The Journal of Chemical Physics

RESEARCH ARTICLE | NOVEMBER 14 2016

Global Langevin model of multidimensional biomolecular dynamics

Norbert Schaudinnus; Benjamin Lickert; Mithun Biswas ; Gerhard Stock

J. Chem. Phys. 145, 184114 (2016)
https://doi.org/10.1063/1.4967341

AIP Publishing

Global Langevin model of multidimensional biomolecular dynamics

Norbert Schaudinnus,[a] Benjamin Lickert,[a] Mithun Biswas, and Gerhard Stock[b]
Biomolecular Dynamics, Institute of Physics, Albert Ludwigs University, 79104 Freiburg, Germany

(Received 12 September 2016; accepted 26 October 2016; published online 14 November 2016)

Molecular dynamics simulations of biomolecular processes are often discussed in terms of diffusive motion on a low-dimensional free energy landscape $F(\boldsymbol{x})$. To provide a theoretical basis for this interpretation, one may invoke the system-bath ansatz á la Zwanzig. That is, by assuming a time scale separation between the slow motion along the system coordinate $\boldsymbol{x}$ and the fast fluctuations of the bath, a memory-free Langevin equation can be derived that describes the system's motion on the free energy landscape $F(\boldsymbol{x})$, which is damped by a friction field and driven by a stochastic force that is related to the friction via the fluctuation-dissipation theorem. While the theoretical formulation of Zwanzig typically assumes a highly idealized form of the bath Hamiltonian and the system-bath coupling, one would like to extend the approach to realistic data-based biomolecular systems. Here a practical method is proposed to construct an analytically defined global model of structural dynamics. Given a molecular dynamics simulation and adequate collective coordinates, the approach employs an "empirical valence bond"-type model which is suitable to represent multidimensional free energy landscapes as well as an approximate description of the friction field. Adopting alanine dipeptide and a three-dimensional model of heptaalanine as simple examples, the resulting Langevin model is shown to reproduce the results of the underlying all-atom simulations. Because the Langevin equation can also be shown to satisfy the underlying assumptions of the theory (such as a delta-correlated Gaussian-distributed noise), the global model provides a correct, albeit empirical, realization of Zwanzig's formulation. As an application, the model can be used to investigate the dependence of the system on parameter changes and to predict the effect of site-selective mutations on the dynamics. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4967341]

I. INTRODUCTION

A well-established strategy to describe the dynamics of a complex system is to partition the problem into a relevant "system" and its "environment" and subsequently derive reduced equations of motion that govern the system degrees of freedom $\boldsymbol{x} = \{x_i\}$ in the presence of the environment. In classical statistical mechanics, for example, this approach leads to the Fokker-Planck equation describing the deterministic time evolution of the system's probability distribution and to the Langevin equation (LE) which accounts for the stochastic time evolution of a trajectory $\boldsymbol{x}(t)$ of the system. Here we focus on the common case that a time scale separation exists between the slow motion of the system and the fast bath fluctuations (which can always be enforced at the expense of a higher dimension of the system). Within this Markov-type approximation, a memory-free Langevin equation (LE) can be derived,[1,2]

$$\mathcal{M}\ddot{\boldsymbol{x}}(t) = \boldsymbol{f} - \Gamma\dot{\boldsymbol{x}}(t) + \mathcal{K}\boldsymbol{\xi}(t), \tag{1}$$

which contains a Newtonian force term $\boldsymbol{f} = -\nabla F$ with $F(\boldsymbol{x})$ being the potential of mean force, a Stokes' term $-\Gamma\dot{\boldsymbol{x}}$ with the friction field $\Gamma(\boldsymbol{x})$, and a stochastic force $\mathcal{K}\boldsymbol{\xi}(t)$ with amplitude $\mathcal{K}(\boldsymbol{x})$ and "noise" $\boldsymbol{\xi}(t)$ that usually is assumed to be of zero mean, $\langle\boldsymbol{\xi}\rangle = 0$, delta-correlated, $\langle\xi_i(t)\xi_j(t')\rangle = \delta_{ij}\delta(t-t')$, and Gaussian distributed. In this way, the quantities $\boldsymbol{f}$, Γ, and $\mathcal{K}$, henceforth referred to as *Langevin fields*, completely define the LE for an open system (just as the Hamiltonian defines the equation of motion of a closed system).

As a prime application, the LE approach has been employed to describe the structural dynamics of biomolecules. To illustrate the significance of LE (1) for this case, we first recall that the potential of mean force can be interpreted as the "free energy landscape" of the system[2,3]

$$F(\boldsymbol{x}) = -k_B T \ln P(\boldsymbol{x}), \tag{2}$$

where $P(\boldsymbol{x})$ denotes the probability distribution along system coordinate $\boldsymbol{x}$. Revealing the metastable conformational states (the minima of F) and the transition states (the barriers of F) of the system, the free energy landscape may directly exhibit the main pathways of a biomolecular process.[4–6] The friction field $\Gamma(\boldsymbol{x})$, on the other hand, accounts for the—in general position-dependent—diffusivity along these pathways and reports on the "roughness" of the energy landscape.[7,8] As usual, we assume that friction field $\Gamma(\boldsymbol{x})$ and noise amplitude $\mathcal{K}(\boldsymbol{x})$ are related via a fluctuation-dissipation theorem[1]

$$\mathcal{K}(\boldsymbol{x})\mathcal{K}^T(\boldsymbol{x}) = 2k_B T\,\Gamma(\boldsymbol{x}), \tag{3}$$

which in equilibrium ensures the balance between fluctuating forces and frictional dissipation caused by the environment. Moreover, we for now assume that the mass tensor $\mathcal{M}$ is known and that above mentioned properties of the noise are satisfied. Taken together, we thus find that the dynamics described by LE (1) is completely characterized by the free energy landscape $F(\boldsymbol{x})$ and the friction field $\Gamma(\boldsymbol{x})$.

[a]N. Schaudinnus and B. Lickert contributed equally to this work.
[b]E-mail: stock@physik.uni-freiburg.de

Besides producing the conformational dynamics of a given model for $F(\boldsymbol{x})$ and $\Gamma(\boldsymbol{x})$, the LE may also be employed for the inverse problem, that is, to construct these Langevin fields for given input data provided, e.g., by all-atom molecular dynamics (MD) simulations.[9–16] Referred to as data-driven Langevin equation (dLE), the approach may serve as a "post-simulation" model to analyze and interpret the ever growing amount of MD data,[17] in a similar way as the popular Markov state models.[18–20] While seeming somewhat more involved than a Markov state model, the dLE avoids the often ambiguous definition of Markov states and represents a more "physical" model, as it is defined in coordinate space (rather than in an abstract mathematical space). Moreover, it uses well-established physical observables (such as $F(\boldsymbol{x})$ and $\Gamma(\boldsymbol{x})$) and contains the temperature as a driving force. It should be stressed, however, that the applicability of LE (1) to describe molecular kinetics crucially depends on the chosen collective coordinate $\boldsymbol{x}$. While a suitable representation of the energy landscape should (at least) reproduce the correct number, energy, and location of the metastable states and barriers, these basic quantities often get lost when the energy landscape is projected on a low-dimensional subspace.[21,22]

As a solution of this problem, we have recently suggested to use a combination of systematic dimensionality reduction methods[23–29] (that identify in a controlled manner adequate system degrees of freedom) and a multidimensional dLE[16,30,31] (that accounts for dimensionality of the collective coordinate). The resulting dLE model is able to quantitatively reproduce dynamical observables (such as time correlation functions and first passage times) that can be directly compared to the original MD data. Moreover the underlying assumptions of the model (such as the time scale separation) can directly be checked.[15,31] Rather than assuming that some intuitively chosen free energy curve may explain a dynamical process, a multidimensional dLE model thus aims to demonstrate that we get the right answers for the right reasons.

In this work, we wish to go one step beyond the above described numerical Langevin modeling of an MD time series. By adopting a suitable functional form of the free energy landscape and the friction field, we aim to construct a *global analytical model* of the dynamics. On the one hand, an analytically defined model of biomolecular dynamics allows us to investigate the effect of theoretical features. This includes the dimensionality of the system and the number, character, and connectivity of metastable states as well as the distributions of barrier heights and the overall ruggedness of the energy landscape. On the other hand, the model facilitates the study of experimental changes of a specific molecular system, such as those of site-selective mutations. Considering protein folding, for example, we might study the effect of on- and off-route intermediate states on the rate and the structural heterogeneity of the folding process.

Rather than using general fitting methods, here we employ a problem-adapted description, referred to as "empirical valence bond" (EVB) model,[32–34] which was originally designed to describe nonadiabatic reactions such as electron or proton transfer. Adopting alanine dipeptide as a simple standard model of conformational dynamics, we first introduce the basic ideas of the ansatz. To demonstrate the potential of the approach and to discuss various approximations to represent multidimensional fields, we then consider a three-dimensional model of the structural dynamics of heptaalanine.

II. DATA-DRIVEN LANGEVIN EQUATION

To introduce a versatile method to calculate multidimensional Langevin fields, we first review the data-driven Langevin equation (dLE) formulation of Ref. 31. To briefly introduce the main results, we consider a discrete time series $\boldsymbol{x}_n \equiv \boldsymbol{x}(n\delta t)$ with time step δt of the dimensionless coordinate $\boldsymbol{x}$ (obtained, e.g., from a MD simulation). Discretizing the time derivatives of LE (1), we obtain the Euler-Maruyama equation

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \hat{\boldsymbol{f}}_n - \hat{\Gamma}_n(\boldsymbol{x}_n - \boldsymbol{x}_{n-1}) + \hat{\mathcal{K}}_n \boldsymbol{\xi}_n, \tag{4}$$

where we introduced dimensionless fields

$$\begin{aligned} \hat{\boldsymbol{f}}_n &= \mathcal{M}^{-1}\delta t^2 \boldsymbol{f}(x_n), & \hat{\Gamma}_n &= \mathcal{M}^{-1}\delta t \Gamma(x_n) - \mathbb{I}, \\ \hat{\mathcal{K}}_n &= \mathcal{M}^{-1}\delta t^{3/2}\mathcal{K}(x_n), & \boldsymbol{\xi}_n &= \boldsymbol{\xi}(x_n)\sqrt{\delta t}. \end{aligned} \tag{5}$$

Owing to the discretization, the "dLE fields" (denoted by a hat symbol) inherently depend on the time step δt, while the corresponding "physical" Langevin fields f, Γ, and $\mathcal{K}$ do not (within the validity of the underlying assumptions).

Due to the stochastic character of the noise, the fields cannot be obtained directly from the input data, but need to be calculated by invoking a coordinate-dependent statistical average over the noise.[16] Considering some function $g(x)$ that is to be locally averaged at a reference point x_n, we obtain

$$\langle g(x_n)\rangle = \frac{1}{k}\sum_m g(x_m), \tag{6}$$

where the sum goes over the k nearest neighbors of reference point $\boldsymbol{x}_n$. The neighborhood size k (typically $k \sim 10^2$) should be small enough to yield local (i.e., coordinate-dependent) averages, but also large enough to obtain statistical convergence.[30]

Using $\Delta\boldsymbol{x}_n = \boldsymbol{x}_n - \boldsymbol{x}_{n-1}$, we obtain the following explicit expressions for the dLE fields:[31]

$$\begin{aligned} \hat{\Gamma}_n &= \mathcal{C}(\Delta\boldsymbol{x}_{n+1}, \Delta\boldsymbol{x}_{n-1}) \cdot \mathcal{C}^{-1}(\Delta\boldsymbol{x}_{n-1}, \Delta\boldsymbol{x}_{n-1}), \\ \hat{\boldsymbol{f}}_n &= \langle \Delta\boldsymbol{x}_{n+1}\rangle + \hat{\Gamma}_n \langle \Delta\boldsymbol{x}_n\rangle, \\ \hat{\mathcal{K}}_n \hat{\mathcal{K}}_n^{\mathcal{T}} &= \mathcal{C}(\Delta\boldsymbol{x}_{n+1}, \Delta\boldsymbol{x}_{n+1}) - \hat{\Gamma}_n \mathcal{C}(\Delta\boldsymbol{x}_{n-1}, \Delta\boldsymbol{x}_{n+1}), \end{aligned} \tag{7}$$

where $\mathcal{C}(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}\boldsymbol{y}^T\rangle - \langle \boldsymbol{x}\rangle\langle \boldsymbol{y}^T\rangle$. Finally, the noise amplitude matrix $\hat{\mathcal{K}}$ is calculated from $\hat{\mathcal{K}}\hat{\mathcal{K}}^{\mathcal{T}}$ via a Cholesky decomposition.[30] We note that the original derivation[31] of these fields assumed "dense sampling," which allows for the additional approximation $\langle \boldsymbol{x}_n\rangle \approx \boldsymbol{x}_n$. This leads to the replacement $\Delta\boldsymbol{x}_k \to \boldsymbol{x}_k$ in the above covariance matrices, which facilitates their inversion.[31] Avoiding this approximation, on the other hand, the resulting expressions in Eq. (7) show a better convergence with respect to the neighborhood size.[35] The dLE program can be downloaded from http://www.theochem.uni-frankfurt.de/ hegger/langevin.tar.gz.

Apart from the choice of the neighborhood size k in Eq. (6), the only other parameter of the dLE simulation to

fix is the time step δt.[15,30] It is bounded from below by the discretization of the given input data, while the upper bound is determined by the convergence properties of the Euler-Maruyama integration scheme in Eq. (4). In practice, we need to find the optimal δt to reproduce the high-dimensional dynamics in the projected low-dimensional space of the collective variables. That is, δt should be large enough to warrant a delta-correlated Gaussian-distributed realization of the noise, but small enough to obtain accurate Langevin fields.[30]

To facilitate the treatment of multidimensional systems, the dLE method proposed in Ref. 31 performs the calculation of the Langevin fields "on the fly," i.e., at every propagation step of the dLE. As outlined in the following, the idea of this work is to use this *local* information obtained from the dLE to subsequently construct a *globally defined* model of the Langevin fields.

III. CONSTRUCTION OF GLOBAL MODEL

To introduce the basic idea, we adopt MD simulations of alanine dipeptide, which are considered as reference calculations for this small system. We then introduce an EVB-type description to model the free energy landscape, run dLE simulation to determine the other Langevin fields, and compare the conformational dynamics produced by the resulting model to the MD results.

A. MD simulations

A 160 ns long MD trajectory of "alanine dipeptide" (Ac-Ala-NHCH$_3$, AlaD) at $T = 300$ K was obtained using standard methods (Amber ff99SB/TIP3P[36] on GROMACS,[37] see the supplementary material). The peptide contains only a single pair of flexible backbone dihedral angles (ϕ, ψ), of which only the ψ angle is considered, as it accounts for the essential dynamics of the system (see Fig. S1 of the supplementary material). To minimize circular effects of the angular coordinate, we furthermore introduce the shifted coordinate $x = \psi - 70$, see Fig. S1 of the supplementary material. The resulting free energy landscape $F(x)$ shown in Fig. 1(a) shows two minima corresponding to the α_R helical ($x \approx -83$) and β/P_{II} extended ($x \approx 85$) conformational states, which are separated by a barrier at $x \approx 10$.

As a standard quantity to account for the conformational dynamics of the one-dimensional system, we consider the position autocorrelation function $C(t) = \langle \delta x(t) \delta x(0) \rangle / \langle \delta x^2 \rangle$ with $\delta x = x - \langle x \rangle$. Figure 1(b) reveals that $C(t)$ decays on a time scale of ≈ 100 ps, which is related to the mean value of the *first passage time* τ_{FP}, i.e., the average time that passes between the first leaving of the initial state and the reaching of the final state. (See the supplementary material for details of the calculation.) The distribution of τ_{FP} in Fig. 1(c) shows an exponential behavior which is typical for Poisson processes.[38] To characterize the transition itself, Fig. 1(d) depicts the distribution of the, in general, much shorter *transition path time* τ_{tp},[25,39,40] which indicates that a transition between the two wells takes on average of ≈ 1 ps. We note that biomolecular transition path times quite recently have become amenable to experimental observation.[41]

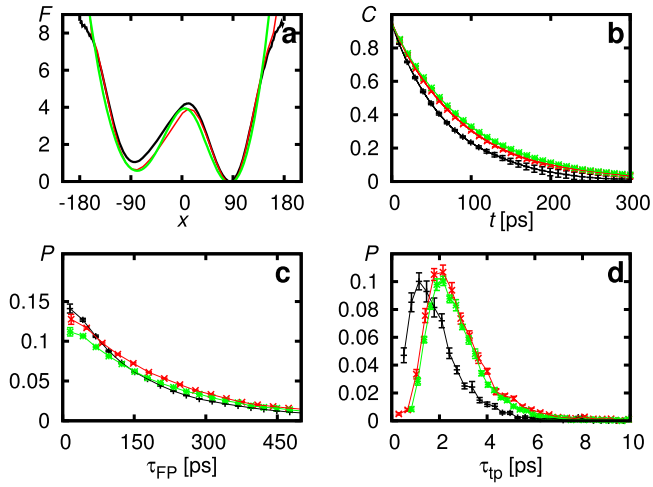

FIG. 1. Conformational dynamics of the one-dimensional model of alanine dipeptide (AlaD), showing (a) free energy landscape $F(x)$ (in units of $k_B T$), (b) position autocorrelation function $C(t)$, (c) distribution of first passage times, and (d) distribution of transition path times. The distributions are averaged over forward and backward transitions. Compared are results obtained by MD simulations (black), dLE simulations (red), and the EVB model (green). Bars indicate standard deviations obtained from calculations using the first and second half of the respective trajectory.

B. EVB model of the free energy landscape

Inspired by the "empirical valence bond" (EVB) model,[32–34,42–45] we describe the minima of the free energy surface by simple functional forms V_{ii}, representing metastable conformational states. These states interact via couplings V_{ij} accounting for the energy barriers. In the example of AlaD, we make a harmonic ansatz for the two minima centered at $x^{(1)} = 85$ and $x^{(2)} = -83$,

$$V_{ii}(x) = \epsilon_i + \frac{\omega_i}{2}(x - x^{(i)})^2, \tag{8}$$

where ϵ_i denotes the minimal energy and $\omega_i = 1/\sigma_i^2$ reflects the width σ_i of minimum i, respectively. Assuming that the couplings V_{ij} between the two states are constant and symmetric, we construct a "diabatic" energy matrix,[46] whose lower eigenvalue represents an intuitive model of the free energy landscape (Fig. 2(a)). To this end, we solve the secular equation of the resulting eigenvalue problem

$$\begin{vmatrix} V_{11} - F & V_{12} \\ V_{21} & V_{22} - F \end{vmatrix} = 0, \tag{9}$$

and obtain the desired free energy function as the ground-state solution

$$F(x) = \frac{1}{2}(V_{11} + V_{22}) - \frac{1}{2}\sqrt{(V_{11} - V_{22})^2 + 4V_{12}^2}. \tag{10}$$

We use a simple iterative scheme to determine the parameters of the EVB model. First, we perform local harmonic fits of the two minima, which give initial values for the harmonic parameters $\epsilon_i, x^{(i)}$ and σ_i^2. To obtain the couplings V_{ij} between both states, we then minimize the quantity

$$\Delta_{ij} = \left(E_{i\to j} - F_{i\to j}\right)^2 + \left(E_{j\to i} - F_{j\to i}\right)^2, \tag{11}$$

where $E_{i\to j}$ and $F_{i\to j}$ denote the barrier heights of EVB model and MD reference, respectively. Insertion of the resulting couplings V_{ij} in Eq. (10) yields a first estimate of $F(x)$, which

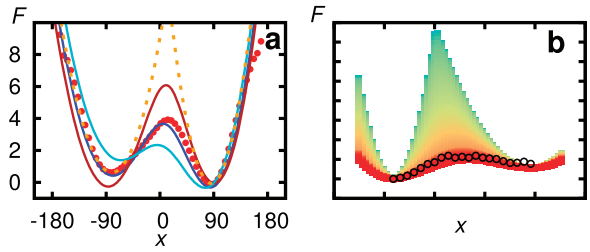

FIG. 2. (a) Free energy landscape $F(x)$ (in units of k_BT) of AlaD, as obtained by dLE simulations (red circles) and various versions of the EVB model [Eq. (10)]. Starting with uncoupled harmonic potentials $V_{ii}(x)$ of the two minima [Eq. (8), dashed orange lines], the coupling V_{12} is varied to approximate the reference results. The cyan and brown lines show examples of too large and too small couplings and the blue line shows the optimized result. (b) Illustration of the fitting procedure of the EVB model to the dLE data (black circles). With increasing number of iterations (and increasing coupling V_{12}), the model results change from green to red.

in general requires some readjustment of the harmonic fits. After a number of iterations of adjusting V_{ij} and the harmonic parameters, we get the optimal EVB representation of the free energy.

In the case of AlaD, we obtain for the positions of the minima $x_1 = 98$ and $x_2 = -94$, for the state energies and width parameters (in units of k_BT) $\epsilon_1 = 4.17$, $\omega_1 = 0.0046$, and $\epsilon_2 = 2.75$, $\omega_2 = 0.0028$ and for the barrier parameter $V_{12} = V_{21} = 17.38$. While the harmonic ansatz in Eq. (8) is appealing due to its simplicity, the approach is readily generalized to more complicated functions (e.g., Morse potentials). The generalization of the method to systems with many minima in several dimensions is given below.

C. Global model of structural dynamics

To devise a global model of the dynamics as motivated in the Introduction, we next need to construct a model for the friction Γ. To this end, we run a dLE simulation (see Sec. II), using the 160 ns MD trajectory of AlaD as input data. We choose a time step $\delta t = 0.02$ ps to resolve the sub-ps dynamics of the system, which yields $8 \cdot 10^6$ input data points for the dLE. Moreover, we choose $k = 200$ neighbor points to evaluate the local averages [Eq. (6)] and propagate the dLE in Eq. (4) for $8 \cdot 10^6$ steps. As shown in Fig. S2 of the supplementary material, the resulting noise model fulfills the assumptions of being delta-correlated and of zero mean and unit variance.

To validate the dLE simulation, we use the resulting dLE time series $x(t)$ to calculate the free energy landscape $F(x)$ and the autocorrelation function $C(t)$ as well as the distributions of first passage times and transition path times. Figure 1 reveals that all observables are in good agreement with the reference MD data, although the overall dynamics (as reflected by the autocorrelation function and the first passage times) of the dLE is slightly slower than the MD results. This finding is in line with the somewhat higher barrier of the dLE free energy curve, see Fig. 1(a). Moreover, we find that the mean transition path time of the dLE is a factor 2 slower than for the MD (Fig. 1(d)). This may be due to a *local* breakdown of the Markov approximation in the barrier region, although *on average* the noise model is well fulfilled (Fig. S2 of the supplementary material). Revealing transition path times about 1 ps, the assumption of a time scale separation between system motion and bath fluctuations comes to its limits.

The dLE simulation yields the drift $\hat{f}(x)$, friction $\hat{\Gamma}(x)$, and noise amplitude $\hat{\mathcal{K}}(x)$ shown in Fig. 3, which completely characterize the system. To facilitate a simple and physically appealing definition of the global model, on the other hand, we wish to employ Eq. (2) to calculate the drift $\hat{f}$ from the free energy landscape F via $\hat{f}(x) = -\mathcal{M}^{-1}\delta t^2 \nabla F(x)$. Moreover, we want to use the fluctuation-dissipation theorem in Eq. (3) to calculate the noise amplitude $\hat{\mathcal{K}}(x)$ from the friction $\hat{\Gamma}(x)$. These relations require explicit knowledge of the mass $\mathcal{M}$, which is implicitly contained in the Langevin fields [Eq. (5)]. When we request the fluctuation-dissipation theorem in Eq. (3) to hold, we obtain

$$\mathcal{M} = m_0 \hat{\mathcal{K}}^{-2}(x)[\hat{\Gamma}(x) + 1], \tag{12}$$

where $m_0 \equiv 2k_BT\delta t^2$. Using units such that $k_BT = 38$ ps^{-1}, we find $m_0 = 0.0304$ ps.

Figure 3(d) shows that the resulting mass slightly varies with coordinate x, which may be a consequence of the fact that x describes the rotation along the backbone dihedral angle ψ of AlaD. Accounting for the moment of inertia, $\mathcal{M}$ may therefore depend in a complicated way on the coordinates of the peptide (and even of the solvent). Generally speaking, we note that masses associated with collective coordinates are often found to be difficult to interpret and are therefore considered as auxiliary parameters in the dLE description. In what follows, we consider the coordinate dependence as negligible and approximate the mass by its mean value, $\mathcal{M}/m_0 = 0.036$. We note that this approximation is a necessary condition for the

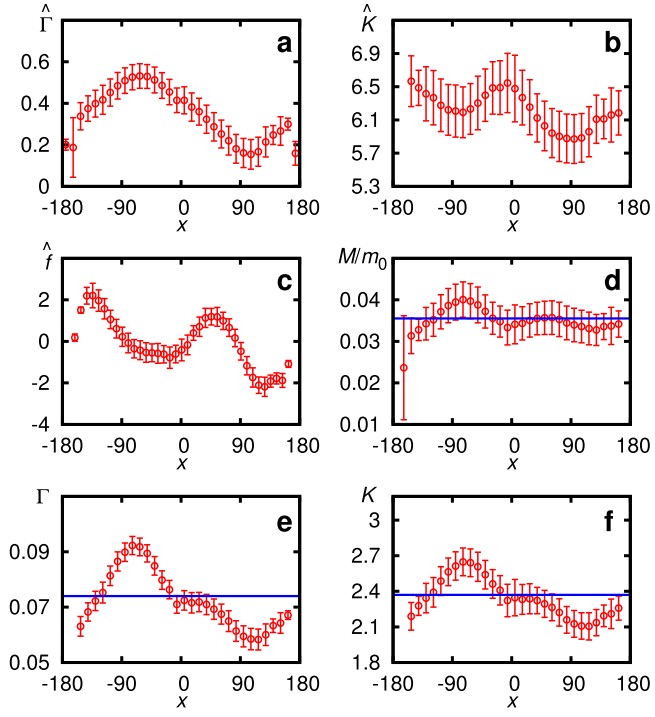

FIG. 3. Langevin fields of the one-dimensional model of AlaD. The dLE simulation yields the dimensionless fields: (a) friction $\hat{\Gamma}(x)$, (b) noise amplitude $\hat{\mathcal{K}}(x)$, and (c) drift $\hat{f}(x)$. By calculating the mass $\mathcal{M}/m_0$ shown in (d), we obtain the "physical Langevin fields" $\Gamma(x)$ (e) and $\mathcal{K}(x)$ ((f), in units of ps$^{1/2}$). Horizontal lines refer to mean values and bars indicate the standard deviation.

validity of the equipartition theorem,[47] which represents the basis for the fluctuation-dissipation theorem we want to use.

Inserting the result for $\mathcal{M}$ in Eq. (5), we finally obtain the desired Langevin fields $f(x)$, $\Gamma(x)$, and $\mathcal{K}(x)$ which are shown in Fig. 3. As a consistency check, we notice that the resulting drift $f(x)$ perfectly matches the negative derivative of the free energy $F(x)$ calculated from Eq. (2) (data not shown). In contrast to the pronounced position dependence of the drift, the friction $\Gamma(x)$ as well as the noise amplitude $\mathcal{K}(x)$ of AlaD is seen to vary only little with x. This is in line with previous studies on AlaD,[48] which suggests that the calculation of the friction tensor via Eq. (7) is a versatile alternative to standard methods,[9–15] in particular for multidimensional systems. For simplicity of the model, we therefore approximate these fields by their mean values, giving $\Gamma = 0.074$ and $\mathcal{K} = 2.371$. At least for the present case, this approximation does not affect the results obtained for the resulting global model discussed below (data not shown).

D. Model-based Langevin equation

The above obtained values for the mass and the friction together with the above reported parameters of the EVB model [Eq. (10)] of the free energy landscape constitute the desired global model of the conformational dynamics of AlaD. By solving LE (1) for this model, henceforth referred to as model-based Langevin equation (mLE), we generate a time series $x(t)$ which exhibits the same statistical properties as the dLE input data. In particular, the global model produces the same distribution (by construction of $F(x)$) and recovers the autocorrelation function $C(t)$ as well as the distributions of first passage times and transition path times, see Fig. 1.

Several comments are in order:

- The mLE is much faster to solve than the corresponding dLE, as the latter needs to perform an extensive search through the MD input data at every time step. In the present example with $\approx 10^7$ MD data points and $\approx 10^7$ LE time steps, the speed-up factor is about 10^4.
- Using precalculated physical fields [Eq. (5)], the mLE can be integrated using a different time step than used in the preceding dLE. This is because the time step of the dLE determines the Langevin fields, while the mLE time step in principle can be chosen arbitrarily (as long as the numerical integration of the equation of motion converges).
- The dLE fields exhibit relatively large standard deviations (indicated by the bars in Fig. 3), which means that at every time step of the dLE propagation, the fields are estimated with significant uncertainty. This is a consequence of the local average [Eq. (6)], whose variance vanishes in the limit of large neighborhood sizes at constant point density.[30] The mLE, on the other hand, uses the mean value of the fields, which is given within a standard error (standard deviation divided by the square root of the number of points in a bin) which is virtually zero. In this sense the global model represents a refinement of the original data.
- Recalling Zwanzig's approach to derive LEs from a microscopic system-bath ansatz,[1] we note that a position-independent friction indicates a system-bath coupling that is linear in the system coordinate but may be nonlinear in the bath coordinate.[49–52] Generally speaking, the finding of a position-independent friction represents the hallmark of a suitable collective coordinate,[7,53,54] which facilitates a direct interpretation of the structural dynamics in terms of the free energy landscape.[4–6] Rather than postulating constant friction, the above described Langevin analysis allows us to identify good collective coordinates by calculating $\Gamma(x)$.

IV. MULTIDIMENSIONAL FORMULATION

A. General EVB model

It is straightforward to extend the above described EVB ansatz for a one-dimensional double well to a model for a N-state system with a d-dimensional collective variable $\boldsymbol{x}$. To this end, we generalize the harmonic ansatz in Eq. (8) to the $N \times N$ matrix $\mathcal{V} = \{V_{ij}\}$ with

$$V_{ii}(\boldsymbol{x}) = \epsilon_i + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^{(i)})^T \Omega_i (\boldsymbol{x} - \boldsymbol{x}^{(i)}), \tag{13}$$

where $\boldsymbol{x}^{(i)}$ denotes the position of state i and $\Omega_i = \{\delta_{nm}/(\sigma_n^2)\}$ is a $d \times d$ matrix that accounts for its width. For simplicity, we again assume that the off-diagonal elements V_{ij} are constant and symmetric. Diagonalization of $\mathcal{V}$ leads to the eigenvalue problem $\mathcal{V}\mathcal{U} = \mathcal{F}\mathcal{U}$ with eigenvector matrix $\mathcal{U}$ and eigenvalue matrix $\mathcal{F} = \{\delta_{ij}F_i\}$. The desired free energy function is obtained as the lowest eigenvalue, giving

$$F(\boldsymbol{x}) = \left(U^\dagger(\boldsymbol{x})\mathcal{V}(\boldsymbol{x})U(\boldsymbol{x})\right)_{11}. \tag{14}$$

To introduce a practical method to determine the parameters of the multidimensional EVB model, our starting point is again a suitable time series $\boldsymbol{x}(t)$ (that describes the conformational dynamics of the system) and a suitable clustering of the data[55–57] (to identify the N metastable conformational states of the system). Assuming first that all off-diagonal elements V_{ij} are zero, it is straightforward to calculate the harmonic parameters ϵ_i, $\boldsymbol{x}^{(i)}$, and Ω_i of Eq. (13). Next we wish to get a first estimate of the $N^2/2-N$ off-diagonal elements V_{ij} determining the barriers between states i and j. To this end, we consider only pairs of states that are directly connected by a single barrier along their connecting vector $\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(i)}$ and set all other off-diagonal elements $V_{ij} = 0$. This yields K pairs (i,j) of connected states. Defining $\boldsymbol{r}_l = \boldsymbol{x}^{(i)} + l/(L-1)(\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(i)})$ with $l \in \{0, \ldots, L-1\}$, we calculate the free energy profile along the connecting vector

$$F_{ij}(\boldsymbol{r}_l) = -k_\mathrm{B}T \ln \sum_n \Theta\left(|\boldsymbol{r}_l - \boldsymbol{x}_n| < \mu\right), \tag{15}$$

where the sum is taken over all frames $\boldsymbol{x}_n$ of the input trajectory using a fixed value for μ. From this estimate, we obtain the barrier height $\Delta F_{i\to j} = \max(F_{ij}(\boldsymbol{r}_l)) - F_{ij}(\boldsymbol{r}_0))$.

Once the connectivity of the states is established and the barrier heights between all connected states are computed, we iteratively adjust the coupling constants V_{ij} and the harmonic parameters $\boldsymbol{x}^{(i)}$ and Ω_i until the quantity Δ_{ij} [Eq. (11)] is minimized. To this end, we start with some small initial value (e.g., $0.001\ k_\mathrm{B}T$) for all couplings V_{ij} and calculate the resulting values for Δ_{ij}. Since the coupling introduces a shift of the

position and energy of the two connected states, we need to readjust the harmonic parameters such that the two minima retain their reference positions $(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$ and energies (ϵ_i, ϵ_j), respectively. Denoting the present position of state i by $\tilde{\boldsymbol{x}}^{(i)}$, this is achieved by

$$\tilde{\boldsymbol{x}}^{(i)} \to [\tilde{\boldsymbol{x}}^{(i)} + \boldsymbol{x}^{(i)}]/2, \tag{16}$$

and similarly for the off-set energies ϵ_i. This step is repeated until $\tilde{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}$ is below a fixed threshold. After performing this adjustment for each pair of states, the couplings V_{ij} are increased and new harmonic parameters are calculated. This procedure is repeated until the average deviation, $K^{-1} \sum_{ij} \Delta_{ij}$, is below a given convergence value [typically 0.01 $(k_B T)^2$], see Fig. 2(b).

Apart from the representation of $F(\boldsymbol{x})$, we also need to consider the coordinate dependence of the friction (and possibly of the mass tensor). To obtain sufficiently smooth and well sampled fields, in a first step a smoothing procedure is applied to the data (see the supplementary material). Since the coordinate dependence of the friction often correlates with the free energy,[7,53] again a EVB modeling could be applied. In the cases discussed below, however, it proved sufficient to approximate these fields as constant.

B. Conformational dynamics of Ala_7

To demonstrate the potential of the above introduced methodology, we now construct a three-dimensional model of the conformational dynamics of heptaalanine (Ala_7). To this end, we adopt the 800 ns trajectory at 300 K provided by Altis *et al.*,[21] which used the GROMOS 45A3 force field,[58] SPC explicit water,[59] and a sampling rate of 1 frame/ps. Since the motion of the terminal residues is hardly correlated to the motion of the inner residues, we focus on the three inner residues. As for AlaD, we consider the dynamics along the ψ_i (i = 1,2,3) dihedral angles, which discriminate α_R helical and β/P_{II} extended conformational states. Hence the resulting free energy landscape is expected to exhibit $2^3 = 8$ metastable states, which are labeled by $\alpha\alpha\alpha$, $\alpha\alpha\beta$, $\alpha\beta\alpha$, $\beta\alpha\alpha$, $\alpha\beta\beta$, $\beta\alpha\beta$, $\beta\beta\alpha$, and $\beta\beta\beta$. For convenience, we again shift all coordinates, $\tilde{x}_i = \psi_i - 70$ and perform a 45° rotation of the coordinate system, $\boldsymbol{x} = \mathcal{R}\tilde{\boldsymbol{x}}$, such that the rotated coordinates x_i resolve all eight states in the x_1-x_2 projection of the three-dimensional space (see the supplementary material).

Figure 4(a) shows the resulting free energy landscape $F(x_1, x_2)$ obtained from the MD simulations. The eight metastable states are seen to lie on the corners of a three-dimensional cube, whose sides are indicated in red. Interstate transitions along the sides of the cube involve the conformational change of a single residue only and therefore occur quite frequently (≈1 ns). Space diagonals of the cube correspond to transitions that change the conformation of all three residues, which are relatively rare (≈100 ns). All transition rates are comprised in Fig. 4(d). The autocorrelation function $C(t)$ for coordinate x_1 in Fig. 4(e) is seen to decay within 1 ns, which qualitatively matches the time scale of the interstate transitions of single residues. Lastly, Fig. 4(e) shows the distribution of the corresponding transition path times, which is peaked at ≈ 10 ps.

Using a time step δt = 2 ps and a neighborhood size of k = 200, we run 5 dLE simulations of each 8×10^7 steps. Figure S4 of the supplementary material shows that the resulting noise model fulfills the assumptions of being delta-correlated and of zero mean and unit variance. Due to the 10-fold increase of data points, the resulting free energy landscape shown in Fig. 4(c) is somewhat smoother than the MD result (Fig. 4(a)), especially in the barrier regions. To assess the quality of the dLE simulations, Fig. 4(d) compares the MD and dLE rates of all interstate transitions. While the overall agreement is quite good, in particular for the well-sampled transitions with high rates, the dLE is seen to generally overestimate the rates. In line with this finding, the decay of the autocorrelation function is somewhat faster for the dLE. As for AlaD, the dLE also exhibits a larger mean transition path time (Fig. 4(f)), reflecting problems of the dLE to accurately account for the details of rare barrier crossing events.

C. Global model

Using the procedure described above, we construct a three-dimensional EVB model of the free energy landscape $F(\boldsymbol{x})$ of Ala_7. While the model can be parameterized using either MD or dLE simulation data, we used dLE data as they provide a better sampling of the system, especially in the barrier regions. The supplementary material comprises all EVB parameters obtained in this way. The resulting free energy landscape shown is found to closely resemble the dLE reference (data not shown).

As explained above, a global model furthermore requires us to determine the friction Γ and the mass $\mathcal{M}$ of the system. As representative examples of the coordinate dependence of these fields, Fig. S3 of the supplementary material shows the first diagonal element $\Gamma_{11}(x_1, x_2)$ and the off-diagonal element $\Gamma_{12}(x_1, x_2)$ of the friction matrix as obtained from the dLE simulations. The figure reveals that the friction is largely constant and mainly shows deviations in areas which are hardly sampled by the system (cf. Fig. 4(a)). By calculating mean and standard deviation of the matrix elements, we obtain

$$\Gamma = \begin{pmatrix} 0.35 \pm 0.03 & 0.002 & 0.001 \\ 0.001 & 0.43 \pm 0.04 & 0.001 \\ 0.001 & 0.001 & 0.41 \pm 0.03 \end{pmatrix}. \tag{17}$$

The off-diagonal elements Γ_{ij} are found to fluctuate around zero and can safely be neglected (standard deviations are not shown). The diagonal elements Γ_{ii} are quite similar and show standard deviations of about 10%. As discussed for the Langevin fields of AlaD in Fig. 3, the corresponding standard *errors* of the matrix elements are almost zero.

To calculate the dLE drift field $\hat{f}(x)$ from the free energy landscape $F(\boldsymbol{x})$, we need to determine the mass tensor $\mathcal{M}$. Although more general forms of $\mathcal{M}$ may be constructed,[60] for our purposes it proved sufficient to treat $\mathcal{M}$ as a diagonal constant matrix. Calculating the mass tensor via Eq. (12), we get

$$\mathcal{M}/m_0 = \mathrm{diag}(1.9 \pm 0.4, 2.3 \pm 0.5, 2.2 \pm 0.3) \cdot 10^{-3} \tag{18}$$

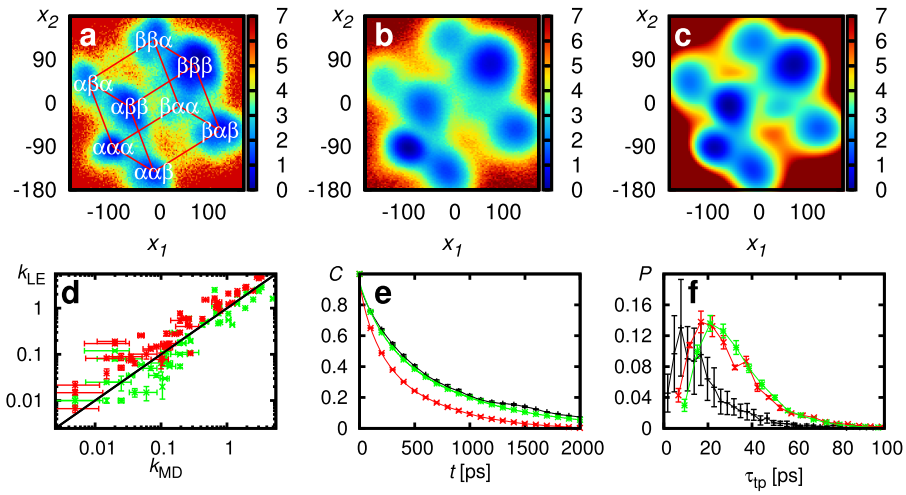

FIG. 4. (a) Free energy landscape $F(x_1, x_2)$ (in units of k_BT) of the three-dimensional model of heptaalanine (Ala_7) as obtained from the MD simulation. The employed x_1-x_2 projection allows us to separate the eight conformational states of the model, which are indicated by a three-letter code. The states are seen to lie on the corners of a three-dimensional cube whose sides are indicated in red. Panels (b) and (c) show the corresponding free energy landscapes obtained from the dLE and mLE, respectively. (d) Transition rates (in units of ns^{-1}), (e) x_1 autocorrelation function, and (f) transition path time distribution as obtained from MD (black), dLE (red), and mLE (green) simulations, respectively. Bars indicate standard deviations obtained from calculations using the first and second half of the respective trajectory.

with m_0 = 304 ps. Equations (17) and (18) together with the above described EVB model constitute the desired global model of the conformational dynamics of Ala_7.

Using the model, we propagated the mLE for 20 μs. By construction, the resulting free energy landscape perfectly recovers the dLE results (Fig. 4(c)). Comparing transition rates and autocorrelation functions obtained for the mLE and the original MD data, we find good agreement (Figs. 4(d) and 4(e)). Interestingly, the mLE results reproduce in both cases the reference MD data even better than the corresponding dLE results. This may be a consequence of the fact that the fields are better represented by the global model than by the noisy dLE data.

D. Modifications of the model

Site-specific mutation studies are a standard tool to investigate the significance of individual residues for the function of a protein. For example, a suitable mutation may introduce a steric hindrance to the system, which effectively blocks some specific motion. Given a global mLE model of a system, we can achieve similar effects by blocking the motion of some specific degrees of freedom. As a simple example, here we localize the motion of the central torsion angle ψ_2, by increasing the four barriers over which transitions of the central angle occur, i.e., $a\alpha b \leftrightarrow a\beta b$, where $a, b \in \{\alpha, \beta\}$. Recalling that the conformational states of the model can be viewed as the corners of a cube in the free energy landscape (Fig. 4(a)), this modification restricts the motion to a plane of the cube which exhibits the same central conformation (either α or β, see Fig. S5 of the supplementary material).

Considering the process $\alpha a\alpha \rightarrow \beta b\beta$ ($a, b \in \{\alpha, \beta\}$) where both outer residues undergo an $\alpha \rightarrow \beta$ transition, we wish to calculate the mean first passage times τ_{org} and τ_{mod} for the original and modified system, respectively. As listed in Table I, this process can be realized via four pathways in the original system, but only via two in the modified one. *A priori* it is not obvious which system produces the faster dynamics: the original system which exhibits more pathways to accomplish the process or the modified system whose dimensionality is reduced to only two degrees of freedom. Interestingly, we find that the mean first passage times of the individual pathways are shorter for the modified system with reduced dimensionality. On the other hand, the original system using four (instead of two) pathways exhibits shorter *total* mean first passage times (shown in the last line of Table I). For the backward-reaction $\alpha a\alpha \leftarrow \beta b\beta$, we find similar results.

TABLE I. Mean first passage times (in units of ns) of the various pathways of the process $\alpha a\alpha \leftrightarrow \beta b\beta$ ($a, b \in \{\alpha, \beta\}$) as well as of the total process. Compared are the results for the forward-reaction ($\tau_{\rightarrow}$) and the backward-reaction ($\tau_{\leftarrow}$), obtained for the original and the modified model of Ala_7.

	$\tau_{\rightarrow}^{org}$	$\tau_{\rightarrow}^{mod}$	$\tau_{\leftarrow}^{org}$	$\tau_{\leftarrow}^{mod}$
$\alpha\alpha\alpha \leftrightarrow \beta\alpha\beta$	3.1	2.2	2.8	0.8
$\alpha\alpha\alpha \leftrightarrow \beta\beta\beta$	1.6	–	2.7	–
$\alpha\beta\alpha \leftrightarrow \beta\alpha\beta$	3.3	–	3.4	–
$\alpha\beta\alpha \leftrightarrow \beta\beta\beta$	1.2	0.7	2.7	2.0
$\alpha a\alpha \leftrightarrow \beta b\beta$	1.1	1.4	1.4	1.5

V. CONCLUDING REMARKS

In an attempt to extend the paradigmatic system-bath ansatz of Zwanzig to realistic data-based biomolecular systems, we have outlined a practical method to construct an analytically defined global model of structural dynamics. Given a MD simulation and adequate collective coordinates to account for the system's essential dynamics, the approach employs a problem-adapted EVB ansatz for the multidimensional free energy landscape as well as an approximate description of the friction field of the dynamics. Used in a nonoverdamped LE, the resulting model has been shown to semiquantitatively

reproduce the results of the underlying all-atom MD simulation. It may therefore serve as a "post-simulation" model to analyze and interpret the ever growing amount of MD data. Moreover, the analytically defined ansatz allows us to investigate the dependence of the system on parameter changes and to predict the effect of site-selective mutations on the dynamics.

The main practical result of the present investigation is maybe the plain conclusion that the approach works. That is, the low-dimensional global model (respectively the corresponding mLE) satisfies the presupposed conditions (such as memory-free noise) and reproduces correctly the dynamics of the underlying high-dimensional MD simulation. Recalling that low-dimensional projections of the free energy landscape often yield incorrect state connectivities and spurious dynamics, this is a nontrivial result, which also represents a stringent test of the applied collective variables. Furthermore, we note that the proposed modeling approach may be employed to refine the (often poorly sampled) original Langevin fields. This is achieved by replacing data-based fields with a large standard deviation by model-based fields with a small standard error. Using the methodology developed in this paper, we currently consider the mLE description of more complex processes, such as the folding of villin headpiece and the functional dynamics of lysozyme.

SUPPLEMENTARY MATERIAL

See supplementary material for MD simulation details, definitions of reaction coordinates, calculations of first passage times, transition rates and transition path times, properties of the noise model, representation of multidimensional fields, EVB model parameters, and modifications of the Ala_7 model.

ACKNOWLEDGMENTS

We thank Björn Bastian and Rainer Hegger for numerous instructive and helpful discussions, as well as Fabian Thielemann for providing the MD data of AlaD as part of his bachelor project. This work has been supported by the Deutsche Forschungsgesellschaft.

[1] R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University, Oxford, 2001).
[2] H. J. C. Berendsen, *Simulating the Physical World* (Cambridge University Press, Cambridge, 2007).
[3] C. Hijon, P. Espanol, E. Vanden-Eijnden, and R. Delgado-Buscalioni, "Mori-Zwanzig formalism as a practical computational tool," Faraday Discuss. **144**, 301 (2010).
[4] J. N. Onuchic, Z. L. Schulten, and P. G. Wolynes, "Theory of protein folding: The energy landscape perspective," Annu. Rev. Phys. Chem. **48**, 545 (1997).
[5] K. A. Dill and H. S. Chan, "From Levinthal to pathways to funnels: The "new view" of protein folding kinetics," Nat. Struct. Biol. **4**, 10 (1997).
[6] D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, 2003).
[7] R. B. Best and G. Hummer, "Coordinate-dependent diffusion in protein folding," Proc. Natl. Acad. Sci. U. S. A. **107**, 1088 (2010).
[8] J. C. F. Schulz, L. Schmidt, R. B. Best, J. Dzubiella, and R. R. Netz, "Peptide chain dynamics in light and heavy water: Zooming in on internal friction," J. Am. Chem. Soc. **134**, 6273 (2012).
[9] J. E. Straub, M. Borkovec, and B. J. Berne, "Calculation of dynamic friction on intramolecular degrees of freedom," J. Phys. Chem. **91**, 4995 (1987).
[10] J. Gradišek, S. Siegert, R. Friedrich, and I. Grabec, "Analysis of time series from stochastic processes," Phys. Rev. E **62**, 3146 (2000).
[11] J. Timmer, "Parameter estimation in nonlinear stochastic differential equations," Chaos, Solitons Fractals **11**, 2571 (2000).
[12] R. B. Best and G. Hummer, "Diffusive model of protein folding dynamics with Kramers turnover in rate," Phys. Rev. Lett. **96**, 228104 (2006).
[13] O. F. Lange and H. Grubmüller, "Collective Langevin dynamics of conformational motions in proteins," J. Chem. Phys. **124**, 214903 (2006).
[14] I. Horenko, C. Hartmann, C. Schütte, and F. Noe, "Data-based parameter estimation of generalized multidimensional Langevin processes," Phys. Rev. E **76**, 016706 (2007).
[15] C. Micheletti, G. Bussi, and A. Laio, "Optimal Langevin modeling of out-of-equilibrium molecular dynamics simulations," J. Chem. Phys. **129**, 074105 (2008).
[16] R. Hegger and G. Stock, "Multidimensional Langevin modeling of biomolecular dynamics," J. Chem. Phys. **130**, 034106 (2009).
[17] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, UK, 1997).
[18] K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, "Msmbuilder2: Modeling conformational dynamics on the picosecond to millisecond scale," J. Chem. Theory Comput. **7**, 3412 (2011).
[19] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noe, "Markov models of molecular kinetics: Generation and validation," J. Chem. Phys. **134**, 174105 (2011).
[20] G. R. Bowman, V. S. Pande, and F. Noe, *An Introduction to Markov State Models* (Springer, Heidelberg, 2013).
[21] A. Altis, M. Otten, P. H. Nguyen, R. Hegger, and G. Stock, "Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis," J. Chem. Phys. **128**, 245102 (2008).
[22] S. V. Krivov and M. Karplus, "Hidden complexity of free energy surfaces for peptide (protein) folding," Proc. Natl. Acad. Sci. U. S. A. **101**, 14766 (2004).
[23] M. A. Rohrdanz, W. Zheng, and C. Clementi, "Discovering mountain passes via torchlight: Methods for the definition of reaction coordinates and pathways in complex macromolecular reactions," Annu. Rev. Phys. Chem. **64**, 295 (2013).
[24] O. F. Lange and H. Grubmüller, "Generalized correlation for biomolecular dynamics," Proteins **62**, 1053 (2006).
[25] R. B. Best and G. Hummer, "Reaction coordinates and rates from transition paths," Proc. Natl. Acad. Sci. U. S. A. **102**, 6732 (2005).
[26] J. S. Hub and B. L. de Groot, "Detection of functional modes in protein dynamics," PLoS Comput. Biol. **5**, e1000480 (2009).
[27] G. Stock, A. Jain, L. Riccardi, and P. H. Nguyen, "Exploring the energy landscape of small peptides and proteins by molecular dynamics simulations," in *Protein and Peptide Folding, Misfolding and Non-Folding*, edited by R. Schweitzer-Stenner (Wiley, New York, 2012), p. 57.
[28] S. V. Krivov, "On reaction coordinate optimality," J. Chem. Theory Comput. **9**, 135 (2013).
[29] G. Perez-Hernandez, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noe, "Identification of slow molecular order parameters for Markov model construction," J. Chem. Phys. **139**, 015102 (2013).
[30] N. Schaudinnus, A. J. Rzepiela, R. Hegger, and G. Stock, "Data driven Langevin modeling of biomolecular dynamics," J. Chem. Phys. **138**, 204106 (2013).
[31] N. Schaudinnus, B. Bastian, R. Hegger, and G. Stock, "Multidimensional Langevin modeling of nonoverdamped dynamics," Phys. Rev. Lett. **115**, 050602 (2015).
[32] A. Warshel, "Electrostatic basis of structure-function correlation in proteins," Acc. Chem. Res. **14**, 284 (1981).
[33] Y.-T. Chang and W. H. Miller, "An empirical valence bond model for constructing global potential energy surfaces for chemical reactions of polyatomic molecular systems," J. Chem. Phys. **94**, 5884 (1990).
[34] H. Chen, P. Liu, and G. A. Voth, "Efficient multistate reactive molecular dynamics approach based on short-range effective potentials," J. Chem. Theory Comput. **6**, 3039 (2010).
[35] B. Bastian, N. Schaudinnus, and G. Stock, "Data driven Langevin equations" (unpublished).
[36] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple Amber force fields and development of improved protein backbone parameters," Proteins: Struct., Funct., Bioinf. **65**, 712 (2006).
[37] S. Pronk *et al.*, "Gromacs 4.5: A high-throughput and highly parallel open source molecular simulation toolkit," Bioinformatics **29**, 845 (2013).
[38] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam, 1997).

[39]P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, "Transition path sampling: Throwing ropes over rough mountain passes, in the dark," Annu. Rev. Phys. Chem. **53**, 291 (2002).
[40]W. K. Kim and R. R. Netz, "The mean shape of transition and first-passage paths," J. Chem. Phys. **143**, 224108 (2015).
[41]H. S. Chung and W. A. Eaton, "Single-molecule fluorescence probes dynamics of barrier crossing," Nature **502**, 685 (2013).
[42]A. Amadei, B. de Groot, M.-A. Ceruso, M. Paci, A. Di Nola, and H. Berendsen, "A kinetic model for the internal motions of proteins: Diffusion between multiple harmonic wells," Proteins: Struct., Funct., Genet. **35**, 283 (1999).
[43]K. Okazaki, N. Koga, S. Takada, J. N. Onuchic, and P. G. Wolynes, "Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations," Proc. Natl. Acad. Sci. U. S. A. **103**, 11844 (2006).
[44]P. Maragakis and M. Karplus, "Large amplitude conformational change in proteins explored with a plastic network model: Adenylate kinase," J. Mol. Biol. **352**, 807 (2005).
[45]W. Zheng, B. R. Brooks, and G. Hummer, "Protein conformational transitions explored by mixed elastic network models," Proteins: Struct., Funct., Genet. **69**, 43 (2007).
[46]W. Domcke and G. Stock, "Theory of ultrafast nonadiabatic excited-state processes and their spectroscopic detection in real time," Adv. Chem. Phys. **100**, 1 (1997).
[47]A. Jain, I. Park, and N. Vaideh, "Equipartition principle for internal coordinate molecular dynamics," J. Chem. Theory Comput. **8**, 2581–2587 (2012).
[48]G. Hummer, "Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations," New J. Phys. **7**, 34 (2005).
[49]B. Carmeli and A. Nitzan, "Theory of activated rate processes: Position dependent friction," Chem. Phys. Lett. **102**, 517 (1983).
[50]B. J. Berne, M. E. Tuckerman, J. E. Straub, and A. L. R. Bug, "Dynamic friction on rigid and flexible bonds," J. Chem. Phys. **93**, 5084 (1990).
[51]G. R. Haynes, G. A. Voth, and E. Pollak, "A theory for the activated barrier crossing rate constant in systems influenced by space and time dependent friction," J. Chem. Phys. **101**, 7811 (1994).
[52]E. Neria and M. Karplus, "A position dependent friction model for solution reactions in the high friction regime: Proton transfer in triosephosphate isomerase (TIM)," J. Chem. Phys. **105**, 10812 (1996).
[53]M. Hinczewski, Y. von Hansen, J. Dzubiella, and R. R. Netz, "How the diffusivity profile reduces the arbitrariness of protein folding free energies," J. Chem. Phys. **132**, 245103 (2010).
[54]A. Berezhkovskii and A. Szabo, "Time scale separation leads to position-dependent diffusion along a slow coordinate," J. Chem. Phys **135**, 074108 (2011).
[55]B. Keller, X. Daura, and W. F. van Gunsteren, "Comparing geometric and kinetic cluster algorithms for molecular simulation data," J. Chem. Phys. **132**, 074110 (2010).
[56]A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," Science **344**, 1492 (2014).
[57]F. Sittel and G. Stock, "Robust density-based clustering to identify metastable conformational states of proteins," J. Chem. Theory Comput. **12**, 2426–2435 (2016).
[58]L. D. Schuler, X. Daura, and W. F. van Gunsteren, "An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase," J. Comput. Chem. **22**, 1205 (2001).
[59]H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, "Interaction models for water in relation to protein hydration," in *Intermolecular Forces*, edited by B. Pullman (D. Reidel Publishing Company, Dordrecht, 1981), pp. 331–342.
[60]M. Fixman, "Classical statistical mechanics of constraints: A theorem and application to polymers," Proc. Natl. Acad. Sci. U. S. A. **71**, 3050 (1974).