

Unbiased free energy estimates in fast nonequilibrium transformations using Gaussian mixtures

Piero Procacci

Citation: *The Journal of Chemical Physics* **142**, 154117 (2015); doi: 10.1063/1.4918558

View online: <http://dx.doi.org/10.1063/1.4918558>

View Table of Contents: <http://scitation.aip.org/content/aip/journal/jcp/142/15?ver=pdfcov>

Published by the [AIP Publishing](#)

Articles you may be interested in

[Estimating equilibrium ensemble averages using multiple time slices from driven nonequilibrium processes: Theory and application to free energies, moments, and thermodynamic length in single-molecule pulling experiments](#)

J. Chem. Phys. **134**, 024111 (2011); 10.1063/1.3516517

[Density-dependent analysis of nonequilibrium paths improves free energy estimates](#)

J. Chem. Phys. **130**, 204102 (2009); 10.1063/1.3139189

[Bayesian estimates of free energies from nonequilibrium work data in the presence of instrument noise](#)

J. Chem. Phys. **129**, 024102 (2008); 10.1063/1.2937892

[Adaptively biased molecular dynamics for free energy calculations](#)

J. Chem. Phys. **128**, 134101 (2008); 10.1063/1.2844595

[Exploring the free-energy landscape of a short peptide using an average force](#)

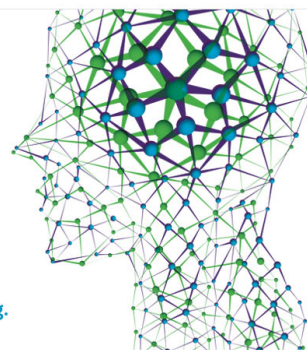
J. Chem. Phys. **123**, 244906 (2005); 10.1063/1.2138694

How can you **REACH 100%**
of researchers at the Top 100
Physical Sciences Universities? (TIMES HIGHER EDUCATION RANKINGS, 2014)

With *The Journal of Chemical Physics*.

AIP | The Journal of
Chemical Physics

THERE'S POWER IN NUMBERS. Reach the world with AIP Publishing.



Unbiased free energy estimates in fast nonequilibrium transformations using Gaussian mixtures

Piero Procacci

Dipartimento di Chimica, Università di Firenze, Via della Lastruccia 3, I-50019 Sesto Fiorentino, Italy and Centro Interdipartimentale per lo Studio delle Dinamiche Complesse (CSDC), Via Sansone 1, I-50019 Sesto Fiorentino, Italy

(Received 19 January 2015; accepted 3 April 2015; published online 20 April 2015)

In this paper, we present an improved method for obtaining unbiased estimates of the free energy difference between two thermodynamic states using the work distribution measured in nonequilibrium driven experiments connecting these states. The method is based on the assumption that any observed work distribution is given by a mixture of Gaussian distributions, whose normal components are identical in either direction of the nonequilibrium process, with weights regulated by the Crooks theorem. Using the prototypical example for the driven unfolding/folding of deca-alanine, we show that the predicted behavior of the forward and reverse work distributions, assuming a combination of only two Gaussian components with Crooks derived weights, explains surprisingly well the striking asymmetry in the observed distributions at fast pulling speeds. The proposed methodology opens the way for a perfectly parallel implementation of Jarzynski-based free energy calculations in complex systems. © 2015 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4918558>]

I. INTRODUCTION

The Jarzynski¹ and the Crooks² theorems relate the free energy difference, $\Delta G = G_B - G_A$, between two thermodynamic states A and B to the external work done on the system in an ensemble of driven nonequilibrium (NE) paths connecting A to B all done with a common time schedule. In these NE experiments, an external potential, acting on some collective variable λ , drives the system from initial (equilibrium) configurations of the state A, canonically sampled at $\lambda = \lambda_A$, to a final state B with $\lambda = \lambda_B$ producing a work W . The Jarzynski theorem then reads

$$e^{-\beta\Delta G} = \langle e^{-\beta W} \rangle = \int e^{-\beta W} P_{AB}(W) dW, \quad (1)$$

where the mean $\langle e^{-\beta W} \rangle$ is taken over these NE realizations and $P_{AB}(W)$ is the associated work distribution function for the $A \rightarrow B$ nonequilibrium process. Equation (1) is valid for virtually any quantum or classical system no matter how fast the transformation from A to B is carried on. For infinitely slow (quasi-static) realizations, the work W is always equal to ΔG , while for instantaneous processes, one recovers the Zwanzig free energy perturbation formula³ $e^{-\beta\Delta G} = \langle e^{-\beta(H_B - H_A)} \rangle_A$.

In the last decade, with the advent of massively parallel high performance computing (HPC), the Jarzynski relation has had a significant impact^{4–11} in the free energy calculations on large scale biological system and in condensed phases in general. This is due to the fact that the parallel computation of the average Eq. (1) is not subject to the Amdahl law¹² in the sense that, once the starting equilibrium states of A have been prepared, all NE trajectories may proceed independently and concurrently with zero fraction of serial code and with no need for communication whatsoever. The free energy may then be recovered with a simple and basically instantaneous

post-processing operation of the work data acquired after the parallel computation had finished. The bad news is that the Jarzynski formula, Eq. (1), is based on an exponential average over a distribution, i.e., a statistics that is both inherently noisy and biased, even if the spread of the work data is only moderately larger than $k_B T$. The weight factor $e^{-\beta W}$ in Eq. (1) makes the average crucially dependent on the behavior at the left tail of the distribution, which, by definition, is not well sampled, resulting in both high variance and bias. Previous studies have explored and demonstrated the poor behavior of exponential averaging for small sample sizes.^{11,13–18} In all these studies it was consistently shown that the statistical bias is more acute when the process is conducted at high pulling speed, i.e., when the spread of the distribution (related to the mean dissipation) increases. In other words, the parallel computation of the Jarzynski exponential average poses an inescapable conflict between accuracy and computational speed and/or computational CPU resources.¹⁷

One possibility to partially cure this pathology is that of performing a small amount of *longer* less dissipative trajectories, hence diverting part of the parallel resources in speeding up these slow trajectories (double layer parallelism). This is the approach commonly used in the potential of mean force based (PMF) calculation of drug-receptor system^{4–9} where the drug is mechanically driven along a selected collective coordinate λ from the binding pocket to the bulk solvent in a time in the order of nanoseconds. The bias in this case is mitigated but is by no means suppressed. Besides, parallelization is less efficient since, while trajectories do not communicate in the Jarzynski parallel layer, communication/synchronization must occur in the force decomposition layer within one single long trajectory.

An alternative way to alleviate the problem of the exponential bias is that of using the *bidirectional* approach with

application of the Crooks theorem² to the distribution of the forward process (equilibrium of A to B), $P_{AB}(W)$, and to the mirror distribution of the reverse process (equilibrium of B to A), $P_{BA}(-W)$, i.e.,

$$\frac{P_{AB}(W)}{P_{BA}(-W)} = e^{\beta(W-\Delta G)}. \quad (2)$$

In Eq. (2), the reverse process is assumed to be done with inverted time schedule. The free energy $\Delta G = G_B - G_A$ lies at the crossing point of the forward and reverse distributions and such point may be determined using the standard Bennett Acceptance Ratio (BAR).¹⁹ Several studies^{16,20,21} have demonstrated the accuracy of the BAR estimate of the free energy with the forward and reverse process being carried on at a much faster rate than that used in unidirectional approaches. Bidirectional experiments thus restore perfect parallelism, but they are cumbersome since they imply the preparation of two equilibrium thermodynamic states and are often non resolutive. To see why, let us take as an example the unfolding/refolding process in a poli-peptide. The NE unfolding from the native state A to a, e.g., *completely extended* or unfolded state B is a straightforward and moderately dissipative process but the refolding from the extended state in B to the native state in A is a difficult and highly dissipative task even at low pulling speeds and with a careful choice of the driven collective variables. Another important example is the driven undocking/docking in drug-receptor systems. For pulling speed of the order of 0.1:1 nm/ns, the dissipation in the undocking process is in general small since the end state of the drug is in the bulk solvent for all trajectories. In the inverted re-docking process from the bulk, the drugs may instead end up in a manifold of unfavorable nonequilibrium poses on the protein before reaching the binding pocket, thus producing a work distribution with a large spread. These typically asymmetrical situations will set the more dissipative distribution $P_{BA}(-W)$ far from $P_{AB}(W)$ rendering difficult and inaccurate the calculation of ΔG via the Bennett formula at high speeds.

There are finally a number of studies aimed at making an important (path) sampling on the NE trajectories^{18,22-25} in order to avoid the problem of the bias in the exponential average Eq. (1). In the adaptive steered molecular dynamics (ASMD), the $A \rightarrow B$ process is divided in a number of segments such that the important path in the nonequilibrium reaction of the N trajectories is generated resetting the starting the N configurations for the so-called subset of pre-defined “environment variables” at the end of each segment to the *single one* that had the work value closest to the exponential average of the work ensemble of the segment.^{22,26} The technique has been recently refined²⁵ in the multiple branching (MB-ASMD) variant whereby from a set of optimal generators taken at the end of each segment and selected with a probabilistic criterion, a fixed number of new trajectories are produced in the subsequent segment. Oberhofer and Dellago evaluated a free energy dependent optimal weighting function for minimizing the bias in the Jarzynski average.¹⁸ Chelli *et al.*²³ adopted a Monte Carlo scheme to eliminate trajectories on the fly that diverge significantly from the same work average. Bussi and coworkers^{24,27} used

a reweighting scheme where the weighted histogram method is used to combine snapshots obtained at different stages of the pulling, deriving again as in Ref. 18 a free energy dependent weight factor. In all these techniques, that involve a certain extent of tweaking,^{26,28} the A to B process is partitioned in a number of sub-processes whose precise sequence requires a decision over the whole data set, introducing a trajectory dependence on a global level that destroys the inherently parallel nature of the Jarzynski formula.

In this contribution, we show that the free energy difference between the end points in typical pathologically asymmetric cases such as the unfolding/folding of a small protein can be computed with great accuracy using relatively few *completed*, *non-communicating*, and *untweaked* NE trajectories with a generalization of the *unbiased* Jarzynski estimate based on the second order cumulant expansion. This is achieved on the assumption that any observed work distribution in NE processes is given by a mixture of Gaussian distributions. More in detail, using the prototypical example for the driven unfolding/folding of deca-alanine,^{11,25,29-31} we show that the predicted behavior of the forward and reverse distributions based on the assumption of the existence of one single metastable state in the folded state characterized by λ_A of the helix explains surprisingly well the observed asymmetry in the forward and reverse work distributions. We also show that the free energy estimate based on the mixture with two Gaussian components provides, by a simple constrained fit of the individual μ_i and σ_i^2 cumulants of the mixture, an unbiased unidirectional estimate of the free energy in *either* direction for pulling speed as fast as 15:20 nm/ns with an accuracy comparable or even superior to that of the bidirectional BAR method. The proposed algorithm removes the bias that plagues the free energy estimate via exponential averages fully preserving the inherent parallel nature of the Jarzynski approach.

II. THEORY

Park and Schulten in their ten year old seminal paper on steered molecular dynamics of deca-alanine,¹¹ showed that, if the driven process is Markovian (i.e., the process may be described by an overdamped Langevin equation), then the work distribution is Gaussian. In this case, the exponential average is replaced by the *unbiased* estimate

$$\Delta G = \mu - \beta \frac{\sigma^2}{2}, \quad (3)$$

where $\mu = \langle W \rangle$ and σ are the mean value and the standard deviation of the normal distribution. It should be stressed that Eq. (3) is an *exact* result if the underlying work distribution is Gaussian. If the forward process is Gaussian, the Crooks theorem imposes that the reverse process is also Gaussian with a *symmetrical* distribution $P_{BA}(-W)$ with respect to $W = \Delta G$ bearing the *same* variance σ and with *identical dissipation* $W_d = \frac{1}{2}\beta\sigma^2$ as in $P_{AB}(W)$. Hence, the forward and reverse mean works in $P_{AB}(W)$ and $P_{BA}(-W)$ are related by $\langle W \rangle_{BA} = \langle W \rangle_{AB} - \beta\sigma^2$. The utility of these equations has been severely undermined, in principle, by the outcomes of several recent bidirectional studies on real systems, all characterized

by non Gaussian or strongly asymmetric Gaussian look-alike distributions.^{20,30,32,33} In practice, however, in many instances where the ideal Markovian distribution was manifestly violated, Eq. (3) has been proved to be surprisingly accurate, even at high pulling speed, when applied on a particular direction of the process as in the NE double *annihilation* method³⁴ for evaluating the binding free energies via fast switching alchemical transformations,²¹ and in the prototypical *unfolding* of deca-alanine.^{11,29,31,32} This is somewhat puzzling given that in the opposite direction, e.g., in the NE refolding of deca-alanine, Eq. (3) may be completely inaccurate even at low pulling speeds.^{25,32} Clearly, in this strongly asymmetric case, these two distributions only “look” Gaussian but they are not, and all their cumulants, according to Marcinkiewicz theorem³⁵ must all be different from zero. Of course, such strongly asymmetrical distributions still obey to the Crooks theorem. A possible explanation for the striking asymmetry observed for seemingly Gaussian forward and reverse distributions may lie in the fact that the driven process in either direction cannot be described by an overdamped Langevin process, as assumed in Ref. 11. An alternative explanation may be the one first suggested by Feng and Crooks³⁶ and formalized later on in Ref. 30, i.e., the observed univariate work distribution is a *superposition of Gaussian processes* related to competitive free energy basins, “felt” during the NE experiments. To understand the implications of such an assumption, let us use again deca-alanine as paradigmatic example. We start from the helix folded state, assuming that this state has overwhelming probability to occur at equilibrium. In the NE experiment, one usually externally drives the λ end-to-end distance from the λ_A corresponding to the end-to-end distance in the helix up to a value λ_B corresponding to that of the all-*trans* totally unfolded state. In the reverse process, also for moderate pushing speed, at the final point λ_A , the helix is rarely seen with deca-alanine populating a manifold of misfolded states. The Boltzmann weights at equilibrium of these misfolded states are negligible, but in the λ driven refolding NE experiment these misfolded states can be explored and detected because of the extra energy provided by the dissipation. To put all this more formally, consider a thermodynamic state A at given value of a reaction coordinate λ_A characterized at equilibrium by a *manifold* of free energy basins, and a state B at λ_B with a *single* free energy basin. We now make the assumption that the distribution of the work for the λ driven *forward* NE process from A (equilibrium) to B is given by a superposition of N Gaussian distributions with weights c_1, \dots, c_N arising from the N metastable states of A. The weights c_i composing the overall work distribution depend on the corresponding starting Boltzmann weights (not necessarily disparate), all at $\lambda = \lambda_A$, and on the pulling speed, i.e.,

$$P_{AB}(W) = \sum_i^N c_i N(W, \mu_i, \sigma_i). \quad (4)$$

In Eq. (4), $N(W, \mu, \sigma) = \frac{1}{\sigma(2\pi)^{1/2}} e^{-(W-\mu)^2/(2\sigma^2)}$ is the normal distribution and the coefficients c_i are such that $\sum_i^N c_i = 1$. Both the coefficients c_i , and the moments μ_i, σ_i of the distributions are a function of the pulling speed. Now,

the Crooks theorem, Eq. (2), imposes that the distribution for the reverse process [from B(equilibrium) to A] must be also a superposition of normal distributions such that

$$P_{BA}(-W) = \sum_{i=1}^N c_i e^{\beta(\Delta G - \mu_i + \frac{\beta\sigma_i^2}{2})} N(W, \mu_i - \beta\sigma_i^2, \sigma_i) \\ = e^{\beta\Delta G} \sum_{i=1}^N c_i e^{-\beta(\mu_i - \frac{\beta\sigma_i^2}{2})} N(W, \mu_i - \beta\sigma_i^2, \sigma_i), \quad (5)$$

where $N(W, \mu_i - \beta\sigma_i^2, \sigma_i) = \frac{1}{\sigma(2\pi)^{1/2}} e^{-(W - \mu_i + \beta\sigma_i^2)/(2\sigma_i^2)}$ has half-width equal to σ_i and mean equal to $\langle W \rangle = \mu_i - \beta\sigma_i^2$.

According to Eq. (5), the coefficients of the reverse distribution, say d_i , are related to the c_i coefficient through the equation

$$d_i = e^{\beta\Delta G} e^{-\beta(\mu_i - \frac{\beta\sigma_i^2}{2})} c_i. \quad (6)$$

By integrating Eq. (5) over W on both sides, the normalization condition on $P_{BA}(-W)$ imposes that

$$1 = e^{\beta\Delta G} \sum_{i=1}^N c_i e^{-\beta\Delta g_i}, \quad (7) \\ \Delta G = -k_B T \ln \sum_i^N c_i e^{-\beta\Delta g_i}$$

where we have defined the quantities

$$\Delta g_i = \mu_i - \frac{\beta\sigma_i^2}{2}. \quad (8)$$

We first note that for $N = 1$ (i.e., only one state in A at λ_A), Eq. (7) reduces to Eq. (3). Equation (7) can therefore be considered as the extension of Eq. (3) to the case of a work distribution given by a mixture of normal distributions and relates the free energy difference between the single state B at λ_B and the state A, characterized by a manifold of competing sub-states all at λ_A , to the moments of the distributions referring to these states and to the combination coefficients c_i . The estimate of ΔG via Eq. (7) is in principle *unbiased* since *all* trajectories must be used to compute the moments μ_i, σ_i , and the coefficients c_i of the underlying normal distributions composing the overall distribution of the work. The energies Δg_i in the $N > 1$ case should be strictly *related* to the individual free energies of the basins in A although they are not, in general, the true free energies of the basins. As stated above, μ_i and σ_i are a function of the pulling speed and the same should apply (except for the $N = 1$ case) to the quantities Δg_i . This is so since, during the NE process in either direction, there can be a mixing between the paths to B originated from the metastable states of A and *viceversa* as schematized in Figure 1 for the case of folding/unfolding of deca-alanine.

The individual free energies of the metastable states obey the *equilibrium* relation

$$e^{\beta\Delta G} = \sum_i e^{\beta\Delta g_i}, \quad (9)$$

where the sense of the process is that of the unfolding, i.e., that for which ΔG is *positive*. Equation (7) is a *nonequilibrium* relation that must coincide with the equilibrium Eq. (9)

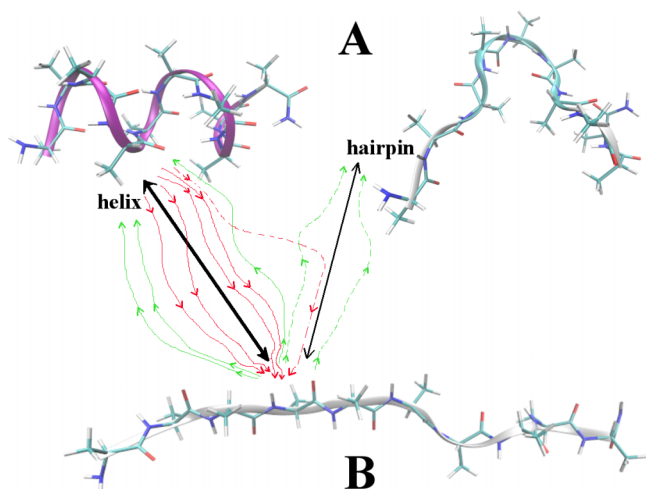


FIG. 1. Folded/unfolded transformation in deca-alanine. The double-headed black line connects A (state 1 and state 2) and B via quasi-static (reversible) transformations. The red lines connect the helix state of A ($\lambda = 1.55$ nm) to the extended state in B ($\lambda = 3.15$ nm) via NE paths. The green lines connect the extended state in B to the two states in A. Dashed lines refer to processes following paths (in part or always) connecting B to the metastable state basin in A.

when the duration τ of the driven experiment is infinite, i.e., for quasi-static transformations. It then follows, comparing Eqs. (7)–(9), that the weights obey to the relation

$$\lim_{\tau \rightarrow \infty} c_i(\tau) = \frac{e^{-\beta \Delta G_i}}{e^{-\beta \Delta G_i(\tau)}}, \quad (10)$$

where $\Delta G_i \equiv G_B - G_{A,i}$ is the true free energy difference referred to the i th metastable state and where we have made explicit the dependence of Δg_i and c_i from the duration of the NE experiment. In the limit $\tau \rightarrow \infty$, i.e., for a reversible process, there is by definition perfect mixing right at the start of the process where all metastable states are sampled with the correct Boltzmann weight, even if one launches the experiment from the configurations referring to a single free energy basin. In this case, one actually has $\mathcal{N} = 1$, as metastability is a void concept for infinite duration of the experiment. In this situation, one can still use the now cumbersome multimodal formula if and only if we set

$$\Delta g_i(\infty) = \Delta G_i \quad (11)$$

such that, according to Eq. (10), all the coefficients in the combination becomes identical. This must occur because, for quasi-static processes, one cannot isolate a subset of starting configurations referring to a single free energy basin: mixing becomes perfect after an infinitesimal advance of the driven reaction. Therefore, when the experiment is done reversibly at infinitely slow speed, all possible subsets of starting configurations of the metastable states should be lumped in a single macrostate with their corresponding Boltzmann weights.

In deriving Eq. (7), we have made no assumption on the Boltzmann weights of the free energy basins in the state A. Let us now consider the case where just one of these basins has a disproportionate Boltzmann weight with respect to the others, all lying several $k_B T$ above the principal state. As stated in Sec. I, the folded (A)/unfolded (B) and the

docked (A)/undocked (B) pairs can be considered two important examples of such situation. We term these underlying, low probability metastable states, that are not included in the pool of initial configurations for the forward A (manifold) to B (single state) process, as *shadow states*. Equation (7) tells us that these shadow states should nonetheless affect the work distribution in the NE experiments from A to B yielding a mixture of Gaussian processes for non zero rates. The existence of the shadow states, as we shall show in Sec. IV, becomes manifest when performing the transformation in the reverse direction, i.e., from the single state B to the manifold A.

What are then the consequences of Eq. (7) in the practice of NE experiments when A has a principal stable state and a manifold of metastable shadow states? More explicitly, since the shape of the reverse distribution is implied in that of the forward distribution via Eqs. (7) and (6), can the observed asymmetry of $P_{AB}(W)$, $P_{BA}(-W)$ at fast pulling speeds for the prototypical case of the unfolding/refolding of deca-alanine be rationalized on the basis of the overall distributions given by a mixture of normal distributions?

We shall limit the following discussion to the case with $\mathcal{N} = 2$, i.e., one principal state and one metastable in A at λ_A . As we shall see later on in Sec. IV, such a simple assumption explains with surprising accuracy the striking asymmetry observed in the folding/unfolding process of a poli-peptide at fast rates.

For $\mathcal{N} = 2$, we have only one independent coefficient $c_1 = c$ so that $c_2 = 1 - c$. Equation (7) in this case becomes

$$1 = e^{\beta \Delta G} \left[c \left(e^{-\beta(\mu_1 - \frac{\beta \sigma_1^2}{2})} - e^{-\beta(\mu_2 - \frac{\beta \sigma_2^2}{2})} \right) + e^{-\beta(\mu_2 - \frac{\beta \sigma_2^2}{2})} \right] \quad (12)$$

recovering Eq. 8 of Ref. 30. Using Eq. (6), we find that the ratio between the c coefficient and the corresponding d coefficient for the reverse distribution is given by

$$r = \frac{c}{d} = \frac{1}{e^{\beta \Delta G - (\mu_1 - \frac{1}{2} \beta \sigma_1^2)}}. \quad (13)$$

Note that this ratio depends on the difference between the true overall free energy and that computed using Eq. (3) referring to the mean and variance of just one of the two Gaussian distributions. With no loss of generality, we now assume that state 1 in A has a much higher Boltzmann weight than the other, so that $\Delta g_1 \gg \Delta g_2$. It follows that in the generation of the NE trajectories from A to B, the starting equilibrium configurations of A should all be sampled from this principal free energy basin at $\lambda = \lambda_A$. This is the typical case of deca-alanine *in vacuo* where at $\lambda = 1.55$ nm the helix is by far the deepest free energy basin among all other states at the same value of λ . At slow pulling speed, i.e., for nearly reversible processes, we find that $\Delta G \approx \mu_1 - \beta \sigma^2/2$ such that the ratio r in Eq. (13) is close to 1. In this situation, the shadow metastable state is basically not perceived neither in the forward nor in the reverse process from B to A, because the latter process is so slow that the system has the time to gradually relax to the lowest state of the A twofold system in virtually all independent return trajectories. Viewed from another perspective, for slow pulling/pushing back processes, the shadow metastable state does not surface

out because there is simply not enough extra energy in the form of dissipation to overcome the barriers around the helix funnel. The system behaves as if $N = 1$ and Eq. (3), rather than Eq. (7), applies. On the other end, when the pulling speed is very fast, then the simple Gaussian estimate Eq. (3) becomes inaccurate, typically *overestimating* the free energy in the *forward* direction,^{11,14,15,25,29,32} i.e., $(\mu_1 - \frac{1}{2}\beta\sigma_1^2) > \Delta G$. In this case, the denominator in Eq. (13) is a number that can be sensibly < 1 , so that $c \gg d$, i.e., $c_1 \approx d_2 \approx 1$. Accordingly, $N(W, \mu_1, \sigma_1)$, related to the principal basin in A, is still the dominant Gaussian contribution in the overall forward distribution $P_{AB}(W)$, but the distribution with standard deviation σ_2 and mean $\mu_2 - \beta\sigma_2^2$, related to the metastable state of A, remarkably becomes dominant in shaping the observed *reverse* distribution. At intermediate pulling speeds, while $P_{AB}(W)$ is always dominated by $N(W, \mu_1, \sigma_1)$, $P_{BA}(-W)$ has comparable contribution from both $N(W, \mu_1 - \beta\sigma_1^2, \sigma_1)$ and $N(W, \mu_2 - \beta\sigma_2^2, \sigma_2)$.

III. METHODS

The set up of the deca-alanine system is identical to that already used in past studies.^{30,32,37} Briefly, molecular dynamics simulations are done *in vacuo* using a Nosé-Hoover thermostat at $T = 300$ K. The Hamiltonian of the system is given by

$$H = H_0 + \frac{1}{2}k(\lambda(x) - \lambda(t))^2, \quad (14)$$

where $\frac{1}{2}k(\lambda(x) - \lambda(t))^2$ is the external potential of the driving external device coupled to the $\lambda(x)$ reaction coordinate corresponding to the distance between the N_1 and the N_{10} ammidic nitrogen atoms. The force constant k is set to $100 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$, i.e., we work in the so-called stiff spring approximation^{11,37} so that

$$\begin{aligned} \mathcal{G}(\lambda) &= \int dx e^{-\beta H_0(x)} \delta[\lambda - \lambda(x)] \\ &\approx G(\lambda) = \int dx e^{-\beta[H_0(x) + \frac{1}{2}k(\lambda(x) - \lambda(t))^2]}, \end{aligned} \quad (15)$$

with $\mathcal{G}(\lambda)$ being the free energy along $\lambda(x)$ of the unconstrained system. As thoroughly discussed in Ref. 37, a force constant of $100 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ for the case of the unfolding/refolding of deca-alanine *in vacuo* at $T = 300$ K is largely within the so-called “stiff spring regime,” that is, largely above the threshold minimum value ($\approx 0.6 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$) for the insurgence of a bi-stability. 512 initial configurations in the A state are prepared with a simulation of ≈ 1 ns time span of the *helical* state with fixed value of the reaction coordinate at $\lambda = 1.55 \text{ nm}$ by sampling the atomic coordinates at regular interval of 2 ps. The 512 initial configurations for the B state are prepared in a similar manner by setting $\lambda = 3.15 \text{ nm}$. It should be stressed here that *all* the 512 equilibrium configurations in the A state at $\lambda = 1.55 \text{ nm}$ refer to the α -helical state, i.e., the state with overwhelming Boltzmann weight in the given thermodynamic conditions. Starting from the corresponding equilibrium configurations, 512 NE trajectories are then produced in the forward (A to B) and in reverse direction (B to A) at various pulling/pushing speeds ranging from 76.2 m/s (for a duration τ of 0.021 ns) down to 0.38 m/s (for a duration τ of 4.2 ns). All calculations were done using the parallel version of the ORAC program.³⁸

IV. RESULTS

As noted in past studies,^{30,32} the misfolded end states in A sampled in the NE refolding of deca-alanine are actually more than one and their number and distribution depend on the pulling speeds. In Figure 1, for example, along with the stable helical state, we show a hairpin configuration. Such state is one of the most probable ending states in the $B \rightarrow A$ NE process for moderate pushing speeds. If the NE $B \rightarrow A$ experiment is done at a slower rate, then the ending misfolded states are mostly α -helix or incomplete helical types with less than three turns. In general, the faster the speed is, the more disordered the sample of the final NE state at $\lambda = 1.55 \text{ nm}$ will be. In Figure 2, we show the representative structures of the most populated clusters observed in the final configurations for the NE folding of deca-alanine using the fastest pushing

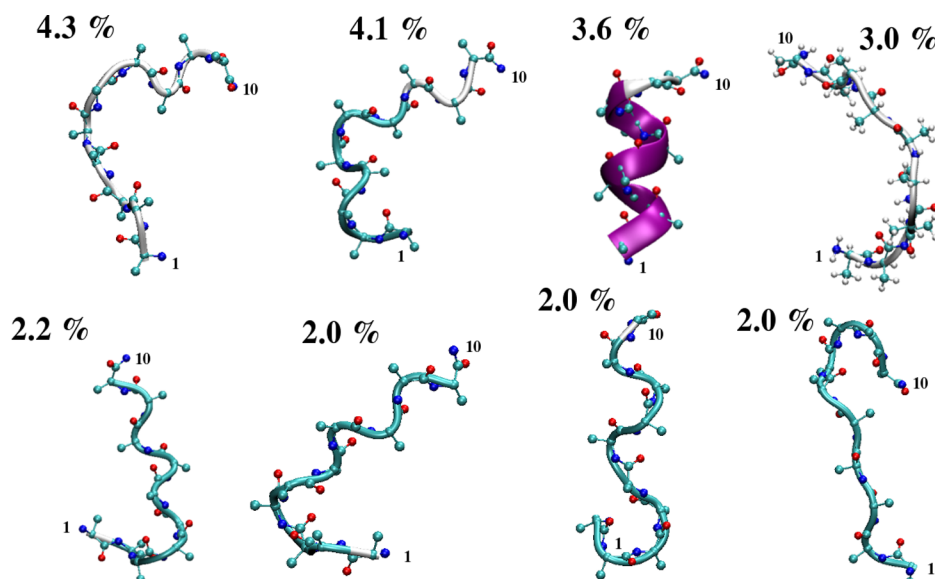


FIG. 2. Representative structures of the most populated clusters of deca-alanine in the $B \rightarrow A$ NE process re-folding as obtained by the quality threshold clustering algorithm using a threshold tolerance of 4.0 \AA (Refs. 39 and 40) on 512 final configurations. The first and last residues of the chain are indicated for each cluster representative. In all structures, $\lambda = 1.55 \text{ nm}$.

speed (76.2 m/s). At such speed, the first eight clusters account for only about 20% of the total population with most of the clusters composed of one single representative, in spite of the rather loose clustering threshold used. This disordered configurational distribution at the end point is due to the huge dissipation (about 80 kJ mol⁻¹) of the NE refolding process that allows to explore a true manifold of high free energy metastable states.

On the basis of these results, in order to evaluate ΔG at a given speed using Eq. (7), we should in principle use a mixture of type Eq. (4) containing a number of terms that is proportional to the number of metastable states observed in the refolding process. In practice, we shall assume for all speeds only *one* metastable state ($N = 2$ in Eq. (7)) with an unknown degeneracy. We will see that such rather drastic assumption, besides being straightforward from a computational standpoint, will turn out to be surprisingly accurate at all speeds for unidirectional and bidirectional estimates.

A. Determination of the normal distributions in the mixture

As stated in the introduction, the present study is aimed at computing the ($N = 2$) mixture-based unbiased unfolding free energy estimates in NE processes with work distribution asymmetry, such as the helix-coil transition in deca-alanine, using Eq. (12), without any tweaking, interruption, or reweighting of the NE trajectories, i.e., using the standard NE approach. The production of NE works in either directions can therefore be done with no communication between parallel instances each running a single NE trajectory, hence, fully exploiting the inherent parallel nature of the Jarzynski methodology. The following describes the subsequent post-processing analysis on the produced works aimed at recovering ΔG via Eq. (12).

Given a distribution $P(W)$ (either for the forward or reverse process), the cumulants $\mu_{1,2}$, $\sigma_{1,2}^2$, and the coefficient c appearing in ΔG estimate Eq. (12) could be computed by evaluating the first five central moments of the observed distribution either by maximum likelihood methods^{41,42} or analytically by finding the negative root of the so-called Pearson nomic equation.⁴³ The intricacy of these techniques, even for the simple case of a mixture with two components, could be in effect quite disheartening. We shall therefore adopt a straightforward fitting procedure based on standard minimization tools such as the Powell⁴⁴ method for the five-parameter function

$$F(\mu_1, \mu_2, \sigma_1, \sigma_2, c) = \sum_k [\tilde{P}(W_k, \mu_1, \mu_2, \sigma_1, \sigma_2, c) - P(W_k)]^2. \quad (16)$$

In Eq. (16), $\tilde{P}(W_k, \mu_1, \mu_2, \sigma_1, \sigma_2, c) = cN(W_k, \mu_1, \sigma_1) + (1 - c)N(W_k, \mu_2, \sigma_2)$ and $P(W_k)$ are the normalized histograms of the calculated and of the observed distribution, respectively, with a speed dependent (i.e., spread dependent) bin size set to $(W_{\max} - W_{\min})/N_{\text{bin}}$ where W_{\max} and W_{\min} are the maximum and minimum work in $P(W_k)$ and N_{bin} is the number of bins (set in our case to 50). Except for the c coefficient, that must be in the interval $[0, 1]$, the fit is performed using very weak

constraints on the values of the free parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$. The central moments μ_1 and μ_2 , that should be related to the individual ΔG_1 and ΔG_2 referred to the principal and secondary basins of the folded state A (see Eq. (10)), as well as the unfolding free energy ΔG , are all allowed to vary between 0 and 200 kJ mol⁻¹, i.e., within the maximum and minimum value of the work observed, at *any* speed, in the $P_{AB}(W)$ and $P_{BA}(-W)$ distributions, respectively. This is actually an unnecessary large domain for deca-alanine, given that the unknown free energy ΔG , because of the second principle, can never exceed the mean value of the work in the forward process or be less than the mean value of the work (with inverted sign) for the reverse transformation. Such mean values are found in the intervals [95,170] and [5,93] kJ mol⁻¹, respectively (see Figure 3). The width of the normal components, σ_1 and σ_2 , is also allowed to vary in the range 0-20 kJ mol⁻¹, corresponding to a maximum dissipation of ≈ 80 kJ mol⁻¹, largely above the widest spread observed in the principal component of any forward or reverse distributions (see Figure 3). The function $F(\mu_1, \mu_2, \sigma_1, \sigma_2, c)$ in Eq. (16) can be minimized using only one of the observed distributions, either the forward distribution ($P(W_k) = P_{AB}(W_k)$), or the reverse distribution ($P(W_k) = P_{BA}(-W_k)$), or using simultaneously both the forward and reverse distributions (i.e., $P(W_k) = P_{AB}(W_k) + P_{BA}(-W_k)$). In the latter case, we have a bidirectional estimate of ΔG while the former case provides an unidirectional estimate that allows to *predict*, according to Eqs. (4) and (5), the shape of the unknown partner distribution. The global fitting is done by coupling the Powell minimization method for local minimum search with random generation of initial points uniformly sampled in the parameters domain. The number of Powell minimization steps with random initial parameters is set to 300. The errors in the fitted distributions as a function of the work (thin lines in Figure 3) are computed by block bootstrapping the 512 forward and reverse works into 50 samples of half size. Taking into account the 50 bootstrap samples and the three possible fitting methods (based on the forward, reverse, and the cumulative observed distribution), in total $300 \times 50 \times 3 = 45\,000$ Powell minimizations are produced for each speed. Although quite dreadful to tell, such post-processing analysis can be done on a laptop computer in a matter of *minutes*. Most importantly, the computational burden of this step depends only on the number of NE trajectories, being of course totally independent of the size of the system.

In Figure 3, on the right we show the best fitting forward $\tilde{P}_{AB}(W_k, \mu_1, \mu_2, \sigma_1, \sigma_2, c)$ (red, thick lines) and reverse $\tilde{P}_{BA}(-W_k, \mu_1, \mu_2, \sigma_1, \sigma_2, d)$ (on the left, green thick lines) obtained by minimization of the function F of Eq. (16) using *both* the observed distributions $P_{AB}(W_k)$ and $P_{BA}(-W_k)$ (reported in black color using solid and dashed lines, respectively). The calculated $\tilde{P}_{AB}(W_k, \mu_1, \mu_2, \sigma_1, \sigma_2, c)$ and $\tilde{P}_{BA}(-W_k, \mu_1, \mu_2, \sigma_1, \sigma_2, d)$ are related via Eq. (6), that is in turn a trivial consequence of Eq. (2). We must stress here once again that the d coefficient, shaping the reverse distribution, is *not a free parameter*, being firmly intertwined via the Crooks theorem to the c parameter. The assumed bimodality of the two partner distributions $\tilde{P}_{AB}(W_k, \mu_1, \mu_2, \sigma_1, \sigma_2, c)$ and $\tilde{P}_{BA}(-W_k, \mu_1, \mu_2, \sigma_1, \sigma_2, d)$ appears to explain, with remarkable correspondence indeed, the observed asymmetry of the

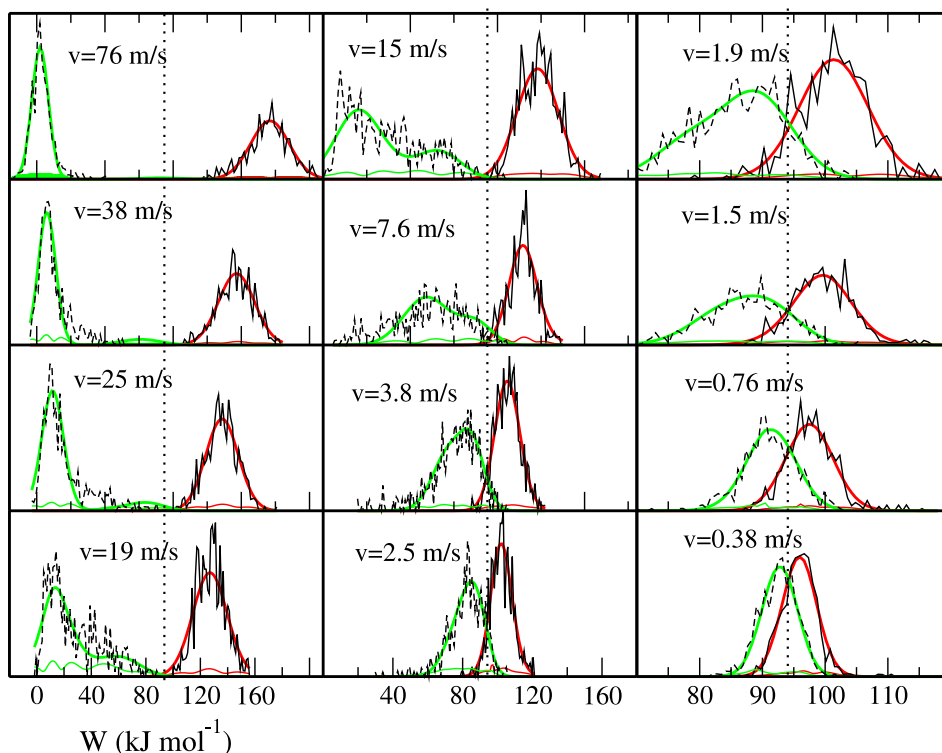


FIG. 3. Forward (right, red) and reverse (left, green) work distributions fitted on the cumulative observed distribution $P_{AB}(W) + P_{BA}(-W)$ using Eq. (4) with $N=2$ at various speeds for the folding-unfolding NE process in deca-alanine *in vacuo*. The observed distributions $P_{AB}(W)$ and $P_{BA}(-W)$ are the black solid and dashed lines, respectively. The dotted vertical line crosses the x-axis at $W = \Delta G$ (reference value computed with BAR).

forward and reverse distributions $P_{AB}(W_k)$ and $P_{BA}(-W_k)$. We also note that the errors in the fitted distributions heavily (and expectedly) depend on the spread, rather than the speed, of the corresponding observed distributions. A large spread produces in fact noisy observed work distribution, thus yielding larger errors on the fitted distributions. So, the errors are minimal at the highest speed of $v = 76$ m/s and for speeds less than $v = 1.5$ m/s. The largest errors are seen in the reverse distribution $\hat{P}_{BA}(-W_k, \mu_1, \mu_2, \sigma_1, \sigma_2, d)$ for $v = 19$ m/s, and $v = 15$ m/s, i.e., when the corresponding observed overall spread exceeds 80 kJ mol^{-1} .

In Figure 4 we plot, as a function of the pulling speed, the fitted c and d coefficients, along with fitted value of Δg_1 and Δg_2 defined in Eq. (8). As discussed in Sec. II, the Δg_1 and Δg_2 quantities are found to be strongly speed dependent

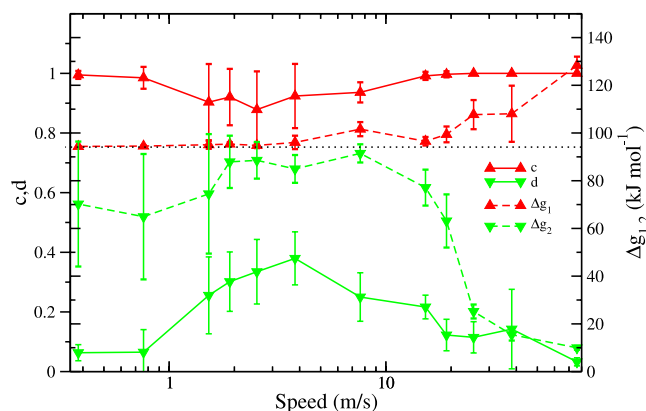


FIG. 4. Left scale: Fitted c coefficient and calculated d coefficient (Eq. (6)) for the folding-unfolding NE process of deca-alanine *in vacuo* as a function of speed. Right scale: Δg_1 , Δg_2 quantities (dashed lines, Eq. (8)).

ranging from 95 to 130 kJ mol^{-1} for Δg_1 and from 10 to 92 for Δg_2 . At low speeds, the two partner distributions are basically two symmetrical (with respect to the free energy $W = \Delta G$) Gaussian distributions satisfying Eq. (3) such that the metastable states cannot simply be discerned at the given resolution. At low speed, in fact, basically all the 512 ending configurations in the NE refolding process are of α -helical type. In this condition, as shown in Figure 4, the unrestrained fit yields $\Delta g_1 = \Delta g_2 \approx \Delta G$ so that the c and d coefficient and Eq. (6) become irrelevant for the ΔG estimate, Eq. (7), yielding a huge variance (especially for the d coefficient). The metastable states are felt more clearly at high pulling speed where Δg_1 and Δg_2 have very different values and where, as discussed in Sec. II, there is, in effect, an inversion in the weight factors (i.e., $c/d \gg 1$) for the forward and reverse distribution. The ending NE process, carried on at high speed, senses indeed a true manifold of metastable states of A (see Figure 2), all these states having basically an insignificant Boltzmann weight. $\Delta g_2 \approx 10 : 30$ can be connected to the (speed-dependent) average level of unfolding free energy starting from of these misfolded metastable states of A. Note also that at intermediate speeds, as anticipated in Sec. II, while the c coefficient in the forward distribution stays close to one (sensing mostly the principal state, as schematized in Figure 1), the d coefficient rises up to ≈ 0.4 indicating a more mixed character of the overall reverse distribution (in accord with Eq. (6)).

In Figure 5, we finally show the resulting forward and reverse distributions at two relatively high speeds when the fitting procedure is done using the work data in *one direction only*. In Figure 5, the thick dashed trait refers to the *predicted* (i.e., non fitted) distributions. Remarkably, the fitting procedure using the reverse distribution provides an estimate of the unknown forward distribution (red dashed curves) that

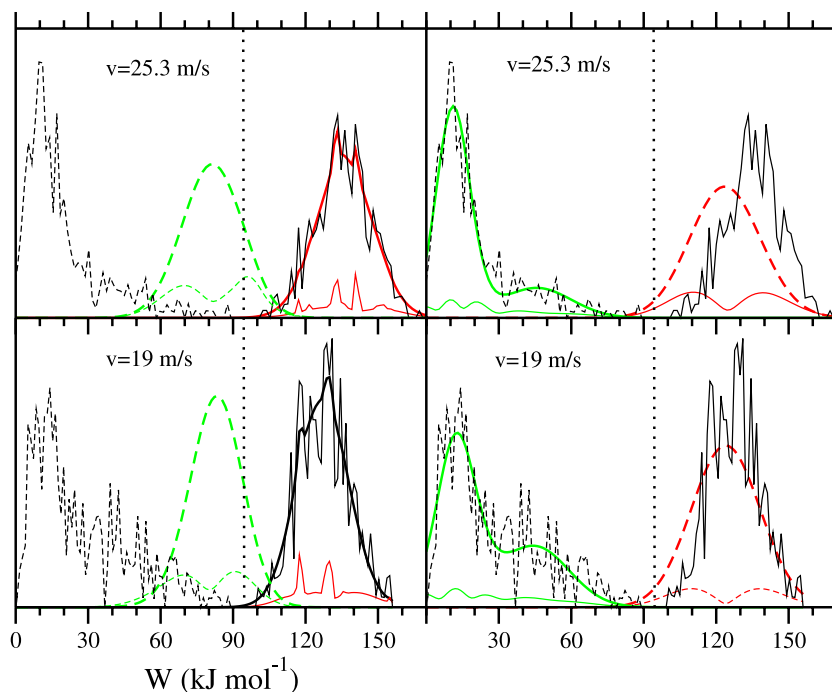


FIG. 5. Left: Forward (on the right, red solid trait) and predicted reverse (on the left, green dashed trait) work distributions fitted on the observed forward distribution $P_{AB}(W)$ (black, solid) at two representative speeds for the folding-unfolding NE process in deca-alanine *in vacuo*. Right: reverse (on the left, green solid trait) and predicted forward (on the right, red dashed trait) work distributions fitted on the observed reverse distribution $P_{BA}(-W)$ at the same speeds. The dotted vertical line crosses the x-axis at $W = \Delta G$ (reference value computed with BAR).

is much more accurate than that obtained for the unknown reverse distribution when fitting using forward data. This is due to the fact that, even if the mean value of the work in the reverse distribution is far from ΔG , the shape of $P_{BA}(-W)$ bears more information regarding the states in A, with respect to the more regular, at all speeds, $P_{AB}(W)$. This brings along, as we shall see later on, a less biased estimate of ΔG at fast pulling speeds when the data are fitted on the refolding work values.

B. Free energy estimates

The free energies ΔG_{AB} are calculated using the following unidirectional method: (i) the Jarzynski exponential average,

Eq. (1); (ii) the $\mathcal{N} = 1$ Gaussian estimate, Eq. (3); (iii) the estimate provided by Eq. (7), based on a mixture of only two normal distributions. For all three unidirectional methods, the ΔG_{AB} estimates are obtained using in turn the forward and the reverse distributions $P_{AB}(W)$, $P_{BA}(-W)$. In all cases, the errors on the ΔG_{AB} estimates are evaluated by block bootstrapping using 50 (forward and reverse) sets each containing 256 works randomly sampled out of the 512 works. The results for the free energy estimates as a function of the speed, using the three methods based on Eqs. (1), (3), and (12), are collected in Figure 6. The curves in red (triangle up) and green colors (triangles down) refer to *unidirectional* estimates based on the forward and reverse work distribution, respectively. The (red, triangles up and green, triangles down)

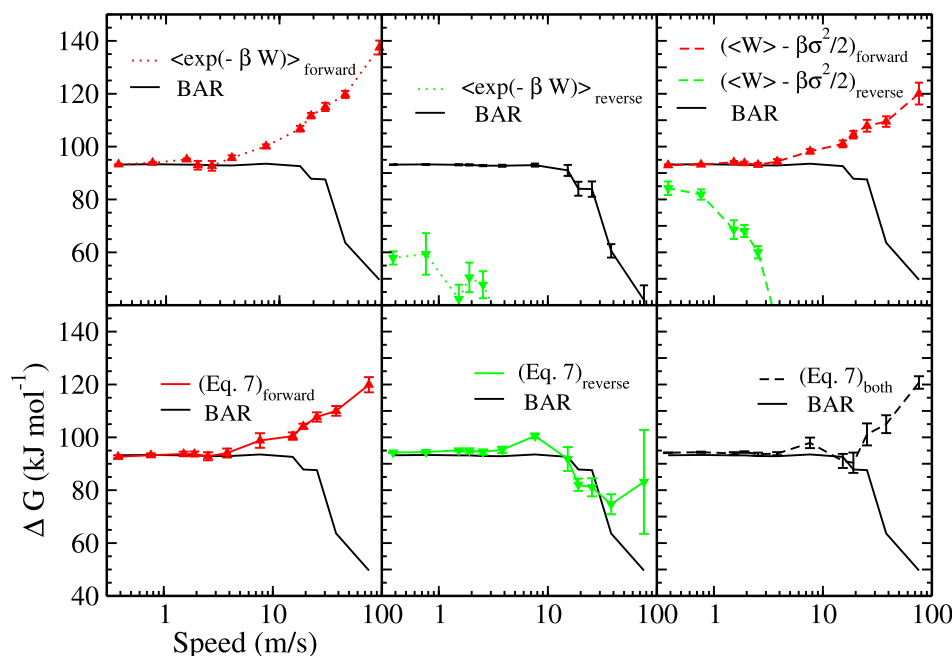
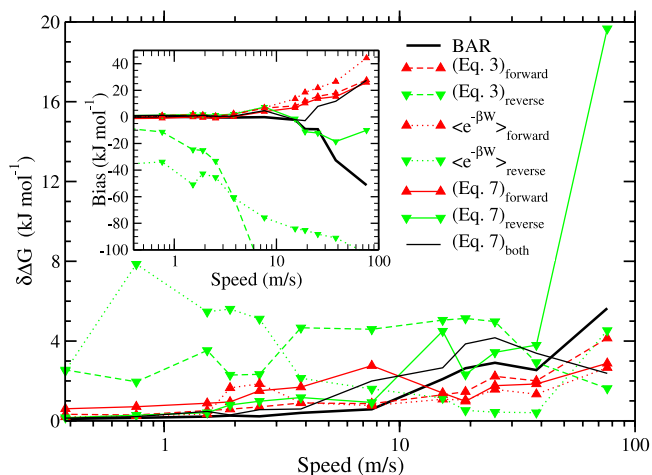


FIG. 6. Free energy as a function of the pulling speed using various methods. (see text).

FIG. 7. Error and Bias (inset) in ΔG_{AB} estimates.

solid traits refer to Eq. (12)-based estimates, while the dashed and dotted lines (red, triangles up and green, triangles down) indicate Eqs. (3) and (1) estimates, respectively. The black curves, finally, are related to *bidirectional* estimate, either using the BAR method (solid trait) or Eq. (12)-based method (dashed trait). The BAR estimate at low speed provides the reference maximum likelihood estimate of the true unfolding free energy ΔG_{AB} . The data concerning the bias and error in the various methods as a function of the speed are finally collected in Figure 7.

As shown in many studies,^{11,14,15,17,25,29,30,32} the Jarzynski unidirectional exponential average yields, at high speeds, a strongly biased ΔG estimate. The bias in the forward direction is positive and grows exponentially already starting from speed of $v \approx 2$ m/s. In the reverse direction, Eq. (1) estimates are plagued by huge errors and a strong negative bias even at low speed.

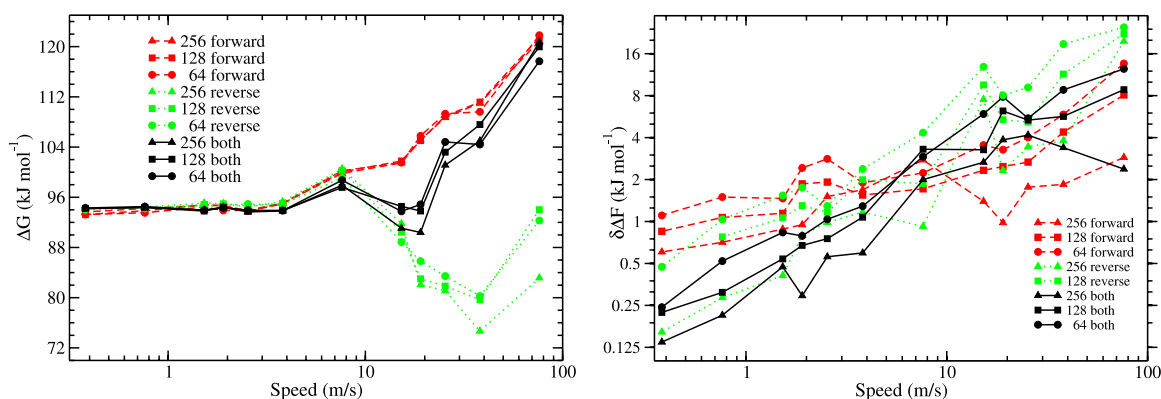
Simple Gaussian expression Eq. (3), provides, in the *forward* direction, a better estimate of ΔG compared to that obtained with Eq. (1), differing at most by 5% with respect to the reference BAR value up to speeds as fast as $v \approx 3 : 4$ m/s (i.e., for a duration of the NE experiment of less than half nanosecond). The positive bias, in this case, grows with speed at a much lower rate than that observed for the exponential average. For the reverse process, the unidirectional estimate based on Eq. (3) consistently yields a huge *negative* bias

that is even more important, at high pulling speeds, than that observed for Jarzynski estimate Eq. (1) (see inset in Figure 7), yielding completely unreliable estimate of ΔG . This is so since in the Gaussian estimate based on the first two cumulants, Eq. (3), the right tail of the distribution yields no contribution to ΔG_{AB} . When using Eq. (1), the low work values (i.e., those that are closer to the reversible work ΔG_{AB}) have instead a disproportionate weight in the integral.

We now come to Eq. (12)-based estimates, i.e., the object of the present study. Regarding the forward direction, we first note that Eq. (12) closely follows, both in the free energy estimates and in the errors, the simple forward Gaussian estimate with $N = 1$. In the unfolding process, the metastable states are not sensed and path mixing is rare (see Figure 1), so that the system behaves as if $N = 1$ at all speeds. On the other hand, in the reverse refolding process, the metastable states are a probable outcome of the NE experiment already at moderate speeds, and the secondary normal components in Eq. (4) make their appearance widening the distribution that exhibits a typical long right tail (see Figure 3). Therefore, Eq. (12)-based free energy estimates using the refolding works remarkably exhibit a small to moderate bias and error *all speeds*. The bias at the highest speeds is even much *lower* than obtained with the BAR bidirectional method. The same holds true for Eq. (12) bidirectional estimate where the positive bias tends to coincide, at high speeds, with that of the corresponding forward estimate (either using Eq. (3) or Eq. (12)). Such high speed bias, however, is much lower, in absolute value, than that obtained with the BAR approach.

C. Free energy estimate and effect of the sample size

We finally analyze the effect of the sample size (that is, the number of independent NE trajectories) on the free energy estimates based on Eq. (12). As previously stated, the accuracy of Eq. (12) depends only on the number of NE trajectories used to construct the histograms of the reference work distributions and not on the system size. We hence estimated the free energy using the fitting procedure described in Sec. IV A by evaluating the observed distributions $P(W_k)$ using 128 and 64 work values out of 512 at constant bin size. The results for the fitted free energy estimates based on the forward $P(W_k) = P_{AB}(W_k)$, reverse $P(W_k) = P_{BA}(-W_k)$, and cumulative

FIG. 8. Left: sample size and ΔG_{AB} estimates; Right: sample size and observed bootstrap errors $\delta \Delta G_{AB}$.

$P(W_k) = P_{AB}(W_k) + P_{BA}(-W_k)$ observed distributions are reported in Figure 8(left). Errors have been computed by block bootstrapping the 512 forward and reverse final work values into 50 samples with 64, 128, and 256 work values. Quite surprisingly, in spite of the noisy observed distributions using 64 and 128 works only, the corresponding free energy estimates appear to be quite robust exhibiting, at all speeds, no significant degradation with respect to the 256 work values. Expectedly, as shown in Figure 8(right), the bootstrap error increases steadily at all speeds as the number of sampled trajectories decreases.

Based on the data shown in the Figures 6–8, we may hence conclude that the bidirectional and the reverse unidirectional approaches based on Eq. (12) in deca-alanine provide an excellent and *unbiased* estimates up to pulling speeds as fast as 11 m/s (for a duration of the NE experiment of about 150 ps). Remarkably, the estimates are still accurate for a number of sampled trajectories as small as 64. At higher speed, Eq. (12) starts to provide an unstable estimate of ΔG , that is in any case always within 10% of the reference value, over-performing the BAR maximum likelihood bidirectional estimate at the fastest speed of 76 m/s, i.e., for duration of the pulling/pushing as short as 21 ps.

V. CONCLUSIONS

The Jarzynski theorem provides in principle a perfectly a parallel algorithm for estimating the free energy difference between any two given thermodynamic states that can be connected via NE driven transformations. All NE trajectories connecting these states may in fact proceed independently and concurrently with zero fraction of serial code and with no need for communication whatsoever. In practice, Jarzynski-based free energy estimates are strongly biased and affected by large errors because of the inherent noisy statistics of the exponential average over a work distribution, crucially depending on the poorly sampled tails. Second order cumulant expansion for seemingly Gaussian work distributions provides a better estimate of the free energy at low pulling speed. At higher speeds, in systems that are characterized by strongly asymmetrical forward and reverse work distributions, unidirectional approaches based on the Gaussian approximation fail, yielding again strongly biased and unreliable estimates, especially when the direction of the NE experiment envisages the entrance in a funnel, like in the folding of a small polypeptide or in the docking of a drug on a receptor.

In this contribution, we have shown that the free energy difference between the end points in these important asymmetric cases can be computed with great accuracy using relatively few *completed, non-communicating, and untweaked* NE trajectories with a generalization of the *unbiased* Jarzynski estimate based on the assumption that any observed work distribution is the result of a mixture of Gaussian distributions. These normal components of the overall observed distribution are related to the existence of metastable sub-states in one of the two final thermodynamic states. If a work distribution in a given direction is the result of a combination of N normal distributions, then the Crooks theorem imposes that the distribution associated to the inverted process must

also be given by a combination of the same normal components, although with different weights. Using the prototypical example for the driven unfolding/folding of deca-alanine, we have shown that the predicted behavior of the forward and reverse work distributions, described by a combination of only two Gaussian components with Crooks derived weights, explains surprisingly well the striking asymmetry in the observed work distributions, providing at the same time a reliable and unbiased unidirectional and bidirectional estimate of the free energy difference at all speeds. Remarkably, the unidirectional estimate based on the Gaussian mixture using the folding work distribution yields an accurate and stable estimate of the free energy also for speeds as fast as 15:20 nm/ns where the maximum likelihood BAR method fails. The proposed algorithm fully preserves the inherent parallel nature of the Jarzynski approach.

¹C. Jarzynski, "Nonequilibrium equality for free energy differences," *Phys. Rev. Lett.* **78**, 2690–2693 (1997).

²G. E. Crooks, "Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems," *J. Stat. Phys.* **90**, 1481–1487 (1998).

³R. W. Zwanzig, "High-temperature equation of state by a perturbation method. I. Nonpolar gases," *J. Chem. Phys.* **22**, 1420–1426 (1954).

⁴J. S. Patel, A. Berteotti, S. Ronsisvalle, W. Rocchia, and A. Cavalli, "Steered molecular dynamics simulations for studying proteinligand interaction in cyclin-dependent kinase 5," *J. Chem. Inf. Model.* **54**(2), 470–480 (2014).

⁵F. Colizzi, R. Perozzo, L. Scapozza, M. Recanatini, and A. Cavalli, "Single-molecule pulling simulations can discern active from inactive enzyme inhibitors," *J. Am. Chem. Soc.* **132**, 7361–7371 (2010).

⁶J. S. Patel, D. Branduardi, M. Masetti, W. Rocchia, and A. Cavalli, "Insights into ligand-protein binding from local mechanical response," *J. Chem. Theory Comput.* **7**, 3368–3378 (2011).

⁷F. Marty Ytreberg, "Absolute fkb binding affinities obtained via nonequilibrium unbinding simulations," *J. Chem. Phys.* **130**(16), 164906 (2009).

⁸T. Bastug, P.-C. Chen, S. M. Patra, and S. Kuyucak, "Potential of mean force calculations of ligand binding to ion channels from jarzynski equality and umbrella sampling," *J. Chem. Phys.* **128**(15), 155104 (2008).

⁹T. Bastug and S. Kuyucak, "Application of jarzynski equality in simple versus complex systems," *Chem. Phys. Lett.* **436**(46), 383–387 (2007).

¹⁰J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, L. Skeel, and K. Schulten, "Scalable molecular dynamics with namd," *J. Comput. Chem.* **26**, 1781–1802 (2005).

¹¹S. Park and K. Schulten, "Calculating potentials of mean force from steered molecular dynamics simulations," *J. Chem. Phys.* **120**(13), 5946–5961 (2004).

¹²G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," in *Proceedings of the April 18–20, 1967, Spring Joint Computer Conference, AFIPS '67 (Spring)* (ACM, New York, NY, USA, 1967), pp. 483–485.

¹³G. Hummer, "Fast-growth thermodynamic integration: Error and efficiency analysis," *J. Chem. Phys.* **114**, 7330–7337 (2001).

¹⁴J. Gore, F. Ritort, and C. Bustamante, "Bias and error in estimates of equilibrium free-energy differences from nonequilibrium measurements," *Proc. Natl. Acad. Sci. U. S. A.* **100**(22), 12564–12569 (2003).

¹⁵D. Wu and D. A. Kofke, "Model for small-sample bias of free-energy calculations applied to Gaussian-distributed nonequilibrium work measurements," *J. Chem. Phys.* **121**(18), 8742–8747 (2004).

¹⁶M. R. Shirts and V. S. Pande, "Comparison of efficiency and bias of free energies computed by exponential averaging, the bennett acceptance ratio, and thermodynamic integration," *J. Chem. Phys.* **122**(14), 144107 (2005).

¹⁷H. Oberhofer, C. Dellago, and P. L. Geissler, "Biased sampling of nonequilibrium trajectories: Can fast switching simulations outperform conventional free energy calculation methods?," *J. Phys. Chem. B* **109**(14), 6902–6915 (2005).

¹⁸H. Oberhofer and C. Dellago, "Optimum bias for fast-switching free energy calculations," *Comput. Phys. Commun.* **179**(13), 41–45 (2008), special issue based on the Conference on Computational Physics 2007 {CCP} 2007.

¹⁹C. H. Bennett, "Efficient estimation of free energy differences from Monte Carlo data," *J. Comput. Phys.* **22**, 245–268 (1976).

- ²⁰M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, "Equilibrium free energies from nonequilibrium measurements using maximum likelihood methods," *Phys. Rev. Lett.* **91**, 140601 (2003).
- ²¹P. Procacci and C. Cardelli, "Fast switching alchemical transformations in molecular dynamics simulations," *J. Chem. Theory Comput.* **10**, 2813–2823 (2014).
- ²²G. Ozer, E. F. Valeev, S. Quirk, and R. Hernandez, "Adaptive steered molecular dynamics of the long-distance unfolding of neuropeptide y," *J. Chem. Theory Comput.* **6**(10), 3026–3038 (2010).
- ²³R. Chelli, C. Gellini, G. Pietraperzia, E. Giovannelli, and G. Cardini, "Path-breaking schemes for nonequilibrium free energy calculations," *J. Chem. Phys.* **138**, 214109 (2013).
- ²⁴F. Colizzi and G. Bussi, "Rna unwinding from reweighted pulling simulations," *J. Am. Chem. Soc.* **134**(11), 5173–5179 (2012).
- ²⁵G. Ozer, T. Keyes, S. Quirk, and R. Hernandez, "Multiple branched adaptive steered molecular dynamics," *J. Chem. Phys.* **141**(6), 064101 (2014).
- ²⁶In the ASMD method,²² for example, the proof of the algorithm (Eqs. 8-17) involves a number of shaky points such as: (i) the assumed ergodicity of the system *along with* the assumption of "a single local basin of attraction" at the intermediate step; (ii) the implication inherent in the "formal relaxing stage" of *different time schedules* for each of the trajectories; (iii) the assumption of bath decoherence that cannot be taken for granted in fully atomistic system subject to, e.g., microsolvation effects.
- ²⁷T. N. Do, P. Carloni, G. Varani, and G. Bussi, "Rna/peptide binding driven by electrostatics: insight from bidirectional pulling simulations," *J. Chem. Theory Comput.* **9**(3), 1720–1730 (2013).
- ²⁸In ubiquitous driven Gaussian processes, the free energy can be *exactly* computed with great accuracy via unbiased estimate Eq. (3) using *few hundreds* of *fast* and *complete* independent non-communicating simulations lasting for *minutes* in systems far more complex³⁴ than the simple helix-coil transition of deca-alanine *in vacuo* used by Chelli *et al.* as a typical case study. Path Breaking (PBS) applied to a Gaussian process obliges one either to use the biased exponential average on the survived trajectories or to extrapolate the first non zero cumulants of a Gaussian distribution using only the completed (least dissipative) trajectories, i.e., introducing artificially a statistical bias with a disastrous impact on the accuracy of the method. This dramatic shortcoming of the PBS is openly admitted in the paper²³ (Sec. II B page 5) where in the discussion of the toy model for a Gaussian process it is stated that "the free energy difference obtained from Eq. (4) (i.e., PBS) using a set of (as much as) 10 000 work samples is -4.58 , which is, as expected, overestimated with respect to the theoretical value (by approximately 10%)."
- ²⁹H. Oberhofer, C. Dellago, and S. Boresch, "Single molecule pulling with large time steps," *Phys. Rev. E* **75**, 061106 (2007).
- ³⁰P. Procacci and S. Marsili, "Energy dissipation asymmetry in the non equilibrium folding/unfolding of the single molecule alanine dcapeptide," *Chem. Phys.* **375**, 8–15 (2010).
- ³¹G. Ozer, S. Quirk, and R. Hernandez, "Adaptive steered molecular dynamics: Validation of the selection criterion and benchmarking energetics in vacuum," *J. Chem. Phys.* **136**(21), 215104 (2012).
- ³²P. Procacci, S. Marsili, A. Barducci, G. F. Signorini, and R. Chelli, "Crooks equation for steered molecular dynamics using a nosé-hoover thermostat," *J. Chem. Phys.* **125**, 164101 (2006).
- ³³S. Kim, Y. W. Kim, P. Talkner, and J. Yi, "Comparison of free-energy estimators and their dependence on dissipated work," *Phys. Rev. E* **86**, 041130 (2012).
- ³⁴R. B. Sandberg, M. Banchelli, C. Guardiani, S. Menichetti, G. Caminati, and P. Procacci, "Efficient nonequilibrium method for binding free energy calculations in molecular dynamics simulations," *J. Chem. Theory Comput.* **11**, 423–435 (2015).
- ³⁵J. Marcinkiewicz, "Sur une propri  t   de la loi de gauss," *Math. Z.* **44**, 612–618 (1939).
- ³⁶E. H. Feng and G. E. Crooks, "Length of time's arrow," *Phys. Rev. Lett.* **101**, 090602 (2008).
- ³⁷S. Marsili and P. Procacci, "Free energy reconstruction in bidirectional force spectroscopy experiments: The effect of the device stiffness," *J. Phys. Chem. B* **114**(7), 2509–2516 (2010).
- ³⁸S. Marsili, G. F. Signorini, R. Chelli, M. Marchi, and P. Procacci, "Orac: A molecular dynamics simulation program to explore free energy surfaces in biomolecular systems at the atomistic level," *J. Comput. Chem.* **31**, 1106–1116 (2010).
- ³⁹L. J. Heyer, S. Kruglyak, and S. Yooseph, "Exploring expression data: Identification and analysis of coexpressed genes," *Genome Res.* **9**, 1106–1115 (1999).
- ⁴⁰C. Guardiani, G. F. Signorini, R. Livi, A. M. Papini, and P. Procacci, "Conformational landscape of N-glycosylated peptides detecting autoantibodies in multiple sclerosis, revealed by Hamiltonian replica exchange," *J. Phys. Chem. B* **116**(18), 5458–5467 (2012).
- ⁴¹M. C. Pardo, "A comparison of some estimators of the mixture proportion of mixed normal distributions," *J. Comput. Appl. Math.* **84**(2), 207–217 (1997).
- ⁴²A. Clifford Cohen, "Estimation in mixtures of two normal distributions," *Technometrics* **9**, 15–28 (1967).
- ⁴³K. Pearson, "Contributions to the mathematical theory of evolution," *Philos. Trans. R. Soc., A* **185**, 71–110 (1894).
- ⁴⁴M. J. D. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives," *Comput. J.* **7**, 155–162 (1964).