Efficient Numerical Reconstruction of Protein Folding Kinetics with Partial Path Sampling and Pathlike Variables

J. Juraszek,[1] G. Saladino,[1] T. S. van Erp,[2] and F. L. Gervasio[1,3,*]

[1]Spanish National Cancer Research Centre (CNIO), calle Melchor Fernandez Almagro 3, 28029, Madrid, Spain
[2]Department of Chemistry, Norwegian University of Science and Technology, 7941 Trondheim, Norway
[3]Department of Chemistry and Institute of Structural and Molecular Biology, University College London, London WC1E 6BT, United Kingdom

(Received 4 April 2012; published 8 March 2013)

Numerically predicting rate constants of protein folding and other relevant biological events is still a significant challenge. We show that the combination of partial path transition interface sampling with the optimal interfaces and free-energy profiles provided by path collective variables makes the rate calculation for practical biological applications feasible and efficient. This methodology can reproduce the experimental rate constant of Trp-cage miniprotein folding with the same level of accuracy as transition path sampling at a fraction of the cost.

DOI: 10.1103/PhysRevLett.110.108106 PACS numbers: 87.15.A−, 82.20.Pm, 87.15.hm

Proteins rely on events spanning several time scales to perform their functions. Protein folding, conformational changes, binding, and unbinding are all rare events when compared to time scales of local fluctuations of protein structure. For this reason, the modeling of protein dynamics is a very demanding task. Even though a lot of progress has recently been made with unbiased molecular dynamics (MD) simulations [1,2], today's fastest purpose-built supercomputers reach at most 10^{13} MD steps [1], corresponding to tens of milliseconds, while the biological time scales can easily span 10^{19} time steps. In this work, we introduce a new practical approach to compute the kinetics of complex biological phenomena. We combine free-energy methods, path sampling, and path collective variables (PCVs) in a single, efficient methodology. The combined approach makes the calculation of protein folding, ligand binding, and other rare events feasible on standard computer clusters.

For decades, the standard way of computing the kinetics associated with rare events in simulations has been the *reactive flux* (RF) approach [3]. First, the free energy as a function of a single reaction coordinate (RC) is determined using an importance sampling technique. The maximum of the so-obtained free-energy profile is then used to define an approximate transition state (TS). The transmission coefficient is calculated by releasing dynamical trajectories from the top of the TS. The main problem of this approach is that, when the transmission coefficient is small or the diffusion to the product or reactant state is slow, the method is very inefficient. What is more, in protein folding, the RC is often unknown. Transition path sampling (TPS) was introduced as a valid alternative to the free-energy-based approach [4–6]. TPS was further improved by formulations that use the concept of interface crossing probabilities [7–10] like the transition interface sampling (TIS) method. Compared to RF, the advantage of TIS is that it is based on an importance sampling of the dynamical factor, the overall crossing probability, which is efficiently determined without any approximation. The partial path TIS (PPTIS) [9] approach is even faster, reducing the average path length in the importance sampling based on a semi-Markovian approximation. In PPTIS, trajectories no longer have to start at the first interface and end at the last but only have to span the region that is enclosed by three consecutive interfaces. The overall crossing probability is reconstructed from the short range crossing probabilities that are determined in these ensembles. Still, in the case of steep and deep free-energy minima, even PPTIS can be very time consuming. What is more, in most events of biological relevance, the knowledge of the free-energy landscape as a function of a set of relevant collective variables (CVs) is as important as the calculation of the kinetics. In these cases, combining PPTIS with a method that calculates the free-energy profile along an optimal CV is very beneficial. Here, we show how PPTIS can be combined with pathlike CVs and metadynamics to determine the kinetics and thermodynamics associated with rare biological events without the need of prohibitively long trajectories. We call our new approach transition state PPTIS (TS-PPTIS). Our approach can be adapted to milestoning [11] or boxed MD [12] that are similar to PPTIS. The memory-loss assumption, however, is more critical for these methods that describe the process using history-independent crossing probabilities. The definition and optimization of the reaction coordinate could be performed by methods similar to PCV and in some cases it could be substituted by geometry-optimization-based approaches. Our specific choice, however, is particularly well suited to complex multidimensional systems.

Like RF, TS-PPTIS expresses the rate as a reactive flux through the TS (λ_0): $k = P_A(\lambda_0) \times R(\lambda_0)$. $P_A(\lambda_0)$ relates to the free energy of the TS: $P_A(\lambda_0) = e^{-\beta G(\lambda_0)} / \int_{-\infty}^{\lambda_0} e^{-\beta G(\lambda)}$. $R(\lambda_0)$ is the (unnormalized)

transmission coefficient of the TS which is normally obtained by releasing many trajectories from the TS backward and forward in time. According to the effective positive flux expression [13], each trajectory is given a value equal to its starting velocity $\dot{\lambda}$; whenever this value is positive, the backward trajectory ends up in the reactant state without recrossing the TS, and the forward trajectory ends up in the product state. In all other cases, the trajectory is counted as zero. Problems arise whenever very few trajectories have a nonzero contribution or when it takes a relatively long time before the trajectories drop off the barrier to reach the stable states. TS-PPTIS solves these two issues by sampling standard PPTIS path ensembles at the barrier region. Introducing the PPTIS memory-loss assumption, the statistics of the long trajectories are then obtained from a series of short trajectories that are confined to certain overlapping regions. At sufficient distances from the TS, the probability of recrossing will become negligible so that the calculations can be stopped. As a result, TS-PPTIS will be much more computationally efficient.

As shown in the Supplemental Material [14], $R(\lambda_0)$ can be obtained using successive approximations (where TST stands for transition state theory) $R_0 = R^{\mathrm{TST}}$, $R_1 = \frac{1}{2}R^{\mathrm{TST}}(p_0^- p_0^{\pm} + p_0^+ p_0^{\mp}), \ldots,$

$$R_m = \frac{\frac{1}{2}R^{\mathrm{TST}}(p_0^- p_0^{\pm} + p_0^+ p_0^{\mp})A_m \bar{A}_m}{p_0^{\pm} A_m + p_0^{\mp}\bar{A}_m + (1 - p_0^{\pm} - p_0^{\mp})A_m\bar{A}_m} \qquad (1)$$

as a function of the PPTIS short distance crossing probabilities $(p^+, p^-, p^{\pm}, \ldots)$ of interfaces on top of the barrier $(\lambda_{-m}, \ldots, \lambda_0, \ldots, \lambda_m)$. A_m and $\bar{A}_m$ can be calculated from recursive relations (see the Supplemental Material [14]), e.g.,

$$A_{m+2} = \frac{p_m^{=} p_{m+1}^{\pm} A_m A_{m+1}}{(p_m^{\mp} p_{m+1}^{=} + p_m^{=} p_{m+1}^{\pm})A_m - p_m^{\mp} p_{m+1}^{=} A_{m+1}}.$$

R^{TST} is calculated from the average crossing velocity of the λ_0 interface (see the Supplemental Material [14]). p_0^+ and p_0^- are like the standard PPTIS conditional crossing probability $p_0^{\mp}$ and $p_0^{\pm}$ without the history dependence. The efficiency of our new method depends on how many interfaces need to be sampled in order to calculate the R_m factor, e.g., how large m is. To calculate the folding rates of a miniprotein (Trp cage), $m = 4$ was already sufficient.

The efficiency of PPTIS is determined by the selection of the RC [$\lambda(x)$ function] defining the PPTIS interfaces. What is more, in the case of TS-PPTIS, the free-energy profile along the chosen RC must be determined. An optimal and natural choice is the PCV [15]. In short, the transition between the initial state A and the final state B is described in terms of intermediate microstates $\mathbf{S}(l)$, with $\mathbf{S}(1) = \mathbf{S}_A$ and $\mathbf{S}(P) = \mathbf{S}_B$. We assume that the transition from A to B can be described by a set of collective variables $\mathbf{S}(\mathbf{R})$ which are, in general, nonlinear functions of the microscopic variables $\mathbf{R}$. The PCVs $s(\mathbf{R})$ and $z(\mathbf{R})$ are then defined as

$$s(\mathbf{R}) = \frac{1}{P-1}\frac{\sum_{l=1}^{P}(l-1)e^{-\lambda\|\mathbf{S}(\mathbf{R})-\mathbf{S}(l)\|^2}}{\sum_{l=1}^{P} e^{-\lambda\|\mathbf{S}(\mathbf{R})-\mathbf{S}(l)\|^2}}, \qquad (2)$$

$$z(\mathbf{R}) = -\frac{1}{\lambda}\ln\left(\sum_{l=1}^{P} e^{-\lambda\|\mathbf{S}(\mathbf{R})-\mathbf{S}(l)\|^2}\right), \qquad (3)$$

where λ is a constant and the information between the double vertical bars stands for the metric that defines the distance between configurations (e.g., root-mean-square deviation or RMSD). Reference [15] presents an efficient procedure improving the initially guessed $\mathbf{S}(t)$ until it lies on a minimum free-energy path connecting A with B.

The reaction rates, obtained by RF, TIS, and similar methods, do not depend on which RC is chosen as long as it distinguishes between the initial and final states. However, a poorly chosen RC can cause severe problems in RF due to hysteresis and a low transmission coefficient. TIS has proven to be less sensitive to these issues [16], but this cannot be generalized to the TIS variations like PPTIS and forward flux sampling (FFS) [13]. PCV-based RCs are, therefore, perfectly suited for these methods.

We tested our methodology on the folding of the Trp-cage (sequence NLYIQ WLKDG GPSSG RPPPS, PDB ID 1L2Y) miniprotein [17]. Despite being rather small, the Trp cage has a compact hydrophobic core and secondary structure elements, making it similar to functional proteins. It exhibits a fast, two-state folding kinetics with a folding time of 4.1 μs [18].

The native structure of a Trp cage is presented in Fig. 1 and consists of an α helix (residues 2–8), a 3_{10} helix (residues 11–14), and a polyproline II helix (residues 17–19), forming a hydrophobic pocket for the tryptophan side chain. The Trp cage has been extensively studied *in silico*, both in implicit [2,19–23] and explicit solvents [24–29]. Explicit solvent replica-exchange molecular dynamics studies and bias-exchange simulations [30]

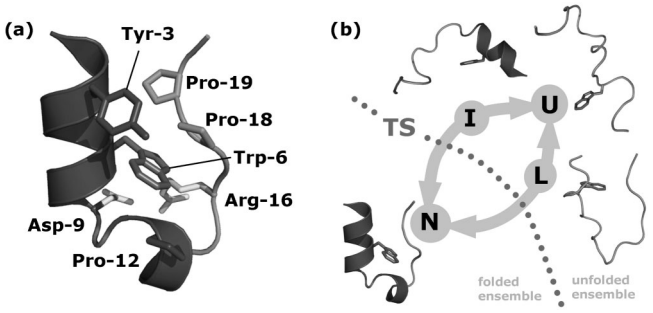

FIG. 1. (a) Structure of the Trp-cage miniprotein (PDB ID 1L2Y). The tryptophan residue and surrounding amino acids are plotted as sticks. (b) Conformational states of the Trp cage (un)folding from Ref. [25]. The TS dividing surface is schematically shown as a dotted line to indicate that both intermediates L and I are committed to the unfolded ensemble.

confirmed the Trp cage as a fast two-state folder, with an an intermediate state containing two hydrophobic cores. In a TPS study [25] using the optimized potentials for liquid simulations all-atom (OPLSAA) force field and explicit solvation, two major (un)folding routes of the Trp cage were identified (see Fig. 1). One route follows a diffusion-collision-like channel (an N-I-U path); the other resembles a nucleation-condensation mechanism (an N-L-U path). The N-I-U route fully agrees with the recent experimental elucidation of the transition state of the Trp cage [31]: The TS contains a fully formed α helix, and the salt bridge may be formed or broken. In another study [26], the TIS technique was employed to calculate the rates of the N-L path. We performed TS-PPTIS and PPTIS calculations using PCV-defined interfaces on a Trp-cage (un)folding route, to which we refer to as the N-I route (see Fig. 1), where N corresponds to the native state and I belongs to the basin of attraction of the unfolded state. All MD simulations were performed in GROMACS [32] using an OPLSAA force field and simple point charge (SPC) water. The electrostatics was treated with particle-mesh Ewald, and we used cubic periodic boundary conditions. More details can be found in the Supplemental Material [14]. To define the PCV for the N-I transition, we steered the native structure along the C_α RMSD.

The resulting initial path first solvated the hydrophobic core and then unfolded the α helix, and thus followed the N-I route. From this trajectory, we selected 26 equidistant frames such that the sum of RMSDs between them was minimal (the first and last frames were kept fixed). This path was then subject to free-energy minimization [15] and resulted in the final PCVs.

The open-source plug-in for free-energy calculations PLUMED [33] was used for metadynamics and PCV calculations. In Fig. 2, we show the free-energy projection on the $s(\mathbf{R})$ obtained using well-tempered metadynamics [34]. To focus the reaction on the N-I-U route, we placed a wall on the $z(\mathbf{R})$. The wall reduces the conformational freedom in highly entropic I and U states. This needs to be corrected for when calculating the true free energies, indicated in the figure with I_{corr} and U_{corr}. Thus, we compared the structures sampled during the metadynamics run to those sampled by the unrestrained MD simulations and calculated the free-energy correction using cluster analysis as $\Delta G_{\mathrm{corr}}(I) = k_B T \ln[N_{\mathrm{MD}}^{\mathrm{clust}}(I_{\mathrm{unbiased}})/N_{\mathrm{META}}^{\mathrm{clust}}(I_{\mathrm{biased}})]$ (where META stands for metadynamics). Since the value of the correction might depend on the cutoff and on other details of the clustering method of choice, we checked that, with two different clustering methods, in the limit of cutoff $\to 0$, it converged to the same value of about $2k_BT$ (see the Supplemental Material [14]).

We first calculated the rates with PPTIS with the PCV-defined interfaces and then compared the results with the full TS-PPTIS approach. We used 16 interfaces. For $s \in (2.5, 4)$ the interfaces were set closer to one another due to the steepness of the free-energy profile. The PPTIS ensembles were sampled on average for 45 ns with the acceptance ratio 50%–60%. We calculated the flux factors by running a 100 ns MD in the initial and final states and counted positive effective recrossings as in Ref. [26]. For more details, see the Supplemental Material [14]. We obtained $f = 1.5\ \mathrm{ns}^{-1}$ for the unfolding flux through the interface $s = 1.85$. The folding flux through $s = 15$ is $f = 0.17\ \mathrm{ns}^{-1}$. Crossing probabilities for folding and unfolding transitions converge after crossing the TS ($s \simeq 7$, Fig. 3). The rate constants are calculated by multiplying the flux factor by the crossing probability, yielding $k_{\mathrm{IN}}^{\mathrm{PPTIS}} = (0.2\ \mu\mathrm{s})^{-1}$ for folding and $k_{\mathrm{NI}}^{\mathrm{PPTIS}} = (5.4\ \mu\mathrm{s})^{-1}$ for unfolding (see Table I). The memory-loss assumption was verified in two ways. First, we monitored the overlap of the end point velocity distribution of different path ensembles as

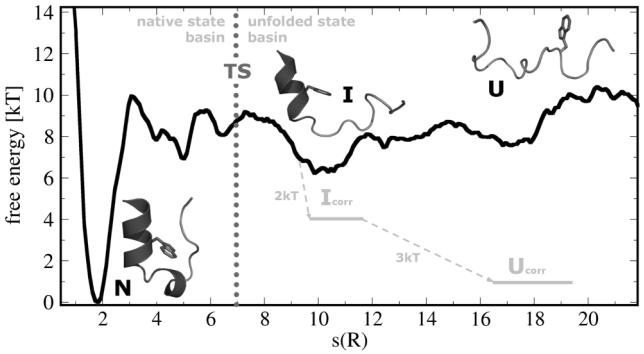

FIG. 2. Free-energy landscape of the Trp-cage N-I transition as a function of the s path collective variable. Sample N, I, and U structures are indicated in the plot as in Fig. 1. The approximate TS used for the TS-PPTIS is indicated with a dotted line. I_{corr} and U_{corr} are the corrected free-energy minima.

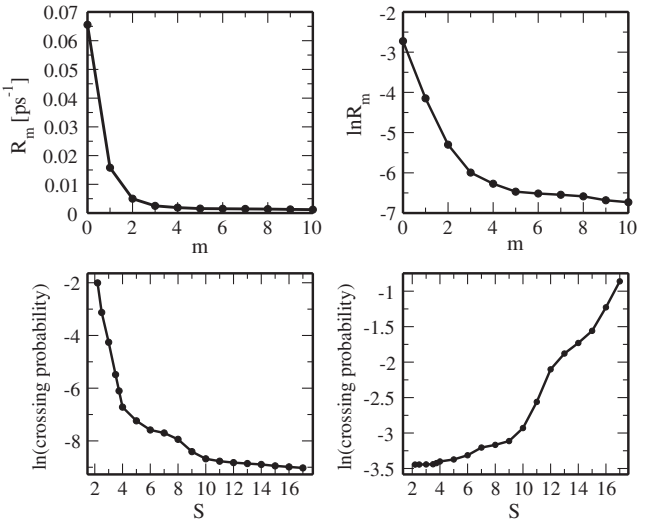

FIG. 3. Top: R_m as a function of m. Bottom: crossing probabilities in a function of the $s(\mathbf{R})$ for N-I (left) and I-N (right) transitions.

TABLE I. Comparison of the Trp-cage rate constants calculated in this work with PPTIS and TS-PPTIS versus previously calculated values using TIS [26], FFS [35], and experimental results [18].

Method	A	B	$\tau_{AB} \approx \tau_{NU}$ [μs]	τ_{BA} [μs]	$\tau_{UN}^{\mathrm{corr}}$ [μs]
Experiment	N	U	13	4.1	4.1
TIS	N	L	1.2	0.4	1.8
FFS	N	L	8	...	...
PPTIS	N	I	5.4	0.2	4.0
TS-PPTIS	N	I	3.3	0.3	6.0

proposed in Ref. [9]. Then, we repeated the calculations with a larger spacing for the interfaces. In both cases, the assumption was verified (see the Supplemental Material [14]).

We then repeated the calculation with TS-PPTIS. From the free-energy profile, the TS appears to be around $s = 7$. Therefore, we only used the interfaces closest to $s = 7$ to calculate the rates (7 out of 16). The unfolding rate hardly depends on the free energy of the I state, and we obtain $k_{NI}^{\mathrm{TS\ PPTIS}} \approx k_{NU}^{\mathrm{TS\ PPTIS}} = (3.3\ \mu\mathrm{s})^{-1}$. To obtain the folding rate, we increased the folding barrier by the calculated correction of $2k_BT$. This gave us $k_{IN}^{\mathrm{TS\ PPTIS}} = (0.3\ \mu\mathrm{s})^{-1}$. The convergence of TS-PPTIS is depicted in Fig. 3. R_m flattens out around $m = 4$. Adding further interfaces does not influence the resulting rate constants. The results are summarized in Table I and compared to the values calculated using TIS [26] and FFS [35] and to the experimental results [18]. The folding time $\tau_{UN}^{\mathrm{corr}}$ for the PPTIS and TS-PPTIS simulations were calculated from the formula $\tau_{UN}^{\mathrm{corr}} = \tau_{IN}/e^{-\Delta G_{IU}}$ with $\Delta G_{IU} = 3k_BT$. Both PPTIS and TS-PPTIS are in agreement with our previous TIS calculations, albeit using only a fraction of the computing cost. The expected slight disagreement with the experimental unfolding times is known to be due to the OPLSAA force field [26]. The folding times τ_{UN} (see Table I) are in near-total agreement with the experimentally measured values. Path collective variables are an important addition to the PPTIS methodology, allowing one to channel the pathways toward the right transition. By employing PCVs, which by construction follow an optimal free-energy path, we solve the problem of PPTIS related to the choice of RC. The advantage of the combination of PPTIS with PCV over TIS is a significant gain in efficiency. In the TIS study [26], the aggregate CPU time necessary to calculate the (un)folding rate constants was 26 μs. PPTIS only required a total sampling time of less than 1 μs. Since we estimate their respective accuracies to be the same, this suggests about a 20-fold improvement in terms of CPU time. The TS-PPTIS rate constant calculation is even more efficient. In the case of the Trp cage, whose free-energy landscape is rather flat, representing a worse-case scenario for the approach, the gain in the total computing time (including the free-energy calculation) is of only 20% between PPTIS and TS-PPTIS. In more complex biological systems, however, whenever deep free-energy minima are present in the landscape (as in the case of ligands binding with a long residence time), very significant efficiency gains are expected. A typical example of the kinetics of a drug binding to a pharmaceutically relevant target (the heat shock protein 90) is reported in the Supplemental Material [14]. The deterioration of the efficiency in steep regions of the free energy with the consequent need for many closely spaced interfaces and the risk of violating the memory-loss assumption makes use of PPTIS and similar approaches that are impractical or even impossible. Thus, when deep minima are found in the free-energy landscape, TS-PPTIS PCV has clear advantages over similar approaches.

We acknowledge the PRACE Research Infrastructure Resource (Tier-1) HECToR at EPCC, United Kingdom, and (Tier-0) SuperMUC at LRZ, Germany, for computational resources. J. J. and G. S. contributed equally to this work.

*f.l.gervasio@ucl.ac.uk

[1] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, Science **334**, 517 (2011).
[2] C. D. Snow, B. Zagrovic, and V. Pande, J. Am. Chem. Soc. **124**, 14548 (2002).
[3] D. Frenkel and B. Smit, *Understanding Molecular Simulation* (Academic, San Diego, 2002), 2nd ed.
[4] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, J. Chem. Phys. **108**, 1964 (1998).
[5] C. Dellago, P. G. Bolhuis, and P. L. Geissler, Adv. Chem. Phys. **123**, 1 (2002).
[6] L. R. Pratt, J. Chem. Phys. **85**, 5045 (1986).
[7] T. S. van Erp, D. Moroni, and P. G. Bolhuis, J. Chem. Phys. **118**, 7762 (2003).
[8] R. Allen, C. Valeriani, and P. ten Wolde, J. Phys. Condens. Matter **21**, 463102 (2009).
[9] D. Moroni, P. G. Bolhuis, and T. S. van Erp, J. Chem. Phys. **120**, 4055 (2004).
[10] T. van Erp and P. Bolhuis, J. Comput. Phys. **205**, 157 (2005).
[11] A. K. Faradjian and R. Elber, J. Chem. Phys. **120**, 10880 (2004).
[12] D. R. Glowacki, E. Paci, and D. V. Shalashilin, J. Phys. Chem. B **113**, 16603 (2009).
[13] T. S. van Erp, Adv. Chem. Phys. **151**, 27 (2012).
[14] See Supplemental Material http://link.aps.org/supplemental/10.1103/PhysRevLett.110.108106 for full derivation of the TS-PPTIS method, application to a realistic protein ligand binding example, and tests of the memory loss assumption.
[15] D. Branduardi, F. Gervasio, and M. Parrinello, J. Chem. Phys. **126**, 054103 (2007).
[16] T. S. van Erp, J. Chem. Phys. **125**, 174106 (2006).
[17] J. Neidigh, R. Fesinmeyer, and H. Andersen, Nat. Struct. Biol. **9**, 425 (2002).
[18] L. Qiu, S. Pabit, A. Roitberg, and S. Hagen, J. Am. Chem. Soc. **124**, 12952 (2002).

[19] C. Simmerling, B. Strockbine, and A. Roitberg, J. Am. Chem. Soc. **124**, 11258 (2002).
[20] M. Ota, M. Ikeguchi, and A. Kidera, Proc. Natl. Acad. Sci. U.S.A. **101**, 17658 (2004).
[21] J. Pitera and W. Swope, Proc. Natl. Acad. Sci. U.S.A. **100**, 7587 (2003).
[22] S. Chowdhury, M. Lee, and Y. Duan, J. Phys. Chem. B **108**, 13855 (2004).
[23] W. Zheng, E. Gallicchio, N. Deng, M. Andrec, and R. Levy, J. Phys. Chem. B **115**, 1512 (2011).
[24] R. Zhou, Proc. Natl. Acad. Sci. U.S.A. **100**, 13280 (2003).
[25] J. Juraszek and P. G. Bolhuis, Proc. Natl. Acad. Sci. U.S.A. **103**, 15859 (2006).
[26] J. Juraszek and P. G. Bolhuis, Biophys. J. **95**, 4246 (2008).
[27] D. Paschek, H. Nymeyer, and A. Garcia, J. Struct. Biol. **157**, 524 (2007).
[28] C. Jimenez-Cruz, G. Makhatadze, and A. Garcia, Phys. Chem. Chem. Phys. **13**, 17056 (2011).
[29] R. Day, D. Paschek, and A. E. Garcia, Proteins **78**, 1889 (2010).
[30] S. Piana and A. Laio, J. Phys. Chem. B **111**, 4553 (2007).
[31] R. Culik, A. Serrano, M. Bunagan, and F. Gai, Angew. Chem., Int. Ed. Engl. **50**, 10884 (2011).
[32] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).
[33] M. Bonomi *et al.*, Comput. Phys. Commun. **180**, 1961 (2009).
[34] M. Bonomi, A. Barducci, and M. Parrinello, J. Comput. Chem. **30**, 1615 (2009).
[35] F. E. C. Velez-Vega and E. E. Borrero, J. Chem. Phys. **133**, 105103 (2010).