Fig. 3 *C*, *Inset* by filled and closed arrows and indicated in Fig. 3*B* by corresponding filled and open colored spheres. We see that reducing the memory time increases all MFPTs; in other words, memory accelerates barrier crossing (36). As expected, the GLE results approach the overdamped Markov limit, denoted by the horizontal lines in the corresponding color and calculated from the exact expression in Eq. **10**, without adjustable parameters as α tends to zero. Interestingly, for folding to the barrier (open green circles), the MFPT for $\alpha = 1$ and the Markovian limit for $\alpha \to 0$ differ only by a factor of around 2.5. On the other hand, for unfolding to the barrier (filled red circles), the $\alpha \to 0$ and $\alpha = 1$ MFPTs differ by a factor of around nine. This means that even when treating the total friction coefficient $\bar{\gamma}$ as a free parameter, the Markovian overdamped theory Eq. **10**, because it is linear in the friction, can reproduce either the MD folding or unfolding times to the barrier but not both simultaneously. This is not due to inertial effects since the overdamped Markovian theory works perfectly for $\alpha \to 0$, as seen in Fig. 3*B*. Rather, memory effects influence the times of folding and unfolding to the barrier top differently. This is demonstrated by the plot of MFPT ratios as a function of α in Fig. 3*C*, where it is seen that the ratio of the folding and unfolding times to the barrier top $\tau_{\mathrm{MFPT}}(q_R, q_B)/\tau_{\mathrm{MFPT}}(q_L, q_B)$, denoted by open green and filled red spheres, depends sensitively on α. In contrast, the ratios of reciprocal MFPTs (i.e., MFPTs with interchanged start and final positions), denoted by red, green, and blue lines with identically colored open and filled circles, do not depend on α, which shows that the memory dependence of ratios of MFPTs depends on the precise MFPT definition and by no means indicates a breakdown of the detailed balance or the law-of-mass action. In *SI Appendix*, section 6, we demonstrate that the memory-induced speedup is even more pronounced for transition-path times compared with folding and unfolding times, in agreement with previous findings (15, 31, 32).

The high accuracy of GLE simulations is furthermore reflected by the good agreement of the mean-square displacement $\langle \Delta q(t)^2 \rangle = \langle (q(t'+t) - q(t'))^2 \rangle$ from MD and GLE simulations in Fig. 2*C*, which exhibits pronounced subdiffusive behavior with an exponent 0.4 for times between 1 ps and 1 ns. Anomalous diffusion is often modeled by fractional theories

Table 1. Fitted memory function parameters from Eq. 3

n	γ_n (u/ps)	τ_n (ps)
1	$2.2 \cdot 10^3$	0.007
2	$1.2 \cdot 10^4$	4.6
3	$4.2 \cdot 10^4$	40.3
4	$2.4 \cdot 10^5$	399
5	$5.7 \cdot 10^4$	4,970
$\bar{\gamma} = \sum_n \gamma_n$	$3.5 \cdot 10^5$	

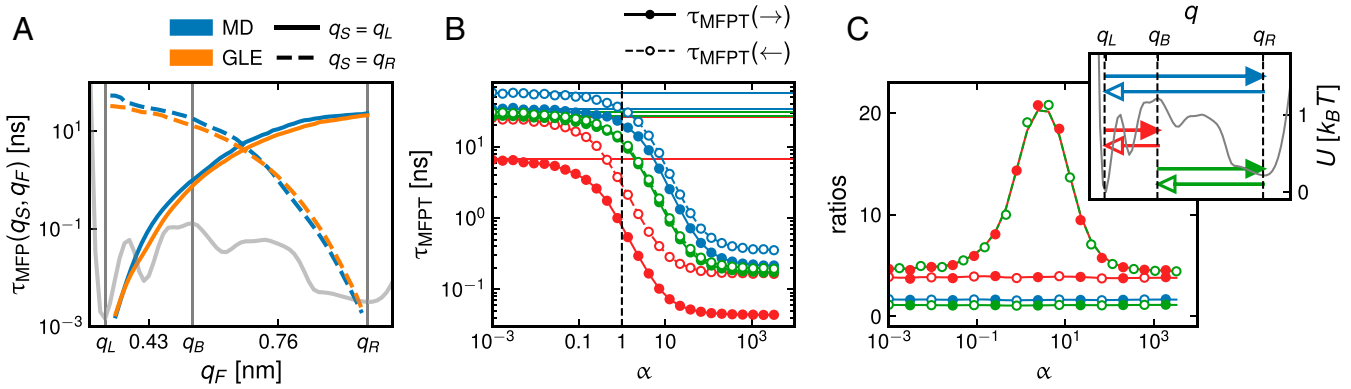

Fig. 3. (*A*) Comparison of unfolding and folding MFPTs $\tau_{MFPT}(q_S, q_F)$ from MD (blue) and GLE (orange) simulations as a function of the final position q_F for start positions $q_S = q_L = 0.32$ nm (solid lines) and $q_S = q_R = 0.99$ nm (broken lines). The gray curve shows the folding free energy $U(q)$. (*B*) Dependence of different MFPTs from GLE simulations on the memory time rescaling factor α; the corresponding start and final positions are illustrated in *C*, *Inset*. Open and filled circles correspond to open and filled arrows, respectively, in *C*, *Inset*. The colored horizontal lines denote corresponding results for the overdamped Markov limit from Eq. **10**. (*C*) Ratios of the MFPTs shown in *B*. Ratios of reciprocal MFPTs do not depend on α (red, green, and blue lines that connect colored circles); only the ratio of the folding and unfolding times to the barrier top, $\tau_{MFPT}(q_R, q_B)/\tau_{MFPT}(q_L, q_B)$ (open green and filled red spheres), depends on α.

(31, 42). Fig. 2*C* shows that it is accurately reproduced by multiexponential memory and that it disappears when memory effects are eliminated, in line with recent theoretical analysis (43). The overall good agreement between MD and GLE simulation results shows that the GLE in the form of Eq. **1** describes the kinetics of Ala_9 very accurately. This is not due to our specific choice of reaction coordinate, as demonstrated in *SI Appendix*, section 7, where we present a similar GLE-based analysis using the Ala_9 end-to-end distance as reaction coordinate.

Reaction Coordinate–Dependent Friction. We so far demonstrated that the GLE in the form of Eq. **1** reproduces the MD simulation kinetics and that memory effects are significant. We now investigate whether reaction coordinate–dependent friction effects, which are not included in the GLE, are relevant. The Markovian LE that incorporates a friction function $\gamma(q)$ has been amply used to describe protein-folding dynamics (29, 30, 44). In the underdamped version, it reads

$$m\ddot{q}(t) = -U'(q) - \gamma(q)\dot{q}(t) + \sqrt{k_B T\gamma(q)}\,\eta(t), \qquad [4]$$

which for general $U(q)$, unfortunately is analytically intractable. The overdamped version

$$0 = -U'(q) - \gamma(q)\dot{q}(t) - \frac{k_B T}{2}\frac{\gamma'(q)}{\gamma(q)} + \sqrt{k_B T\gamma(q)}\,\eta(t) \qquad [5]$$

is much more useful since the MFPTs can be calculated analytically. In these expressions, the random force $\eta(t)$ has vanishing mean, and its correlator is given by $\langle\eta(t)\eta(t')\rangle = 2\delta(t-t')$. For constant friction, the underdamped LE (Eq. **4**) can be derived from the GLE (Eq. **1**) by a systematic expansion of the integral kernel (*SI Appendix*, section 8). The overdamped LE (Eq. **5**) follows from Eq. **4** by neglecting the inertia term; the term proportional to the gradient $\gamma'(q)$ cancels a spurious drift term and follows by mapping on the Fokker–Planck equation (*SI Appendix*, section 9) (6). In fact, from the overdamped LE with constant friction, an arbitrary friction profile $\gamma(q)$ can be created by a nonlinear transformation of the reaction coordinate (29, 30) (*SI Appendix*, section 10); this suggests that spatially dependent friction is related to nonlinearities in the reaction coordinate that are not straightforwardly captured by the projection techniques used to derive the GLE Eq. **1**.

Various methods to extract $\gamma(q)$ from experimental or simulated trajectories have been proposed; a systematic approach involves the KMCs, which for the overdamped case and for finite lag time Δt, read

$$D_k(q) = \frac{1}{k!}\frac{1}{\Delta t}\left\langle (q(t+\Delta t) - q(t))^k \right\rangle_{q(t)=q}. \qquad [6]$$

The Fokker–Planck equation for the time-dependent probability distribution $P(q,t)$ in terms of the KMCs follows in the limit $\Delta t \to 0$ as (7)

$$\frac{\partial P(q,t)}{\partial t} = \sum_{k=1}^{\infty}\frac{\partial^k}{\partial q^k}\left[D_k(q)P(q,t)\right], \qquad [7]$$

and the underdamped case is treated in *SI Appendix*, section 11. According to Pawula's theorem, for a Markovian process, all KMCs with $k > 2$ vanish for $\Delta t \to 0$, and Eq. **7** takes the standard form of a second-order partial differential equation (7). For a non-Markovian process [i.e., if the memory function $\Gamma(t)$ in Eq. **1** has a finite range], all KMCs with $k > 1$ vanish for $\Delta t \to 0$, and thus, the stochastic properties of the process cannot be described by a partial differential equation for $P(q,t)$ at all (*SI Appendix*, section 12).

For the underdamped LE, the relation between the second-order velocity KMC D_{vv} and the friction profile $\gamma_{UD}(q)$ reads (7)

$$D_{vv}(q) = \frac{1}{2\Delta t}\langle (v(t+\Delta t) - v(t))^2\rangle_{q(t)=q} = k_B T\frac{\gamma_{UD}(q)}{m^2}. \qquad [8]$$

For the overdamped LE, $\gamma_{OD}(q)$ follows from the second-order position KMC D_{qq} as

$$D_{qq}(q) = \frac{1}{2\Delta t}\langle (q(t+\Delta t) - q(t))^2\rangle_{q(t)=q} = \frac{k_B T}{\gamma_{OD}(q)} \qquad [9]$$

(*SI Appendix*, section 9). For the numerical computation of the KMCs, we use kernel-density estimators (45) (*SI Appendix*, section 13). In Fig. 4*A*, we show the friction profiles $\gamma_{UD}(q)$ (circles) and $\gamma_{OD}(q)$ (lines) computed from the KMCs for different lag times Δt; a number of points are noteworthy. 1) We find no significant deviations between the friction profiles extracted from MD (solid lines and filled circles) and GLE (broken lines and open circles) trajectories; this reverberates that the GLE describes the protein dynamics very faithfully. 2) The underdamped and overdamped friction profiles $\gamma_{UD}(q)$ and $\gamma_{OD}(q)$

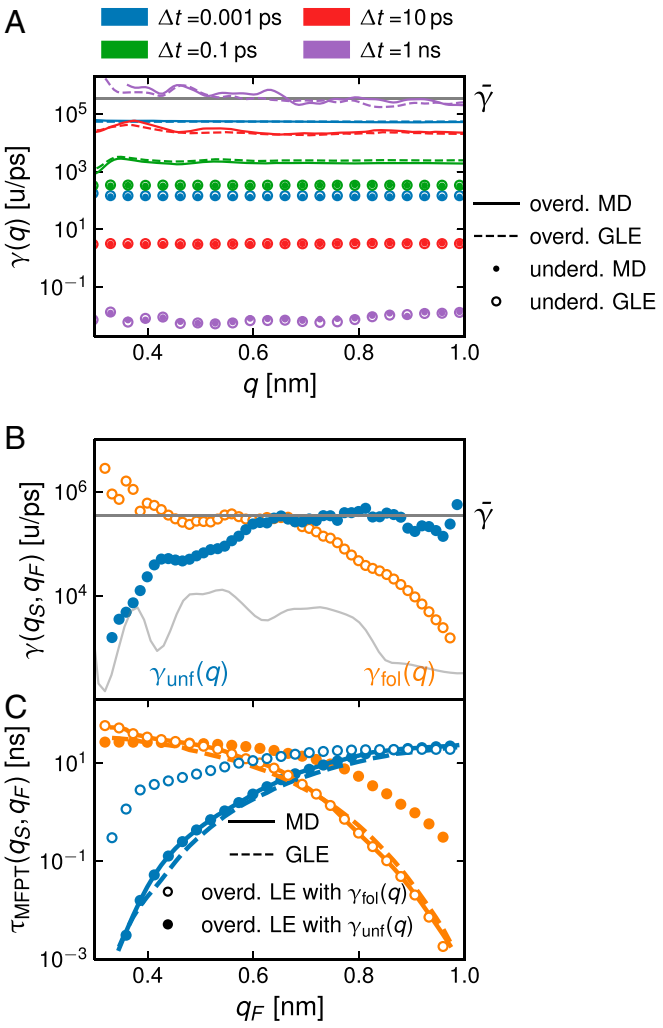

Fig. 4. (*A*) Friction coefficient profiles $\gamma(q)$ from KMC analysis for different lag times Δt (different colors) for the underdamped (underd.) Langevin model, Eq. **8**, from MD (filled circles) and GLE simulations (open circles) and for the overdamped (overd.) Langevin model, Eq. **9**, from MD (solid lines) and GLE simulations (broken lines). The gray horizontal line shows the total friction coefficient $\bar{\gamma}$ extracted from MD simulations. (*B*) Friction profiles computed from the MD MFPT profiles in Fig. 3*A* using Eq. **11**. $\gamma_{\mathrm{unf}}(q_F)$ follows from the unfolding MFPTs for start position $q_S = 0.32$ nm, and $\gamma_{\mathrm{fol}}(q_F)$ follows from folding MFPTs for $q_S = 0.99$ nm. The gray horizontal line denotes the friction coefficient $\bar{\gamma}$ extracted from MD simulations. The gray curve in the background shows the folding free energy $U(q)$. (*C*) MFPTs from MD and GLE simulations are compared with overd. Markovian predictions according to Eq. **10** using $\gamma_{\mathrm{unf}}(q_F)$ and $\gamma_{\mathrm{fol}}(q_F)$ from *B*.

disagree for all lag times Δt, which very clearly demonstrates an inconsistency in the Markovian description of protein folding. In fact, in the limit $\Delta t \to 0$, both D_{qq} and D_{vv} vanish; thus, $\gamma_{\mathrm{OD}}(q)$ diverges, while $\gamma_{\mathrm{UD}}(q)$ goes to zero (*SI Appendix*, section 12). 3) While the underdamped friction $\gamma_{\mathrm{UD}}(q)$ never reaches a realistic value close to $\bar{\gamma}$, the overdamped friction $\gamma_{\mathrm{OD}}(q)$ approaches $\bar{\gamma}$ for $\Delta t \approx 1$ ns. This shows that lag times of the order of the longest memory time have to be used in order to generate realistic friction values. 4) The friction profiles extracted from the GLE simulations are position dependent, seen most clearly in $\gamma_{\mathrm{OD}}(q)$ for $\Delta t = 1$ ns (purple broken line); this is clearly a spurious effect since the GLE has no position-dependent friction. We conclude that the mapping of a non-Markovian process onto a Markovian LE produces spurious position-dependent friction effects. Presumably, the effective friction of proteins will in general exhibit a dependence on the reaction coordinate, but the extraction of friction profiles would have to account for memory effects in order to avoid spurious effects. The capability of the GLE Eq. **1** to very accurately reproduce the MD simulation kinetics suggests that for the present case of Ala_9, the spatial dependence of friction is negligible.

An alternative way to determine a friction profile $\gamma(q)$ in the overdamped limit uses the one-to-one relation between the MFPT profiles in Fig. 3*A* and $\gamma(q)$. From the expressions for the folding and unfolding times (Eq. **10**), $\gamma(q)$ follows by inversion according to Eq. **11** (30). In Fig. 4*B*, we show $\gamma_{\mathrm{unf}}(q_F)$ and $\gamma_{\mathrm{fol}}(q_F)$ computed from unfolding and folding MFPTs from MD simulations for start positions $q_S = q_L$ and $q_S = q_R$, respectively. Not surprisingly, the profiles $\gamma_{\mathrm{unf}}(q_F)$ and $\gamma_{\mathrm{fol}}(q_F)$ are rather close to $\bar{\gamma}$ extracted from the MD simulations, which is shown as a gray horizontal line in Fig. 4*B*, but differ significantly from each other. This suggests that a single friction profile cannot describe folding and unfolding of Ala_9 simultaneously. In fact, the values of $\gamma_{\mathrm{unf}}(q_F)$ and $\gamma_{\mathrm{fol}}(q_F)$ go down as q_F moves to the respective start positions (i.e., as the folding and unfolding times become shorter). This reflects that memory effects particularly accelerate fast transitions (36–38).

To demonstrate the limitations of the friction profiles in Fig. 4*B*, we show in Fig. 4*C* folding and unfolding MFPT profiles that are calculated according to Eq. **10** from $\gamma_{\mathrm{unf}}(q)$ (filled circles) and $\gamma_{\mathrm{fol}}(q)$ (open circles). By construction, the MFPTs using $\gamma_{\mathrm{unf}}(q)$ reproduce the unfolding simulation data, while the MFPTs using $\gamma_{\mathrm{fol}}(q)$ reproduce the folding simulation data. In contrast, the MFPTs using $\gamma_{\mathrm{unf}}(q)$ fail to reproduce the simulated folding times, and the MFPTs using $\gamma_{\mathrm{fol}}(q)$ fail to reproduce the simulated unfolding times, in particular when the folding/unfolding times become smaller than about 10 ns. In contrast, the GLE model (broken lines) reproduces both folding and unfolding MD dynamics (solid lines). This underlines that there is no consistent way of describing the complete folding/unfolding dynamics with a Markovian model.

Conclusions

By extracting the time-dependent friction from MD simulations for the polypeptide Ala_9 from explicit-water MD simulations, we demonstrate that the resulting GLE model can be straightforwardly integrated numerically and reproduces the folding and unfolding kinetics of the MD simulations very accurately. Our findings are not restricted to a reaction coordinate based on the summed distances between native H bonds. As we show in *SI Appendix*, section 7, the same analysis of the Ala9 end-to-end distance leads to similar results. Decreasing the memory time in the GLE while keeping the friction coefficient (i.e., the integral over the memory function) constant, the folding kinetics changes significantly for folding and unfolding events. This shows that memory effects are important even for the formation kinetics of a single α-helix.

In contrast, the Markovian LE cannot reproduce the full Ala_9 reconfiguration dynamics, even with a fitted friction profile; this follows from the comparison of the folding and unfolding kinetics, which would need to be modeled with different friction profiles in order to reproduce the MD simulation kinetics.

We have mostly used the GLE model as a diagnostic tool to understand and quantify non-Markovian effects; since non-Markovian simulations are rather inexpensive, they can also be used as an efficient tool to simulate the response of proteins to environmental changes (e.g., externally applied forces). In fact, our extraction technique for the memory function can in principle also be applied to trajectories from single-molecule experiments (13–15), which would enable us to perform non-Markovian GLE simulations on experimental systems directly, without the need of atomistic MD simulations. Because of the limited time resolution of typical

experimental data, suitable extraction techniques would have to be used (24, 46).

Methods

MD and GLE Simulation Details. We use the all-atom Amber03 force field (47) with extended simple point-charge (SPC/E) water (48). The cubic simulation box has side lengths of 4.95 nm and contains 4,023 water molecules. The Lennard–Jones interactions are cut off after 1.0 nm. For long-range electrostatic interactions, we use the particle Mesh Ewald method (49). The simulation time step is 1 fs, and the total simulation time is 10 μs. All simulations are performed in the NVT ensemble using the Gromacs 2019 MD package (50). Further details are given in *SI Appendix*, section 1. In the GLE simulations, we used the same time step and simulation time as in the MD simulations. Input files of the MD simulations are available for download under (http://dx.doi.org/10.17169/refubium-29935). Our Python scripts for the numerical extraction of the memory kernel, for performing a GLE simulation, and computing MFPTs can be found in GitHub (https://github.com/lucastepper/memtools).

From MFPTs to Friction Profiles. The MFPT is defined as the mean time needed to reach the final position q_F for the first time when starting from a position q_S. For the overdamped LE in Eq. **5**, it reads for $q_S < q_F$ (51),

$$\tau_{\mathrm{MFPT}}(q_S, q_F) = \beta \int_{q_S}^{q_F} \mathrm{d}q\, e^{\beta U(q)} \gamma(q) \int_{q_{min}}^{q} \mathrm{d}q'\, e^{-\beta U(q')} \qquad [10a]$$

and for $q_S > q_F$,

$$\tau_{\mathrm{MFPT}}(q_S, q_F) = \beta \int_{q_F}^{q_S} \mathrm{d}q\, e^{\beta U(q)} \gamma(q) \int_{q}^{q_{max}} \mathrm{d}q'\, e^{-\beta U(q')}. \qquad [10b]$$

Taking the derivative of Eq. **10** w.r.t. q_F gives the friction profile $\gamma(q_F)$ as (30)

$$\gamma_{\mathrm{unf}}(q_F) = k_B T \frac{e^{-\beta U(q_F)}}{Z_1} \frac{\partial \tau_{\mathrm{MFPT}}}{\partial q_F} \qquad \text{for } q_S < q_F, \qquad [11a]$$

$$\gamma_{\mathrm{fol}}(q_F) = -k_B T \frac{e^{-\beta U(q_f)}}{Z_2} \frac{\partial \tau_{\mathrm{MFPT}}}{\partial q_F} \qquad \text{for } q_S > q_F, \qquad [11b]$$

where $Z_1 = \int_{q_{min}}^{q_F} \mathrm{d}q\, e^{-\beta U(q)}$ and $Z_2 = \int_{q_F}^{q_{max}} \mathrm{d}q\, e^{-\beta U(q)}$.

Data Availability. Derivations that support the findings of this study are included in *SI Appendix*. Simulation input files data have been deposited in Institutional Repository (http://dx.doi.org/10.17169/refubium-29935). Our codes for extracting the memory kernel, running GLE simulations, and for computing MFPTs are available in GitHub (https://github.com/lucastepper/memtools).

ACKNOWLEDGMENTS. We acknowledge discussions with W. A. Eaton; support by Deutsche Forschungsgemeinschaft Grant CRC 1114 "Scaling Cascades in Complex System," Project 235221301, Project B03; and support by European Research Council Advanced Grant NoMaMemo Grant 835117. Work was funded in part by the European Research Council under the EU's Horizon 2020 Program, Grant 740269.

1. J. D. Bryngelson, J. N. Onuchic, N. D. Socci, P. G. Wolynes, Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins Structure Function Genetic* **21**, 167–195 (1995).
2. Y. Levy, J. N. Onuchic, Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 389–415 (2006).
3. K. A. Dill, J. L. MacCallum, The protein-folding problem, 50 years on. *Science* **338**, 1042–1046 (2012).
4. R. Zwanzig, Memory effects in irreversible thermodynamics. *Phys. Rev.* **124**, 983–992 (1961).
5. H. Mori, Transport, collective motion, and Brownian motion. *Prog. Theor. Phys.* **33**, 423–455 (1965).
6. N. G. Van Kampen, Elimination of fast variables. *Phys. Rep.* **124**, 69–160 (1985).
7. H. Risken, "Fokker-Planck equation" in *The Fokker-Planck Equation: Methods of Solution and Applications*, H. Risken, Ed. (Springer Series in Synergetics, Springer, Berlin, Germany, 1996), pp. 63–95.
8. H. Grabert, P. Hänggi, P. Talkner, Microdynamics and nonlinear stochastic processes of gross variables. *J. Stat. Phys.* **22**, 537–552 (1980).
9. O. F. Lange, H. Grubmüller, Collective Langevin dynamics of conformational motions in proteins. *J. Chem. Phys.* **124**, 214903 (2006).
10. T. Kinjo, S. Hyodo, Equation of motion for coarse-grained simulation based on microscopic description. *Phys. Rev.* **75**, 051109 (2007).
11. C. Hijón, P. Español, E. Vanden-Eijnden, R. Delgado-Buscalioni, Mori–Zwanzig formalism as a practical computational tool. *Faraday Discuss* **144**, 301–322 (2010).
12. B. Schuler, W. A. Eaton, Protein folding studied by single-molecule FRET. *Curr. Opin. Struct. Biol.* **18**, 16–26 (2008).
13. H. Yu *et al.*, Energy landscape analysis of native folding of the prion protein yields the diffusion constant, transition path time, and rates. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 14452–14457 (2012).
14. M. Hinczewski, J. C. M. Gebhardt, M. Rief, D. Thirumalai, From mechanical folding trajectories to intrinsic energy landscapes of biopolymers. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 4500–4505 (2013).
15. K. Neupane *et al.*, Direct observation of transition paths during the folding of proteins and nucleic acids. *Science* **352**, 239–242 (2016).
16. R. Hegger, G. Stock, Multidimensional Langevin modeling of biomolecular dynamics. *J. Chem. Phys.* **130**, 034106 (2009).
17. R. B. Best, G. Hummer, Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 6732–6737 (2005).
18. F. Noé, I. Horenko, C. Schütte, J. C. Smith, Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J. Chem. Phys.* **126**, 155102 (2007).
19. J. D. Chodera, N. Singhai, V. S. Pande, K. A. Dill, W. C. Swope, Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* **126**, 155101 (2007).
20. J. E. Straub, M. Borkovec, B. J. Berne, Calculation of dynamic friction on intramolecular degrees of freedom in. *J. Phys. Chem.* **91**, 4995–4998 (1987).
21. I. Horenko, C. Hartmann, C. Schütte, F. Noé, Data-based parameter estimation of generalized multidimensional Langevin processes. *Phys. Rev. E* **76**, 016706 (2007).
22. E. Darve, J. Solomon, A. Kia, Computing generalized Langevin equations and generalized Fokker–Planck equations. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10884–10889 (2009).
23. F. Gottwald, SD. Ivanov, O. Kühn, Vibrational spectroscopy via the Caldeira-Leggett model with anharmonic system potentials. *J. Chem. Phys.* **144**, 164102 (2016).
24. G. Jung, M. Hanke, F. Schmid, Iterative reconstruction of memory kernels. *J. Chem. Theor. Comput.* **13**, 2481–2488 (2017).
25. J. O. Daldrop, J. Kappler, F. N. Brünig, R. R. Netz, Butane dihedral angle dynamics in water is dominated by internal friction. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 5169–5174 (2018).
26. H. S. Lee, S. H. Ahn, E. F. Darve, The multi-dimensional generalized Langevin equation for conformational motion of proteins. *J. Chem. Phys.* **150**, 174113 (2019).
27. H. S. Chung, K. McHale, J. M. Louis, W. A. Eaton, Single-molecule fluorescence experiments determine protein folding transition path times. *Science* **335**, 981–984 (2012).
28. H. S. Chung, S. Piana-Agostinetti, D. E. Shaw, W. A. Eaton, Structural origin of slow diffusion in protein folding. *Science* **349**, 1504–1510 (2015).
29. R. B. Best, G. Hummer, Coordinate-dependent diffusion in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 1088–1093 (2010).
30. M. Hinczewski, Y. von Hansen, J. Dzubiella, R. R. Netz, How the diffusivity profile reduces the arbitrariness of protein folding free energies. *J. Chem. Phys.* **132**, 245103 (2010).
31. R. Satija, A. Das, D. E. Makarov, Transition path times reveal memory effects and anomalous diffusion in the dynamics of protein folding. *J. Chem. Phys.* **147**, 152707 (2017).
32. R. Satija, D. E. Makarov, Generalized Langevin equation as a model for barrier crossing dynamics in biomolecular folding. *J. Phys. Chem. B* **123**, 802–810 (2019).
33. R. F. Grote, J. T. Hynes, The stable states picture of chemical reactions. II. Rate constants for condensed and gas phase reaction models. *J. Chem. Phys.* **73**, 2715–2732 (1980).
34. P. Hanggi, F. Mojtabai, Thermally activated escape rate in presence of long-time memory. *Phys. Rev.* **26**, 1168–1170 (1982).
35. E. Pollak, H. Grabert, P. Hänggi, Theory of activated rate processes for arbitrary frequency dependent friction: Solution of the turnover problem. *J. Chem. Phys.* **91**, 4073–4087 (1989).
36. J. Kappler, J. O. Daldrop, F. N. Brünig, M. D. Boehle, R. R. Netz, Memory-induced acceleration and slowdown of barrier crossing. *J. Chem. Phys.* **148**, 014903 (2018).
37. J. Kappler, V. B. Hinrichsen, R. R. Netz, Non-Markovian barrier crossing with two-time-scale memory is dominated by the faster memory component. *Euro. Phys. J. E* **42**, 119 (2019).
38. L. Lavacchi, J. Kappler, R. R. Netz, Barrier crossing in the presence of multi-exponential memory functions with unequal friction amplitudes and memory times. *Europhys. Lett.* **131**, 40004 (2020).
39. G. S. Jas, W. A. Eaton, J. Hofrichter, Effect of viscosity on the kinetics of α-helix and β-hairpin formation. *J. Phys. Chem. B* **105**, 261–272 (2001).
40. B. Kowalik *et al.*, Memory-kernel extraction for different molecular solutes in solvents of varying viscosity in confinement. *Phys. Rev.* **100**, 012126 (2019).
41. H. Flyvbjerg, "Error estimates on averages of correlated data" in *Advances in Computer Simulation*, J. Kertész, I. Kondor, Eds. (Lecture Notes in Physics, Springer, Berlin, Germany, 1998), pp. 88–103.
42. R. Metzler, J. H. Jeon, A. G. Cherstvy, E. Barkai, Anomalous diffusion models and their properties: Non-stationarity, non-ergodicity, and ageing at the centenary of single particle tracking. *Phys. Chem. Chem. Phys.* **16**, 24128–24164 (2014).
43. B. G. Mitterwallner, L. Lavacchi, R. R. Netz, Negative friction memory induces persistent motion. *Eur. Phys. J. E* **43**, 67 (2020).

44. J. Chahine, R. J. Oliveira, V. B. P. Leite, J. Wang, Configuration-dependent diffusion can shift the kinetic transition state and barrier height of protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 14646–14651 (2007).
45. L. R. Gorjão, F. Meirinhos, kramersmoyal: Kramers–Moyal coefficients for stochastic processes. *J. Open. Source Softw.* **4**, 1693 (2019).
46. B. G. Mitterwallner, C. Schreiber, J. O. Daldrop, Rädler, R. R. Netz, Non-Markovian data-driven modeling of single-cell motility. *Phys. Rev. E* **101**, 032408 (2020).
47. Y. Duan *et al.*, A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **24**, 1999–2012 (2003).
48. H. J. C. Berendsen, J. R. Grigera, T. P. Straatsma, The missing term in effective pair potentials. *J. Phys. Chem.* **91**, 6269–6271 (1987).
49. T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
50. M. J. Abraham *et al.*, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25 (2015).
51. G. H. Weiss, "First passage time problems in chemical physics" in *Advances in Chemical Physics*, I. Prigogine, Ed. (John Wiley & Sons, Ltd, 2007), vol 13, pp. 1–18.