The Journal of
Chemical Physics

All-atom calculation of protein free-energy profiles

Cite as: J. Chem. Phys. **147**, 152724 (2017); https://doi.org/10.1063/1.5006039
Submitted: 26 May 2017 • Accepted: 20 September 2017 • Published Online: 06 October 2017

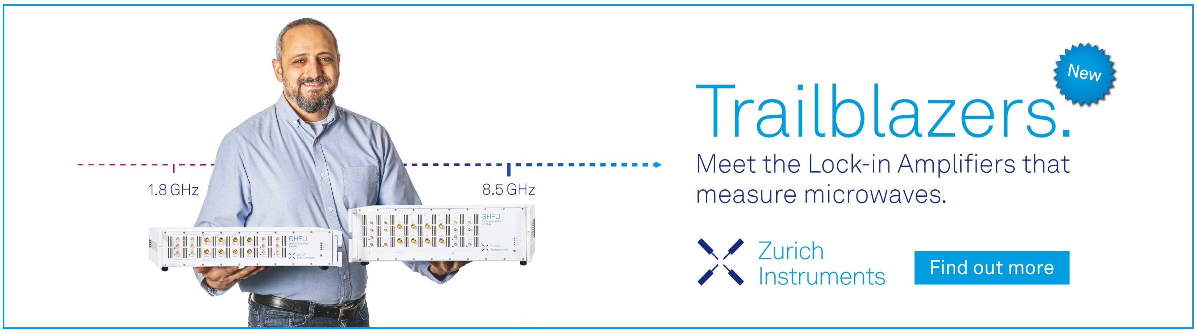
S. Orioli, A. Ianeselli, G. Spagnolli, et al.

ARTICLES YOU MAY BE INTERESTED IN

Erratum: “All-atom calculation of protein free-energy profiles” [J. Chem. Phys. 147, 152724 (2017)]
The Journal of Chemical Physics **150**, 059901 (2019); https://doi.org/10.1063/1.5063534

Self-consistent calculation of protein folding pathways
The Journal of Chemical Physics **147**, 064108 (2017); https://doi.org/10.1063/1.4997197

Transition path sampling of rare events by shooting from the top
The Journal of Chemical Physics **147**, 152716 (2017); https://doi.org/10.1063/1.4997378



AIP Publishing

All-atom calculation of protein free-energy profiles

S. Orioli,[1,2] A. Ianeselli,[1,3] G. Spagnolli,[1,3] and P. Faccioli[1,2,a)]
[1]*Dipartimento di Fisica, Università degli Studi di Trento, Via Sommarive 14, Povo (Trento) I-38123, Italy*
[2]*INFN-TIFPA, Via Sommarive 14, Povo (Trento) I-38123, Italy*
[3]*Centre for Integrative Biology (CIBIO), Università degli Studi di Trento, Via Sommarive 14,
Povo (Trento) I-38123, Italy*

(Received 26 May 2017; accepted 20 September 2017; published online 6 October 2017)

The Bias Functional (BF) approach is a variational method which enables one to efficiently generate ensembles of reactive trajectories for complex biomolecular transitions, using ordinary computer clusters. For example, this scheme was applied to simulate in atomistic detail the folding of proteins consisting of several hundreds of amino acids and with experimental folding time of several minutes. A drawback of the BF approach is that it produces trajectories which do not satisfy microscopic reversibility. Consequently, this method cannot be used to directly compute equilibrium observables, such as free energy landscapes or equilibrium constants. In this work, we develop a statistical analysis which permits us to compute the potential of mean-force (PMF) along an arbitrary collective coordinate, by exploiting the information contained in the reactive trajectories calculated with the BF approach. We assess the accuracy and computational efficiency of this scheme by comparing its results with the PMF obtained for a small protein by means of plain molecular dynamics. *Published by AIP Publishing.* https://doi.org/10.1063/1.5006039

I. INTRODUCTION

Thermally activated conformational transitions are involved in many biological functions performed by proteins and other biomolecules. The number of amino acids participating in these structural changes can vary significantly, ranging from a few units to even several hundreds, as in some large allosteric transitions or in protein folding.

From a theoretician's perspective, the problem of investigating the dynamics of protein thermally activated structural reactions involves two distinct main tasks. The first challenge consists in generating an ensemble of statistically significant reactive trajectories connecting the reactant and product states, in configuration space. The second challenge involves reducing this large amount of data, in order to extract the relevant physico-chemical information. This second problem includes, for example, identifying and structurally characterizing long-lived metastable states and estimating the rate limiting free energy barriers.

Tackling both such challenges requires an extensive use of computational resources. A first reason is the large number of degrees of freedom present in proteins and in their hydration shells. A second reason is that the relevant time scales of large conformational reactions are many orders of magnitude longer than the short time scales associated with fast atomic vibrations or even local rearrangements of the polypeptide chain.

For example, in protein folding (for recent reviews, see, e.g., Refs. 1 and 2), the longest relevant time scale is the folding time (or inverse folding rate), i.e., the average time it takes for the chain to reach the native state for the first time, starting from the unfolded state. According to the Kramers-Arrhenius theory, this time scale increases exponentially with the height of the barrier separating the two states, $\tau = \tau_0\ \exp[\Delta G/k_B T]$, where the pre-factor τ_0 is typically in the μs time scale. Since the folding energy barriers vary from a few to many units of thermal energy $k_B T$, the folding times span over many orders of magnitude, ranging from ms to even minutes. In contrast, elementary local rearrangements of the chain, such as the rotation of a dihedral angle or the formation of a hydrogen bond, usually occur over time scales ranging from several ps to a few ns.

Fortunately, in order to gain microscopic insight into the reaction mechanism, one does not necessarily need to simulate the time evolution of the system for times as long as the mean-first-passage time. Indeed, productive reaction pathways are very rapid events: it has been shown that the so-called transition path time (TPT)—i.e., the time it takes for a system to reach the product *along productive reactive trajectory*—scales only logarithmically with the height of the barrier, $\tau_{TPT} \simeq \tau_0 \log[\alpha \Delta G/k_B T]$.[3,5] This result explains why proteins with widely different folding times have comparable TPTs, typically in the few μs range.[4] In view of these considerations, it is clear that the most efficient way to investigate the reaction mechanism consists in sampling directly the productive reaction pathways, without wasting computational time to simulate uninteresting thermal oscillations in the reactant or generate unsuccessful reactive attempts, as one would do in plain molecular dynamics (MD) simulations.

Many advanced methods and algorithms have been proposed in order to lower the computational cost of generating reaction pathways for rare biomolecular transitions (for a recent review, see, e.g., Refs. 6). In this context, the

[a)]pietro.faccioli@unitn.it

path integral (PI) formalism of stochastic dynamics (which is briefly reviewed in the Appendix) offers an attractive theoretical framework because it enables to express the conditional probability to perform the conformational transition in a given time interval t as a sum over all possible reactive trajectories connecting reactants and products, in the molecule's configuration space. In particular, within the framework of Langevin dynamics, it is possible to compute the statistical weight of each of such reaction pathways.

Based on the path integral formalism, a stochastic algorithm called Transition Path Sampling (TPS) was developed which enables to rigorously sample the ensemble of reactive trajectories.[7] Unfortunately, applications of TPS to protein folding or to other comparably complex protein conformational reactions are computationally very expensive. This limitation has triggered the development of several approximation schemes, aiming to further lowering the computational load associated with sampling the stochastic PI. Some of these methods focus on the most probable reaction pathways.[8–12] All these techniques are based on applying global optimization algorithms in order to explore the space of reaction pathways starting from a given initial trial trajectory and maximize a target functional of the reactive path, e.g., related to its probability to occur in a Langevin dynamics.

A main limitation of all these path optimization schemes is that they typically produce results which are strongly correlated with the initial trial guess. Indeed, even using state-of-the art global optimization algorithms, the exploration of the path space is restricted to the functional neighbourhood of the initial trajectory. Furthermore, this type of calculations can only be performed using implicit solvent models. The reason is that, in an explicit solvent calculation, the target functional to be optimized would be dominated by the solvent degrees of freedom and basically insensitive to the reaction pathway undertaken by the solute.

The Bias Functional (BF) approach[13,14] was developed in order to overcome these two limitations. This method is based on combining a special kind of biased MD called Ratchet-and-Pawl MD (rMD)[15,16] with a new variational principle, rigorously derived from the PI representation of Langevin dynamics.[14] Namely, the BF approach is based on the following two-step procedure: first, rMD is used to generate many independent *biased* reaction pathways. Next, the variational principle is applied to this ensemble of trial paths, to identify the biased trajectories which have the largest probability to occur in an *unbiased* simulation. This way it is possible to explore larger regions of the space of reaction pathways by comparing many statistically uncorrelated trajectories. We also emphasize that in the rMD, the effect of the bias is kept to a minimum, since no bias is applied to the system, as long as it spontaneously progresses towards the product state. A harmonic history-dependent force is introduced only to discourage spontaneous backtracking towards the reactant.

A second attractive feature of the BF approach is that it can be used to perform explicit solvent calculations. Indeed, the variational principle of the BF approach is based on a target functional which does not depend at all on the solvent degrees of freedom.

The BF technique and its closely related precursor, called dominant reaction pathway, have been successfully applied to investigate very slow and complex protein folding or conformational transitions, using both implicit and explicit all-atom force fields, providing results in good agreement with experiments.[17–20] A remarkable example is provided by the simulation of folding[17] and latency transition[18] of serpin proteins, which consist of nearly 400 amino acids and have folding times as long as tens of minutes.

An important limitation of the BF approach is that the rMD biasing force depends on the path "history" and therefore violates the requirement of microscopic reversibility. As a consequence, even after applying the variational condition, in the BF scheme, it is impossible to extract thermodynamical and kinetic information *directly* from the reactive trajectories.

In this work, we tackle this limitation of the BF scheme by introducing a rigorous method to efficiently compute the potential of mean-force (PMF) along an arbitrarily chosen collective coordinate. A number of sophisticated techniques have been developed to compute the free energy as a function of one or few collective variables,[21–25] to directly extract information about the reaction kinetics,[26–28] or to achieve a low-resolution representation of reaction kinetics based on Markov state models.[29–32] Unfortunately, all these methods are in general very computationally demanding when applied to the characterization of complex structural reactions, such as protein folding. The method we introduce in this work is much less computationally expensive because it exploits the information contained in the reactive trajectories calculated by the BF approach, thus focusing on the region of configuration space which is visited by the reactive pathways.

To illustrate and validate our algorithm, we apply it to investigate the folding of the FIP35 WW-domain. The dynamics of this small protein has been extensively investigated by means of ultra-long plain MD simulations performed using an all-atom force field on the Anton special purpose supercomputer[33] and by means of world-wide distributed computing.[34] In particular, using the long MD trajectories generated by Anton, it is possible to elucidate the folding mechanism and profile the PMF as a function of an arbitrary collective variable. First, we report on BF protein folding simulations performed at the same temperature, using the same force field, and we compare the folding mechanisms in the two simulations. Next, we compute the PMF of the fraction of native contacts and compare it with the same function calculated from a frequency histogram of equilibrium MD trajectories, again finding consistent results. It should be stressed that our simulations were performed in a few days on a small computer cluster.

The paper is organized as follows. In Sec. II, we provide short description of the BF approach and the related rMD algorithm. In Sec. III, we introduce our algorithm to evaluate the PMF along a reaction coordinate and we discuss the comparison to the Anton data. The main conclusions are summarized in Sec. IV.

II. THE BF APPROACH

In this section, we briefly review how the BF algorithm is used to sample the ensemble of reaction pathways of a protein. The underlying theoretical foundation is reported in the Appendix. Even though we shall focus on applications to protein folding, this algorithm can be applied to simulate arbitrary conformational transitions in which the product state is structurally characterized with atomic resolution.

The BF algorithm consists in the following multi-step procedure:

1. **Sampling of the denatured state**: An initial denatured configuration is generated, for example, by running short MD simulations at high temperature, starting from the crystal native structure (thermal unfolding).
2. **Generating many trial folding pathways**: Many biased simulations initiated from the initial configuration generated at step 1 are used to produce several independent folding trajectories. In this dynamics, an unphysical biasing force is introduced in order to discourage backtracking towards the unfolded state,[15]

   $$\mathbf{F}_i(X,t) = -k_R\, \nabla_i z(X)\, (z(X) - z_m(t))\, \theta(z(X) - z_m). \qquad (1)$$

   Here, $\theta(x)$ is the Heaviside step-function and $z(X)$ is a collective coordinate[16] defined as

   $$z(X) = \sum_{|i-j|>35}^{N} [C_{ij}(X) - C_{ij}^0]^2, \qquad (2)$$

   which measures the distance between the instantaneous contact map $C_{ij}(X)$ and the native contact map $C_{ij}^0 = C_{ij}(X_N)$. The entries of the contact maps are continuous and given by

   $$C_{ij}(X) = \frac{1 - \left(\frac{|\mathbf{x}_i - \mathbf{x}_j|}{r_0}\right)^6}{1 - \left(\frac{|\mathbf{x}_i - \mathbf{x}_j|}{r_0}\right)^{10}}. \qquad (3)$$

   The constraint $|i - j| > 35$ in the summation in Eq. (2) is introduced in order to exclude the contribution from neighbouring atoms, while the constant r_0 = 7.5 Å in Eq. (3) provides a reference distance for native contacts. In addition, a cutoff is usually introduced that sets $C_{ij}(r_{ij}) = 0$ for atomic distances larger than a threshold, $r_{ij} > r_c \simeq 1.2$ nm. This ensures that the computational cost of the simulation scales linearly with the number of atoms in the protein. The function $z_m(t)$ in Eq. (1) represents the minimum value assumed by the collective variable z along the rMD trajectory, up to time t.

   We emphasize that the biasing force (1) is *not* active whenever the chain spontaneously evolves towards more native-like configurations, along the direction defined by z, i.e., when $z[X(t + \Delta t)] < z_m(t)$. Conversely, a force proportional to $z[X(t)] - z_m(t)$ sets in when the chain attempts to backtrack towards the unfolded state, i.e., for $z[X(t + \Delta t)] > z_m(t)$.
3. **Selecting the most likely rMD trial path**: The set of trial folding trajectories generated at step 2 are scored according to the following functional:

   $$T[X] = \int_0^t d\tau \sum_i \frac{1}{4k_B T m_i \gamma_i} |\mathbf{F}_i(X,\tau)|^2, \qquad (4)$$

   which we shall refer to as the bias functional. The best scoring trajectory is referred to as the *least biased* one. In the Appendix, we explicitly show that the rMD trajectory with the least value of this target functional coincides with the one with the largest probability to occur in an *unbiased* Langevin simulation. Therefore, restricting to such a minimum bias path provides a variational approximation of the folding trajectory.
4. **Generating an ensemble of least biased trajectories**: Steps 1 through 3 are repeated for different initial unfolded conditions.

Using this algorithm, it is feasible to compute protein folding trajectories of proteins consisting of even a few hundred amino acids, using relatively modest computer resources, typically consisting of $\mathcal{O}(10^2)$ cores. This computational efficiency is due to the fact that for each initial condition, one needs to generate about 20-40 trial rMD independent folding trajectories, which are only less than 1 ns long—for a discussion of the convergence criteria in the variational search, see Ref. 14.

In the following, we report on the results of BF simulations of the folding of FIP35 WW domain (pdb code: pin1), which is one of the smallest and fastest folding proteins. We used the AMBER99FS-ILDN all-atom force field[36] in explicit TIP3P water, to allow for a direct comparison with the results obtained for the same system by means of plain MD simulations, using the Anton supercomputer.[33]

We produced 8 different initial conditions by thermal unfolding simulations at the temperature T = 800 K, initiated after energy relaxing the crystal native structure shown in Fig. 1. From each of such unfolded configurations, we generated 20 trial folding trajectories by means of 0.5 ns of rMD at T = 395 K, with initial momenta sampled from a Maxwell-Boltzmann distribution. The system was coupled to the Nosé-Hoover thermostat and to the Parrinello-Rahman barostat, in order to replicate the same conditions of the simulation carried out in Ref. 33. The coupling constant of the rMD bias was set to $k_R = 3 \times 10^{-4}$ kJ/mol. With this value, the square modulus of the total biasing potential is always at least two orders of magnitude smaller than the physical potential energy. For each initial condition, the least biased trajectory was selected out of the 20 trial rMD pathways using the minimum BF criterium.

The eight BF folding trajectories projected onto the plain selected by the Root Mean Square Distance (RMSD) of the two hairpins to their native structure is shown in the left panel of Fig. 1. The heat-map in the background describes the free energy landscape calculated from a frequency histogram of the Anton MD trajectories. In agreement with our previous finding,[14] we see that the BF trajectories reach the native state by traveling along low free energy regions and describe the correct folding mechanism, i.e., one in which the two hairpins fold in sequence. A more quantitative analysis demonstrating the consistency between the folding mechanism predicted by BF and MD simulations was presented in Refs. 14

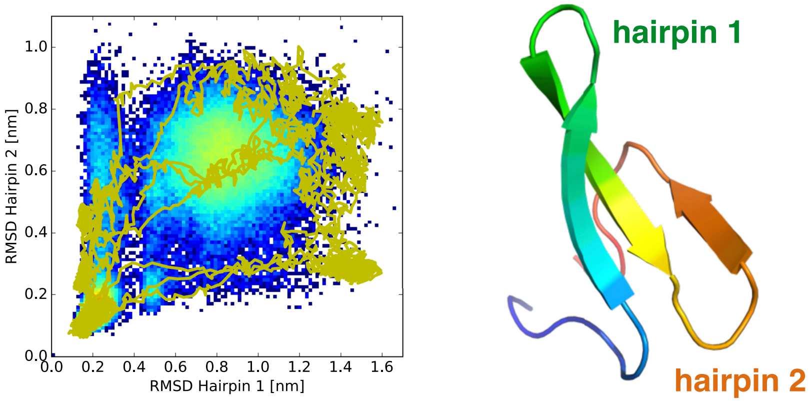


FIG. 1. The left panel reports the 8 folding pathways obtained with the BF approach, projected onto the plane identified by the RMSD to native of the two hairpins of FIP35 (crystal structure shown in the right panel). The heat map in the background is the free energy as a function of the same collective variables obtained from a histogram of plain MD simulations performed on the Anton supercomputer.

and 35. The folding mechanism of much larger polypeptide chains predicted by BF was validated against experiment in Refs. 17–19.

We conclude this section by discussing the main drawbacks of the BF approach. Like any variational approximation, this method may suffer from systematic errors related to the choice of the trial space. In particular, low accuracy is generally expected whenever the quality of the rMD trajectories is poor. This scenario is realized if the collective variable z given in Eq. (2) is not a good reaction coordinate. This problem was solved in a recent work,[35] by developing an improved iterative rMD algorithm which enables to correct the collective coordinate in a self-consistent way. We demonstrated that the trial paths obtained with this new type of rMD dynamics are biased along the average *unbiased* folding trajectory and that the corresponding biasing collective coordinate provides a stochastic estimate of the reaction coordinate.

A second important limitation of the BF approach arises from the fact that the rMD trajectories do not satisfy the microscopic reversibility condition. As a consequence, the corresponding time scales do not have a direct physical interpretation and the BF trajectories cannot be used to extract in a straightforward way the information about the relevant metastable states and free energy barriers involved in the folding transition. In Sec. III, we present an algorithm to tackle this limitation.

III. COMPUTING FREE ENERGY PROFILES FROM BF SIMULATIONS

A commonly adopted strategy to gain insight into reaction mechanisms in complex molecular systems consists in projecting the very high dimensional configuration space into a single collective variable, which is assumed to approximate the reaction coordinate. Using Zwanzig-Mori projection formalism, it can be shown that this collective variable evolves according to a generalized Langevin equation.[37,38] In addition, if the characteristic relaxation time scales of this collective variable are much longer than that those of all internal degrees of freedom in the system, then the generalized Langevin equation can be replaced by a standard over-damped Langevin equation, which depends only on the diffusion coefficient and on the PMF of the collective variable.

In principle, the PMF $G(Q)$ can be estimated from an ensemble of equilibrium configurations, e.g., sampled from long MD trajectories. Indeed, if $P_{eq}(Q)$ is the probability of observing a value Q at equilibrium (which can be estimated from a frequency histogram), then $G(Q)$ is defined by

$$G(Q) = -k_B T \log P_{eq}(Q). \tag{5}$$

This method is in principle exact, but computationally extremely expensive. Indeed, it requires simulating the dynamics for a time scale sufficiently long to attain complete thermal equilibrium. For most polypeptide chains of biophysical or biological interest, the sampling of the equilibrium distribution $P_{eq}(Q)$ remains a formidable computational challenge, even using advanced Monte Carlo algorithms or more sophisticated methods.[23,24]

In the following, we devise an alternative scheme which exploits the results of BF simulations and enables us to compute $G(Q)$ in a very computationally efficient way. For the sake of simplicity, to illustrate the approach, we shall assume that the slow dynamics of the collective variable Q can be effectively described by an over-damped Langevin equation with a uniform diffusion coefficient D_0,

$$\dot{Q} = -\frac{D_0}{k_B T} G'(Q) + \eta(t). \tag{6}$$

The probability distribution generated by integrating Eq. (6) evolves according to the Fokker–Planck equation,

$$\frac{\partial}{\partial t} P(Q,t) = \hat{F} P(Q,t), \tag{7}$$

where

$$\hat{F} = D_0 \frac{d}{dQ} \left(\frac{d}{dQ} + \frac{1}{k_B T} G'(Q) \right). \tag{8}$$

In such a framework, an arbitrary initial probability density $\rho_0(Q)$ changes in time according to an evolution operator defined by

$$P(Q,t) = e^{-\hat{F} t} \rho_0(Q). \tag{9}$$

Equivalently, Eq. (9) can be written in terms of the conditional probability to perform a transition form Q' to Q in time t (i.e., the propagator),

$$P(Q,t) = \int dQ' \; P(Q,t|Q',0) \; \rho_0(Q'). \tag{10}$$

Some general properties of the dynamics defined by Eq. (7) are in order. Even though the $\hat{F}$ operator is not Hermitian, its left and right eigenfrequencies coincide and are real,

$$\overset{\rightarrow}{\hat{F}} \; R_k(Q) = \lambda_k \; R_k(Q), \tag{11}$$

$$L_k(Q)\,\overleftarrow{\hat{F}} = \lambda_k\, L_k(Q), \tag{12}$$

while left and right eigenstates are related by a non-unitary transformation,

$$L_k(Q) = e^{\beta G(Q)} R_k(Q), \tag{13}$$

and obey the orthogonality condition

$$\int dQ L_k(Q) R_m(Q) = \delta_{km}. \tag{14}$$

In particular, the lowest eigenfrequency vanishes, $\lambda_0 = 0$. The corresponding right-eigenstate is the Gibbs distribution, $R_0(Q) = \frac{1}{Z} e^{-\beta G(Q)}$, while the lowest left-eigenstate is the identity, $L_0 = 1$. It can be also shown that, for a two-state system, the function $L_1(Q)$ is monotonic and is closely related to the committor probability.[39]

The probability distribution $P(Q,t)$ evolving according to Eq. (9) can be expanded in terms of the right and left eigenstates,

$$P(Q,t) = \sum_k c_k\, e^{-\lambda_k t} R_k(Q), \qquad \left(c_k = \int dQ L_k(Q) \rho_0(Q) \right). \tag{15}$$

The long-time evolution is governed by the lowest frequency modes,

$$\lim_{t\to\infty} P(Q,t) = \frac{1}{Z} e^{-\beta G(Q)} + c_1 R_1(Q)\, e^{-\lambda_1 t} + c_2 R_2(Q)\, e^{-\lambda_2 t} \ldots, \tag{16}$$

where the dots represent terms which are exponentially suppressed at large times, i.e., for $t \gg 1/(\lambda_3 - \lambda_2)$. Equation (16) implies that, in order to attain thermal equilibrium starting from an arbitrary initial distribution $\rho_0(Q)$, the system needs to evolve for a time $t \gg \lambda_1^{-1}$. Thus, λ_1 is interpreted as the inverse thermal relaxation time scale.

Higher eigenfrequencies $\lambda_2, \lambda_3, \ldots$ are associated with faster local relaxation processes, e.g., those within the local metastable states. In particular, if the free energy surface displays two local minima separated by a single thermally activated barrier ΔG, then $\lambda_1 \sim e^{-\frac{\Delta G}{k_B T}}$ and decouples from all other eigenfrequencies, i.e., $\frac{\lambda_1}{\lambda_2} \ll 1$.

We recall that we are assuming that Eq. (7) describes at the effective level the evolution of the collective coordinate evaluated along *microscopic* trajectories in configuration space. Let us consider the case in which the dynamics of our molecular system is generated starting from an ensemble of initial microscopic configurations characterized by some probability distribution $P_0(X)$ and let

$$\rho_0(Q) = \int dX\, P_0(X)\, \delta(Q - f_Q(X)), \tag{17}$$

where $f_Q(X)$ is the function of the molecular configuration X which defines the collective coordinate Q.

From (16) it follows that if the distribution $\rho_0(Q)$ is such that $c_1 = 0$, then the convergence to the thermal equilibrium distribution starting from the ensemble of configurations with probability $P_0(X)$ would take a time $\sim 1/\lambda_2$, *exponentially* shorter than the thermal relaxation time, $1/\lambda_1$. This would provide a computationally efficient scheme to obtain the PMF $G(Q)$ from *short* MD simulations, of length $t \gtrsim 1/\lambda_2$.

How can we sample an ensemble of initial conditions with a distribution of collective variables $\rho_0(Q)$ such that $c_1 \simeq 0$? To address this problem, let us focus on a two-state system and imagine to evaluate the time-dependent probability distribution $P(Q,t)$ obtained evolving some initial distribution $\rho_L(Q)$ *entirely localised in the reactant*. We conventionally choose the reactant to be the leftmost minimum of $G(Q)$. Let $J(Q,t)$ be the corresponding probability current,

$$J(Q,t) = D_0 \left(\frac{d}{dQ} + \beta U'(Q) \right) P(Q,t). \tag{18}$$

We now specialize even further, by choosing to consider time intervals t much shorter than thermal relaxation time scale, yet much longer than the time scale associated with local relaxation within each metastable states, i.e.,

$$\frac{1}{\lambda_2} \ll t \ll \frac{1}{\lambda_1}. \tag{19}$$

In such a time regime, the reactive current $J(Q,t)$ is nearly steady. This can be shown by applying expansion (15),

$$J(Q,t) = \sum_{k>0} c_k\, J_k(Q)\, e^{-\lambda_k t}, \tag{20}$$

where

$$J_k(Q) \equiv D_0 \left(\frac{d}{dQ} + \beta U'(Q) \right) R_k(Q). \tag{21}$$

In the intermediate time regime (19), only the contribution proportional to $R_1(Q)$ in Eqs. (20) and (21) survives. Indeed, the contribution from $R_0(Q)$ vanishes identically, while all terms with $k > 1$ are exponentially suppressed. Furthermore $t \ll 1/\lambda_1$ so $\exp(-t\lambda_1) \simeq 1$. Thus we have

$$J(Q,t) \simeq c_1 J_1(Q), \qquad \text{for } \frac{1}{\lambda_2} \ll t \ll \frac{1}{\lambda_1}. \tag{22}$$

We note that the function $J_1(Q,t)$ does not change sign, since $J_1(Q) = D_0 \frac{d}{dQ} L_1(Q)$ and $L_1(Q)$ is a monotonic function. Thus, this function defines a probability density.

It has also be shown that, in the time regime (19), the probability density $J_1(Q)$ measures the probability of observing the value of Q in the *in the reactive pathways*—see, e.g., Eq. (99) of Ref. 41. The physical reason behind this result is that, in the steady reactive current regime (19), the probability current flowing across the transition state is dominated by single barrier crossing transitions. Therefore, the distribution of values of Q in reactive trajectories yields $J_1(Q)$.

Finally, we note that $J_1(Q)$ satisfies the desired property

$$\begin{aligned} c_1 \propto \int dQ L_1(Q) J_1(Q) &= \int dQ L_1(Q) \left(\frac{d}{dQ} + \beta \frac{d}{dQ} G(Q) \right) \\ &\quad \times e^{-\beta G(Q)} L_1(Q) \\ &= \int dQ L_1(Q) \frac{d}{dQ} L_1(Q) \\ &= \frac{1}{2} \int dQ \frac{d}{dQ} L_1^2(Q) = 0. \end{aligned} \tag{23}$$

Therefore, a Fokker-Planck time evolution of the initial distribution $J_1(Q)$ is expected to very rapidly attain thermal equilibrium, within a time scale $\sim \lambda_2^{-1} \ll \lambda_1^{-1}$.

At this point, the benefit coming from the application of the BF method becomes evident: This variational approximation of the reaction pathways enables us to obtain an

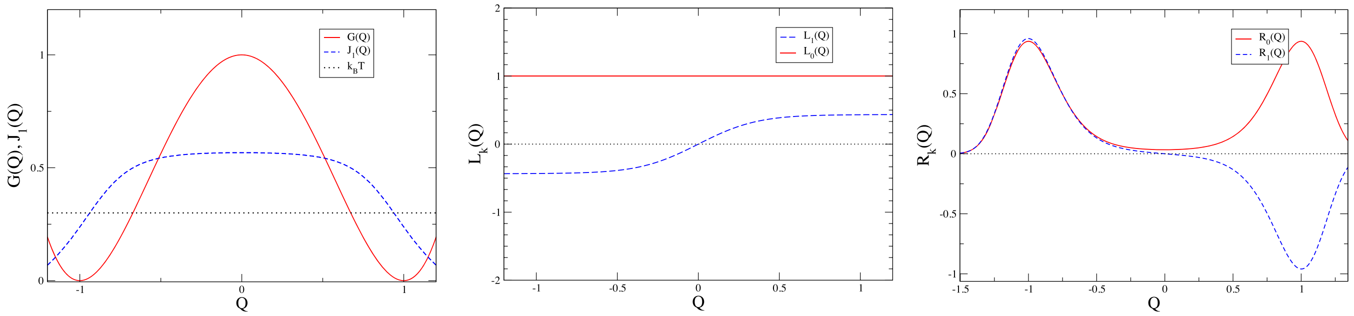


FIG. 2. Left panel: double well free energy surface of the toy model and corresponding $J_1(Q)$ distribution (see text). Central and right panels: lowest two left eigenstates and right eigenstate of the $\hat{F}$ operator.

approximation of the $J_1(Q)$ distribution, from a frequency histogram of the values of Q attained by the reactive trajectories. According to Eq. (10), the time evolution of the $J_1(Q)$ distribution can be implemented by running many plain MD simulations, initiated from protein configurations with the same values of Q. If such an evolution lasts for a time $\sim\lambda_2^{-1}$, then $P(Q,t)$ will converge to its equilibrium distribution $P_{eq}(Q)$, from which one immediately obtains the PMF through Eq. (5).

A. Illustration in a toy model

It is instructive to first analyze how this fast thermal relaxation works in a simple one dimensional model. We imagine a molecular system undergoing a two-state conformational reaction described by some collective variable Q with PMF given by

$$G(Q) = G_0(Q^2 - 1)^2, \tag{24}$$

with $G_0 = 1$, $D_0 = 1$, and $k_BT = 0.3$, in some appropriate units (see the left panel of Fig. 2). In the central and right panels of Fig. 2, we show, respectively, the corresponding two lowest right and left eigenstates, assuming a thermal energy $k_BT = 0.3$. We note that, as expected, the $L_1(Q)$ function is monotonic, locally odd in the transition region, and null at the transition state. $J_1(Q)$ is even in the transition region and approximately constant near the transition state.

According to the arguments discussed above, we expect that evolving in time the initial distribution $\rho_0(Q) \propto J_1(Q)$ should very rapidly lead to the thermal equilibrium distribution. This feature is clearly illustrated in Fig. 3 which shows the time evolution of $P(Q,t)$ evaluated by integrating the Langevin equation starting from initial conditions sampled from two different initial distributions: (i) a sharp Gaussian, peaked near one of the metastable minima [for which $c_1 \sim o(1)$] and (ii) a flat distribution in the transition region (which

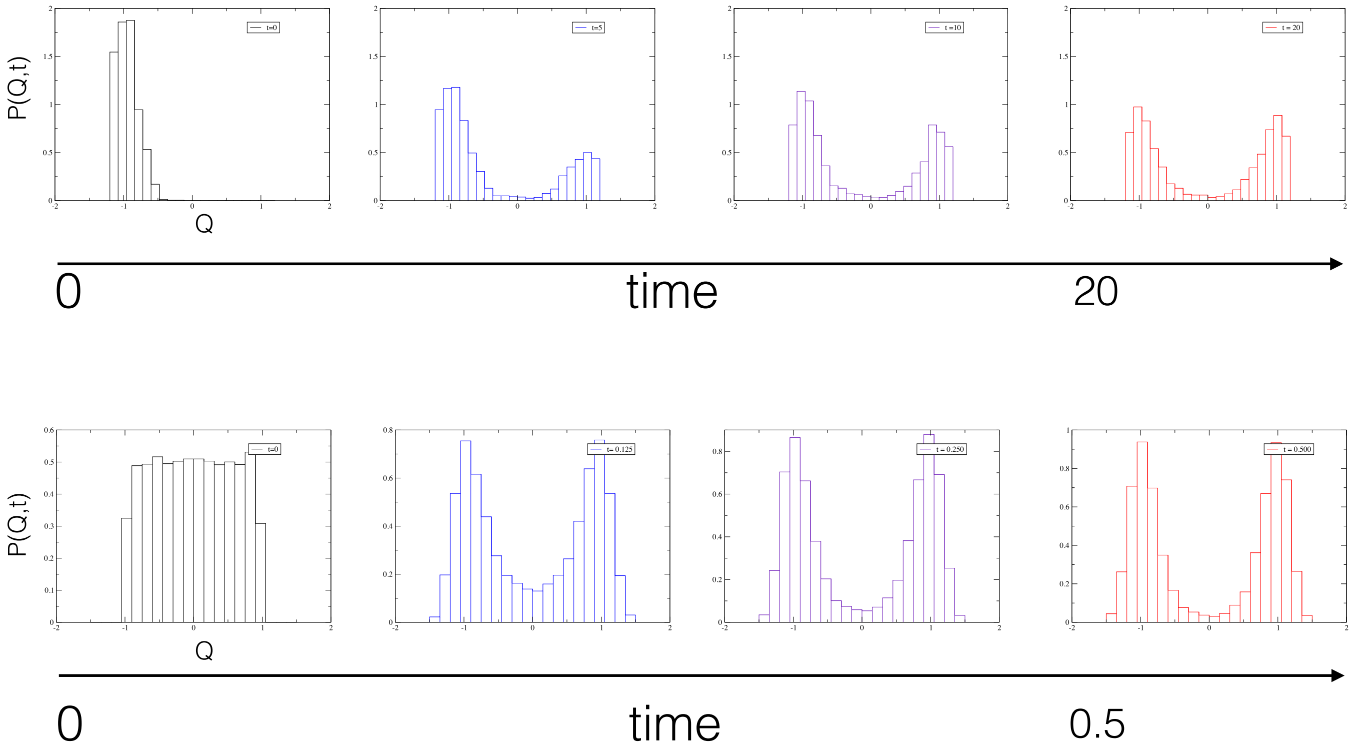


FIG. 3. Relaxation to thermal equilibrium in the double-well energy surface, starting from two different initial distributions.

models the transition current distribution and obeys $c_1 = 0$). As expected, the first distribution relaxes to equilibrium about 40 times more slowly than the second distribution. This ratio of relaxation time scales is consistent with the fact that, for this system, $\lambda_2/\lambda_1 \sim 50$.

B. Realistic application

Let us now see how this result can be used to profile the PMF of protein FIP35, along a specific reaction coordinate. In particular, we choose the fraction of native contacts, which can be defined as follows:

$$Q(X) = \frac{\sum_{|i-j|>4} \theta(r_0 - r_{ij}(X))\, \theta(r_0 - r_{ij}(X_N))}{\sum_{|i-j|>4} \theta(r_0 - r_{ij}(X_N))}. \tag{25}$$

In this equation, r_{ij} is the distance between the ith and jth C_α atoms and $r_0 = 7.5$ Å is a typical reference value for a native contact. Note that the theta-function $\theta(r_0 - r_{ij}(X_N))$ restricts the summation to the pairs of C_α atoms which are in contact in the native state, while the denominator contains the total number of native contacts. The constraint $|i-j| > 4$ in the summation excludes the contribution of the amino acids which are topologically close along the polypeptide chain. In Ref. 40, it was shown that Q correlates relatively well with the committor probability for small globular proteins. Furthermore, in the same work, it was shown that the diffusion coefficient depends rather weakly on Q, so the approximation $D(Q) \simeq D_0$ is reasonable.

In practice, we implemented our method to profile the PMF, by adopting the following procedure:

1. We evaluated $J_1^{BF}(Q)$ [the BF estimate for $J_1(Q)$] from a frequency histogram of values of Q visited along the 8 reaction pathways computed with the BF approach.
2. We clustered the frames visited by the 8 folding trajectories according to their value of Q, which was defined on a discrete mash with bin size $\Delta Q = 0.02$.
3. We randomly picked 6 frames from each of such clusters of configurations and used them as starting point for 5 ns of plain MD simulation. MD trajectories have been computed following the same simulation setup of the rMD simulations.
4. We performed a weighted histogram of the values of Q visited during such MD simulations, using the distribution $J_1^{BF}(Q)$ to re-weight. According to Eq. (10), this is equivalent to evolve for a time interval t an initial distribution $J_1^{BF}(Q)$. Once the resulting distribution $P(Q,t)$ stops evolving with time t, the PMF $G(Q)$ was extracted using Eq. (5).

The PMF calculated according to this procedure is shown in the left panel of Fig. 4, where it is compared with the exact calculation of $G(Q)$ obtained from an histogram of the Anton equilibrium MD trajectories. These ultra-long MD simulations include about a dozen unfolding-refolding events. We see that our calculation of $G(Q)$ is in overall good agreement with the exact result. In particular, the height of the barrier is accurately estimated, the transition state is correctly located, and even the double hump structure of the barrier top (which is due to the sequential folding of the individual hairpins) is well reproduced. A minor discrepancy—of the order of 0.5 k_BT—is observed only in the highly denatured region, $Q \lesssim 0.2$. We note that an insufficient sampling of the denatured state (which is plausible with only 8 initial conditions) would lead to an underestimate of its entropic contribution, thus providing a possible explanation for overshooting the PMF in this region.

The comparison with the negative logarithm of the initial distribution of values of Q (central panel of Fig. 4) shows that the short time evolution has significantly improved the quality the estimate of $G(Q)$, with respect to a naive histogram analysis of the reactive trajectories. In the right panel of Fig. 4, we show the estimates of $G(Q)$ obtained evolving for different time intervals. We see that after about 5 ns, our estimate for the PMF stops to sizeably depend on t, suggesting that equilibrium has been attained.

We can check *a posteriori* that the observed thermalization time scale of 5 ns is quite reasonable. Indeed, we expect this value to be of the same order of the time scale

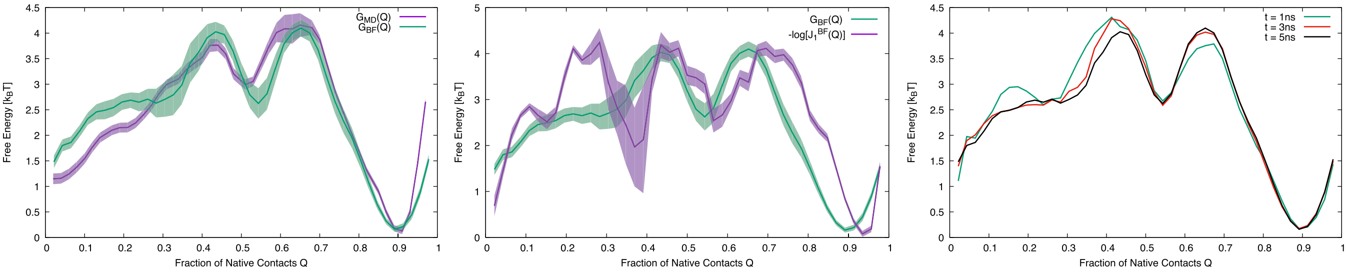


FIG. 4. Left panel: PMF along the fraction of native contacts Q computed from a frequency histogram of equilibrium MD trajectories—$G_{MD}(Q)$—and from the short time evolution based on BF simulations developed in this work—$G_{BF}(Q)$. Center panel: Comparison of the results for $G(Q)$ obtained before and after the time evolution. Right panel: predictions for $G(Q)$ obtained after time evolving the initial distribution $J_1(Q)$ calculated from BF simulations, for different time intervals. Statistical errors in these curves have been estimated according to the following jackknife procedure. The set of 300 MD trajectories was split in six subsets of 50 trajectories each. For six times, one group of trajectories was excluded from the calculation of the frequency histogram and the jackknife histogram was computed as the average between the six results, equipped with the corresponding variance. Free energy was computed from the jackknifed histogram and the statistical error was propagated correspondingly. The error on the free energy obtained from the Anton calculations was obtained in a similar fashion: the two ~100 μs trajectories were split in 20 trajectories of approximately ~10 μs each. Frequency histograms were computed by excluding one short trajectory from the set and iterating until all the trajectories were excluded once. Free energy was obtained by averaging on the corresponding frequency histograms and propagating the corresponding variance.

associated with local thermalization within the metastable states. We recall that diffusion in a harmonic oscillator thermalizes in a time scale $\lambda^{\mathrm{HO}} = \frac{D_0}{k_B T} G''(Q_0)$. By applying the harmonic approximation to the PMF near the native state with a typical value $D_0 \sim 1\ \mu\mathrm{s}^{-1}$, we obtain $1/\lambda_1^{\mathrm{HO}} \sim 10$ ns, which is of the same order of magnitude of the equilibration time scale of our simulations. Furthermore, by applying Kramers escape rate formula, we obtain a folding time in the same order of magnitude of that observed in MD simulations.

A final remark about systematic errors is in order. In general, the initial distribution $J_1^{BF}(Q)$ calculated from an histogram of trajectories obtained in BF simulations is expected to provide only variational approximation to the exact $J_1(Q)$ distribution (i.e., the one related to the reaction probability current generated by the original MD). As a consequence, the coefficient c_1 of the BF distribution is not expected to be strictly null. A small value of c_1 implies that a residual discrepancy between the BF estimate for $G(Q)$ and the corresponding exact result should persist even at times $t \gtrsim 1/\lambda_2$, ultimately fading out only at time scales of the order of the thermalization time. On the other hand, the agreement with MD suggests that this systematic error is actually small.

Of course, an additional systematic error may arise from the incomplete thermalization of the degrees of freedom orthogonal to the collective variable Q. However, we emphasize that the short MD simulations have been initiated from several configurations sampled from *independent* reactive trajectories, which visit an extended region of the orthogonal subspace. Finally, we stress that our results have been obtained by assuming that the dynamics of the collective coordinate is diffusive and the diffusion coefficient is uniform.

IV. CONCLUSIONS

In this work, we have introduced an algorithm to compute the PMF for an arbitrary collective variable Q. This method is based on running many very short (5 ns) MD simulations, starting from configurations harvested from the reactive pathways generated by means of the BF approach. The main idea underlying this approach is to generate short relaxation trajectories sampling an initial distribution such that the overlap with the first eigenvalue of Fokker-Planck operator is null. Under such a condition, the relaxation to thermal equilibrium of the distribution of collective coordinate $P(Q, t)$ is attained within a time scale comparable to that associated with local relaxation within the metastable states. For protein FIP35, a complete characterization of the free-energy profile is required to simulate 1.5 μs of MD, a time comparable with the TPT of a single folding event.

While computationally very efficient, this scheme is based on a number of assumptions, thus potentially prone to some systematic errors. However, direct comparison with the results of plain MD simulations suggests that it may lead to result which is sufficiently accurate to estimate free energy barrier and locate the transition state.

As a final remark, we emphasize that the spectral decomposition of the time-dependent probability density upon which this method is based is also used for dimensional reduction of Markov state models, e.g., via the Perron cluster analysis[42] or renormalisation group.[43] It plays a central role also in the diffusion map formalism,[44,45] in view of the fact that eigenstates of the backward Fokker-Planck operator define a convenient Euclidean system of coordinates, directly related to the diffusive distance.

ACKNOWLEDGMENTS

We thank *DES Research* for making their MD trajectories for FIP35 folding available to us. Part of the calculations was performed on Tier-0 Marconi at the CINECA supercomputing facility. S. Orioli, A. Ianeselli, and G. Spagnolli equally contributed to this work.

APPENDIX: PATH INTEGRAL FORMULATION OF LANGEVIN DYNAMICS AND DERIVATION OF THE BF VARIATIONAL CONDITION

In this appendix, we briefly review the PI formulation of stochastic dynamics and we sketch the derivation of this variational theorem underlying the BF approach. More details on the mathematical proof of this result can be found in the original publication.[14]

Our starting point is the Langevin description of the dynamics in the solute and solvent,

$$m_i \ddot{\mathbf{x}}_i = -m_i \gamma_i \dot{\mathbf{x}}_i - \nabla_i U(X, Y) + \eta_i(t), \tag{A1}$$

$$\mu \ddot{\mathbf{y}}_i = -\mu\,\sigma \dot{\mathbf{y}}_i - \nabla_i U(X, Y) + \xi_i(t). \tag{A2}$$

$X = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ and $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_{N_s})$ denote the collection of Cartesian coordinates, specifying, respectively, the configuration of the solute and solvent. m_i and μ, respectively, denote the masses of the atoms in the protein and of the water molecules in the solvent, while γ_i and σ are viscosity coefficients. $-\nabla_i U(X, Y)$ is the atomistic force field, while η_i and ξ_i are delta-correlated stochastic forces, obeying the standard fluctuation-dissipation relationship,

$$\langle \eta_i(t) \cdot \eta_j(t') \rangle = 6\ m_i \gamma_i k_B T\ \delta_{ij}\ \delta(t - t'), \tag{A3}$$

$$\langle \xi_i(t) \cdot \xi_j(t') \rangle = 6\ \mu\,\sigma k_B T\ \delta_{ij}\ \delta(t - t'). \tag{A4}$$

Within the stochastic dynamics defined by Eqs. (A1) through (A4), the conditional probability density $p(X_f, t|X_i)$ for the protein initially in configurations X_i in the reactant state to visit configuration X_f in the product at time t can be written as follows:

$$p(X_f, t|X_i) = \int dY_f \int dY_i \int_{Y_i}^{Y_f} \mathcal{D}Y \int_{X_i}^{X_f} \mathcal{D}X\ e^{-S[X,Y]}\, \frac{e^{-\beta U(X_i, Y_i)}}{Z}, \tag{A5}$$

where $\beta = 1/k_B T$. In this equation, the last exponential factor is the Boltzmann distribution of the solvent molecules around the initial protein's configuration X_i, Z is the system's partition function, while $S[X, Y\,]$ is the so-called Onsager-Machlup (OM) functional,

$$S[X, Y] \equiv \frac{\beta}{4} \int_0^t d\tau \left[\sum_{i=1}^{N} \frac{1}{\gamma_i m_i} (m_i \ddot{\mathbf{x}}_i + m_i \gamma_i \dot{\mathbf{x}}_i + \nabla_i U(X, Y))^2 + \sum_{k=1}^{N_s} \frac{1}{\sigma \mu} (\mu \ddot{\mathbf{y}}_k + \mu \sigma \dot{\mathbf{y}}_k + \nabla_k U(X, Y))^2 \right]. \tag{A6}$$

The probability for the protein to perform a transition from the reactant to the product within a given time interval t is obtained by integrating the point-to-point conditional probability density (A5) over the set of protein configurations in the product state (e.g., the native state) and by averaging over the initial conditions in the reactant state (e.g., the unfolded state),

$$P_{R\to P}(t) = \int dX_P h_P(X_f) \int dX_i h_R(X_i)\, p(X_f, t|X_i)\, \rho_R(X_i), \tag{A7}$$

where $h_R(X)$ and $h_P(X)$ are the characteristic functions which, respectively, define the reactant and product state and $\rho_R(X)$ is the density of configurations in the reactant. The stochastic sampling of the transition probability (A7) for time intervals t of the order of the TPT would provide an exact description of the reactive dynamics. In the next paragraphs, we discuss the BF approach, which provides a variational estimate of the path integral (A5).

From Eq. (A5), it is immediate to read off a functional distribution $\mathcal{P}[X]$ which measures the probability to fold according to a specific folding pathway $X(\tau)$ in an *unbiased* Langevin simulation started from a given initial condition X_U and unconstrained initial momenta,

$$\mathcal{P}[X] = \int dY_f \int dY_i \int_{Y_i}^{Y_f} \mathcal{D}Y\, e^{-S[X,Y]}\, \frac{e^{-\beta\, U(X_U,Y_i)}}{Z}. \tag{A8}$$

We emphasize that in this expression, the solute trajectory $X(\tau)$ is the argument of the path probability density functional, while a path integral is performed over all solvent trajectories $Y(\tau)$.

Statistically significant folding pathways are close to the functional maximum of the path probability distribution $\mathcal{P}[X]$ and thus obey

$$\frac{\delta}{\delta X}\mathcal{P}[X] \simeq 0. \tag{A9}$$

After applying the standard reweighing trick, this extremum condition can be expressed in terms of quantities which can be extracted from *biased* simulations,

$$\frac{\delta}{\delta X}\mathcal{P}[X] = \frac{\delta}{\delta X}\left(\mathcal{P}_{bias}[X]\, \langle e^{-(S-S_{bias})}\rangle_{bias}\right) \simeq 0. \tag{A10}$$

In this equation, $\mathcal{P}_{bias}[X]$ is the path probability density in the biased dynamics, while the exponent in the second average contains the difference between the OM functionals of the biased and unbiased dynamics, respectively, averaged over the solvent trajectories—cf. Eq. (A5).

Let us now restrict the sampling to a specific subspace of folding pathways generated by the biased dynamics, which therefore defines model subspace of our variational approach. By definition, typical trajectories in this ensemble satisfy $\frac{\delta}{\delta X}\mathcal{P}_{bias} \simeq 0$. Consequently, imposing the stationary condition (A10) reduces to

$$\frac{\delta}{\delta X}\langle e^{-(S-S_{bias})}\rangle_{bias} \simeq 0. \tag{A11}$$

Furthermore, we can exploit the convexity of the path probability distribution (Feynman-Kac theorem),

$$\langle e^{-(S-S_{bias})}\rangle_{bias} \geq e^{-\langle (S-S_{bias})\rangle_{bias}}. \tag{A12}$$

So, the optimal biased folding pathway is one for which the path average

$$\frac{\int dY_f \int dY_i \int_{Y_i}^{Y_f} \mathcal{D}Y (S[X,Y] - S_{bias}[X,Y]))\; e^{-S_{bias}[X,Y]}\; e^{-\beta U(X_U,Y_i)}}{\int dY_f \int dY_i \int_{Y_i}^{Y_f} \mathcal{D}Y\; e^{-S_{bias}[X,Y]}\; e^{-\beta\; U(X_U,Y_i)}} \tag{A13}$$

is least.

Now, we recall that the rMD biasing force acts only on the solute degrees of freedom. As a consequence, even though the biased and unbiased OM actions individually depend on the solvent and solute trajectories—$X(\tau)$ and $Y(\tau)$, the so-called bias functional $T[X] \equiv (S[X,Y] - S_{bias}[X,Y]))$ depends only on the solute trajectory X. Thus, the path integrals in the numerator and denominator of Eq. (A13) cancel out, leading to the variational condition,

$$\min T[X] = \min\, (S[X,Y] - S_{bias}[X,Y]). \tag{A14}$$

Finally, in the original paper,[14] it was shown that, as long as the trial paths are generated using the biased dynamics, the BF functional $T[X]$ can be well approximated as follows:

$$T[X] \simeq \int_0^t d\tau \sum_i \frac{1}{4k_B T m_i \gamma_i} |\mathbf{F}_{bias}[X,\tau]|^2. \tag{A15}$$

In conclusion, the *biased* trajectories which have the largest probability to occur in an *unbiased* simulation—cf. condition (A9)—are those for which the biasing force has acted the least, in the sense defined by the BF functional.

[1] S. W. Englander and L. Mayne, Proc. Natl. Acad. Sci. U. S. A. **111**, 15873 (2014).

[2] A. K. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, Annu. Rev. Biophys. **37**, 289 (2008).

[3] G. Hummer and A. Szabo Proc, Proc. Natl. Acad. Sci. U. S. A. **98**, 3658 (2001).

[4] H. S. Chung, K. McHale, J. M. Louis, and W. A. Eaton, Science **335**, 981 (2012).

[5] P. Faccioli and F. Pederiva, Phys. Rev. E **86**, 061916 (2012).

[6] R. Elber, J. Chem. Phys. **144**, 060901 (2016).

[7] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).

[8] R. Elber and D. Shalloway, J. Chem. Phys. **112**, 5539 (2000).

[9] A. Ghosh, R. Elber, and H. Scheraga, Proc. Natl. Acad. Sci. U. S. A. **99**, 10394 (2002).

[10] P. Eastman, N. Gronbech-Jensen, and S. Doniach, J. Chem. Phys. **114**, 3823 (2001).

[11] P. Faccioli, M. Sega, F. Pederiva, and H. Orland, Phys. Rev. Lett. **97**, 108101 (2006).

[12] M. Sega, P. Faccioli, F. Pederiva, G. Garberoglio, and H. Orland, Phys. Rev. Lett. **99**, 118102 (2007).

[13] S. a. Beccara, T. Skrbic, R. Covino, and P. Faccioli, Proc. Natl. Acad. Sci. U. S. A. **109**, 2330 (2012).

[14] S. a. Beccara, L. Fant, and P. Faccioli, Phys. Rev. Lett. **114**, 098103 (2015).

[15] E. Paci and M. Karplus, J. Mol. Biol. **288**, 441 (1999).

[16] C. Camilloni, R. A. Broglia, and G. Tiana, J. Chem. Phys. **134**, 045105 (2011).

[17] F. Wang, S. Orioli, A. Ianeselli, G. Spagnolli, S. a. Beccara, A. Gershenson, P. Faccioli, and P. L. Wintrode, preprint arXiv:1707.05019.

[18] G. Cazzolli *et al.*, Proc. Natl. Acad. Sci. U. S. A. **111**, 15414 (2014).

[19] W. Wang, G. Cazzolli, P. Wintrode, and P. Faccioli, J. Phys. Chem. B **120**, 9297 (2016).

[20] S. a. Beccara, T. Skrbic, R. Covino, C. Micheletti, and P. Faccioli, PLoS Comput. Biol. **9**, e1003002 (2013).

[21] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).

[22] S. Kumar, J. M. Rosemberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, J. Comput. Chem. **13**, 1011–1021 (1992).

[23] A. Barducci, G. Bussi, and M. Parrinello, Phys. Rev. Lett. **100**, 020603 (2008).

[24] L. Maragliano, A. Fischer, E. Vanden Eijnden, and G. Ciccotti, J. Chem. Phys. **125**, 024106 (2006).

[25] L. Rosso, P. Minary, Z. Zhu, and M. E. Tuckerman, J. Chem. Phys. **116**, 4389 (2002).

[26] C. T. Leahy, R. D. Murphy, G. Hummer, E. Rosta, and N.-V. Buchete, J. Phys. Chem. Lett. **7**, 2676 (2016).

[27] J. M. Bello-Rivas and R. Elber, J. Chem. Phys. **142**, 094102 (2015).

[28] T. S. Van Erp, D. Moroni, and P. Bolhuis, J. Chem. Phys. **118**, 7762 (2003).

[29] V. Pande, K. Beauchamp, and G. R. Bowman, Methods **52**, 99 (2010).

[30] G. R. Bowman, V. S. Pande, and F. Noe, *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* (Springer, 2014).

[31] J.-H. Prinz, B. Keller, and F. Noé, Phys. Chem. Chem. Phys. **13**, 16912 (2011).

[32] J.-H. Prinz *et al.*, J. Chem. Phys. **134**, 174105 (2011).

[33] D. E. Shaw *et al.*, Science **330**, 341 (2010).

[34] T. J. Lane, G. R. Bowman, K. Beauchamp, V. A. Voeltz, and V. S. Pande, J. Am. Chem. Soc. **133**, 18413–18419 (2011).

[35] S. Orioli, S. a. Beccara, and P. Faccioli, J. Chem. Phys. **147**, 064108 (2017).

[36] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis *et al.*, Proteins: Struct., Funct., Bioinf. **78**, 1950 (2010).

[37] R. Zwanzig, Phys. Rev. **124**, 983 (1961).

[38] H. Mori, Prog. Theor. Phys. **33**(3), 423 (1964).

[39] W. E, W. Ren, and E. Vanden-Eijnden, Chem. Phys. Lett. **413**, 242 (2005).

[40] R. B. Best and G. Hummer, Proc. Natl. Acad. Sci. U. S. A. **107**, 1088–1093 (2010).

[41] S. Tanase-Nicola and J. Kurchan, J. Stat. Phys. **116**, 1201 (2004).

[42] F. Noé and C. Clementi, J. Chem. Theory Comput. **11**, 5002 (2015).

[43] S. Orioli and P. Faccioli, J. Chem. Phys. **145**, 124120 (2016).

[44] R. R. Coifman and S. Lafron, Appl. Comput. Harmonic Anal. **21**, 53 (2006).

[45] L. Boninsegna, G. Gobbo, F. Noé, and C. Clementi, J. Chem. Theory Comput. **11**, 5947 (2015).